Quality of Experience in Immersive Video Technologies

THÈSE N^{O} 6971 (2016)

PRÉSENTÉE LE 15 AVRIL 2016

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

GROUPE EBRAHIMI

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Philippe HANHART

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Prof. T. Ebrahimi, directeur de thèse
Prof. P. Schelkens, rapporteur
Dr A. Smolic, rapporteur
Prof. P. Frossard, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

Do or do not.
There is no try.
— Yoda, The Empire Strikes Back (1980)

To my family…

Acknowledgements

You are about to read the tale of a quest that took me four years to complete. It was a memorable journey, with many challenges to overcome, each requiring a sophisticated mix of magic and sword skills to release the key to the next level. However, without the help, inspiration, encouragement, and support from many wonderful people, this quest would have remained a pure myth.

First of all, I would like to express my sincere gratitude to Lord Prof. Touradj Ebrahimi for giving me the opportunity to serve at his command and for giving me so much freedom in solving my quest. I appreciated the many opportunities he had offered me to take part in different tournaments across the world to meet other knights of the round table and share the tales of my adventures.

Then, I would like to thank Lords Prof. Pascal Frossard, Prof. Peter Schelkens, and Dr. Aljoscha Smolic for accepting to present as judges for my ultimate challenge, for the thorough reading of this manuscript, and for the valuable commenting that helped improving its quality. Prof. Jean-Philippe Thiran was a great wizard and made sure everything went smoothly during this rite of passage.

My appreciation goes to my fellow knights and obliging squires for riding along with me during these adventures: Dr. Pavel Korshunov, Dr. Martin Rerabek, Dr. Eleni Kroupi, Dr. Jong-Seok Lee, Mr. Hiromi Nemoto, Dr. Francesca De Simone, Dr. Emilie Bosc, Prof. Patrick Le Callet, Mr. Marco Bernardo, Prof. Manuela Pereira, Prof. Antonio Pinheiro, Dr. Alessandro Artusi, Dr. Rafal Mantiuk, Dr. Thomas Richter, Dr. Naeem Ramzan, Mr. Vittorio Baroncini, Mr. Alexandre Chappuis, and Mr. Carmelo di Nolfo.

Special thanks go to all other current and former companions who have served under the same gonfalon: Lin, Anne-Flore, He, Margherita, and others, for the feasts that we have organized together inside and outside the castle EPFL.

Last but not least, my most important acknowledgments are dedicated to my parents Lady Anne-Claude and Sir Bernard, as well as their life partners Lady Gilberte and Sir Marcel, and my grand mother Lady Juliette, for their love, encouragement, advice, and unconditional support throughout my whole life. I am particularly thankful to them for keeping me away from the hassles of everyday life during all these years spent at the keep, which allowed me to dedicate my time and my energy for a more noble cause, as the monk who dedicated his life to religion. This manuscript is dedicated to my family.

Lausanne, 31 March 2016 P. H.

Abstract

Over the last decades, several technological revolutions have impacted the television industry, such as the shifts from black & white to color and from standard to high-definition. Nevertheless, further considerable improvements can still be achieved to provide a better multimedia experience, for example with ultra-high-definition, high dynamic range & wide color gamut, or 3D. These so-called immersive technologies aim at providing better, more realistic, and emotionally stronger experiences.

To measure quality of experience (QoE), subjective evaluation is the ultimate means since it relies on a pool of human subjects. However, reliable and meaningful results can only be obtained if experiments are properly designed and conducted following a strict methodology. In this thesis, we build a rigorous framework for subjective evaluation of new types of image and video content. We propose different procedures and analysis tools for measuring QoE in immersive technologies.

As immersive technologies capture more information than conventional technologies, they have the ability to provide more details, enhanced depth perception, as well as better color, contrast, and brightness. To measure the impact of immersive technologies on the viewers' QoE, we apply the proposed framework for designing experiments and analyzing collected subjects' ratings. We also analyze eye movements to study human visual attention during immersive content playback.

Since immersive content carries more information than conventional content, efficient compression algorithms are needed for storage and transmission using existing infrastructures. To determine the required bandwidth for high-quality transmission of immersive content, we use the proposed framework to conduct meticulous evaluations of recent image and video codecs in the context of immersive technologies.

Subjective evaluation is time consuming, expensive, and is not always feasible. Consequently, researchers have developed objective metrics to automatically predict quality. To measure the performance of objective metrics in assessing immersive content quality, we perform several in-depth benchmarks of state-of-the-art and commonly used objective metrics. For this aim, we use ground truth quality scores, which are collected under our subjective evaluation framework.

To improve QoE, we propose different systems for stereoscopic and autostereoscopic 3D displays in particular. The proposed systems can help reducing the artifacts generated at the visualization stage, which impact picture quality, depth quality, and visual comfort. To demonstrate the effectiveness of these systems, we use the proposed framework to measure

viewers' preference between these systems and standard 2D & 3D modes.
In summary, this thesis tackles the problems of measuring, predicting, and improving QoE in immersive technologies. To address these problems, we build a rigorous framework and we apply it through several in-depth investigations. We put essential concepts of multimedia QoE under this framework. These concepts not only are of fundamental nature, but also have shown their impact in very practical applications. In particular, the JPEG, MPEG, and VCEG standardization bodies have adopted these concepts to select technologies that were proposed for standardization and to validate the resulting standards in terms of compression efficiency.

Key words: quality of experience, immersive video technology, ultra-high-definition, high dynamic range, 3D, subjective quality assessment, subjective evaluation, pair comparison, Thurstone Case V model, visual quality, evaluation protocol, crowdsourcing, eye tracking, visual attention, objective quality metric, objective quality assessment, performance evaluation, HEVC, VP9, JPEG, JPEG 2000, JPEG XT, coding efficiency, Bjøntegaard model, crosstalk, pseudoscopy, vergence-accommodation rivalry, stereoscopic display, multiview autostereoscopic display

Résumé

Au cours des dernières décennies, plusieurs révolutions technologiques ont eu un impact sur le secteur de la télévision, tels que les passages du noir & blanc à la couleur et à la haute définition. Néanmoins, d'autres améliorations considérables peuvent encore être atteintes pour fournir une meilleure expérience multimédia, par exemple avec la ultra-haute définition, high dynamic range & wide color gamut ou la 3D. Ces technologies dites immersives visent à fournir de meilleures expériences, plus réalistes et plus fortes émotionnellement.

Pour mesurer la qualité d'expérience (QoE), les évaluations subjectives sont le moyen ultime car elles reposent sur un pool de sujets humains. Cependant, des résultats fiables et significatifs ne peuvent être obtenus que si les expériences sont correctement conçues et réalisées suivant une méthodologie rigoureuse. Dans cette thèse, nous construisons un cadre rigoureux pour l'évaluation subjective de nouveaux types de contenus image et vidéo. Nous proposons différentes procédures et outils d'analyse pour mesurer la QoE dans les technologies immersives.

Comme les technologies immersives capturent plus d'informations que les technologies conventionnelles, elles ont la capacité de fournir plus de détails, une amélioration de la perception de la profondeur, ainsi qu'un meilleur affichage des couleurs, du contraste et de la luminosité. Pour mesurer l'impact de ces technologies sur la QoE des téléspectateurs, nous appliquons le cadre proposé pour concevoir des expériences et pour analyser les scores recueillis auprès des sujets. Nous analysons également les mouvements oculaires afin d'étudier l'attention visuelle lors du visionnement de contenus immersifs.

Puisque les contenus immersifs capturent plus d'informations que les contenus conventionnels, des algorithmes de compression efficaces sont nécessaires pour le stockage et la transmission en utilisant les infrastructures existantes. Pour déterminer la bande passante requise pour la transmission de contenus immersifs de haute qualité, nous utilisons le cadre proposé pour effectuer des évaluations minutieuses de codecs image et vidéo récents dans le contexte des technologies immersives.

Les évaluations subjectives demandent beaucoup de temps, sont coûteuses et ne sont pas toujours réalisables. Par conséquent, les chercheurs ont développé des métriques objectives afin de prédire automatiquement la qualité. Pour mesurer la performance des métriques objectives à évaluer la qualité de contenus immersifs, nous effectuons plusieurs benchmarks détaillés de métriques objectives de pointe et couramment utilisées. Dans ce but, nous utilisons des scores de qualité de vérité terrain, recueillis à l'aide du cadre proposé pour les évaluations subjectives.

Afin d'améliorer la QoE, nous proposons différents systèmes pour les écrans 3D stéréoscopiques et autostéréoscopiques en particulier. Les systèmes proposés peuvent aider à réduire les artefacts générés lors de la visualisation, ce qui impacte la qualité d'image, la qualité de la profondeur et le confort visuel. Pour démontrer l'efficacité de ces systèmes, nous utilisons le cadre proposé pour mesurer la préférence des téléspectateurs entre ces systèmes et les modes 2D et 3D standards.
En résumé, cette thèse aborde les problèmes de la mesure, de la prédiction et de l'amélioration de la QoE dans les technologies immersives. Pour résoudre ces problèmes, nous construisons un cadre rigoureux et nous l'appliquons à travers plusieurs études approfondies. Nous avons mis des concepts essentiels de la QoE dans le multimédia sous ce cadre. Ces concepts ne sont pas seulement de nature fondamentale, mais ont aussi montré leur impact dans des applications très pratiques. En particulier, les organismes de normalisation JPEG, MPEG et VCEG ont adopté ces concepts pour sélectionner les technologies qui ont été proposées pour la normalisation et pour valider les normes qui en résultent, en termes d'efficacité de compression.

Mots clefs : qualité d'expérience, technologie vidéo immersive, ultra-haute définition, high dynamic range, 3D, assessment subjectif de la qualité, évaluation subjective, comparaison par paire, modèle de Thurstone, qualité visuelle, protocole d'évaluation, crowdsourcing, oculométrie, attention visuelle, métrique de qualité objective, assessment objectif de la qualité, évaluation de la performance, HEVC, VP9, JPEG, JPEG 2000, JPEG XT, efficacité de codage, modèle de Bjøntegaard, crosstalk, pseudoscopie, rivalité convergence-accommodation, écran stéréoscopique, écran autostéréoscopique multi-vues

Contents

Acknowledgements i

Abstract iii

Résumé v

List of Abbreviations xiii

1 Introduction 1
- 1.1 Immersive Video Technologies . . . 1
  - 1.1.1 Ultra High Definition . . . 2
  - 1.1.2 High Frame Rate . . . 4
  - 1.1.3 High Dynamic Range and Wide Color Gamut . . . 5
  - 1.1.4 3D . . . 9
- 1.2 Quality of Experience . . . 12
- 1.3 Contributions . . . 13
  - 1.3.1 Measuring Quality of Experience . . . 14
  - 1.3.2 Predicting Quality of Experience . . . 19
  - 1.3.3 Improving Quality of Experience . . . 21
- 1.4 Organization . . . 23

I Measuring Quality of Experience 25

2 Design and Analysis of Subjective Experiments 27
- 2.1 Viewing Conditions . . . 28
- 2.2 Test Material . . . 29
- 2.3 Subjects . . . 31
- 2.4 Test Methods . . . 31
  - 2.4.1 Single Stimulus and Absolute Category Rating . . . 32
  - 2.4.2 Double Stimulus Impairment Scale and Degradation Category Rating . . . 34
  - 2.4.3 Stimulus Comparison and Pair Comparison . . . 35
  - 2.4.4 Double Stimulus Continuous Quality Scale . . . 37
  - 2.4.5 Subjective Assessment Methodology for Video Quality . . . 38

2.4.6 Single Stimulus Continuous Quality Evaluation and Simultaneous Double Stimulus for Continuous Evaluation . . . 38
2.5 Test Design . . . 39
2.5.1 Training Session . . . 39
2.5.2 Test Session . . . 40
2.6 Data Processing . . . 40
2.6.1 Outlier Detection . . . 41
2.6.2 Mean Opinion Scores and Confidence Intervals . . . 41
2.6.3 The Bradley–Terry–Luce and Thurstone Case V Models . . . 43
2.6.4 Relationship Between Estimated Mean Values . . . 48
2.7 Comparing MOS Values of Different Experiments . . . 48
2.7.1 Mapping Subjective Scores of Two Experiments . . . 49
2.7.2 Statistical Evaluation Metrics . . . 50
2.7.3 Estimation Errors . . . 52
2.7.4 Classification Errors . . . 52
2.7.5 Comparing Paired Comparison Data of Different Experiments . . . 53
2.8 Analysis of Eye Tracking Data . . . 53
2.8.1 Computation of Fixation Density Maps . . . 54
2.8.2 Statistical Evaluation Metrics . . . 54
2.9 Conclusion . . . 55

3 Calculation of Coding Efficiency 57
3.1 The Bjøntegaard Model . . . 58
3.2 Extension for Two-Layer Coding Systems . . . 59
3.2.1 Proposed Model . . . 60
3.2.2 Applications and Discussions . . . 64
3.3 Extension for Calculation Based on Subjective Quality Scores . . . 66
3.3.1 Proposed Model . . . 67
3.3.2 Applications and Discussions . . . 75
3.4 Conclusion . . . 81

4 Performance Analysis of Image and Video Compression 83
4.1 Evaluation of HEVC Video Compression . . . 85
4.1.1 Dataset . . . 86
4.1.2 Methodology . . . 89
4.1.3 Results . . . 92
4.2 Evaluation of HEVC Image Compression . . . 95
4.2.1 Dataset . . . 96
4.2.2 Methodology . . . 97
4.2.3 Results . . . 99
4.3 Evaluation of VP9 Video Compression . . . 104
4.3.1 Dataset . . . 105
4.3.2 Methodology . . . 107

4.3.3 Results . . . 109
4.4 Evaluation of JPEG XT HDR Image Compression . . . 112
4.4.1 Dataset . . . 113
4.4.2 Methodology . . . 117
4.4.3 Results . . . 119
4.5 Towards HDR Extensions of HEVC . . . 122
4.5.1 Dataset . . . 123
4.5.2 Methodology . . . 124
4.5.3 Results . . . 127
4.6 Cross-lab Evaluation of MVC+D and 3D-AVC 3D Video Compression . . . 129
4.6.1 Dataset . . . 130
4.6.2 Methodology . . . 131
4.6.3 Results . . . 132
4.7 Conclusion . . . 137

5 Investigation of Alternative Evaluation Protocols 139
5.1 A Quality Assessment Protocol for Free-Viewpoint Video Sequences . . . 140
5.1.1 Dataset . . . 142
5.1.2 Methodology . . . 144
5.1.3 Results . . . 146
5.2 Crowd-based Evaluation of Multiview Video plus Depth Coding . . . 154
5.2.1 Dataset . . . 155
5.2.2 Methodology . . . 155
5.2.3 Results . . . 158
5.3 Crowdsourcing Evaluation of HDR Image Compression . . . 160
5.3.1 Dataset . . . 161
5.3.2 Methodology . . . 162
5.3.3 Results . . . 166
5.4 Conclusion . . . 172

6 Evaluation of Immersive Video Technologies 175
6.1 Sensation of Reality in 3DTV . . . 176
6.1.1 Dataset . . . 176
6.1.2 Methodology . . . 177
6.1.3 Results . . . 179
6.2 Evaluation of Higher Dynamic Range Video . . . 181
6.2.1 Dataset . . . 182
6.2.2 Methodology . . . 183
6.2.3 Results . . . 187
6.3 Conclusion . . . 194

7 Visual Attention in Immersive Video Technologies 197
7.1 Impact of Ultra High Definition on Visual Attention . . . 198
7.1.1 Dataset . . . 199
7.1.2 Methodology . . . 199
7.1.3 Results . . . 203
7.2 Visual Attention in LDR and HDR Images . . . 208
7.2.1 Dataset . . . 208
7.2.2 Methodology . . . 210
7.2.3 Results . . . 212
7.3 Conclusion . . . 214

II Predicting Quality of Experience 217

8 Objective Quality Metrics 219
8.1 Image Quality Metrics . . . 220
8.1.1 Full-Reference Metrics . . . 221
8.1.2 No-Reference Metrics . . . 223
8.2 Video Quality Metrics . . . 223
8.3 HDR Quality Metrics . . . 225
8.4 3D Quality Metrics . . . 226
8.5 Conclusion . . . 229

9 Procedures for Statistical Evaluation of Objective Quality Metrics 231
9.1 Mapping Objective Values to Subjective Data . . . 232
9.2 Performance Indexes . . . 233
9.2.1 Pearson Correlation Coefficient . . . 233
9.2.2 Spearman Rank Order Correlation Coefficient . . . 233
9.2.3 Root Mean Square Error . . . 234
9.2.4 Outlier Ratio . . . 234
9.3 Statistical Significance Evaluation . . . 234
9.3.1 Significance of the Difference between the Correlation Coefficients . . . 235
9.3.2 Significance of the Difference between the Root Mean Square Errors . . . 236
9.3.3 Significance of the Difference between the Outlier Ratios . . . 236
9.4 Resolving Power . . . 237
9.5 Classification Errors . . . 238
9.6 Conclusion . . . 240

10 Performance Evaluation of Objective Quality Metrics 241
10.1 Benchmarking of Objective Metrics on Asymmetric Stereo Pairs . . . 242
10.1.1 Methodology . . . 243
10.1.2 Results . . . 246
10.2 Benchmarking of Objective Metrics on Symmetric Stereo Pairs . . . 253
10.2.1 Methodology . . . 254

10.2.2 Results . . . 254
10.3 Benchmarking of Objective Metrics on Free-Viewpoint Video Sequences . . . 257
10.3.1 Methodology . . . 258
10.3.2 Results . . . 258
10.4 Benchmarking of Objective Metrics for HDR Image Quality Assessment . . . 266
10.4.1 Methodology . . . 267
10.4.2 Results . . . 269
10.5 Effectiveness of Objective Metrics for HDR Video Quality Assessment . . . 279
10.5.1 Methodology . . . 280
10.5.2 Results . . . 281
10.6 Conclusion . . . 284

11 Predicting 3D Quality based on Content Analysis 287
11.1 Proposed Model . . . 288
11.1.1 Feature Extraction . . . 288
11.1.2 Mapping Function . . . 289
11.1.3 Feature Selection . . . 289
11.2 Performance Evaluation . . . 290
11.2.1 Selected Features . . . 290
11.2.2 Anchors . . . 291
11.2.3 Results . . . 292
11.3 Conclusion . . . 292

III Improving Quality of Experience 295

12 Improving 3D Quality of Experience 297
12.1 Improving 3D QoE on Stereoscopic Displays . . . 299
12.1.1 System Description and Implementation . . . 300
12.1.2 Subjective Evaluation . . . 305
12.1.3 Results . . . 309
12.2 Improving 3D QoE on Mobile Autostereoscopic Displays . . . 311
12.2.1 Display Characterization . . . 312
12.2.2 System Description . . . 316
12.2.3 Subjective Evaluation . . . 321
12.2.4 Results . . . 323
12.3 Improving 3D QoE on Multiview Autostereoscopic Displays . . . 323
12.3.1 Display Characterization . . . 324
12.3.2 System Description . . . 328
12.3.3 Subjective Evaluation . . . 332
12.3.4 Results . . . 333
12.4 Conclusion . . . 334

13 Conclusion **337**
13.1 Technical Contributions 337
13.1.1 Measuring Quality of Experience 338
13.1.2 Predicting Quality of Experience 342
13.1.3 Improving Quality of Experience 343
13.2 Contributions to Reproducible Research 344
13.3 Outlook for Future Research 346

A Maximum Likelihood for Two Options **349**

Bibliography **351**

Curriculum Vitae **387**

List of Abbreviations

3D-AVC 3D video extension of AVC.
3D-HEVC 3D video extension of HEVC.
3DTV 3D television.

ACR absolute category rating.
ACR-HR absolute category rating with hidden reference.
ANOVA analysis of variance.
AVC H.264/MPEG-4 Part 10 Advanced Video Coding.

BD-PSNR Bjøntegaard delta PSNR.
BD-Rate Bjøntegaard delta rate.

CDF cumulative distribution function.
CfE Call for Evidence.
CfP Call for Proposals.
CI confidence interval.
CRT cathode ray tube.
CTC common test conditions.

DCR degradation category rating.
DIBR depth-image-based rendering.
DMOS differential mean opinion score.
DSCQS double stimulus continuous quality scale.
DSIS double stimulus impairment scale.
DSLR digital single-lens reflex.

EOTF electro-optical transfer function.

FC forced choice.
FDM fixation density map.
FOV field of view.
FR full-reference.
FTV free viewpoint television.
FVV free viewpoint video.

GOP group of pictures.

HD high-definition.
HDR high dynamic range.
HDTV high-definition television.
HEVC H.265/MPEG-H Part 2 High Efficiency Video Coding.
HFR high frame rate.
HM HEVC reference software.
HVS human visual system.

IPD interpupillary distance.
ITU International Telecommunication Union.

JCT-3V Joint Collaborative Team on 3D Video Coding.
JCT-VC Joint Collaborative Team on Video Coding.
JM AVC reference software.
JPEG Joint Photographic Experts Group.

KLD Kullback-Leibler divergence.

LCD liquid-crystal display.
LD low delay.
LDR low dynamic range.
LED light-emitting diode.

MAE mean absolute error.
MFC multi-resolution frame-compatible.
MOS mean opinion score.
MPEG Moving Picture Experts Group.
MSE mean squared error.
MV-HEVC multiview video coding extension of HEVC.
MVC multiview video coding (multiview video coding extension of AVC).
MVC+D MVC plus depth (depth enhanced extension of MVC).
MVD multi-view video plus depth.

NR no-reference.
NTSC National Television System Committee.

OQR objective quality rating.
OR outlier ratio.
OVD optimal viewing distance.

PC pair comparison.
PCA principal component analysis.
PCC Pearson correlation coefficient.
PDF probability density function.
PQ perceptual quantizer.

PSF point spread function.
PSNR peak signal-to-noise ratio.
PU perceptually uniform.

QFHD quad full high definition.
QoE quality of experience.
QoS quality of service.
QP quantization parameter.

R^2-D rate-rate-distortion.
R-D rate-distortion.
RA random access.
RMSE root-mean-square error.
RR reduced-reference.

SAMVIQ subjective assessment methodology for video quality.
SC stimulus comparison.
SCENIC subjective comparison of encoders based on fitted curves.
SDR standard dynamic range.
SDSCE simultaneous double stimulus for continuous evaluation.
SHVC scalability extensions of HEVC.
SI spatial perceptual information.
SROCC Spearman's rank correlation coefficient.
SS single stimulus.
SSCQE single stimulus continuous quality evaluation.
SSIM structural similarity.
SVC scalable video coding (scalability video coding extensions of AVC).

TI temporal perceptual information.
TMO tone mapping operator.

UHD ultra-high-definition.
UHDTV ultra-high-definition television.

VCEG Video Coding Experts Group.
VQEG Video Quality Experts Group.

WCG wide color gamut.

ZDP zero disparity plane.

1 Introduction

According to Sandvine Global Internet Phenomena Reports[1], real-time entertainment (audio and video) accounts for more than 45% and 70% of all downstream traffic during peak hours on fixed access networks in Europe and North America, respectively. In North American, video streaming has increased at a rapid pace to the point that real-time entertainment traffic doubled in five years. In Europe, YouTube (24.4%), BitTorrent (6.1%), and Netflix (4.8%) are the top three multimedia services consuming most of the bandwidth. In North America, Netflix accounts for over 37% of the bandwidth, whereas YouTube and Amazon Video are the second and third most demanding services with 17.8% and 3.1% of downstream traffic, respectively. The most popular video-on-demand service, i.e., Netflix, is relatively new in Europe, as it entered the market in 2012 and became available in some European countries only in 2013 or 2014, which explains the large difference. Regarding mobile network, real-time entertainment represents more than 35% and 40% of peak downstream traffic in Europe and North America, respectively. In both cases, YouTube is the most bandwidth consuming service with about 21% of downstream bandwidth. Considering that 4 billion videos are viewed on YouTube everyday, with 300 hours of additional video uploaded every minute, these figures demonstrate that we should pay special care to providing high quality multimedia services. Indeed, the quality of experience (QoE) provided by multimedia systems and services will greatly impact how much we will use and interact with these technologies. Considering the increasing popularity of immersive video technologies such as 3D, ultra-high-definition (UHD), and high dynamic range (HDR) thanks to the recent developments in capture, storage, compression, and display technologies and content availability, it is essential to conduct research on QoE in immersive video technologies.

1.1 Immersive Video Technologies

Over the last decades, several technological revolutions have impacted the television industry, such as the shifts from black & white to color and from standard to high-definition (HD).

[1] Sandvine Global Internet Phenomena Reports: available at https://www.sandvine.com

Nevertheless, further considerable improvements may still be achieved to provide a better multimedia experience and a better picture quality, for example with ultra-high-definition (UHD) (more pixels), high frame rate (HFR) (faster pixels), high dynamic range (HDR) & wide color gamut (WCG) (better pixels), or 3D (volumetric pixels). These so-called immersive video technologies aim at providing better, more realistic, and emotionally stronger experiences.

1.1.1 Ultra High Definition

Since the invention of television in the late 19th century, researchers have always been trying to increase resolution. Earlier broadcasting television systems that were based on mechanical systems had only about 30 lines of resolution. The first fully electronic television system, i.e., the Marconi-EMI 405-line system, was introduced with the BBC Television Service in 1936. This system offered an actual image resolution of 377 lines high, which was a big step over the best mechanical system, i.e., Baird 240-line sequential scan. However, the first electronic systems used interlacing, whereas mechanical systems were progressive. The US National Television System Committee (NTSC) 525-line system was introduced in 1941, whereas the French 819-line system, which was introduced in 1949, is often considered as the first high-definition (HD) television system with its 737 active lines.

The first color system was introduced in 1953 by the US NTSC and had a resolution of 525 lines for compatibility reasons with existing B&W systems. In Europe, the PAL and SECAM color systems were added to the monochrome 625-line broadcasts in the 1960s. The NTSC and PAL/SECAM had a 4:3 aspect ratio and actual image resolution of 480 and 576 lines, respectively, which is commonly referred to as standard definition.

The Japan Broadcasting Corporation, NHK, began conducting research to “unlock the fundamental mechanism of video and sound interactions with the five human senses” after the Tokyo Olympics, in 1964. In 1979, NHK developed the MUSE system, also marketed as Hi-Vision (a contraction of HIgh-definition teleVISION), a 1125-line standard, with 1035 active lines, 60 Hz refresh rate, and 5:3 aspect ratio. Based on this standard, work began on imaging systems, recording devices, transmission systems, and large-screen displays.

Since 1972, the International Telecommunication Union (ITU) tried to create a standard for high-definition television (HDTV). The efforts finally paid off in the 1980s, with the settlement on 16:9 aspect ratio, which was a compromise between the 5:3 format used in MUSE and the common 1.85 widescreen cinema format. Additionally, the first version of the ITU-R BT.709 (2015) recommendation was approved in 1990. This recommendation includes the 16:9 aspect ratio, a specified colorimetry, and the scan modes 1080i (interlaced) and 1080p (progressive).

NHK started to explore a next-generation television system for HDTV as early as 1995. They developed the first UHD system, nicknamed Super Hi-Vision, with 4000 scanning lines and a 22.2 channel multichannel sound system (Sugawara et al., 2003). The prototype was demonstrated in 2003 and used an array of 16 HDTV recorders with a total capacity of almost 3.5 TB,

which could capture only 18 min of test footage. The camera was built using four CCDs (two for green and one each for red and blue), each with a resolution of 3840 × 2048 pixels, to reach the resolution of 7680 × 4320 pixels.

The Society of Motion Picture and Television Engineers (SMPTE) first released the ST 2036 Standard series for ultra-high-definition television (UHDTV) in 2007 and included two levels: UHDTV1 (3840 × 2160 or 4K UHDTV) and UHDTV2 (7680 × 4320 or 8K UHDTV). The ITU recommendation ITU-R BT.2020 (2015) was published in 2012 and is the equivalent of ITU-R BT.709 (2015) for UHDTV. This recommendation specifies the picture spatial and temporal characteristics, system colorimetry, signal format, and digital representation for new TV systems, including displays.

As the angular resolution of the human visual system (HVS) is fixed (at about 1 arcmin for normal vision), higher image resolution increases the field of view (FOV). Standard definition only offered a FOV of 11° to 13° and full HD resolution corresponds to 31° FOV. However, the FOV is increased to 58° and even 96° with 4K and 8K resolutions, respectively. To investigate the impact of FOV on viewers, Emoto et al. (2006) conducted a subjective experiment with still images acquired with a FOV of 60° and 100°. The images were presented at different resolutions using a UHD projector, which resulted in FOV ranging from 30° to 100°. They used a Likert scale to evaluate 'presence', 'powerfulness', 'comfortableness', and 'depth'. While the results for 'comfortableness' and 'depth' had a tendency to saturate as the FOV increased, 'presence' and 'powerfulness' almost monotonously increased as the FOV increased.

The 4K and 8K UHD resolutions contain 4 and 16 times the number of pixels of HD resolution, respectively. Hence, the increase in resolution is at the cost of the amount of data that has to be transmitted. If the same video coding format is used for 4K and 8K UHDTV as for HDTV, the bandwidth capacity must be increased to preserve the same visual quality. Terrestrial broadcasting typically uses a bit rate of 18 Mbit/s to carry the audiovisual data, whereas Blu-ray disks have a maximum data transfer rate of 54 Mbit/s for both audio and video data.

The H.262/MPEG-2 Part 2 video compression standard, which was standardized in 1996, is still widely used for video broadcasting, even for HDTV. Its successor, i.e., H.264/MPEG-4 Part 10 Advanced Video Coding (AVC) (Wiegand et al., 2003a), showed a 50% bit rate reduction for the same visual quality (Oelbaum et al., 2004). The latest standard developed by the Joint Collaborative Team on Video Coding (JCT-VC), named H.265/MPEG-H Part 2 High Efficiency Video Coding (HEVC) (Sullivan et al., 2012), also shows a 50% bit rate reduction over AVC (Weerakkody et al., 2014). The performance of HEVC is mainly due to better flexibility and adaptability, which is achieved with a larger block size (up to 64 × 64 for inter-frame coding and up to 32 × 32 for intra-frame coding) when compared to previous standards (up to 16 × 16 for inter-frame coding and 4 × 4 or 8 × 8 for intra-frame coding). Thus, HEVC is a perfect candidate for UHD video compression. Note that HEVC supports resolutions up to 8192 × 4320 pixels.

1.1.2 High Frame Rate

The HVS can perceive 10 to 12 frames per second (fps) as individual images. However, beyond this limit, persistence of vision may create an illusion of continuity and the impression of motion may be perceived from a sequence of still images. Early silent films had frame rates between 14 and 26 fps, but the motion was often perceived as jerky or uneven as the film was hand-cranked while recording. Moreover, during playback, the film was also often hand-cranked or played at a different (typically higher) speed by the projection system.

With the introduction of sound film in 1926, where sound was inserted as an optical track on the filmstrip alongside the image, variations in film speed were no longer tolerated, as humans are more sensitive to changes in audio frequency. Since film is an expensive medium, the movie industry settled for the slowest frame rate possible for producing intelligible sound, i.e., 24 fps for 35 mm sound film.

With the advent of television, new frame rates were introduced for broadcasting. Indeed, the first TV units used cathode ray tube (CRT) displays, which required a refresh rate at (multiple of) AC line frequency. In particular, in the Americas and parts of Asia, the AC line frequency is 60 Hz, whereas 50 Hz is used in large parts of the world. This lead to the adaption of 30 fps, interlaced (60i), for NTSC formats and 25 fps, interlaced (50i), for PAL and SECAM formats. With interlaced scan, two video fields are flashed one after the other to make up one frame. This format was used to double the perceived frame rate, which improves motion and reduces flicker, without the need to increase bandwidth. However, current display technologies, e.g., liquid-crystal display (LCD), do not require to refresh the pixels anymore. These displays use progressive scan, where each frame is scanned sequentially in its entirety.

Thanks to the development in camera technology, higher frame rates have appeared, e.g., 48, 50, 60, 72, 100, 120, and 240 fps. The Hobbit film series from Peter Jackson was shot in 3D at 48 fps and screened in this format in selected theaters starting from December 2012. Recommendation ITU-R BT.2020 (2015) specifies frame rates of 100 and 120 fps (among others) for UHDTV, which shows that UHD and high frame rate (HFR) are profoundly linked. BBC Research also made some tests with frame rates up to 300 fps (Armstrong et al., 2009), which can be easily down-converted to 50 and 60 fps for compatibility with existing standards. Note that very high speed cameras with frame rates of 1000 fps and higher exist, but the captured video cannot be played back in real time.

HFR reduces motion blur and allows to display a clearer image, which can be particularly beneficial for fast motion content such as sport. Emoto et al. (2014) investigated the degree of improvement in video sequences recorded and displayed at different frame rates. They used 12 HD video sequences (mostly sport content) recorded and displayed at 60, 120, and 240 fps. Results showed that the improvement from 60 to 120 fps (0.46 on a five point scale) was higher than from 120 to 240 fps (0.23). The improvements were content dependent and varied from 0 to 1.4. Moreover, the authors observed a bandpass type relationship between the angular velocity and the degree of improvement. This relationship can be due to many factors, e.g.,

accumulation of blur in filming, response time of LCD, motion blur caused by eye movements, and visual characteristics in pursuit and saccadic eye movements.

Increasing frame rate also increases the amount of data that has to be stored and transmitted. However, motion between successive frames should be reduced with higher frame rates, which means that a better temporal estimation could be achieved, also as the rigidity and constant luminosity constraints are more likely to be met. Thus, the necessary bit rate is most likely not going to be proportional to the frame rate. HEVC introduced a better signaling of the motion information , which can be also beneficial for HFR content. Note that the maximum frame rate supported by HEVC is 300 fps.

1.1.3 High Dynamic Range and Wide Color Gamut

An important part of our impressions and understanding about our surroundings are based on sight. The HVS is capable of adapting to lighting conditions that span about ten orders of magnitude (Ferwerda, 2001). The HVS can take up to 20 min to adapt from sunlight (typically 10^3 cd/m^2) to starlight (typically 10^{-3} cd/m^2). However, once the HVS is adapted to a scene, it functions over a range of about five orders of magnitude simultaneously (Reinhard et al., 2005).

Since the beginning of photography, people have been trying to capture representations of a scene that are as close as possible as what the HVS can see. However, the first photographic equipment had very poor light sensitivity and required a rather long exposure time to produce a result that captured only a black & white image with limited contrast. Through the years, there has been a lot of progress in photographic films to capture color images and with a wide contrast (for example, 8,000:1 for the Kodak VISION3 film). Nowadays, high-end digital single-lens reflexs (DSLRs) cameras, e.g., the Nikon D810, can capture about 14.8 stops, also referred to as exposure values (EVs), which corresponds to 28,500:1 contrast ratios.

To capture a dynamic range wider than that of the camera, the idea is to capture multiple shots of the same scene with different exposure times and to recompose these shots into a single image considering the transfer characteristics of the camera (Debevec and Malik, 2008; Mann and Picard, 1995; Mitsunaga and Nayar, 1999; Robertson et al., 1999). This technique is called *exposure bracketing* and was pioneered by Gustave Le Gray in the 1850s to render seascapes showing both the sky and the sea. Le Gray used two negatives, one for the sky and another one with a longer exposure for the sea, and combined them into one picture. DSLR cameras offer an auto exposure bracketing mode, which typically takes 3, 5, or even 7 shots of the same scene with different exposure times. The main problem with this technique is motion between successive pictures, which results in blur in the composed image. For this reason, a tripod is largely recommended, even if image alignment can be performed in post-processing. The first mathematical theory to construct a HDR image with luminance values from differently exposed pictures acquired via bracketing was proposed by Mann and Picard (1995). Note that nowadays, HDR images can also be acquired using specific image sensors.

A certain time after the development of new imaging technologies, the capabilities of capture system frequently exceeded that of reproduction systems. Regarding photography, the dynamic range captured by negative films is significantly wider than what can be reproduced by positive paper prints. Thus, manual tone mapping was applied during the development process to selectively increase or decrease the exposure of specific regions of the photograph. This process is called dodging and burning and is used to generate a better tonality reproduction. The American photographer Ansel Adams played a lot with this technique and proposed the zone system, a technique for determining optimal film exposure and development (Adams, 1980; Adams, 1981; Adams, 1983). The zone system is based on 11 zones, from 0 to 10, with 0 representing pure black, 5 middle grey, and 10 pure white. Zones 1 to 9 are recommended to represent the darkest and lightest "useful" negative densities, while zones 2 to 8 are meant to convey a sense of texture and the recognition of substance.

Conventional low dynamic range (LDR) displays, e.g., CRT and LCD, can best reproduce a range of luminance values from 1 to 100 cd/m^2, i.e., they can cover two orders of magnitude. Thus, tone mapping operators (TMOs) have been designed to map HDR content into the luminance range and color gamut of conventional LDR displays. Tone mapping is either applied locally or globally over the whole picture. The work from Oppenheim et al. (1968) is the first attempt at tone reproduction in computer graphics. The authors suggested a method for simultaneously reducing dynamic range and enhancing contrast using homomorphic filtering, thus proposing a local operator. Later, several TMOs were proposed (Devlin, 2002). In particular, the TMO proposed by Reinhard et al. (2002) is based on the zone system from Ansel Adams. Several TMOs have also been proposed for HDR video sequences, though a major issue is the temporal coherence of the tone-mapped video sequence (Aydin et al., 2014; Eilertsen et al., 2013). Finally, inverse TMOs have also been developed to expand the dynamic range of LDR content to display legacy content on new HDR monitors (Banterle et al., 2009). However, Akyüz et al. (2007) have found that simply linearly boosting the dynamic range of an LDR image can be preferred over a true HDR image.

Traditional LCD display use a uniform backlight source, e.g., a series of cold cathode fluorescent lamps or an array of white or colored light-emitting diodes (LEDs). The front LCD panel then modulates the backlight to produce the desired light. To create the first HDR display, Seetzen et al. (2003) had the idea of using an active matrix array of ultra high brightness white LEDs for the backlight. The LEDs were driven individually to control the local luminance. As the sampling of the LED backlight is sparer than the definition of the front LCD panel and because the point spread function (PSF) of one LED leaks over neighboring pixels, a compensation has to be applied in the LCD panel to compensate the light leakage to obtain the desired luminance. Therefore, a dual modulation between the LED backlight and front LCD was proposed by Seetzen et al. to correct for the low resolution backlight through compensation in the high resolution LCD panel. The HDR display developed by Seetzen et al. was capable of displaying a luminance range from 0.1 cd/m^2 up to 10000 cd/m^2, i.e., five orders of magnitude, while maintaining the resolution, refresh rate, and image quality found in conventional LCD displays. Seetzen et al. (2004) have also proposed another design based on a video projector

instead of an active LED array for the backlight. This design uses a Fresnel lens and a diffuser to collimate the projected light into a narrow viewing angle for maximum brightness and to avoid color distortion due to diverging light passing through the color filters of the LCD (Seetzen et al., 2004). This system can be constructed with off-the-shelf components. More details regarding the signal processing to create the dual modulation signals can be found in (Seetzen et al., 2004).

Gamma encoding, which was originally developed to compensate for the characteristics of CRT displays, relies on a power law (typically with an exponent between 1.8 and 2.6) electro-optical transfer function (EOTF) to map code values to luminance values to optimize quantization when encoding an image (Poynton, 2012). Under common illumination conditions, the HVS is more sensitive to relative differences between darker than brighter tones. According to Weber's law, the HVS sensitivity approximately follows a logarithm function at high luminance values (Shevell, 2003). However, at the darkest levels, the HVS sensitivity is closer to a square-root behavior, according to Rose–DeVries law (De Vries, 1943; Rose, 1948). Thus, gamma encoding is not optimized for encoding of dark and bright luminance values, as the shape of the EOTF should be adjusted to take into account the Rose–DeVries and Weber laws. For this purpose, Miller et al. (2013) have proposed an new EOTF for HDR content, named perceptual quantizer (PQ), which is derived from the Barten contrast sensitivity function (Barten, 1999). The PQ curve has a square-root and log behavior at the darkest and brightest light levels, respectively, while it exhibits a slope similar to the gamma non-linearities between those extreme luminance regions. The PQ EOTF was approved as SMPTE Standard 2084 and is used in the HEVC HDR10 profile, which is one of the current HDR formats accepted by HDR TV sets. Recommendation ITU-R BT.2020 (2015) specifies a digital representation of 10 or 12 bits per component, which is beneficial for HDR coding (using the PQ EOTF or another transform). This specification shows that UHD and HDR are profoundly linked.

The red, green, and blue primaries of a monitor define the color gamut that can be rendered. The color primaries specified in recommendation ITU-R BT.709 (2015) were defined considering the phosphorus capabilities of the CRT technology. However, the resulting color gamut only covers 33.51% of the visible light that the HVS can perceive (Shevell, 2003). To be able to render a larger portion of the human gamut, new color primaries must be used to obtain a wide color gamut (WCG) display. The color gamut of typical LCD monitors can be extended by using a LED backlight with red, green, and blue LEDs (Kakinuma et al., 2007; Sugiura et al., 2003). The color gamut can be further extended using semiconductor lasers to generate the three primary colors (Someya et al., 2006). This technology was considered to design the colorimetry specifications in recommendation ITU-R BT.2020 (2015) (Masaoka et al., 2010). This specification shows that UHD and WCG are profoundly linked too.

To conduct research on HDR imaging, the only available HDR monitor on the market is the Sim2 HDR47E S 4K monitor, which can reproduce luminance levels from 0.001 cd/m^2 to 4000 cd/m^2, i.e., about 6.6 orders of magnitude. However, a major problem with this display is its color reproduction fidelity, but an appropriate display characterization and pre-processing

can be applied to provide a more accurate color reproduction (J. Liu et al., 2015b). For the consumer market, several HDR&WCG TV sets with 4K UHD resolution have been released in 2015 and at the Consumer Electronics Show in Las Vegas in January 2016, e.g., the Samsung JS9500 Series (1000 cd/m^2 peak luminance and 240 active LED zones) and Vizio Reference Series (800 cd/m^2 peak luminance and 384 active LED zones), both using the Quantum Dot technology, or the Panasonic DX900 Series (1000 cd/m^2 peak luminance and 512 active LED zones), Philips 9000 Series (1000 cd/m^2 peak luminance and 256 active LED zones), Sony X940C/X930C Series, and LG G6 Signature Series (which uses the OLED technology).

An important question for HDR displays is how much dynamic range is needed. Should it be the range of luminance values that can be perceived by the HVS? Probably not, as this would not be feasible from a technical point of view, especially to render a strong sunlight. Most consumer-grade HDR monitors are mainly characterized by their peak luminance. However, HDR is not only about brightness. In particular, the black level, or in other words, the contrast ratio, is as important as the peak white level. Seetzen et al. (2006) investigated the impact of peak luminance and contrast ratio on viewers preferences for peak luminance levels ranging from 400 cd/m^2 to 6400 cd/m^2 and contrast ratios ranging from 2,500:1 to 10,000:1. They found that the optimal contrast ratio increases logarithmically with peak luminance. For an appropriate contrast ratio, they found that image quality also increases logarithmically with peak luminance. However, above 6000-7000 cd/m^2, image quality started to decrease, but this effect might be due to discomfort considering the ambient light conditions. A similar study was conducted by Daly et al. (2013) on a custom built HDR display with a peak luminance of 20000 cd/m^2 and a 5,000,000:1 contrast ratio. They found that for diffuse reflective regions, $[0.1, 650]$ cd/m^2 match the average preferences, whereas $[0.005, 3000]$ cd/m^2 is required to satisfy 90% of the viewers. However, for specular highlights and emissive sources, 2500 cd/m^2 peak luminance is sufficient to match the average preferences, whereas over 20000 cd/m^2 is necessary to satisfy 90% of the viewers.

Both JPEG 2000 (Schelkens et al., 2009; Skodras et al., 2001) and JPEG XR (Dufaux et al., 2009) standards can represent HDR images when used in combination with an appropriate pixel encoding, such as logLuv (Pattanaik and Hughes, 2005; Ward, 1998) or perceptual quantization (Mantiuk et al., 2004; Miller et al., 2013), as they support higher bit-depth. These two standards can also be used to encode directly HDR images in floating point representation, though with less efficiency. Nevertheless, those standards have not been adopted by the digital photography market. As JPEG is currently *de facto* the most popular imaging format, it is believed that an HDR image coding format should be backward compatible with the legacy JPEG format to facilitate its adoption and inclusion in current imaging ecosystems.

First attempts to design a coding system for HDR still images that would also provide backward compatibility were made by Spaulding et al. (2003) and Ward and Simmons (2006). The latter, known as JPEG-HDR, also proposed a software implementation which made it popular for compression of HDR images among some HDR enthusiasts. Minor limitations of that format were the lack of support for WCG and lack of lossless coding. To overcome the lack of a

standard for compression of HDR images that is backward compatible with JPEG format, the Joint Photographic Experts Group (JPEG) Committee created the JPEG XT standard (Artusi et al., 2015). Using this compression standard, HDR images are coded in two layers. A tone-mapped version of the HDR image is encoded using the legacy JPEG format in a base layer, and the extra HDR information is encoded in a residual layer.

Regarding compression of HDR video sequences, backward-compatible compression methods that decompose an HDR video stream into a residual stream and a standard LDR stream have also been proposed (C. Lee and C.-S. Kim, 2008; Mantiuk et al., 2006b). Additionally, Mantiuk et al. (2004) have proposed an extension of MPEG-4 to accommodate HDR video content. Similarly to LogLuv encoding, the algorithm uses an 11-bit perceptually uniform representation for the luma channel and 8-bit for the chroma channels. Garbas and Thoma (2011) have proposed a similar method with 12-bit for the luma channel. However, none of these algorithms have been used in real applications. Recognizing the rise of HDR applications and the lack of a corresponding video coding standard, the Moving Picture Experts Group (MPEG) released in February 2015 a Call for Evidence (CfE) for HDR and WCG video coding (N15083). The purpose of this CfE was to explore whether the coding efficiency and/or the functionality of HEVC Main 10 and Scalable Main 10 profiles can be significantly improved for HDR and WCG content. The results showed that visual quality can be noticeably improved and efforts towards the development of HDR/WCG extensions of HEVC were initiated.

1.1.4 3D

3D can be considered as the oldest immersive video technologies, as the first stereoscopic device, i.e., the stereoscope, was developed in 1838 by Sir Charles Wheatstone. Early attempts were made to show stereo footage in 1915 using anaglyph glasses. Later in the 1950s, many 3D movies were produced by cinema industry as a reaction to the invention of television. Even if this period is called the “golden era” of 3D, the added value was not sufficient to overcome the quality degradations when compared to 2D, which would explain why it did not successfully break through. It is only recently, that 3D seems to have become increasingly successful. 3D reached its climax in 2009 with Avatar, the highest-grossing movie of all time.

Currently, two main technologies are considered for stereoscopic displays to separate the left- and right-eye images (Urey et al., 2011). The first one is passive and relies on light polarization using filters mounted on the glasses to separate the two images. Circular polarization (left-/right handedness) is the most common. Linear polarization is also used, in particular in IMAX theaters (because of patent issues), but crosstalk, i.e., imperfect separation that causes a small proportion of one eye image to be seen by the other eye as well, starts to appear when you lean your head because of the imperfect alignment between the screen and your glasses. With this technology, projection systems require a silver screen, which can reflect light while preserving its polarization. Passive stereoscopic displays rely on a line-interleaved based set of filters, where odd lines use one polarization, whereas the even lines use the other one. The second

technology is active and relies on time multiplexing of the left and right images. It requires glasses equipped with two tiny LCDs instead of the lenses, which are synchronized with the display. With both technologies, the amount of light perceived by each eye is reduce by more than two, but the spatial resolution of active systems is twice better than that of passive systems. Active stereoscopic systems require at least 120 Hz refresh rate, so temporal resolution is usually not a problem. The amount of perceived crosstalk is similar in both technologies as long as the eyes are aligned vertically with the center of the display with passive glasses, but active glasses create flicker, which is mainly influenced by the video content and lighting conditions (Andrén et al., 2012).

Current stereoscopic technologies still require the user to wear bulky glasses. This factor has a significant impact on QoE, especially for users who are already wearing glasses. Autostereoscopic displays can be the solution to this problem. Two-view autostereoscopic displays are the most common types of glasses-free displays. They use either a parallax barrier (Benzie et al., 2007) or a lenticular sheet (Urey et al., 2011) to separate the two views. In the first system, the left and right views are column interlaced. The parallax barrier, defined as a set of vertical apertures placed in front of the screen, allows light to pass only to the desired viewing zone. In the second system, the views are also column interlaced, and a lenticular sheet, i.e., a set of vertical lenses placed in front of the screen, redirects the light to different viewing zones.

Multiview autostereoscopic displays have been developed to allow several users to enjoy 3D at the same time. These displays mimic reality by offering different viewing angles. However, current autostereoscopic display devices suffer from large quantities of crosstalk. Nevertheless, it has the advantage of providing a smooth transition between the different views when moving around the display, which is used to provide a good motion parallax depth cue. However, to enjoy a quality 3D experience, viewers should sit in specific positions relatively to the display, called sweet spots, where the amount of crosstalk is limited and the left and right views are projected to the left and right eyes, respectively. The most common technology uses a slanted lenticular sheet placed on the top of a regular screen (Urey et al., 2011). Each lens covers several pixels horizontally such that the different views are projected to different locations. The slanted system helps to reduce the “picket fence” effect and provides better transition between two adjacent views (Benzie et al., 2007). Most displays have between 5 and 9 views, some may have more than 25. The higher the number of views, the more natural is the motion parallax (Nam et al., 2011). However, the higher the number of views, the lower the resolution of each view since they are spatially multiplexed and the number of pixels on current displays is limited. Even though this technology is not yet mature enough for a wide acceptance in the consumer market, it shows promising results.

Several formats have been proposed for 3D content, e.g., stereoscopic, multiview, 2D-plus-depth (2D+Z), and multi-view video plus depth (MVD) (Smolic et al., 2009a; Vetro et al., 2008). The different formats have different characteristics and each application works best with a particular format. Stereoscopic (left and right) is the “easiest” 3D format, but it also provides very little capabilities. However, this format is the most common and is used for 3D movies,

3D Blu-rays, 3D broadcasting, and current stereoscopic 3D displays.

The multiview format consists of two or more views of the same scene. This format can be used for interactive system that allow the viewer to change the viewpoint to have a look around capability (Huang et al., 2012; Maugey and Frossard, 2011; Maugey and Frossard, 2013; Maugey et al., 2013; Toni et al., 2013). Additionally, the scene can be seen either as monoscopic or stereoscopic. To efficiently encode multiview data, several prediction structures have been proposed to take into account the spatial redundancy between the different views (Khattak et al., 2012; Khattak et al., 2013; Merkle et al., 2006; Merkle et al., 2007c). Multiview extension of AVC has been standardized under the name multiview video coding (multiview video coding extension of AVC) (MVC) (Y. Chen et al., 2009; Vetro et al., 2011), as well as multiview video coding extension of HEVC (MV-HEVC) (Muller et al., 2013; Sullivan et al., 2013).

The 2D-plus-depth format offers the ability to synthesize additional views, for example using depth-image-based rendering (DIBR) (Fehn, 2004a). Since only one view is provided originally, the additional views have to be extrapolated and the missing information, which was occluded in the available view, has to be filled using the available neighboring information. Hence, the view synthesis capabilities with this format are rather limited to a narrow range around the original viewpoint without creating too much artifacts. 2D-plus-depth coding was standardized by MPEG in the MPEG-C Part 3 specification (Daribo et al., 2008).

The MVD format can be seen as the ultimate 3D format, as the other formats discussed above can be considered as subsets of this format. Thanks to the multiple views, the interpolation of virtual views in between two existing views (Smolic et al., 2008) will result in better visual quality than with the 2D-plus-depth format, as the occluded information in one view is visible in the other view. This possibility opens the door to several applications (Kauff et al., 2007; Muller et al., 2008). For example, while watching 3D content on a stereoscopic display, the depth perception can be adjusted by synthesizing a new stereo pair to cope with different viewing preferences, viewing distances, and screen sizes (D. Kim et al., 2011). In multiview autostereoscopic monitors, with the MVD format, the N displayed views can be interpolated from a limited set of $M \ll N$ input views (Smolic et al., 2008; Smolic et al., 2009b; Vetro et al., 2008), as the capture capabilities are limited. Free viewpoint television (FTV) and free viewpoint video (FVV) rely on the MVD format to provide the ability to change the viewpoint and view direction (Smolic et al., 2004; Smolic and Kauff, 2005; Smolic et al., 2006; Smolic, 2011; Tanimoto, 2006). To efficiently encode MVD content, several prediction structures have been proposed to take into account the spatial redundancy between the different views and depth maps, and to use view synthesis prediction from the already coded views and depth maps (Merkle et al., 2007a; Merkle et al., 2007b; Muller et al., 2009). To encode MVD content, extensions of AVC and HEVC have been standardized under the names 3D video extension of AVC (3D-AVC) (Y. Chen and Vetro, 2014) and 3D video extension of HEVC (3D-HEVC) (Muller et al., 2013; Sullivan et al., 2013), as well as extension of MVC, which is referred to as MVC plus depth (depth enhanced extension of MVC) (MVC+D) (Y. Chen et al., 2014).

1.2 Quality of Experience

For many years, the quality assessment of multimedia systems and services was focused on their fidelity and ability to satisfy a set of requirements. In the multimedia field, quality assessment was performed by measuring the quality of service (QoS) of a particular multimedia system. For ITU, QoS is defined as the "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" (ITU-T E.800, 2008). Thus, the definition of QoS is very much focused on telecommunications services. However, since the end of 1990s, the notion of quality of experience (QoE) has gained popularity in different contexts. In particular, regarding communication, the concept of QoS was perceived as not sufficient enough to represent the different aspects of modern communication systems, which are more engaging, more interactive, more user-centered, etc.

The European Network on Quality of Experience in Multimedia Systems and Services, Qualinet (COST Action IC 10032), was initiated in 2011 for a duration of four years. One of the main topic in Qualinet was the discussion and definition of the term QoE and related concepts. One of the major outcome of Qualinet is a White Paper on definitions of QoE, which gives the following definition: "QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state." (Le Callet et al., 2013).

From the Qualinet definition of QoE, three major factors influencing QoE can be identified: human, system, and context factors (Le Callet et al., 2013). The human factors are related to the demographic and socio-economic background of the user, as well as his/her physical and mental constitution and emotional state. The systems factors are related to the technical properties and characteristics that determine the quality produced by an application or a service. In the case of multimedia communication, this include all the aspects related to content, media, network, and device. The context factors are related to the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics.

The notion of QoS is centered around network performance and systems-level parameters, whereas the notion of QoE has a wider scope and is more user-centric. QoE is a multi-dimensional and multi-modal notion that includes important factors such as user characteristics and context of usage. However, these aspects are not considered in the ITU definition of QoS. QoS considers the performance aspects of physical systems, whereas QoE considers the users' assessment of the overall systems performance, which can be influenced by many factors, e.g., context, culture, users' expectations, socio-economic issues, or psychological profiles. The assessment of QoS is very technology-oriented and relies on analytic approaches and empirical or simulative measurements. However, the assessment of QoE requires a multi-disciplinary and multi-methodological approach for its understanding. Nevertheless, QoS and QoE are not opposite notions. On the contrary, QoS can be seen as a subset of QoE and, in many cases, QoE is highly dependent on QoS. Regarding multimedia systems, the technical

aspects can have a significant impact on some dimensions of QoE (Fiedler et al., 2010).

For more details about the many aspects of QoE and different applications of QoE, e.g., multimedia, web browsing, gaming, recognition tasks, or human-computer interaction, the reader is recommended to have a look at the excellent book edited by Möller and Raake (2014).

1.3 Contributions

To measure QoE, subjective evaluation is the ultimate means since it relies on a pool of human subjects. However, reliable and meaningful results can only be obtained if experiments are properly designed and conducted following a strict methodology. In this thesis, we build a rigorous framework for subjective evaluation of new types of image and video content. We propose different procedures and analysis tools for measuring QoE in immersive technologies.

As immersive technologies capture more information than conventional technologies, they have the ability to provide more details, enhanced depth perception, as well as better color, contrast, and brightness. To measure the impact of immersive technologies on the viewers' QoE, we apply the proposed framework for designing experiments and analyzing collected subjects' ratings. We also analyze eye movements to study human visual attention during immersive content playback.

Since immersive content carries more information than conventional content, efficient compression algorithms are needed for storage and transmission using existing infrastructures. To determine the required bandwidth for high-quality transmission of immersive content, we use the proposed framework to conduct meticulous evaluations of recent image and video codecs in the context of immersive technologies.

Subjective evaluation is time consuming, expensive, and is not always feasible. Consequently, researchers have developed objective metrics to automatically predict quality. To measure the performance of objective metrics in assessing immersive content quality, we perform several in-depth benchmarks of state-of-the-art and commonly used objective metrics. For this aim, we use ground truth quality scores, which are collected under our subjective evaluation framework.

To improve QoE, we propose different systems for stereoscopic and autostereoscopic 3D displays in particular. The proposed systems can help reducing the artifacts generated at the visualization stage, which impact picture quality, depth quality, and visual comfort. To demonstrate the effectiveness of these systems, we use the proposed framework to measure viewers' preference between these systems and standard 2D & 3D modes.

In summary, this thesis tackles the problems of measuring, predicting, and improving QoE in immersive technologies. The following subsections describe in details the contributions of the thesis in each of these categories.

1.3.1 Measuring Quality of Experience

To compress an image or video sequence and reduce its file size, compression algorithms typically try to exploit correlation, e.g., spatial and temporal, in the data, for example to predict the current frame from previously encoded frames. Additionally, properties of the HVS are exploited to further reduce the amount of data, for example to adaptively quantize the data. To efficiently compress images and video sequences, i.e., reduce the number of bits used for their representation, lossy processes must be used, which might result in visible quality degradation. Therefore, the visual quality of compressed images and video sequence need to be assessed to determine the range of compression ratio values where acceptable quality can be achieved. As humans are ultimately the end-users of multimedia applications, the coding efficiency of different compression algorithms is best compared by means of subjective quality evaluations, carried out according to common evaluation methodologies defined by experts.

To evaluate the performance of image and video compression, a direct scaling of the different algorithms under study is typically measured using category scaling or magnitude estimation. However, for other scenarios such as the evaluation of different rendering techniques or new display technologies, a direct scaling is often impossible or would introduce too much bias. In this cases, the pair comparison (PC) method is more appropriate has it is similar to the process followed by humans when they have to decide between two products. However, the results of PC experiments are harder to interpret. Thus, PC results are often converted to mean opinion score (MOS) equivalent results, as obtained with direct scaling, using statistical models, e.g., the Bradley-Terry-Luce (Bradley and Terry, 1952; Luce, 1959) and Thurstone Case V models (Thurstone, 1927). Considering that ties convey information about significant differences between two stimuli being compared, we proposed an extension of the Thurstone Case V model to estimate confidence interval (CI) from PC experiments conducted with a ternary scale (Hanhart et al., 2014b).

Comparing results of two subjective experiments conducted with the same test material but with different conditions is essential. One goal can be to investigate the influence of different factors, e.g., viewing distance, lighting conditions, display, test methodology, or rating scale. For this task, it is recommended to compute four statistical evaluation metrics to estimate the linearity, monotonicity, accuracy, and consistency between two groups of MOS values corresponding to two different experiments. However, such a simple analysis may not be sufficient to investigate the possible difference between two experiments. Thus, we proposed new methods to compare MOS values of different experiments (Hanhart and Ebrahimi, 2013c), which were inspired from recommended methods for benchmarking of objective metrics.

To calculate the coding efficiency between two compression algorithms, the Bjøntegaard model (Bjøntegaard, 2001) is commonly used to compute the average peak signal-to-noise ratio (PSNR) and bit rate differences between two rate-distortion (R-D) curves obtained from the PSNR measurement when encoding a content at different bit rates. However, this model considers only one bit rate and thus cannot be used to investigate the impact on quality of the

interaction of the base and enhancement layers bit rates when comparing two-layer coding systems. Therefore, we proposed an extension of the Bjøntegaard model from R-D curve fitting to rate-rate-distortion (R^2-D) surface fitting (Hanhart and Ebrahimi, 2015). Additionally, the Bjøntegaard model might not be an accurate predictor of the true coding efficiency as it relies on PSNR measurements. To estimate a more realistic coding efficiency, subjective quality scores should be considered instead of PSNR measurements. Thus, we proposed a model to calculate the average coding efficiency based on MOSs gathered during subjective evaluations instead of PSNR measurements (Hanhart and Ebrahimi, 2014a).

Several standardization bodies, e.g., MPEG, Video Coding Experts Group (VCEG), and JPEG, are at the roots of the still and moving pictures coding formats used over the past 30 years. When a new activity is initiated by standardization bodies, evidence must be brought to show potential value for a new coding format or extension of an existing coding format. In the past, the standardization bodies have always relied on subjective quality evaluations to prove that considerable coding gains, e.g., 50% bit rate reduction for the same visual quality, can be achieved. Alternatively, standardization activities are also initiated when there is a lack of standard, e.g., for new applications or new image/video formats. In particular, recognizing the rise of HDR applications and the lack of a corresponding video coding standard, MPEG released in February 2015 a CfE for HDR and WCG video coding (N15083). The purpose of this CfE was to explore whether the coding efficiency and/or the functionality of HEVC Main 10 and Scalable Main 10 profiles can be significantly improved for HDR and WCG content. In total, eight companies or aggregations of different companies and one university responded to the CfE and submitted responses to one or more of the different categories. To benchmark the potential coding technologies submitted in response to the CfE, we conducted a subjective quality evaluation to determine whether the proposed technologies could achieve better visual quality than the HEVC Anchor (Hanhart et al., 2015c). The subjective quality evaluation was conducted on 5 HDR video contents encoded at 4 bit rates by each algorithms in competition, leading to a total of 176 paired comparison against the HEVC Anchor. Overall 48 naïve subjects participated in the evaluation to collect a total of 24 ratings per video stimuli. Extensions of HEVC for HDR video coding are still under development and, in December 2015, MPEG and VCEG initiated a joint activity on this topic.

It is also important to conduct subjective evaluations during the development of coding standards to assess the impact of new coding tools, or alternatively, to assess the impact of removing coding tools, and to measure the quality improvements between different versions of the test model. Additionally, the performance of the standard in development is also assessed for new applications. In particular, efforts on the development of HEVC, the successor of AVC, were initiated in October 2004 and the first version of the standard was completed in January 2013. It was expected that HEVC could achieve even better compression efficiency for resolutions beyond HDTV, especially due to increased prediction flexibility and a wider range of block sizes. However, until August 2012, no subjective evaluation, including those performed in the context of the Call for Proposals (CfP) evaluations (De Simone et al., 2011), had been performed on resolutions higher than HDTV, mostly because of hardware limitations

and the lack of high quality uncompressed content. To address this problem, we conducted the first subjective quality evaluation to benchmark the performance of HEVC and AVC on 4K UHD video content (Hanhart et al., 2012b). The subjective quality evaluation was conducted with 36 naïve subjects on 3 4K UHD video contents encoded with AVC and HEVC at 5 bit rates, leading to a total of 30 video stimuli.

Similarly to previous video compression standards, HEVC provides an intra coding mode, where each frame can be encoded separately by considering only intra picture prediction and by disabling inter picture prediction. Thus, HEVC can also be used to compress still images or video sequences without considering any temporal prediction. The coding efficiency of HEVC intra coding for still image compression was investigated in a few studies that compare still images compression standards with HEVC intra coding by using PSNR as an objective metric for visual quality (JCTVC-I0461; JCTVC-I0595). These objective evaluations demonstrated that HEVC can achieve a considerable gain even compared to the state of the art JPEG 2000 compression standard. However, until January 2013, no subjective evaluation had been performed to assess the performance of HEVC intra coding for still image compression. To address this problem, we conducted the first subjective quality evaluation to benchmark HEVC intra coding for still image compression (Hanhart et al., 2013) following the guidelines defined by the JPEG committee for the evaluation of JPEG XR (De Simone et al., 2009b). The subjective quality evaluation was conducted with 22 naïve subjects on 6 high resolution image contents encoded with JPEG, JPEG 2000 (both 4:2:0 and 4:4:4 chroma sampling formats), and HEVC at 6 bit rates, leading to a total of 144 image stimuli. Since its first version in January 2013, HEVC defines a Main Still Picture profile for coding of 8-bits images with 4:2:0 chroma sampling. The second version completed in 2014 defines the Main 4:4:4 Still Picture and Main 4:4:4 16 Still Picture profiles for coding of still images with up to 4:4:4 chroma sampling and up to 16-bits per sample, respectively.

Recognizing the lack of a widely accepted standard for HDR image coding that can be seamlessly integrated into existing products and applications, JPEG issued a CfP in 2012, which led to the initiation of JPEG XT. This standard is meant to compress HDR images while preserving backward compatibility with the original JPEG format. The core part of JPEG XT has been published in June 2015, but the parts related to HDR coding are still to be published. During its development, several objective evaluations of JPEG XT have been performed, mainly by Richter. However, until May 2015, only one subjective evaluation was performed by Mantel et al. (2014), but only for Profile C and only on six different images. To overcome the lack of subjective evaluations of JPEG XT, we conducted the first extensive subject quality assessment of the three main profiles, i.e., profiles A, B, and C (Artusi et al., 2015; Korshunov et al., 2015) The subjective quality evaluation was conducted on 20 HDR image contents encoded with profiles A, B, and C at 4 bit rates, leading to a total of 240 image stimuli. Overall 48 naïve subjects participated in the evaluation to collect a total of 24 ratings per stimuli.

Subjective quality evaluations are also important after the finalization of the standard for verification purposes and to have an idea of the actual performance of the standard as published.

In particular, the Joint Collaborative Team on 3D Video Coding (JCT-3V) of MPEG and VCEG finalized the MVC+D and 3D-AVC amendments of AVC for 3D video coding in January and November 2013, respectively. In November 2013, JCT-3V issued a test plan (JCT3V-F1011) to evaluate the performance of two amendments of these two coding technologies. Three laboratories took part in this verification test: at EPFL in Switzerland, UWS in Scotland, and FUB in Italy (Hanhart et al., 2014c). All laboratories evaluated the same video data, i.e., 4 MVD contents encoded with MVC+D and 3D-AVC at 4 bit rates and rendered on a stereoscopic display considering two different configurations, leading to a total of 64 video stimuli. At the EPFL, 22 naïve subjects participated in the evaluation.

HEVC is the latest video compression standard developed by JCT-VC. However, its commercial use is subject to royalties, as HEVC is protected by several patents. This lead to the development of royalty-free and license-free alternatives, e.g., VP9 (and its successor, VP10). Thus, it is important as well to evaluate the performance of these alternatives in competition with international standards. The developers of VP9 have shown that VP9 has similar compression efficiency when compared to HEVC and a significantly higher compression efficiency when compared to AVC Mukherjee et al. (2013). However, a different studies by Grois et al. (2013) comes to a different conclusion, namely that VP9 is inferior to both AVC and HEVC. Such conflicting conclusions are mainly caused by different usage scenarios assumed in the papers and by different encoding configurations used. These results show that a fair subjective evaluation by a neutral and independent test laboratory is required. To address this problem, we conducted the first subjective quality evaluation to compare the compression efficiency between HEVC, VP9, and AVC assuming a real-time Internet-based streaming scenario (Rerabek et al., 2015b). The subjective quality evaluation was conducted with 26 naïve subjects considering a crowdsourcing environment on 8 high-definition video contents encoded with AVC, HEVC, and VP9 at 4 bit rates, leading to a total of 96 video stimuli.

For more than 40 years, most subjective quality evaluations have been conducted on 2D LDR still images and video sequences. Since then, many technological revolutions have occurred in imaging and display technologies, but the guidelines and methodologies for subjective evaluations have not always been updated to reflect the requirements of new technologies. For example, even if ITU has recently released a new recommendation for the assessment of stereoscopic 3D television (3DTV) systems (ITU-R BT.2021, 2012), there is no recommendation that addresses the specific issue of synthesized views. This lack of recommendation affects the evaluation of FTV systems, which rely on DIBR or alternative methods to allow the user to interactively control the viewpoint of the scene. To overcome the lack of standardize test methodologies for FTV scenarios, we proposed an experimental protocol to evaluate the impact of depth compression on perceived image quality in a FTV scenario (Bosc et al., 2013). A specific use case was considered to allow a reliable comprehension of the impact of depth coding: a smooth camera motion during a time freeze. This protocol is expected to enable the evaluation of different types of depth coding distortions. To illustrate the suitability of the proposed protocol, we conducted a subjective quality evaluation to assess the quality of FVV sequences corresponding to a smooth camera motion during a time freeze, which were

generated through DIBR from 3D content represented in the MVD format. The subjective quality evaluation was conducted with 27 naïve subjects on 6 MVD contents, with depth maps compressed by 7 algorithms at 3 bit rates and processed by 2 more algorithms, and rendered using 2 different view synthesis configurations, leading to a total of 276 video stimuli.

For more than 40 years, most subjective quality evaluations have been conducted in laboratory environments. However, conducting subjective experiments is very time consuming and can be quite expensive. To reduce the costs of subjective evaluations and also to consider more practical environments, researchers are investigating crowdsourcing platforms, which allow employing workers online from around the world. One of the constraints is the limited variety of display devices used by online workers. Due to this limitation, for example, a direct evaluation of 3D or HDR content is impossible, since 2D standard dynamic range (SDR) displays are the most commonly used. Therefore, it is necessary to use alternative representations of 3D and HDR content in crowdsourcing evaluations. To address the problem of crowdsourcing evaluation of 3D content, we investigated two possible approaches to assess the quality of MVD content on 2D displays: by using a virtual view and by using a FVV, which corresponds to a smooth camera motion during a time freeze (Hanhart et al., 2014g). To demonstrate the feasibility of the proposed approaches, the results of a crowd-based evaluation were compared to the ground truth results of a lab-based evaluation on a database of seven MVD contents encoded with 3D-AVC at four bit rates. The reference ground truth was obtained via a subjective evaluation of stereo pairs on a stereoscopic monitor in a laboratory environment with 22 naïve subjects. The two proposed 2D representations were generated for each bit rate and evaluated in a crowdsourcing environment with 20 naïve subjects. To address the problem of crowdsourcing evaluation of HDR content, we investigated the feasibility of using LDR versions of original HDR content obtained with TMOs in crowdsourcing evaluations (Hanhart et al., 2014d). To demonstrate the feasibility of the proposed approach, the results of a crowdsourcing evaluation were compared to the ground truth results of a lab-based evaluations on a database of five HDR image contents encoded with JPEG XT profile A at four bit rates. The reference ground truth was obtained via a subjective quality evaluation of the HDR images on a HDR monitor in a laboratory environment with 18 naïve subjects. The LDR versions were generated for each HDR image using eleven TMOs and evaluated in a crowdsourcing experiment by 18 naïve subjects.

Immersive video technologies aim at providing better, more realistic, and emotionally stronger experiences. An important question however is how significantly these technologies impact the viewers' QoE? To measure the impact of 3D on viewers' QoE, we investigated immersive video presentation experience via explicit subjective rating analysis for 2D and 3D multimedia contents (Kroupi et al., 2014a; Kroupi et al., 2014b; Kroupi et al., 2014c; Kroupi et al., 2015). A subjective experiment was conducted with 16 naïve subjects on 7 video contents presented in 2D and 3D modes, with low and high quality levels, leading to a total of 28 video stimuli. Various QoE-related aspects were investigated and compared. In particular, perceived quality, depth perception, content preference, and sensation of reality are investigated with respect to how they influence each other. To measure the impact of HDR on viewers' QoE, we investigated

the added value of higher dynamic range to viewers' preference using stimulus comparison (SC) with hidden reference and full pair comparison methods (Hanhart et al., 2014a; Hanhart et al., 2014b; Hanhart et al., 2015a). Two subjective experiments were conducted using eight HDR video contents presented at 100, 400, 1000, and 4000 cd/m^2 peak luminance levels, which were displayed side-by-side on a professional reference HDR monitor. The first experiment was conducted with 21 naïve subjects and the full PC method was used with all possible pairs, including 4000 cd/m^2 versus 4000 cd/m^2, leading to a total of 56 paired comparison. The second experiment was conducted with 20 naïve subjects and the SC method was used considering the 4000 cd/m^2 grade as hidden reference, leading to a total of 32 video stimuli.

Since immersive technologies have the ability to provide more details and depth, as well as better color, contrast, and brightness, understanding human attention patterns and viewing strategies for immersive image and video content is important for developing efficient data compression algorithms, as well as accurate objective quality metrics and computational models of visual attention. Although a significant number of public image and video datasets for visual attention exist, there are very few eye tracking datasets for immersive technologies. Regarding HDR content, there is only one dataset for HDR images by Narwaria et al. (2014) and two for HDR video sequences (Dong et al., 2014; Narwaria et al., 2014). To the best of our knowledge, no dataset with eye tracking data is available for UHD content. However, without this subjective data, it is hard to understand what is the impact of immersive technologies on visual attention and whether it is significant for practical applications. To measure the impact of UHD on visual attention, we created the first dataset of 4K UHD images with eye tracking data (Nemoto et al., 2014a; Nemoto et al., 2014b). The eye tracking experiment was conducted with 20 naïve subjects on 45 4K UHD images and their resized HD versions. The fixation density maps (FDMs) computed from the eye tracking data for UHD and HD resolutions were compared using three metrics to understand if there is a difference in visual attention between UHD and HD resolutions. To measure the impact of HDR on visual attention, we conducted the first eye tracking experiment investigating the difference in human visual attention between a HDR image generated from multiple exposure pictures and a single exposure LDR image of the same scene (Nemoto et al., 2015). The eye tracking experiment was conducted with 20 naïve subjects on 46 HDR images and their LDR versions. The FDMs computed from the eye tracking data for HDR and LDR resolutions were compared using the similarity score metric to understand if there is a difference in visual attention between HDR and LDR.

1.3.2 Predicting Quality of Experience

Subjective visual experiments are time consuming, expensive, and not always feasible. Therefore, objective quality metrics are needed to predict perceived visual quality. However, it is known that quality metrics do not always accurately reflect perceived visual quality. For example, it is known that PSNR is quite reliable to tune the performance of a particular codec on a specific content (Huynh-Thu and Ghanbari, 2008), but that it fails at predicting visual

quality when different contents and distortions are considered (Z. Wang et al., 2004). Therefore, it is essential to evaluate the performance of objective quality metrics in predicting perceived visual quality and to determine their scope of validity. For this purpose, ground truth subjective quality scores obtained via subjective visual quality experiments are used to evaluate the performance of objective metrics. For new applications, e.g., FTV, or types of content, e.g., 3D and HDR, it is also fundamental to determine the performance of existing metrics that are widely used.

In this thesis, we evaluated the performance of several objective metrics for different applications. First, we investigated the performance of state-of-the-art 2D metrics for quality assessment of stereo pairs formed from decoded and synthesized views (Hanhart et al., 2012a; Hanhart and Ebrahimi, 2012). A total of 9 metrics were computed considering 5 objective video quality models on a database of 8 MVD contents encoded with 24 compression algorithms at 4 bit rates, leading to a total of 768 video stimuli. The ground truth consisted of MOS and corresponding CI values collected from 18 naïve subjects for each video stimuli during the evaluations of the CfP on 3D Video Coding Technology issued by MPEG (N12036). The objective metrics were evaluated in terms of linearity, monotonicity, accuracy, and consistency with the ground truth. Additionally, the resolving power and classification errors of the metrics were computed.

Next, we evaluated the performance of the same metrics as in the first scenario, but for quality assessment of stereo pairs formed from two synthesized views (Hanhart and Ebrahimi, 2013a). The metrics were computed considering three objective video quality models on the same database. However, the ground truth was obtained for different stereo pairs and was collected from 36 naïve subjects, coming from two different test laboratories, for each video stimuli. The objective metrics were evaluated in terms of linearity, monotonicity, and accuracy with the ground truth.

The third application is also related to 3D, as it considers a FTV application. In particular, we investigated the performance of state-of-the-art 2D metrics for quality assessment of FVV sequences corresponding to a smooth camera motion during a time freeze (Hanhart et al., 2014e). A total of 7 metrics were computed on the database of FVV sequences created in Part I, which is composed of 6 MVD contents, with depth maps compressed by 7 algorithms at 3 bit rates and processed by 2 more algorithms, and rendered using 2 different view synthesis configurations, leading to a total of 276 video stimuli. The ground truth consisted of differential mean opinion score (DMOS) and corresponding CI values collected from 27 naïve subjects. The objective metrics were evaluated in terms of linearity, monotonicity, accuracy, and consistency with the ground truth. Statistical tests were performed to determine whether the difference between two different objective metrics is statistically significant. A PCA was also applied between the DMOSs and objective scores to further investigate the correlation of the objective metrics with perceived quality.

In the fourth application, we investigated the performance of HDR and LDR metrics for HDR

image quality assessment (Hanhart et al., 2015d). In total, 35 metrics (22 full-reference (FR) and 11 no-reference (NR) LDR metrics, as well as 2 HDR FR metrics) were computed on the database of HDR images encoded with JPEG XT created in Part I, which is composed of 20 HDR image contents encoded with profiles A, B, and C at 4 bit rates, leading to a total of 240 image stimuli. The LDR metrics were computed in the linear, logarithm, perceptually uniform (PU), and PQ domains. The ground truth consisted of MOS and corresponding CI values collected from 24 naïve subjects for each image stimuli. The objective metrics were evaluated in terms of linearity, monotonicity, accuracy, and consistency with the ground truth. Statistical tests were performed to determine whether the difference between two different objective metrics is statistically significant.

Finally, we investigated the effectiveness of HDR and LDR metrics to discriminate between quality levels when comparing two HDR video sequences (Hanhart et al., 2015c). In total, 9 metrics (4 LDR metrics computed in the PQ domain, 2 color difference metrics, 1 metric computed using multiple-exposure, and 2 HDR metrics) were computed on database of HDR video sequences created in Part I, which is composed of 5 HDR video contents encoded with HEVC and 9 algorithms in competition at 4 bit rates, leading to a total of 176 paired comparison against the HEVC Anchor. The ground truth consisted of preference scores collected from 24 naïve subjects for each video stimuli. The classification errors of the metrics were computed to assess their performance.

PSNR values below 25 dB and over 40 dB are often considered as bad and excellent quality, respectively. However, the exact relationship between PSNR values and perceived quality has not been established yet. This relationship should consider non-linearities and saturation effect of the HVS. As it was shown that PSNR is strongly content dependent, this relationship should also be determined for each content separately. To predict perceived quality of stereoscopic video sequences, we proposed a model based on content analysis (Hanhart and Ebrahimi, 2013b). A logistic function was used to map the PSNR values to perceived quality. The parameters of the mapping function were predicted using 2D and 3D content features. The model was trained and evaluated on a dataset of stereoscopic video sequences with associated ground truth MOS.

1.3.3 Improving Quality of Experience

Quality assessment in the conventional video processing chain takes into account many characteristic 2D artifacts (Yuen and H. Wu, 2005). When extended to 3D video, the HVS further processes additional monocular and binocular stimuli. Thus, the resulting video quality at the end of the 3D video processing chain depends also on the level of stereoscopic artifacts or binocular impairments affecting the depth perception. In fact, stereo artifacts can cause unnatural changes in structure, motion, and color vision of the scene and distort the binocular depth cues, which result in visual discomfort and eyestrain. Regarding the visualization stage, crosstalk is one of the stereo artifacts with the largest influence on image quality and visual

comfort (Meesters et al., 2004; Seuntiëns et al., 2005). Vergence-accommodation rivalry is believed to increase visual discomfort (Hoffman et al., 2008). Additionally, all the problems related to the sweet spot position in autostereoscopic displays also considerably reduce the overall 3D QoE. In this thesis, we proposed and evaluated different systems to reduce stereo artifacts generated at the visualization stage to improve QoE on 3D displays.

To improve the QoE provided by stereoscopic displays, researchers have proposed to exploit visual attention (Huynh-Thu et al., 2011b). Since two decades, researchers have investigated different solutions based on visual attention to reduce crosstalk and vergence-accommodation rivalry on stereoscopic displays. Several systems have been developed, but they rely on the accuracy of a computational model of visual attention and have not been assessed in a formal subjective evaluation. To address these problems, we proposed and evaluated two different approaches that exploit visual attention: an offline system, which uses a computational model of visual attention to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions (Hanhart and Ebrahimi, 2014b). From the saliency map, which was computed using a 3D visual attention model, the region-of-interest and its disparity were extracted. From the eye tracking measurements, filtered gaze points were used in conjunction with the disparity map to extract the disparity of the object-of-interest. Horizontal image translation was performed to bring the fixated object on the screen plane. The shift was determined based on the extracted disparity values and filtered in time to have smooth transitions that do no create visual discomfort. The user preference between standard 3D mode and the two proposed systems was evaluated in terms of image quality, depth quality, and visual discomfort. A subjective evaluation was conducted with 21 naïve subjects on 8 stereoscopic video contents using the PC method, leading to 24 paired comparisons.

To improve the QoE provided by mobile autostereoscopic displays, researchers have proposed to perform active crosstalk reduction based on user position (Boev et al., 2008; Park et al., 2011). A few systems have been developed and implemented on specific platforms (Boev et al., 2009b; Park et al., 2011), but, to the best of our knowledge, no subjective assessment demonstrating the effectiveness of an active crosstalk reduction system on a mobile device had been reported. To overcome this lack, we proposed and evaluated an active crosstalk reduction system for mobile autostereoscopic displays (Chappuis et al., 2014). The proposed system was implemented on a HTC EVO 3D smartphone. To determine the crosstalk level at each position, a full display characterization was performed. Furthermore, the localization of sweet spot and computation of the viewing freedom was performed. A special Android application was implemented to track the user face and eyes, and to correct artifacts in real-time according to his/her position. The proposed system was designed in the way that it first helps the user to find the sweet spot and then compensates for crosstalk artifacts and/or pseudoscopy. The user preference between standard 2D and 3D modes and the proposed system was evaluated in terms of image quality and depth quality. A subjective evaluation was conducted with 18 naïve subjects on 5 stereoscopic image contents using the PC method, leading to 15 paired comparisons.

To improve the QoE provided by multiview autostereoscopic displays, researchers have proposed to exploit viewer tracking and perform on-the-fly visual optimization to avoid the repetition effect between the lobes and mitigate crosstalk (Boev et al., 2008; Kooima et al., 2010). Most of the previous works only describe a proposed system without evaluating its performance. Except for some specific research on very expensive technologies, e.g., laser projection and low loss transparent display screen, which are far from mass production, most previous works were performed on multiview autostereoscopic displays having a rather limited number of views (typically eight to nine), whereas most advanced multiview autostereoscopic displays, e.g., the Dimenco displays, typically have around 30 views. With fewer views, the separation between the different luminance profiles is more pronounced and crosstalk compensation is relatively easy, whereas this problem is much more difficult as the number of views increases since the overlap between the luminance profiles is more sever. Additionally, none of these works provides a full description and subjective evaluation of a complete active crosstalk reduction system for current multiview autostereoscopic display technology. To address these problems, we proposed and evaluated an active crosstalk reduction system for current and future multiview autostereoscopic display technologies (Hanhart et al., 2015b). The proposed system was implemented considering a 52" full HD 28-view Dimenco BDL5231V autostereoscopic display with slanted lenticular sheet. The display was characterized in terms of luminance distribution and the luminance profiles were modeled using a limited set of parameters. A Kinect sensor was used to determine the viewer position in front of the display. The proposed system performs an intelligent on the fly allocation of the output views to minimize the perceived crosstalk. The user preference between standard 2D and 3D modes and the proposed system was evaluated in terms of image quality, depth quality, and visual discomfort. An informal subjective evaluation was conducted with 5 expert viewers on 4 MVD image contents using the PC method, leading to 12 paired comparisons.

1.4 Organization

The remainder of this thesis is structured as follows. Part I addresses different topics related to the measurement of QoE in immersive video technologies. In particular, Chapter 2 discusses the different aspects that must be taken into account and procedures than can be used when designing and analyzing subjective experiments. Chapter 3 focuses on different models to calculate the coding efficiency in terms of bit rate and quality differences between two codecs. Chapter 4 reports the performance analysis of different coding formats for still image, video, HDR image, HDR video, and 3D video compression. All these performance analyses were mainly performed using subjective quality evaluations to provide a more realistic estimation of the true coding efficiency. Chapter 5 investigates alternative evaluation protocols for subjective quality assessment. In particular, an experimental protocol to evaluate the impact of depth compression on perceived image quality in a FTV scenario is proposed. Additionally, alternative representations of 3D and HDR content are proposed for crowdsourcing evaluations of MVD video and HDR image coding, respectively, on 2D LDR displays. Chapter 6

investigates the impact of 3D and HDR on viewers' QoE via subjective experiments. Chapter 7 quantifies the impact of UHD and HDR on visual attention via eye tracking experiments.

Part II addresses the challenging problem of predicting QoE in immersive video technologies using objective models. In particular, Chapter 8 describes some of the most common quality metrics for still image, video, HDR, and 3D quality assessment. Chapter 9 provides a detailed description of the different procedures available to benchmark objective quality metrics. Chapter 10 reports the results of performance evaluation of state-of-the-art metrics for quality assessment of stereo pairs formed from decoded and synthesized views and from two synthesized views, HDR images, and HDR video sequences. Chapter 11 describes a model to predict perceived quality of stereoscopic video sequences based on content analysis.

Part III focuses on different solutions to improve QoE on 3D displays. Chapter 12 describes and evaluates different systems to reduce stereo artifacts generated at the visualization stage. In particular, two different approaches that exploit visual attention to improve 3D QoE on stereoscopic displays are investigated, as well as active crosstalk reduction systems for mobile autostereoscopic displays and multiview autostereoscopic displays.

Finally, Chapter 13 concludes the thesis with the summary of the main achievements and some outlook for future research.

Measuring Quality of Experience Part I

2 Design and Analysis of Subjective Experiments

In a subjective visual quality experiment, a group of people, referred to as subjects, is presented with a set of images or video sequences, referred to as stimuli, and is asked to judge their aspect, referred to as assessment factor, e.g., overall quality, color rendition, sharpness, etc. The stimuli are presented following a specific procedure and the subjects have to express their judgment using a particular scale, which can be either discrete or continuous. The selection of a particular stimuli presentation procedure and rating scale is referred to as test method. The experiment is conducted under specific viewing conditions, e.g., test environment, viewing distance, monitor peak luminance, ambient lighting, etc.

Subjective experiments are the ultimate means to assess quality of experience as they rely on a pool of human subjects. However, reliable and meaningful results can only be obtained if the experiments are properly designed and conducted following a rigorous methodology. Several international recommendations have been published to provide guidelines for conducting subjective visual quality experiments (ITU-R BT.1788, 2007; ITU-T P.910, 2008; ITU-R BT.500-13, 2012; ITU-R BT.2022, 2012; ITU-R BT.2021, 2012; ITU-T P.911, 1998). The different recommendations cover the selection of the test material, set up of the viewing environment, choice of test method, pre- and post-screening of the subjects, and even analysis of data. These recommendations result from experience gathered by difference groups, e.g., Video Quality Experts Group (VQEG), JPEG, MPEG, VCEG, and some ITU study groups. These recommendations can be considered as a set of best practices and guidelines that should be followed when designing a subjective experiment.

This chapter describes the different aspects that must be taken into account and procedures than can be used when designing and analyzing subjective experiments. This chapter covers the guidelines proposed in the different international recommendations as well as some common practices followed by researchers actively working in the field of subjective quality assessment. Different techniques are presented to analyze results of subjective experiments and to compare results of different experiments. Some of these techniques are coming from international recommendations and scientific publications, whereas other techniques were developed during this thesis. Finally, a brief overview of the common techniques used to

analyze eye movement data recorded with an eye tracker is presented. The different techniques and guidelines described in this chapter were used in the subjective experiments reported in the rest of the thesis.

2.1 Viewing Conditions

The viewing conditions can be decomposed into two main components: the test environment and the monitor. Two different types of test environments are usually considered: laboratory and home environment. The first environment is intended to provide critical viewing conditions, whereas the second is intended to provide a means to evaluate quality at the consumer side of the TV chain. The characteristics of the home environment have changed over the years and can be quite different from one country to another. Moreover, the variety of home environment characteristics is quite large, so it is hard to find the conditions that are representative of most home environments. In the home environment, the viewing distance is usually large (typically about 3.5 m), which prevents the viewers to resolve small details in high resolution content (see optimal viewing distance below), and the lighting conditions are usually quite bright (about 200 lx), which prevents viewers to see details in dark areas. Therefore, the laboratory environment is usually preferred, as it is quite well defined and provides more challenging test conditions. Nowadays, researchers are moving away from well-controlled environments and conducting online crowdsourcing experiments (Hossfeld and Keimel, 2014). In this case, there is almost no control on the viewing conditions and the guidelines described in this section do not apply.

In the laboratory environment, walls and curtains are either black (similar to color grading environment in video production) or mid grey (neutral). The room should not be completely dark, so lights should be placed behind the monitor such that they do not reflect off of the display. It is recommended that the ratio of luminance of background behind monitor to display peak luminance should be around 15%. However, this value was determined for CRT monitors and there is no recommendation for new display technologies and for HDR displays in particular. For HDR content, its was found that high ambient light significantly attenuates the perception of leakage defect (Mantel et al., 2015a), as the adaptation of the human eye to higher luminance values could masks details in dark areas. Additionally, it was shown that the loss of contrast in the dark regions could prompt subjects to elevate brightness settings for higher ambient light levels (Rempel et al., 2009). However, it was reported that visual fatigue is not a serious concern even in dark environments (Rempel et al., 2009). While some studies on HDR content have been conducted with ambient levels of 200 cd/m^2 (Narwaria et al., 2015b), we believe that the ambient light for HDR content should be set similar to LDR content. The motivation is that HDR is also intended to provide deeper black levels, while head room should be reserved for specular, highlights, special effects, etc. and that the average luminance level should not be significantly higher than for LDR content.

The monitor should be calibrated to have a color reproduction as faithful as possible. Profes-

Table 2.1: Optimal horizontal viewing angle and viewing distance in picture heights (H).

Resolution	Optimal horizontal viewing angle	Optimal viewing distance
1280 × 720	21 °	4.8H
1920 × 1080	31 °	3.2H
3840 × 2160	58 °	1.6H
7680 × 4320	96 °	0.8H

sional monitors, e.g., the Eizo ColorEdge series, typically support a 3D look up table, generated via a proper calibration, to provide accurate color reproduction. Most consumer-grade monitors only provide basic controls, e.g., red/green/blue gain, whereas high-end consumer monitors allow setting the red/green/blue/yellow/magenta/cyan colors and white balance at ten different luminance levels. The white point should be calibrated to D65 (6500 K) and the color gamut should be calibrated to Rec. 709 (Rec. 2020 or DCI P3 for WCG displays). According to ITU-R BT.2022 (2012), the peak luminance should be between 70 and 250 cd/m^2, but the value 120 cd/m^2 is typically adopted, as this value is commonly used for reference monitors in a production environment in Europe (100 cd/m^2 in US and Japan) and is the default value in most display calibration software. Obviously, this recommendation does not apply to HDR displays.

The viewing distance, i.e., the distance between the display and the subjects, also plays an important role. If subjects are seated too far away, then they cannot resolve small details and some artifacts could be masked. Most recommendations were drafted at the time CRT monitors were mostly used and their guidelines regarding viewing distance should not be followed for LCD displays. Instead, the viewing distance should be set according to recommendation ITU-R BT.2022 (2012), i.e., the distance at which two adjacent pixels subtend an angle of 1 arcmin at the viewer's eye. This value was selected at it corresponds to normal visual acuity (see Section 2.3). The optimal viewing distance depends on the display resolution and is typically expressed in relative units as a multiple of the display height (active part only). Table 2.1 lists the optimal viewing distance in picture heights (H) and corresponding optimal horizontal viewing angle for the most common resolutions found in today's video formats. As it can be observed, the relative viewing distance decreases as the resolution increases. However, when displaying mixed resolution sources in their native format on the same display, the absolute viewing distance remains the same, as it is determined by the pixel size.

2.2 Test Material

The source images or video sequences should be selected according to the goal of the experiment, but should be of optimum quality for the standard considered. The absence of defects in the source reference is crucial to obtain stable results. The number of sources should be determined according to the goal of the experiment. However, at least four different scenes

should be selected to avoid boring the subjects and to achieve a minimum reliability of the results. The source selection is an important issue, especially when assessing the performance of image and video compression algorithms. In this case, it is of essential importance to select scenes that will challenge the compression algorithms. However, the scenes should also be representative and consistent with the media service that the transmission channel is intended to provide.

The source contents can be described following different characteristics. To characterize the spatial information of a still image or video frame, the spatial perceptual information (SI) measurement is proposed in ITU-T P.910 (2008). The spatial perceptual information (SI) is based on the Sobel filter. The luma component of the still image or video frame is first filtered using a Sobel filter. For still images, the SI value is computed as the standard deviation computed over the pixels of the Sobel-filtered image. For video sequences, this process is repeated for each frame and the SI value is computed as the maximum value across all frames.

To characterize the temporal information of a video sequence, the temporal perceptual information (TI) measurement is proposed in ITU-T P.910 (2008). The temporal perceptual information (TI) is based on the pixel difference between consecutive frames, to estimate motion difference. First, the difference between the luma component of the current frame and that of the previous frame is computed. Then, the standard deviation is computed over all pixels. Finally, the TI value is computed as the maximum standard deviation value across all frames. Note that a higher TI value corresponds to more motion between consecutive frames.

For stereoscopic content, to characterize the depth along the spatial and temporal dimensions, Urvoy et al. (2012) proposed to compute the SI and TI values on the depth maps instead of the texture video, leading to the proposal of depth spatial indicator (DSI) and depth temporal indicator (DTI).

For HDR content, the dynamic range of a still image or video frame is computed as

$$\text{dynamic range} = \log_{10}\left(\frac{L_{max}}{L_{min}}\right) \tag{2.1}$$

where L_{min} and L_{max} are the minimum and maximum luminance values, respectively, computed after excluding 1% of darkest and brightest pixels. For video sequences, the maximum dynamic range across all frames is reported. Another measurement was proposed by Akyüz and Reinhard (2006) and is referred to as key. The key is in the range $[0,1]$ and gives a measure of the overall brightness. The key is computed as

$$\text{key} = \frac{\log L_{avg} - \log L_{min}}{\log L_{max} - \log L_{min}} \tag{2.2}$$

where L_{min}, L_{max}, and L_{avg} are the minimum, maximum, and average luminance values, respectively, computed after excluding 1% of darkest and brightest pixels. Finally, Narwaria et al. (2015b) have also computed the SI and TI values in the PU (Aydın et al., 2008) domain.

These measurements can be used in the content selection process. For example, for assessing the performance of video compression algorithms, one will select sources with low SI and low TI, high SI and low TI, low SI and high TI, and high SI and high TI, as well as some sources with intermediate values. The set of sources should span the range of measurements of interest to users of the devices under test.

2.3 Subjects

Subjects can be classified into two categories: naïve and expert viewers. Expert viewers have expertise in quality assessment and in particular in assessing the artifacts that may be introduced by the system under test. Researchers working on image and video compression or quality assessment are typically considered as expert viewers, but, to have a neutral comparison, subjects should not be, or have been, directly involved in the development of the system under test. On the other hand, naïve viewers have no expertise in quality assessment.

It is recommended that at least 15 subjects should assess each test stimuli, but it is usually worthless to consider more than 40 subjects. For preliminary or pilot experiments carried out before a larger test, a small group of four (absolute minimum for statistical reasons) to eight expert viewers can provide indicative results and are referred to as informal studies.

Prior to the test, each subject must be pre-screened for normal visual acuity or corrected-to-normal acuity and for normal color vision. Visual acuity can be tested using the Snellen or Landolt chart. Normal visual acuity, sometimes referred to as 20/20 vision, means that a human eye with nominal performance is able to separate contours that subtend an angle of 1 arcmin, which approximately corresponds to 1.75 mm apart at a distance of 20 feet. In the Snellen scale, this corresponds to 20/20 and subjects should be rejected if they have a visual acuity below 20/30. Color vision can be tested using the Ishihara chart. In this case, subjects should be rejected if they miss more than 2 plates out of 12. In the context of assessment of 3DTV or auto-stereoscopic systems, it is recommended to also screen subjects for correct binocular vision, which can be tested using the Randot test.

There is no recommendation regarding the age range or gender balance among the subjects. However, subjects between 18 and 30 years old are often preferred because their visual system is fully developed and they have a good visual acuity.

2.4 Test Methods

The selection of a particular test method, i.e., stimuli presentation procedure and rating scale, is mainly determined by the systems under test that the experimenter wants to evaluate. For example, to evaluate the performance of algorithms that introduce distortions, e.g., image and video compression algorithms, the single stimulus (SS) or absolute category rating (ACR) methods are used if the test material is spread over a wide range of quality levels. The double

stimulus impairment scale (DSIS) or degradation category rating (DCR) methods are used if it is necessary to check the fidelity with respect to the reference image or video sequence. On the other hand, if the quality of the source reference image or video sequence is not perfect or the algorithm under test can improve visual quality, e.g., image sharpening and denoising algorithms, then the absolute category rating with hidden reference (ACR-HR) or double stimulus continuous quality scale (DSCQS) methods are preferred. The single stimulus continuous quality evaluation (SSCQE) or simultaneous double stimulus for continuous evaluation (SDSCE) methods are selected if the rating should be made temporally along the video sequence. To compare different rendering algorithms, display technologies, or other algorithms, e.g., TMOs, the SC or PC methods are preferred as they rely on an indirect scaling based on preference instead of a direct scaling based on a rating scale. These methods also have a high discriminatory power, which is of particular value when visual differences between stimuli are small. The test methods are described in details in the following subsections.

2.4.1 Single Stimulus and Absolute Category Rating

The single stimulus (SS) method, also referred to as absolute category rating (ACR) method, is a category judgment where the stimuli are presented one at a time and are rated independently on a category scale. Each test image or video sequence is presented for a particular duration (typically about 10 s). Subjects should be asked to look at the display for the entire presentation and to base their judgment on the overall impression given by the presentation. Subjects should be asked to provide their judgment immediately after each presentation and to express these judgments in terms of the wordings used to define the rating scale. During the voting time, which is typically set to 5 s, the display should be set to mid grey.

The following five-grade quality scale is commonly used for rating overall quality

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

For the assessment of low bit-rate video codecs, the use of rating scales with more than five grades could be beneficial. An extension of the five-grade scale to a nine-grade scale can be used, with labels used for every second grade, as illustrated below

9 Excellent
8
7 Good
6
5 Fair
4
3 Poor
2
1 Bad

A further extension of this scale is shown below, where the endpoints have been verbally defined as anchoring points which are not used for the rating. In this verbal definition, some kind of explicit or implicit reference is used, e.g., the reference image or video sequence for the upper endpoint, and it will be clearly illustrated during the training phase.

		The number 10 denotes a quality of reproduction that is perfectly faithful to the original.
10		No further improvement is possible.
9	Excellent	
8		
7	Good	
6		
5	Fair	
4		
3	Poor	
2		
1	Bad	
0		The number 0 denotes a quality of reproduction that has no similarity to the original.
		A worse quality cannot be imagined.

Furthermore, a continuous scale divided into five segments associated with labels corresponding to those of the five-grade scale can be used. The nine-grade, eleven-grade, and continuous scales can be used if higher discriminative power is required, but it does not necessarily ensure that the differentiation between two conditions is going to be more powerful.

Other dimensions than overall quality, e.g., brightness, contrast, or color reproduction, can also be assessed. These dimensions may be useful for better understanding different perceptual factors when the perceived overall quality is nearly the same, although the systems are clearly perceived as different. For example, to assess stereoscopic 3DTV systems, it is recommended to assess the following dimensions: picture quality, depth quality, and visual comfort. In this case, picture and depth quality can be assessed using the same scale as for overall quality. However, to assess visual comfort, the following labels should be used instead: *Very comfortable*, *Comfortable*, *Mildly uncomfortable*, *Uncomfortable*, and *Extremely uncomfortable*.

These methods are easy and fast to implement, as each stimulus is presented one after the other, with a voting period after each stimulus. Therefore, the presentation time is quite short: the duration of the stimulus plus the duration of the voting time. If replications are required, the stimuli are simply repeated at different points in time. In this case, the total time is multiplied by the number of repetitions.

The absolute category rating with hidden reference (ACR-HR) method is a slight variation of the ACR method in which the source reference images or video sequences are presented and evaluated as any other stimulus. Instead of keeping the individual scores of the test and reference images or video sequences, a differential score is computed between each test image or sequence and its corresponding source reference (see Section 2.6.2). The advantage is that the perceptual impact of the source reference image or video sequence can be removed from the subjective scores. In particular, the influence of content preference, quality of the source reference (e.g., due to camera quality), and monitor (e.g., professional quality versus consumer grade) on the subjective scores can be reduced. Nevertheless, the ACR-HR method should only be used with source reference images and video sequences having visual quality evaluated as good or excellent by expert viewers. Additionally, this method may not be suitable when impairments occur in the first and last 1 s of the video sequence, as the viewers might be unfamiliar with the source reference video sequence due to the stimuli presentation order.

2.4.2 Double Stimulus Impairment Scale and Degradation Category Rating

In the double stimulus impairment scale (DSIS) method, also referred to as degradation category rating (DCR) method, subjects are presented with pairs of images or video sequences, referred to as stimuli A and B. The first stimulus in the pair (stimulus A) is always the unimpaired source reference and the second stimulus (stimulus B) is the same source presented through one of the systems to be evaluated, i.e., the same source impaired. Subjects are asked to rate the impairments of the second stimulus in relation to the first stimulus, and to express these judgments in terms of the wordings used to define the rating scale. The method uses an impairment scale, e.g., the following five-grade impairment scale

5 Imperceptible
4 Perceptible, but not annoying
3 Slightly annoying
2 Annoying
1 Very annoying

The DSIS method is typically considered to evaluate the transmission fidelity with respect to the source signal, which is an important factor in high quality video systems. In this case, the labels associated with the rating scale (imperceptible/perceptible) are valuable when the detection of impairment is an important factor. Similarly to the SS and ACR methods, an extension of the five-grade scale to a nine-grade scale can be used, with labels used for every second grade, as illustrated below

9 Imperceptible
8
7 Perceptible, but not annoying
6
5 Slightly annoying
4
3 Annoying
2
1 Very annoying

Note that it is usually found that the stability of the results is greater for small impairments than for large impairments (ITU-R BT.500-13, 2012). However, it is recommended to use this method with test stimuli covering a full range of impairments rather than a limited range of impairments. If the discrimination of very small impairments is required, the reference and test stimuli should be presented twice in a alternate manner, i.e., reference, test, reference, and test. When the pair is presented only once, the method is referred to as DSIS Variant I, whereas it is referred to as DSIS Variant II when the pair is presented twice. Before each stimulus presentation, the display should be set to mid grey for 1 to 3 s. It is common practice to display the letter A or B in black, at the center of the display, and over the mid gray background, to indicate which of the reference or test stimulus will be presented. For Variant II, a star is typically added next to the A and B letters at the second presentation to indicate that the subjects will have to vote after the presentation of this pair. Alternatively, if display and source resolutions permit, then the reference and test stimuli can be presented simultaneously on the same monitor. In this case, the two stimuli should be displayed as side-by-side on a mid grey background. During the whole test duration, the reference should always be placed on the same side and the subjects must be aware of the positions of the reference and test stimuli. However, the positions can be changed from one group of subjects to another, for example to compensate for imperfect display uniformity. In the case of video content, the two video sequences must be perfectly synchronized. At the end of the presentation, the display should be set to mid grey for the voting time, which is typically set to 5 s.

2.4.3 Stimulus Comparison and Pair Comparison

In the stimulus comparison (SC) method, also referred to as pair comparison (PC) method, subjects are presented with pairs of images or video sequences, referred to as stimuli A and B. The pair of stimuli consists of the same source being presented first through one system under test and then through another system. For each source reference, considering N systems under test, all two-permutations of N should be considered to generate the pairs. This leads to $N(N-1)$ pairs for each source reference, which increases exponentially as the number of systems under test increases. Hence, this method can require a lot of time when the number of systems under test is large. Note that, ideally, all possible orders, e.g., XY, YX, should be considered. However, the length of the test can be reduced by a factor two by spreading

all possible orders over the different subjects, i.e., one half of the subjects will see the pair XY, whereas the other half will see the pair YX. More complex designs, e.g., square design, optimized square design, or adaptive square design have been investigated to further reduce the number of pairs to be evaluated and it was shown that they provide comparable results to that of the full pair comparison (J. Li et al., 2013a; J. Li et al., 2013b). The two stimuli can be presented as side-by-side, either on the same display or on two aligned monitors, or sequentially in time. In the first case, the two stimuli should be perfectly synchronized, whereas the presentation time should be identical if the presentation is sequential. Note that the test duration becomes longer in case of sequential presentation.

The SC method is classified in three types of methods: performance, adjectival categorical judgment, and non-categorical judgment methods. In the performance method, subjects are asked to select which stimulus in the pair is preferred based on some factor, e.g., overall quality, depth quality, visual comfort, etc. To collect the answer, either a binary (*A*, *B*) or a ternary (*A*, *B*, *Same*) scale is used. In the first case, which is referred to as forced choice (FC), the subject is forced to select one or the other stimulus, even when no difference is visible between the two stimuli.

The adjectival categorical judgment method aims at better quantifying the relation between stimuli in a pair. The following scale is used to quantify the overall quality, for example, of stimulus B when compared to stimulus A

+3	Much better
+2	Better
+1	Slightly better
0	The same
-1	Slightly worse
-2	Worse
-3	Much worse

In the non-categorical judgment method, two forms are considered to evaluate the relation between stimuli in a pair. In the first form, a continuous scaling is considered using a scale defined by its two extremes, e.g., *Same-Different* or *Much better-Much worse*. Additional intermediate labels can be added. In the second form, each subject assigns each relation with a number (the range may be constrained or not) that reflects its judgment on a specified dimension, e.g., difference in quality. The number assigned may describe the relation in absolute terms or in terms of that in a standard pair.

The main advantage of the PC method is its high discriminatory power, which is especially interesting when several test stimuli have similar quality levels. This method is also very valuable to assess more abstract dimensions, e.g., immersiveness or sense of presence, to determine whether systems are perceived to differ, or to establish the point at which impairments become visible. When training subjects on how to use the rating scale (see Section 2.5.1), it is of common practice to present conditions representative of the different levels of the rating scale.

With the PC method, when using the binary or ternary scale, training is thus easier and less biased. Indeed, the experimenter can show examples where there are differences, but does not have to relate these differences to a particular grade. The PC method is also particularly suited to assess other systems than compression algorithms, e.g., new display technologies or different rendering algorithms, as this method is similar to the process followed by humans when they have to decide between two products.

2.4.4 Double Stimulus Continuous Quality Scale

In the double stimulus continuous quality scale (DSCQS) method, subjects are presented with pairs of images or video sequences, referred to as stimuli A and B. One of the stimuli in the pair is always the unimpaired source reference, whereas the other stimulus is the same source presented through one of the systems to be evaluated. However, unlike in the DSIS method, the order of the two stimuli is pseudo-random and the subjects does not know which stimulus is the source reference. Subjects are asked to rate the quality of both stimuli using a continuous quality scale divided into five segments (see Figure 2.1).

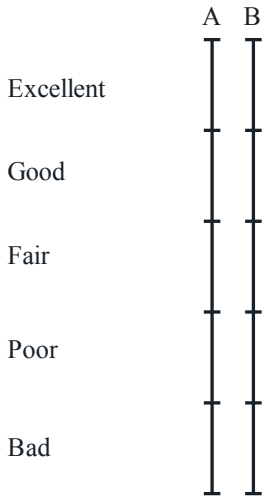


Figure 2.1: DSCQS rating scale.

The two stimuli can be presented following two variants. In Variant I, only one subject is seated in front of the monitor and the subject is free to select between stimulus A and B at each presentation, until he/she has a mental measure of the quality associated with both stimuli. In Variant II, which considers simultaneous subjects, the pair of stimuli is presented one or more times and the presentation order is determined by the experimenter. For still pictures, a 3 to 4 s presentation time with five repetitions is recommended. For video sequences, two presentations are recommended. In all cases, the display should be set to mid grey for 1 to 3 s before each stimulus presentation. It is common practice to display the letter A or B in black, at the center of the display, and over a mid gray background, to indicate which stimulus will be presented. Subjects should vote immediately after the last presentation while the display is

set to mid grey (typically set to 5 s).

The DSCQS method is useful when the test material does not cover the full range of quality. Since both the source reference and system under test are evaluated, the DSCQS method has the same advantages as the ACR-HR method (see Section 2.4.1) regarding the influence of source reference or monitor. This method can also be used to evaluate algorithms that aim at improving visual quality, e.g., image sharpening and denoising.

2.4.5 Subjective Assessment Methodology for Video Quality

The subjective assessment methodology for video quality (SAMVIQ) method (ITU-R BT.1788, 2007) uses a multi-stimuli approach. A graphical user interface presents a single source video sequence, available as explicit reference and at different quality levels (including a hidden reference). The subject is instructed to play the reference source as well as its different versions and to rate their visual quality. For each source, the subject is instructed to compare between all processed versions as well as against the reference, such that the subject can judge the quality of all video sequences. The video sequences can be paused and stopped such that the subject can switch between sequences. Each video sequence can be played as many times as necessary until the subject can rate properly all sequences. Once the subject has made his/her judgment about all sequences, then the next source can be evaluated. The SAMVIQ method uses a continuous five-level quality scale (see Section 2.4.1) ranging from 0 (*Bad*) to 100 (*Excellent*).

Huynh-Thu et al. (2007) have shown that the SAMVIQ method provides similar results to ACR. However, the method can could have some advantages, for example when subjects have difficulties judging quality on a single viewing or when subjects might want to re-watch the reference or change their score on a particular sequence. Nevertheless, the review capability increases the artificiality of the method and the method limits the number of systems under test that can be evaluated.

2.4.6 Single Stimulus Continuous Quality Evaluation and Simultaneous Double Stimulus for Continuous Evaluation

With the methods previously described in this section, only an overall quality score is obtained in the case of video sequences. Therefore, the effects of temporal quality fluctuations cannot be measured. To assess quality along the temporal axis of the video sequence, the single stimulus continuous quality evaluation (SSCQE) method can be used. This method is an extension of the SS method (see Section 2.4.1), but the subjects use a liner slider to evaluate video quality. The continuous five-level quality scale is used and subjects must adjust the slider when they notice changes in the video quality. The slider position is recorded during the whole sequence duration to allow a temporal analysis. While 10 s long video sequence are typically used with the other methods, video sequences should be at least 5 min long in the

SSCQE method. However, this method is commonly used with video sequences of 1 min long only.

If temporal fidelity with respect to the source reference must be assessed, then the simultaneous double stimulus for continuous evaluation (SDSCE) method should be used. This method is an extension of the DSIS method (see Section 2.4.2), but the subjects use a liner slider to evaluate fidelity. The source reference and test video sequences are presented side-by-side, either on the same display or on two aligned monitors, and subjects are aware of which is the reference. A continuous five-level impairment scale is used and subjects must check the differences between the two sequences and assess the fidelity of the test video by moving the slider throughout the whole sequence duration.

2.5 Test Design

Visual experiments typically consist of one training session and one or more test sessions. The goal of the training session is to explain the task to the subjects. Several training samples are presented such that they can get familiar with the presentation methodology and the range of quality or impairments. The test material is then evaluated during one or more test sessions, depending on the number of stimuli and presentation duration of each stimulus. The following subsections describe in details how the training and test sessions should be designed.

2.5.1 Training Session

Before starting with the actual test, a scenario of the intended application of the systems under test should be presented to the subjects. The test methodology, i.e., presentation of the stimuli, timing, method of assessment, types and levels of impairments likely to occur, and grading scale should be carefully described to the subjects. A minimum of five training images or sequences should be presented to the subjects following the same procedure as in the actual test. The training samples should be representative of the material shown later during the test session(s), i.e., they should have similar types of impairments and cover the same range of impairment or quality. It is of common practice to select one training sample representative of each level of the rating scale, such that subjects can better relate each level of the scale with a particular quality or impairment level. The different perceptual attributes, e.g., sharpness, blurriness, blockiness, colors reproduction, etc. that should be considered to evaluate overall quality or other dimensions should be explained and illustrated. Finally, questions about the procedure or the instructions should be answered, but only before the start of the test session.

Regarding the training sources, recommendation ITU-R BT.500-13 (2012) states that “training sequences [...] should be used with illustrating pictures other than those used in the test, but of comparable sensitivity”, whereas recommendation ITU-T P.910 (2008) states that “[training] may contain video sequences other than those used in the actual tests”, which is more flexible.

In general, it is preferred to use different sources for the training session than for the actual test sessions. However, when impairments are difficult to perceive, training with the same sources can improve the subjects discriminative power.

2.5.2 Test Session

In general, a test session should not last more than half an hour. For new technologies, e.g., 3D and HDR imaging, the duration should be reduced to 15-20 min maximum, as they can induce visual discomfort to the subjects. Therefore, depending on the number of images or sequences to evaluate and the methodology, i.e., with or without repetition and with or without the reference, the experiment has to be fragmented into several test sessions. Note that each subject can also take part to only a subset of all test sessions depending on the total test duration, as subjects tends to get bored and less effective after more than 1 to 1.5 hour. The important is that each stimulus should be evaluated by a certain number of subjects (see Section 2.3).

At the beginning of the first session, about five dummy presentations, whose scores are not included in the results, should be included to stabilize the subjects' opinion. It is of common practice to select at least one sample representative of high, low, and mid quality. If the test is split into several sessions, about three dummy presentations should be included at the beginning of the following sessions.

The stimuli order of presentation should be pseudo-random and can be generated using different designs, e.g., randomized design, Latin or Graeco-Latin square designs. The different sources and quality levels should be balanced out from session to session. Different order of presentation should be considered for the different (groups of) subjects to reduce any effect on grading, which could be due to the presentation order (for example, presenting a low quality stimulus after a high quality stimulus) or subject tiredness or adaptation. In any case, the same source should never be shown in two successive presentations, even if the levels of impairment are different.

Some of the stimuli can be presented twice or more, at different time instants during the experiment, to check coherence. These replications can be used to estimate within subject variation or to test subjects' reliability. In the latter case, the some stimuli order of presentation under identical conditions can be used. Otherwise, the resulting variation in the data can also be impacted by the presentation order effect.

2.6 Data Processing

The methods described in Section 2.4 use either a discrete or a continuous scale. In the first case, the range of scores is defined by the number of grades in the rating scale, e.g., 1-5 or 1-9. In the latter case, the scores should be normalized to integer values between 0 (bottom of the

scale, typically corresponding to *Bad* quality) and 100 (top of the scale, typically corresponding to *Excellent* quality). Hence, subjective experiments performed using one of the methods described in Section 2.4 will produce distributions of integer values, for example between 1 and 5 or between 0 and 100, for each test stimulus. To understand how two stimuli compare to each other, their distributions must be analyzed using different statistical tools, as described in the following subsections.

2.6.1 Outlier Detection

The first thing to do before performing any type of analysis is to discard subjects whose scores appear to deviate strongly from others in a test session. When a subject is detected as outlier, all his/her scores are removed from the results of the session. The outlier detection process is applied independently to each test session. Then, the clean scores can be analyzed.

An outlier detection technique is suggested in ITU-R BT.500-13 (2012) for methods where subjects have to provide an overall score. If the kurtosis coefficient is between 2 and 4, then the data roughly follows a normal distribution. Otherwise, the data is considered as not-normal. The methodology counts the number of times the subject fall outside of a specific interval, defined as the mean value plus/minus the associated standard deviation times 2 (if normal) or times $\sqrt{20}$ (if non-normal). If this count is higher than 5% of the scores and the relative absolute difference of occurrences below and above the specified interval is lower than 30%, then the subject is classified as outlier.

Another detection technique was used by De Simone et al. (2011) and is inspired by the Tukey boxplot (Tukey, 1977). In the Tukey boxplot, the lower and upper inner fences are defined by the lower quartile minus 1.5 IQR and the upper quartile plus 1.5 IQR, respectively, where IQR is the interquartile range and is defined as the difference between the upper and lower quartiles. If the data is normally distributed, this range roughly corresponds to ± 2.7 the standard deviation, which covers about 99.3% of the data. A subject is then classified as outlier if more than 20% of his/her scores are outside of the region determined by the lower and upper inner fences.

2.6.2 Mean Opinion Scores and Confidence Intervals

The statistical analysis is based on the assumption that a score s_{ij} given by subject j for the test condition i can be expressed as

$$s_{ij} = \mu_j + \epsilon_{ij} \tag{2.3}$$

where μ_i is the reaction to the test condition and defined by the controlled experimental variables (e.g., source and system under test) and ϵ_{ij} is an error caused by a set of uncontrolled variables (Bech and Zacharov, 2006). This error is related to the subject (e.g., emotional state, mood, expectations, interpretation, bias, etc.) and/or the experiment set-up (e.g.,

lighting conditions, background noise, etc.). The experimental error is often assumed to be normally distributed with a zero mean, $\mathcal{N}(0,\sigma_i^2)$, and thus the subjective scores for each test condition are assumed to be normally distributed, $\mathcal{N}(\mu_i,\sigma_i^2)$, with mean μ_i and standard deviation σ_i. However, this assumption is sometimes not met, for example when the number of subjects is low, or near the extremes of the scale, or because of the discrete nature of the rating scale. Nevertheless, the data will be approximately normally distributed with large number of subjects, regardless of the underlying distribution, according to the central limit theorem.

Based on these characteristics, the subjective scores are commonly characterized using the mean opinion score (MOS), related to the mean of the distribution, and the confidence interval (CI), related to the standard deviation of the distribution. For some methods, the a differential mean opinion score (DMOS) is computed between the source reference and test stimulus instead. These properties are further described in the following parts.

Mean Opinion Scores

The MOS is computed independently for each test condition as

$$\mathrm{MOS}_i = \frac{1}{N}\sum_{j=1}^{N} s_{ij} \tag{2.4}$$

where N is the number of valid subjects and s_{ij} is the score by subject j for the test condition i. The MOS reports the average score of a particular test condition computed based on a sample of the population, i.e., the subjects who took part in the experiment, and is an unbiased estimator of the true mean of the distribution (infinite number of subjects).

Differential Mean Opinion Scores

In some methods, e.g., ACR-HR and DSCQS (see Section 2.4), the source reference must also be graded by the subjects. With these methods, instead of reporting the MOS for both the source reference (SRC) and processed stimuli (PS), a DMOS is reported instead and is computed as:

$$\mathrm{DMOS(PS)} = \mathrm{MOS(PS)} - \mathrm{MOS(SRC)} + \max(\text{rating scale}) \tag{2.5}$$

Note that DMOS values can be higher than the highest grade on the rating scale if the processed stimuli was evaluated better than the source reference. Such condition should be considered as valid.

Confidence Intervals

The sample standard deviation, s, is an unbiased estimator of the true standard deviation of the distribution and is computed as

$$s_i = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(s_{ij} - \mathrm{MOS}_i\right)^2} \tag{2.6}$$

The reader can refer to (Winkler, 2009) for an analysis of the impact of the rating scale and MOS on the standard deviation.

The sample standard deviation is then used to compute the $100 \times (1-\alpha)\%$ confidence interval, which is given by:

$$[\mathrm{MOS}_i - \delta_i, \mathrm{MOS}_i + \delta_i] \tag{2.7}$$

where

$$\delta_i = t(1-\alpha/2, N)\frac{s_i}{\sqrt{N}} \tag{2.8}$$

where $t(1-\alpha/2, v)$ is the t-value corresponding to a two-tailed Student's t-distribution with degrees of freedom v (which is set to $N-1$) and a desired significance level α (equal to 1-degree of confidence). It is recommended to used the 95% CI, corresponding to $\alpha = 0.05$. Note that it is common to use the z-value corresponding to a two-tailed normal distribution instead of the t-value (ITU-R BT.500-13, 2012). For the 95% CI, the z-value is equal to 1.96, whereas the t-value for 15 subjects is equal to 2.14. Thus, the t-value generally leads to larger confidence intervals, but this is the correct approach from a statistical point of view, because the variance is unknown (and has to be estimated from the samples) and the number of samples (i.e., subjects) is generally relatively low. Recommendation ITU-T P.1401 (2012) suggests to use the Student's t-distribution if there are less than 30 subjects and the normal distribution otherwise.

With a probability of 95%, the true mean of the distribution lies within the 95% CI. Thus, when presenting the results, the MOSs should always be reported together with their corresponding CIs.

2.6.3 The Bradley–Terry–Luce and Thurstone Case V Models

To analyze the data of a subjective evaluation, MOSs are generally computed for each test condition. However, when using the PC or SC methods (see Section 2.4.3), only preference scores between pairs of stimuli are recorded. In this case, relative MOSs can be estimated from the preference scores using statistical models, e.g., the Bradley-Terry-Luce (Bradley and Terry, 1952; Luce, 1959) and Thurstone Case V models (Thurstone, 1927). These models take into account the relations between the different stimuli to rank them and to estimate relative

scores. If a full pair comparison design was used, then all combinations are used to rank the stimuli. Otherwise, only the tested combinations can be used, while the missing combinations can be inferred, for example by transitivity (ex: if B is better than A and C is better than B, then C is likely to be better than A).

The major difference between the Bradley-Terry-Luce and Thurstone Case V models is the assumption behind the distribution of the quality difference of two stimuli. The Thurstone model assumes a Gaussian distribution, whereas the Bradley-Terry-Luce model assumes a logistic distribution. Thus, the Bradley-Terry-Luce model was often preferred because the logistic cumulative distribution function (CDF) has a closed-form expression, whereas the Gaussian CDF requires evaluating the error function. Note that these models are only suitable for binary or ternary scales, but scores from other scales can be converted. For example, when considering an adjectival categorical judgment method, one vote can be attributed to *Slightly better*, two votes for *Better*, and three for *Much better*.

Mean Opinion Scores Estimation

Before estimating the MOS values, the winning frequency w_{ij} and the tie frequency t_{ij} (if the option *Same* was used) are computed from the obtained subjective ratings for each pair of stimuli i and j. Note that $t_{ij} = t_{ji}$ and $w_{ij} + w_{ji} + t_{ij} = N$, where N is the number of subjects. This can be done individually for each source or jointly over all sources.

Then, using winning frequencies w_{ij} and the tie frequencies t_{ij}, a count matrix C is constructed. Each element of the count matrix C_{ij} is computed as follow

$$C_{ij} = w_{ij} + \frac{t_{ij}}{2} \tag{2.9}$$

Thus, C_{ij} represents the number of times stimulus i is preferred over stimulus j, where i and j are the row and column of the matrix. Ties are considered as being half way between the two preference options, i.e., they are distributed equally between C_{ij} and C_{ji} (Glickman, 1999).

In the Thurstone model, the quality scores are assumed to follow a Gaussian distribution, $\mathcal{N}(\mu, \sigma^2)$, with mean μ and standard deviation σ. The corresponding probability density function (PDF) is

$$p_X(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu_X}{\sigma_X}\right) \tag{2.10}$$

where ϕ is the standard normal PDF with zero mean and unit variance

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{2.11}$$

If only two stimuli, A and B, are compared, the probability of choosing A over B can be

expressed as

$$P(A > B) = P(A - B > 0) \tag{2.12}$$

Since A and B are two Gaussian random variables, their difference is also a Gaussian random variable following the statistics $\mathcal{N}(\mu_{AB}, \sigma^2_{AB})$, with mean $\mu_{AB} = \mu_A - \mu_B$ and standard deviation $\sigma^2_{AB} = \sigma^2_A + \sigma^2_B - 2\rho_{AB}\sigma_A\sigma_B$, where ρ_{AB} is the correlation between A and B.

The probability of choosing A over B can then be written as

$$\begin{aligned} P(A > B) = P(A - B > 0) &= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2_{AB}}} e^{-\frac{(x-\mu_{AB})^2}{2\sigma^2_{AB}}} \mathrm{d}x = \int_{-\mu_{AB}}^\infty \frac{1}{\sqrt{2\pi\sigma^2_{AB}}} e^{-\frac{x^2}{2\sigma^2_{AB}}} \mathrm{d}x \\ &= \int_{-\infty}^{\mu_{AB}} \frac{1}{\sqrt{2\pi\sigma^2_{AB}}} e^{-\frac{x^2}{2\sigma^2_{AB}}} \mathrm{d}x = \int_{-\infty}^{\mu_{AB}} \frac{1}{\sigma_{AB}} \phi\left(\frac{x}{\sigma_{AB}}\right) \mathrm{d}x = \int_{-\infty}^{\frac{\mu_{AB}}{\sigma_{AB}}} \phi(t)\mathrm{d}t = \Phi\left(\frac{\mu_{AB}}{\sigma_{AB}}\right) \end{aligned} \tag{2.13}$$

where Φ is the standard normal CDF.

The mean quality difference, μ_{AB}, can then be obtained by inverting Equation (2.13)

$$\mu_{AB} = \sigma_{AB}\Phi^{-1}(P(A > B)) \tag{2.14}$$

where Φ^{-1} is the inverse CDF of the standard normal. Thurstone proposed to estimate the probability $P(A > B)$ by the empirical proportion of people preferring A and B

$$\hat{\mu}_{AB} = \sigma_{AB}\Phi^{-1}\left(\frac{C_{AB}}{C_{AB} + C_{BA}}\right) \tag{2.15}$$

where $\hat{\mu}_{AB}$ is an estimator of the true mean difference μ_{AB}. In the Thurstone Case V model, it is further assumed that the two options have equal variance and zero correlation, i.e., $\sigma_A = \sigma_B$ and $\rho_{AB} = 0$. Without any loss of generality, the variances can be set to one half, meaning that the quality score values, μ_A and μ_B, can be estimated as

$$\mu_A - \mu_B = \Phi^{-1}\left(\frac{C_{AB}}{C_{AB} + C_{BA}}\right) \tag{2.16}$$

If multiple stimuli are compared, then a maximum likelihood estimation of the quality score values should be performed to consider the interactions between the different pairs (Tsukida and Gupta, 2011). The log-likelihood function is

$$\mathscr{L}\left(\mu|C\right) = \sum_{i,j} C_{ij} \log\left[\Phi\left(\mu_i - \mu_j\right)\right] \tag{2.17}$$

To find the maximum likelihood solution quality scale values, one must solve

$$\arg\max_{\Delta\mu} \mathscr{L}\left(\Delta\mu|C,\mu\right) \quad \text{subject to} \quad \sum_i \mu_i = 0 \tag{2.18}$$

To help regularize the estimates, a prior of 1 can be added on all the counts, meaning that *a priori* all choices are possible, which corresponds to Laplace smoothing (Tsukida and Gupta, 2011).

In the Bradley-Terry-Luce model, the probability of choosing A over B is defined as

$$P(A > B) = \frac{\pi_A}{\pi_A + \pi_B} \tag{2.19}$$

where π_i satisfying $\pi_i \geq 0$ and $\sum_i \pi_i = 1$ can be considered as the quality score for stimulus i.

By changing variables $\pi_i = e^{\frac{\mu_i}{s}}$, where s is a scale parameter, Equation (2.19) can be rewritten as

$$P_{AB} = \frac{e^{\frac{\mu_A}{s}}}{e^{\frac{\mu_A}{s}} + e^{\frac{\mu_B}{s}}} = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{\mu_A - \mu_B}{2s}\right) = F_{B-A}(0) = 1 - F_{A-B}(0) \tag{2.20}$$

where F is the logistic CDF. Thus, it is consistent to assume that the random quality difference A-B is a logistic random variable with mean $\mu_A - \mu_B$ and scale parameter s.

The mean quality difference, μ_{AB}, can then be obtained by inverting Equation (2.20)

$$\begin{aligned} \mu_{AB} &= 2s\tanh^{-1}(2P(A > B) - 1) = s\left[\ln(P(A > B)) - \ln(1 - P(A > B))\right] \\ &= s\left[\ln(P(A > B)) - \ln(P(B > A))\right] \end{aligned} \tag{2.21}$$

since $\tanh^{-1}(x) = \frac{1}{2}\left[\ln(1+x) - \ln(1-x)\right]$. The probability $P(A > B)$ can be estimated by the empirical proportion of people preferring A and B

$$\hat{\mu}_{AB} = s\left[\ln\left(\frac{C_{AB}}{C_{AB} + C_{BA}}\right) - \ln\left(\frac{C_{BA}}{C_{AB} + C_{BA}}\right)\right] \tag{2.22}$$

where $\hat{\mu}_{AB}$ is an estimator of the true mean difference μ_{AB}.

Similarly to the Thurstone model, a maximum likelihood estimation of the quality score values is performed if more than two stimuli are compared.

Note that for both models, the estimated MOSs are defined up to a scaled factor. Thus, it is common to normalized them to the range $[0, 100]$ for a better representation.

Confidence Intervals Estimation

The preference scores collected during a PC experiment can also be used to estimate CIs. Some techniques for example use the Hessian matrix of the log-likelihood function used to estimate the MOSs. Another technique was proposed by J.-S. Lee et al. (2011) and assumes that ties convey information about significant differences between two stimuli being compared. The original method was proposed for the Bradley-Terry-Luce model. In the following, we described an extension of this technique for the Thurstone Case V model.

First, the lower and upper bounds of the count matrix of stimulus i, C^-_{ij} and C^+_{ij}, are computed as

$$C^-_{ij} = w_{ij} \qquad C^+_{ij} = w_{ij} + t_{ij} \tag{2.23}$$

assuming that the ties have been the preferences of stimulus j or i, respectively.

Then, the CI $\left[\mu_i - \Delta\mu^-_i, \mu_i + \Delta\mu^+_i\right]$ related to the quality score value for stimulus i is estimated based on the Thurstone Case V model. If only two stimuli, A and B, are compared, the lower and upper errors, $\Delta\mu^-$ and $\Delta\mu^+$, are given by

$$\begin{aligned}
\left(\mu_A - \Delta\mu^-_A\right) - \left(\mu_B + \Delta\mu^+_B\right) &= \Phi^{-1}\left(\frac{C^-_{AB}}{C^-_{AB} + C^+_{BA}}\right) \\
\left(\mu_A + \Delta\mu^+_A\right) - \left(\mu_B - \Delta\mu^-_B\right) &= \Phi^{-1}\left(\frac{C^+_{AB}}{C^+_{AB} + C^-_{BA}}\right)
\end{aligned} \tag{2.24}$$

where μ_A and μ_B are the quality score values for stimulus A and B, respectively, estimated considering ties as being half way between the two options.

If multiple stimuli are compared, then a maximum likelihood estimation of the errors is performed. The log-likelihood function is

$$\mathscr{L}\left(\Delta\mu | C, \mu\right) = \sum_{i,j} C^-_{ij} \log\left\{\Phi\left[\left(\mu_i - \Delta\mu^-_i\right) - \left(\mu_j + \Delta\mu^+_j\right)\right]\right\} + \sum_{i,j} C^+_{ij} \log\left\{\Phi\left[\left(\mu_i + \Delta\mu^+_i\right) - \left(\mu_j - \Delta\mu^-_j\right)\right]\right\} \tag{2.25}$$

where μ_i is the quality score values for stimulus i estimated considering ties as being half way between the two options. To find the maximum likelihood solution quality scale values, one must solve

$$\arg\max_{\Delta\mu} \mathscr{L}\left(\Delta\mu | C, \mu\right) \quad \text{subject to} \quad \Delta\mu_i \geq 0 \quad \forall i \tag{2.26}$$

The proof that the modified definitions of Thurstone's Law for the lower and upper counts (see Equation (2.24)) yield the maximum likelihood solution (see Equation (2.25)) for two stimuli is given in Appendix A.

To help regularize the estimates, a prior of 1 is added on all the counts, meaning that *a priori* all choices are possible, which corresponds to Laplace smoothing (Tsukida and Gupta, 2011).

2.6.4 Relationship Between Estimated Mean Values

To determine whether the difference between two MOS values is statistically significant, a two-sample unpooled t-test can be performed as the score distributions have unknown and unequal variances. If the observed value is inside the critical region determined by the 95% two-tailed Student's t-distribution, then the two MOS values are considered to be statistically different at a 5% significance level.

When comparing several groups of MOSs, the chance of incorrectly finding a significant difference would increase with the number of comparisons if a simple t-test is performed for each comparison. To overcome this problem, a multiple comparison procedure should be applied instead (Snedecor and Cochran, 1989). These procedures are designed to provide an upper bound on the probability that any comparison will be incorrectly found significant. An analysis of variance (ANOVA) can be used to compare groups of MOSs, but also to evaluate the significance of the test parameters.

When using the PC and SC methods (see Section 2.4.3), a statistical hypothesis test can be applied on the ratings to determine whether the preference for one stimulus over the other is statistically significant. First, the data need to be arranged in only two classes, for example by splitting ties equally between the two preference options. This data roughly follows a Bernoulli process $B(N, p)$, where N is the number of subjects and p is the probability of success in a Bernoulli trial, which is set to 0.5, considering that, a priori, both options have the same chance of success. The binomial CDF is then used to determine the critical region for the statistical test.

The Barnard's test (Barnard, 1945) can also be used to determine whether preference for one stimulus over the other is statistically significant. This test is a statistical significance test of the null hypothesis of independence of rows and columns in a 2×2 contingency table. It is claimed that the Barnard's test is more powerful than Fisher's exact test for contingency tables. Thus, this statistical test can be used to test whether the preference probability is statistically significantly different from 0.5.

2.7 Comparing MOS Values of Different Experiments

Comparing results of two subjective experiments conducted with the same test material but with different conditions is essential. One goal can be to investigate the influence of different factors, e.g., viewing distance, lighting conditions, display, test methodology, or rating scale. There are plenty of studies investigating these aspects in different scenarios, for example in the context of 3D evaluations (Barkowsky et al., 2013; Brunnström et al., 2013;

Kulyk et al., 2013; J. Li et al., 2013c; Perkis et al., 2012). Recently, with the increasing popularity of crowdsourcing based quality assessment, several studies have also been conducted to investigate the correlation between experiments conducted in laboratory environment and through crowdsourcing (Hossfeld et al., 2014a; Hossfeld and Keimel, 2014; Keimel et al., 2012; Redi et al., 2013; Ribeiro et al., 2011).

Even if the same experiment is conducted on the same subject, with the same conditions and presentation order, the scores given by the subject will never be exactly the same. This can be interpreted as some noise overlaid on the results. Then, short-term context will impact grading; this effect is commonly referred to as presentation order effect. Using a different presentation order for each subject can be used to balance out this effect, but the statistical uncertainty remains. Medium and long-term context will also impact grading. For example, if an experiment contains mainly low quality stimuli, then subjects tend to score them higher, and vice versa. This effect is due the fact that people tend to use the whole range of the rating scale during the experiment, despite the labels associated with the scale. Finally, there are long term dependencies that reflect the general cultural behavior of the subject, e.g., interpretation of the category labels, attitude to quality, or language. Experience with multimedia technologies is also a factor and expectations may change over time. Differences between experiments can be due to these effects, but they can also be minimized by a proper training, well-balanced design and mixed pseudo-random display orders, and by considering enough subjects.

In the following subsections, a procedure to compare MOS values of different experiments is proposed. This procedure is inspired by the standard procedure used to benchmark objective quality metrics described in recommendations ITU-T P.1401 (2012) and ITU-T J.149 (2004). The main difference is that instead of comparing objective to subjective results, two groups of subjective results are compared.

2.7.1 Mapping Subjective Scores of Two Experiments

On top of normal uncertainties described above, systematically observed differences can be classified as

i) Bias (or offset): a bias consists of a constant offset between MOS values and can be due to the overall quality of all stimuli, which can influence subjects to score more pessimistically or more optimistically. For example, if the same experiment with compressed images is performed in a country where people have very fast internet access, e.g., fiber-optic, and very high quality displays, e.g., UHD, then the scores might be generally lower than for subjects coming from a country where internet access is very slow, e.g., dial-up, and with standard monitors, e.g., VGA resolution. Different environments and displays can also be the source of bias. However, an offset is usually observed in conjunction with a gradient difference.
ii) Gradient difference: a gradient difference is observed when scores tend to become more

pessimistic faster in one experiment than in the other. This effect can be observed when the test stimuli do not cover the whole quality range.

iii) Ranking difference: a ranking difference occurs when the ranking of some stimuli is different from one experiment to the other. This is the most sever problem, as the goal of most quality assessment experiment is to determine the relative ranking of the systems under test.

These effects can be observed in a scatter plot showing the MOSs of experiment B versus the MOSs of experiment A (or vice versa). To remove the bias and gradient difference, a simple linear mapping can be applied to align the scores of one experiment to that of the other experiment. However, in general, a third order polynomial mapping is used for the mapping (ITU-T P.1401, 2012), as it will reduce some ranking difference, for example when the data on the scatter plot forms a banana shape.

2.7.2 Statistical Evaluation Metrics

Statistical evaluation metrics are used to estimate the linearity, monotonicity, accuracy, and consistency between two groups of MOS values corresponding to two different experiments. In particular, one group of MOS values corresponds to the MOS values of experiment A, MOS^{ExpA}, while the second group corresponds to the MOS values of experiment B, MOS^{ExpB}, mapped to those of experiment A, $\widehat{MOS}^{ExpB}$, considering a third order polynomial mapping (see Section 2.7.1). As the mapping of MOS^{ExpB} to MOS^{ExpA} yields slightly different results when compared to mapping of MOS^{ExpA} to MOS^{ExpB}, both mappings should be considered and results should be reported for both cases. The Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SROCC) are computed between the two groups of MOS values to estimate linearity and monotonicity, respectively. Accuracy and consistency are estimated using the root-mean-square error (RMSE) and outlier ratio (OR), respectively. Note that none of these metrics takes into account the subjective uncertainty.

Pearson Correlation Coefficient

The Pearson correlation coefficient (PCC) is a measure of the linear correlation between two variables X and Y. The resulting value is in the range $[-1,1]$, where -1 corresponds to a total positive correlation, 0 to no correlation, and 1 to a total positive correlation. The PCC is computed as

$$\mathrm{PCC} = \frac{\sum_{i=1}^{M}\left(X_i-\overline{X_i}\right)\left(Y_i-\overline{Y_i}\right)}{\sqrt{\sum_{i=1}^{M}\left(X_i-\overline{X_i}\right)^2}\sqrt{\sum_{i=1}^{M}\left(Y_i-\overline{Y_i}\right)^2}} \tag{2.27}$$

where M is the total number of points.

The PCC is computed to estimate the linearity between the two groups of MOS values.

Spearman Rank Order Correlation Coefficient

The Spearman's rank correlation coefficient (SROCC) is a nonparametric measure of statistical dependence between two variables X and Y, assessing how well their relationship can be described using a monotonic function. The resulting value is in the range $[-1,1]$, and, if there are no repeated values, a value of ± 1 indicates a perfect monotone function. The SROCC is computed as

$$\text{SROCC} = \frac{\sum_{i=1}^{M} \left(x_i - \overline{x_i}\right)\left(y_i - \overline{y_i}\right)}{\sqrt{\sum_{i=1}^{M} \left(x_i - \overline{x_i}\right)^2} \sqrt{\sum_{i=1}^{M} \left(y_i - \overline{y_i}\right)^2}} \tag{2.28}$$

where x_i and x_i denote the ranked variables and M is the total number of points.

The SROCC is computed to estimate the monotonicity between the two groups of MOS values.

Root Mean Square Error

The root-mean-square error (RMSE) of the absolute prediction error computed between MOS^{ExpA} and $\widehat{MOS}^{ExpB}$ is defined as

$$\text{RMSE} = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} \left(MOS_i^{ExpA} - \widehat{MOS}_i^{ExpB}\right)^2} \tag{2.29}$$

where M is the total number of points. Note that the division by $M-1$ corresponds to the unbiased estimator for the RMSE.

The RMSE is computed to estimate the accuracy between the two groups of MOS values.

Outlier Ratio

The outlier ratio (OR) represents the ratio of the number of outlier-points divided by the total number of points

$$\text{OR} = \frac{\text{total number of outliers}}{M} \tag{2.30}$$

where M is the total number of points and an outlier is defined as a point i for which the 95% CIs do not overlap

$$\left|MOS_i^{ExpA} - \widehat{MOS}_i^{ExpB}\right| > \delta_i^{ExpA} + \delta_i^{ExpB} \tag{2.31}$$

where δ_i^{ExpA} and δ_i^{ExpB} are related to the 95% CIs (see Section 2.6.2) corresponding to MOS_i^{ExpA} and $\widehat{MOS}_i^{ExpB}$, respectively.

The OR is computed to estimate the consistency between the two groups of MOS values.

2.7.3 Estimation Errors

Another way to analyze the results of two experiments is to consider the outcome of one experiment as ground truth and to record the number of times the other experiment underestimates or overestimates the results. For each condition, the two groups of scores are compared according to Section 2.6.4 and the percentage of *Correct Estimation*, *Underestimation*, and *Overestimation* are recorded. Note that it important to align the data following the procedure described in Section 2.7.1, otherwise any systematic error, e.g., offset, will impact the results. The process can be repeated considering the outcome of the other experiment as ground truth.

2.7.4 Classification Errors

In recommendation ITU-T J.149 (2004), it is suggested to compute the classification errors to evaluate the performance of an objective metric. A classification error is made when the objective metric and subjective test lead to different conclusions on a pair of images or video sequences (see Section 9.5). Here, we extend this methodology to the comparison of a pair of subjective tests, *X* and *Y*, evaluated in a reference and test experiments. Three types of error can happen

a) *False Tie*, the least offensive error, which occurs when the reference experiment says that *X* and *Y* are different, whereas the test experiment says that they are identical,
b) *False Differentiation*, which occurs when the reference experiment says that *X* and *Y* are identical, whereas the test experiment says that they are different,
c) *False Ranking*, the most offensive error, which occurs when the reference experiment says that *X* (*Y*) is better than *Y* (*X*), whereas the test experiment says the opposite.

The two groups of scores are compared according to Section 2.6.4 and the percentage of *Correct Decision*, *False Tie*, *False Differentiation*, and *False Ranking* are recorded from all possible distinct pairs of conditions. Note that unlike the estimation error (see Section 2.7.3), the data of the two experiments should not be aligned, as there is no direct comparison between the two experiments. The process can be repeated considering the outcome of the other experiment as ground truth.

2.7.5 Comparing Paired Comparison Data of Different Experiments

In Section 2.6.4, the Barnard's test (Barnard, 1945) is proposed to determine whether the preference probability when comparing two stimuli in a paired comparison fashion is statistically significantly different from 0.5. This statistical test can also be used to determine whether the difference between two preference probabilities, corresponding to the same pair of stimuli but evaluated in different conditions, is statistically significant. The Barnard's test can be applied to all pairs to record the number of times results significantly differ between the two conditions.

To determine whether the difference between the two conditions has a significant impact on the results, J. Li et al. (2013b) proposed to conduct a Monte Carlo simulation. At each simulation, a group of results is randomly permuted between the two conditions and the ratio of significantly different pairs is recorded. With a sufficiently large number of simulations, e.g., 1000, the distribution of the ratio can be estimated. If the observed ratio is higher than the 95th percentile, then it can be assumed that the influence of the difference between the two conditions is statistically significant.

The classification errors (see Section 2.7.4) can also be computed for paired comparison data. In this case, the comparison of two groups of scores is made using the Barnard's test.

2.8 Analysis of Eye Tracking Data

Eye tracking experiments are conducted to record eye movements from individual subjects in various applications. In the context of quality assessment, eye movements are used for example to investigate the impact of visual artifacts, e.g., compression artifacts, on visual attention (Le Meur et al., 2010a; Ninassi et al., 2006) or to improve the performance of quality metrics by considering the probability of watching a specific part of the image or video sequence (Engelke et al., 2011; Le Meur et al., 2010b; H. Liu and Heynderickx, 2011; Ninassi et al., 2007).

Eye movements are classified into two categories: fixations and saccades. Fixations last from about 100 to 600 ms and allow the brain to process the visual information received by the eyes. Saccades are fast jumps between two fixations and last from about 20 to 40 ms, while the eye velocity can be up to 600 °/s. Information from a scene is mainly acquired during fixations, whereas vision is largely suppressed during saccades. Thus, eye movements corresponding to saccades should be removed. Similarly, eye movements recorded during blinks should also be removed.

Typically, gaze points associated with gaze velocity below a fixation threshold are classified as fixation points, while saccades are detected when the gaze velocity lies above the fixation threshold. Blinks can also be detected automatically based on the distance between the two eyelids of each eye. Most modern eye tracking systems automatically classify gaze points as

fixation, saccade, or blink.

From all gaze points labeled as fixation points, a FDM is recorded. This map is an estimate of the probability of watching a particular pixel in the image or frame of a video sequence. The following subsections describe how to compute the FDM and how to compare two FDMs using different statistical evaluation metrics.

2.8.1 Computation of Fixation Density Maps

A fixation density map (FDM) is computed by convolving the recorded gaze points with a Gaussian filter, and then normalizing the result to values between 0 and 1. The FDM is an estimate of the probability of watching a particular pixel in the image or frame of a video sequence. Only gaze points corresponding to fixation points are used to compute a FDM. Gaze points associated with saccades and blinks are not used in the computation. In the case of still images, all gaze points recorded from all subjects during the presentation of a particular image are used to compute the FDM of that particular image. For video sequences, this process is performed for each frame independently and only the gaze points recorded during the presentation of that particular frame are used, which requires a perfect synchronization between the eye tracker and the video playback system.

The fixation points are filtered with a Gaussian kernel to compensate the eye tracker inaccuracies and to simulate the foveal point spread function of the human eye. As suggested in the state of the art (Engelke et al., 2009; Judd et al., 2012), the standard deviation of the Gaussian filter used for computing the FDMs should be set to 1 degree of visual angle. This standard deviation value is based on the assumption that the fovea of the human eye covers approximately 2 degrees of visual angle.

2.8.2 Statistical Evaluation Metrics

Although several statistical evaluation metrics have been proposed to measure the similarity between two FDMs, there is no standardized procedure. Typically, the similarity score and Kullback-Leibler divergence (KLD) are used to measure the similarity between two FDMs. Additionally, the attentional focus (Jermann et al., 2012) can be computed to determine whether subjects look at few objects or more or less uniformly at several objects.

Similarity Score

The similarity score is a distribution-based metric of how similar two FDMs are. The similarity score S between two normalized maps P and Q is computed as

$$S = \sum_{i,j} \min(P_{i,j}, Q_{i,j}) \quad \text{where} \quad \sum_{i,j} P_{i,j} = \sum_{i,j} Q_{i,j} = 1 \tag{2.32}$$

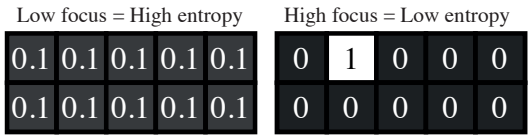


Figure 2.2: Illustration of attentional focus metric.

A similarity score of 1 means that the two FDMs are the same, whereas 0 indicates that there is no overlap between them.

Kullback-Leibler Divergence

The Kullback-Leibler divergence (KLD) is usually used to estimate the dissimilarity between two probability distributions. In the context of FDMs, this is a measure of dissimilarity between two histograms. If, in the corresponding histograms, $p(x)$ and $q(x)$ represent the probabilities of a pixel to have value x, the symmetric KLD is computed as

$$\mathrm{KLD} = \frac{1}{2}\sum_{x}\left[p(x)\log\frac{p(x)}{q(x)} + q(x)\log\frac{q(x)}{p(x)}\right] \tag{2.33}$$

When two probability distributions are strictly equal, the KLD value is 0, and when histograms do not overlap at all, it tends to infinity.

Attentional Focus

The attentional focus (Jermann et al., 2012) is defined as the number of objects that are viewed by the subjects during image observation. The rationale is to distinguish between cases where subjects look at few objects versus cases where they look more or less uniformly at several objects. To compute attentional focus, the FDM is first partitioned into blocks of $N \times N$ pixels. Then, the average intensity is computed for each block. Finally, the attentional focus is computed as the entropy of the normalized intensity across different blocks. Low entropy indicates high attentional focus while high entropy indicates low attentional focus. Figure 2.2 shows a schematic representation of this concept. The size of the blocks is determined so as to match the size of fovea, corresponding to 2 degrees of visual angle.

2.9 Conclusion

This chapter provided a detailed description of the different factors that have to be taken into account when designing subjective experiments. From the selection of the test material to the screening of subjects, we reviewed the guidelines suggested by the relevant international recommendations, as well as some common practices. These factors were considered in the different subjective experiments reported in the rest of the thesis. We have presented the

common procedure used to compute MOSs and CIs, as well as alternatives procedures for paired comparison methods, including a novel method to estimate CIs for the Thurstone Case V model. Statistical tools to compare two groups of subjective scores were described for MOSs and PC data. This chapter also provided a description of the recommended procedure to compare results of subjective experiments and some novel procedures that we proposed, which were inspired from the procedures used to benchmark objective quality metrics. Finally, a brief overview of the common techniques used to analyze eye movement data recorded with an eye tracker was presented. These procedures were used to process and analyze the subjective data collected in the rest of the thesis.

3 Calculation of Coding Efficiency

"If you can not measure it, you can not improve it" (Lord Kelvin). This statement is especially true in the case of image and video compression. To design efficient compression algorithms, it is necessary to benchmark the performance of new algorithms against well-established and state-of-the-art algorithms on a dataset containing different contents. The quality of the compressed images and video sequences can be assessed by means of objective and subjective evaluations. Objective quality assessment relies on the use of objective quality metrics (see Chapter 8), which have been designed to predict the perceived quality of media content. Objective evaluations based on PSNR measurements are widely used by most researchers and coding experts as they are simple and can be performed automatically. However, it is known that PSNR does not accurately reflect human perception of visual quality (Sheikh et al., 2006). Nevertheless, previous studies (Huynh-Thu and Ghanbari, 2008; Huynh-Thu and Ghanbari, 2012; Korhonen and You, 2012) have shown that the PSNR metric is reliable as long as the content is not changed. In the case of subjective quality assessment, the quality of the decoded data is evaluated by a pool of human subjects (typically more than 15 people), following a common methodology (see Chapter 2). Subjective tests are undeniably the most accurate means to evaluate quality, as measurements are performed by human observers. However, they are time consuming, expensive, and not always feasible. Moreover, for codec optimization, where several parameters can be tuned to improve quality, subjective evaluations are impractical.

To calculate the coding efficiency between different codecs based on PSNR measurements, a model was proposed by Gisle Bjøntegaard (2001) during the development of AVC. The Bjøntegaard model is used by various experts to calculate the coding efficiency of compression standards. For example, this model was used during the development of AVC (Wiegand et al., 2003b), MVC (Merkle et al., 2007c), HEVC (Ohm et al., 2012), and MV-HEVC (Vetro and Tian, 2012). The Bjøntegaard model is also widely used by researchers working on image and video compression to benchmark the performance of their algorithms against well-established and state-of-the-art compression algorithms. The Bjøntegaard model is used to calculate the average PSNR and bit rate differences between two R-D curves obtained from the PSNR

measurement when encoding a content at different bit rates. The model reports two values

i) the Bjøntegaard delta PSNR (BD-PSNR), which corresponds to the average PSNR difference in dB for the same bit rate,
ii) the Bjøntegaard delta rate (BD-Rate), which corresponds to the average bit rate difference in percent for the same PSNR.

Section 3.1 describes in details the Bjøntegaard model and how these values are computed.

To investigate the impact on quality of the interaction of the base and enhancement layers bit rates when comparing two-layer coding systems, the simple Bjøntegaard model cannot be used because it considers only one bit rate. Therefore, we propose an extension of the Bjøntegaard model from R-D curve fitting to R^2-D surface fitting. Section 3.2 provides a detailed description of the proposed model and some examples of application.

The Bjøntegaard model might not be an accurate predictor of the true coding efficiency as it relies on PSNR measurements. To estimate a more realistic coding efficiency, subjective quality scores should be considered instead of PSNR measurements. Therefore, we propose a model to calculate the average coding efficiency based on MOSs gathered during subjective evaluations instead of PSNR measurements. Section 3.3 provides a detailed description of the proposed model and some examples of application.

3.1 The Bjøntegaard Model

Gisle Bjøntegaard (2001) has proposed a model to measure the coding efficiency between two different compression algorithms. To approximate a R-D curve given by a set of N bit rate values $(R_1, ..., R_N)$ with corresponding PSNR measurements $(D_1, ..., D_N)$, a third order logarithmic polynomial fitting has been proposed in the Bjøntegaard model, based on experimental observations

$$\hat{D}(R) = a\log^3 R + b\log^2 R + c\log R + d \tag{3.1}$$

where $\hat{D}$ is the fitted distortion in PSNR, R is the bit rate, and a, b, c, and d are the parameters of the fitting function.

To simplify notation, in the rest of the chapter, we use lower case r when referring to the logarithm of the bit rate, i.e., $r = \log R$. Therefore, Equation (3.1) is rewritten as

$$\hat{D}(r) = ar^3 + br^2 + cr + d \tag{3.2}$$

At least four R-D values are required to determine the fitting parameters of Equation (3.2). If more than four values are used, then the R-D values are fitted in a least square sense.

The average PSNR difference between two R-D curves is approximated by the difference

between the integrals of the fitted R-D curves divided by the integration interval (Bjøntegaard, 2001)

$$\Delta D = \mathrm{E}\left[D_2 - D_1\right] \approx \frac{1}{r_H - r_L} \int_{r_L}^{r_H} \left[\hat{D}_2(r) - \hat{D}_1(r)\right] \mathrm{d}r \tag{3.3}$$

where ΔD is BD-PSNR computed between the two fitted R-D curves $\hat{D}_1(r)$ and $\hat{D}_2(r)$, respectively, and the integration bounds, r_L and r_H, are

$$\begin{aligned} r_L &= \max\left\{\min(r_{1,1}, ..., r_{1,N_1}), \min(r_{2,1}, ..., r_{2,N_2})\right\} \\ r_H &= \min\left\{\max(r_{1,1}, ..., r_{1,N_1}), \max(r_{2,1}, ..., r_{2,N_2})\right\} \end{aligned} \tag{3.4}$$

To express the (logarithm of the) rate as a function of the distortion, a third order polynomial fitting has been proposed in the Bjøntegaard model to fit the R-D values

$$\hat{r}(D) = aD^3 + bD^2 + cD + d \tag{3.5}$$

Note that a second fitting process is required to fit the bit rate values and that $\hat{r}(D)$ (see Equation (3.5)) is not the inverse function of $\hat{D}(r)$ (see Equation (3.2)).

The average bit rate difference between two R-D curves is approximated as (Bjøntegaard, 2001)

$$\Delta R = \mathrm{E}\left[\frac{R_2 - R_1}{R_1}\right] = \mathrm{E}\left[\frac{R_2}{R_1}\right] - 1 = \mathrm{E}\left[10^{r_2 - r_1}\right] - 1 \approx 10^{\mathrm{E}[r_2 - r_1]} - 1 \approx 10^{\frac{1}{D_H - D_L} \int_{D_L}^{D_H} [\hat{r}_2(D) - \hat{r}_1(D)] \mathrm{d}D} - 1 \tag{3.6}$$

where ΔR is the BD-Rate computed between the two fitted R-D curves $\hat{r}_1(r)$ and $\hat{r}_2(r)$, respectively, and the integration bounds, D_L and D_H, are

$$\begin{aligned} D_L &= \max\left\{\min(D_{1,1}, ..., D_{1,N_1}), \min(D_{2,1}, ..., D_{2,N_2})\right\} \\ D_H &= \min\left\{\max(D_{1,1}, ..., D_{1,N_1}), \max(D_{2,1}, ..., D_{2,N_2})\right\} \end{aligned} \tag{3.7}$$

Thanks to the logarithmic bit rate scale, the estimation of the average bit rate reduction is also simplified.

3.2 Extension for Two-Layer Coding Systems

In the recent years, layered coding (Ghanbari, 1989) has gained a large popularity in the image and video compression community. Multilayer coding systems partition the information between one base layer and one or more enhancement layers. This approach is typically used for scalable coding, where the enhancement layers can provide spatial, temporal, or quality

improvements when compared to the base layer. Additional scalable features, e.g., bit depth, color gamut, or hybrid coding, can also be implemented. Several standards, e.g., JPEG 2000 (Skodras et al., 2001), scalable video coding (scalability video coding extensions of AVC) (SVC) (Schwarz et al., 2007), and scalability extensions of HEVC (SHVC) (Boyce et al., 2016) rely on layered coding to provide scalability. Backward compatibility is another feature that can be implemented using two-layer coding: the base layer is encoded using a legacy encoder for backward compatibility, whereas the enhancement layer is encoded using a different and optimized coding scheme. JPEG XT (Artusi et al., 2015) and multi-resolution frame-compatible (MFC) stereo coding (Lu et al., 2013) are examples of backward compatible standards using two-layer coding.

X. Li et al. (2010) have proposed an extension of the Bjøntegaard model, referred to as generalized BD-PSNR, to take coding complexity into account. However, neither this model nor the Bjøntegaard model can be used to investigate the impact on quality of the interaction of the base and enhancement layers bit rates when comparing two-layer coding systems. Therefore, we propose an extension of the Bjøntegaard model from R-D curve fitting to R^2-D surface fitting. The proposed model uses a cubic surface as fitting function. While the generalized BD-PSNR model (X. Li et al., 2010) only considers a rectangular domain in the RC-plane to evaluate the delta PSNR, the proposed model uses a more complex characterization of the domain formed by the data points to compute a more realistic estimate of the compression efficiency.

3.2.1 Proposed Model

In this subsection, we propose an extension of the Bjøntegaard model for measuring the compression efficiency between two R^2-D surfaces. First, the function used to fit the R^2-D surfaces is described. Then, the calculation of average PSNR and bit rate differences between two fitted R^2-D surfaces is presented. A MATLAB implementation of the proposed model can be downloaded from: `http://mmspg.epfl.ch/2dbd`

Fitting Function

The Bjøntegaard model uses a cubic function to fit the R-D curve, based on the observation that R-D values expressed in $\big(\log(\text{bit rate}), \text{PSNR}\big)$ do not deviate much from straight lines (Bjøntegaard, 2008). Following the same principle, we propose to use a cubic surface to fit the R^2-D surface. The cubic surface is given by

$$z(x,y) = \sum_{(i,j)\in S} p_{ij} x^i y^j \qquad\qquad S = \big\{(i,j) \in \mathbb{N}^2 | i + j \leq 3\big\} \tag{3.8}$$

where p_{ij} are the parameters of the fitting function.

The cross-terms, i.e., $p_{11}xy$, $p_{21}x^2y$, and $p_{12}xy^2$, allow more flexibility for the fitting of the

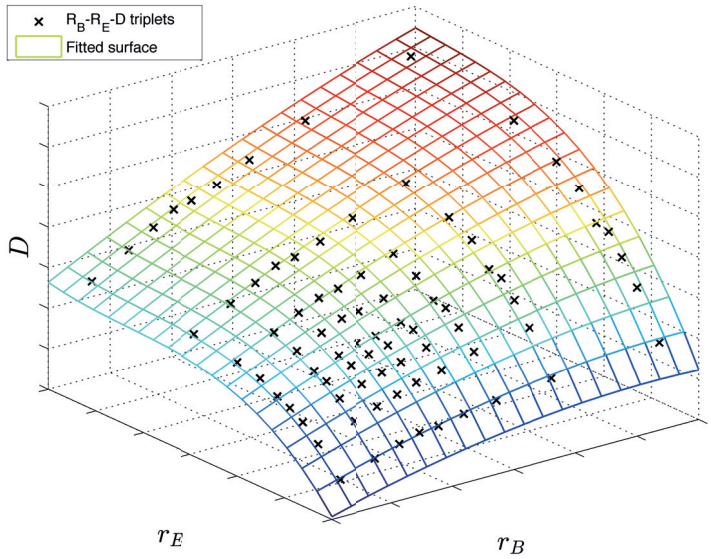


Figure 3.1: R_B-R_E-D surface fitting.

R^2-D surface, which improves the goodness of the fit, but increases the number of required data points. At least ten (x, y, z) triplets are required to determine the fitting parameters of Equation (3.8). If more than ten triplets are used, then the data points are fitted in a least square sense. However, in practice, to obtain a more realistic estimate of the performance evaluation, 16 or more triplets should be used. Figure 3.1 shows the fitting result for one HDR image encoded with JPEG XT. As it can be observed, the fitting accuracy is quite good.

Average PSNR Difference

The R^2-D surface is obtained by varying one parameter of the base and enhancement layers in coding schemes while measuring the PSNR of the reconstructed image or video sequence. Considering M base layer parameter values $(P_{B,1}, ..., P_{B,M})$ and N enhancement layer parameter values $(P_{E,1}, ..., P_{E,N})$, this yields to a set of $M \times N$ base layer bit rate values $(R_{B,11}, ..., R_{B,MN})$ and enhancement layer bit rate values $(R_{E,11}, ..., R_{E,MN})$ with corresponding PSNR values $(D_{11}, ..., D_{MN})$. The corresponding R^2-D surface is fitted in a least square sense using a cubic surface

$$\hat{D}(r_B, r_E) = \sum_{(i,j)\in S} p_{ij} {r_B}^i {r_E}^j \qquad S = \{(i,j) \in \mathbb{N}^2 | i + j \leq 3\} \tag{3.9}$$

where $\hat{D}$ is the fitted distortion in PSNR, r_B and r_E are the logarithms of the base and enhancement layers bit rates, respectively, and p_{ij} are the parameters of the fitting function.

Similarly to the Bjøntegaard model, the average PSNR difference between two R^2-D surfaces is approximated by the difference between the integrals of the fitted R^2-D surfaces divided by

the area of the integration domain

$$\Delta D = \mathrm{E}\left[D_2 - D_1\right] \approx \frac{1}{|A|} \iint_A \left[\hat{D}_2(r_B, r_E) - \hat{D}_1(r_B, r_E)\right] \mathrm{d}r_B \mathrm{d}r_E \tag{3.10}$$

where ΔD is the delta PSNR computed between the two fitted R^2-D surfaces $\hat{D}_1(r_B, r_E)$ and $\hat{D}_2(r_B, r_E)$, respectively, and the integration domain A is given by the intersection of the domains on which the two R^2-D surfaces are fitted

$$A = A_1 \cap A_2 \tag{3.11}$$

Figure 3.2 illustrates the (r_B, r_E, D) triplets projected on the $r_B r_E$-plane. The data points form a domain defined by four corners corresponding to the extrema of P_B and P_E (see Figure 3.2). The domain is delimited by the four contours connecting the four corners. The contour which starts at I and ends at J is defined by the pairs $\big((r_{B,11}, r_{E,11}), ..., (r_{B,M1}, r_{E,M1})\big)$. We propose to fit these pairs with a cubic curve to estimate the contour

$$\hat{r}_E(r_B) = a{r_B}^3 + b{r_B}^2 + c r_B + d \tag{3.12}$$

The same principle is applied to estimate the three remaining contours, with the exception that the contours between I and K and between J and L are expressed as a function of r_E.

The domain delimited by the four contours (represented in gray in Figure 3.2) is thus defined as

$$A = \left\{(r_B, r_E) \in \mathbb{R}^2 | \beta_{\min}(r_E) \leq r_B \leq \beta_{\max}(r_E), \epsilon_{\min}(r_B) \leq r_E \leq \epsilon_{\max}(r_B)\right\} \tag{3.13}$$

where the functions β and ϵ are extensions of the contour fitting functions that simply perform repetition for points that lie outside the range of fitted values (as illustrated by the dashed lines in Figure 3.2). For example, $\epsilon_{\min}$ is the extension of the contour which starts at I and ends at J

$$\epsilon_{\min}(r_B) = \begin{cases} \hat{r}_E(r_{B,11}) & \text{if } r_B < r_{B,11} \\ \hat{r}_E(r_{B,M1}) & \text{if } r_B > r_{B,M1} \\ \hat{r}_E(r_B) & \text{otherwise} \end{cases} \tag{3.14}$$

The same principle is used for the other extensions.

The domain on which the R^2-D surface is fitted is determined independently for both surfaces following the procedure described here above. Then, the integral is evaluated on the intersection of the two domains. Even though the analytical form of the integral can be easily determined, its evaluation would require a complex parameterization of the integration bounds. Therefore, the integral is approximated using a finite sum. Note that in the generalized BD-PSNR (X. Li et al., 2010) model, the integration domain corresponds to a rectangular domain defined by the extreme values (as represented by the hatched area in Figure 3.2).

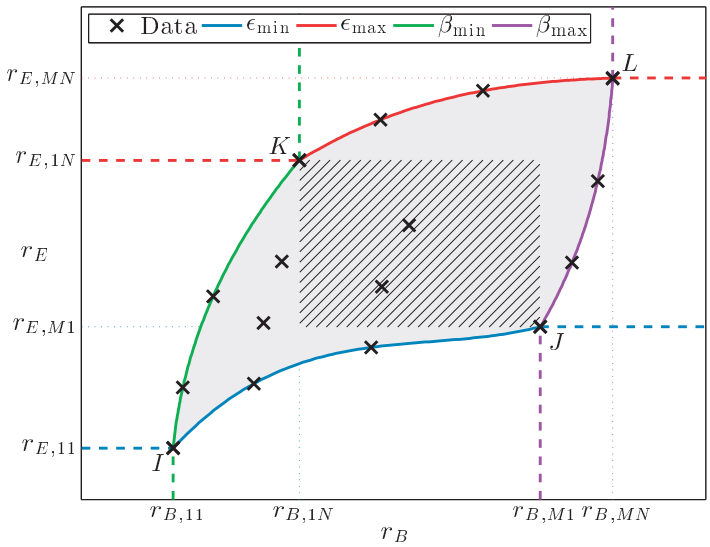


Figure 3.2: Domain on which the R^2-D surface is fitted. The hatched area represents a simple integration domain based on min and max values, as used in (X. Li et al., 2010), while the proposed model integrates over the whole area (represented in gray).

However, this simple integration domain might not be representative of the full domain.

Average Bit Rate Difference

To express the (logarithm of the) base layer bit rate as a function of the enhancement layer bit rate and distortion, a cubic surface is used to fit the R^2-D values

$$\hat{r}_B(r_E, D) = \sum_{(i,j)\in S} p_{ij} {r_E}^i D^j \qquad S = \{(i,j) \in \mathbb{N}^2 | i + j \leq 3\} \tag{3.15}$$

where $\hat{r}_B$ is the fitted base layer bit rate, r_E is the logarithm of the enhancement layer bit rate, D is the distortion in PSNR, and p_{ij} are the parameters of the fitting function.

Similarly to the Bjøntegaard model, a second fitting process is required and Equation (3.15) is not the inverse of Equation (3.9). The inverse function of a cubic function can be determined using Cardano's formula, but fitting a new surface yields better accuracy between measured and fitted values.

Then, the average base layer bit rate difference between two R^2-D surfaces is approximated as

$$\Delta R_B = \mathrm{E}\left[\frac{R_{B,2} - R_{B,1}}{R_{B,1}}\right] \approx 10^{\frac{1}{|A|}\iint_A [\hat{r}_{B,2}(r_E,D) - \hat{r}_{B,1}(r_E,D)]\mathrm{d}r_E\mathrm{d}D} - 1 \tag{3.16}$$

where ΔR_B is the delta base layer rate computed between the two fitted R^2-D surfaces $\hat{r}_1(r_E, D)$ and $\hat{r}_2(r_E, D)$, respectively, and the integration domain is determined following a similar procedure as for the average PSNR difference.

Table 3.1: Goodness of fit: coefficient of determination ρ^2 and RMSE.

cross-terms	$\hat{D}(r_B, r_E)$		$\hat{r}_B(r_E, D)$		$\hat{r}_E(r_B, D)$	
	yes	no	yes	no	yes	no
ρ^2	0.9967	0.9919	0.9885	0.9695	0.8968	0.8759
RMSE	0.1830	0.3202	0.0316	0.0570	0.0901	0.1011

The computation of the average enhancement layer bit rate difference between two R^2-D surfaces is similar to the computation of the average base layer bit rate difference.

3.2.2 Applications and Discussions

In this subsection, we show two examples of application of the proposed model. For this purpose, we used a dataset of 17 HDR image, which were encoded with JPEG XT (Artusi et al., 2015). JPEG XT is based on a two-layer design and is backward compatible with the popular JPEG compression standard. The base layer contains a LDR image, which is a tone-mapped version of the HDR image, accessible to legacy implementations, while the enhancement layer allows recovering the full dynamic range. The three main profiles of JPEG XT were used. For each profile, the quality factor of the base and enhancement layers was varied in the range $[20, 98]$ with a step of 2. Five different TMOs were considered to create the base layer LDR image. The dataset consisted of 17 HDR images × 3 profiles × 40 base layer quality values × 40 enhancement layer quality values × 5 TMOs = 408,000 compressed images with corresponding PSNR values.

Influence of Cross-Terms

Table 3.1 reports the goodness of fit computed between the (R_B, R_E, D) triplets and the fitted values. The coefficient of determination ρ^2 and RMSE are averaged over the 17 HDR images, 3 profiles, and 5 TMOs. Results show that cross-terms increase the coefficient of determination and decrease the RMSE in all cases.

Coding Performance of JPEG XT

Table 3.2a reports the average coding efficiency of the three main profiles of JPEG XT. The values were averaged over the 17 HDR images and 5 TMOs. Results show that Profile C provides a gain of over 3.2dB in PSNR for the same bit rate when compared to profiles A and B, whereas Profile B provides a gain of about 0.84dB over Profile A. On the other hand, for the same PSNR and enhancement layer bit rate, the bit rate of the base layer can be reduced by about 7.46% for Profile A when compared to Profile B. For the same base layer bit rate, i.e., same quality of the LDR image, the enhancement layer bit rate for Profile C can be reduced by about 30% and 20% when compared to profiles A and B, respectively.

Table 3.2: Average coding efficiency of the JPEG XT HDR image compression standard.

(a) Average coding efficiency of the three main profiles, computed over 17 HDR images and 5 TMOs.

Profile	ΔD relative to (dB)			ΔR_B relative to (%)			ΔR_E relative to (%)		
	A	B	C	A	B	C	A	B	C
A	-	−0.84	−3.86	-	−7.46	+52.40	-	+3.23	+66.15
B	+0.84	-	−3.24	+14.51	-	+84.25	+3.27	-	+30.94
C	+3.86	+3.24	-	−29.42	−32.88	-	−30.78	−18.34	-

(b) Influence of the TMO on the coding efficiency of Profile B.

TMO	ΔD relative to (dB)					ΔR_B relative to (%)					ΔR_E relative to (%)				
	d03	γ	*m11*	*m06*	*r02*	*d03*	γ	*m11*	*m06*	*r02*	*d03*	γ	*m11*	*m06*	*r02*
drago03 (d03)	-	+2.8	−2.3	−3.4	+1.1	-	−24.4	+19.3	+34.9	−16.7	-	−32.3	+3.5	+66.4	−13.2
gamma (γ)	−2.8	-	−4.7	−6.2	−1.7	+35.0	-	+73.8	+80.1	+16.8	+58.6	-	+30.8	+101.4	+44.6
mai11 (m11)	+2.3	+4.7	-	−1.2	+3.2	−13.7	−37.8	-	+23.7	−30.2	+0.8	−22.0	-	+44.4	−3.2
mantiuk06 (m06)	+3.4	+6.2	+1.2	-	+4.6	−19.4	−36.8	−11.6	-	−27.7	−31.9	−46.1	−28.6	-	−34.2
reinhard02 (r02)	−1.1	+1.7	−3.2	−4.6	-	+21.5	−13.0	+55.2	+47.1	-	+16.4	−25.5	+6.7	+59.6	-

A negative (positive) delta PSNR ΔD indicates a decrease (increase) of PSNR for the same base and enhancement layer bit rates.
A negative (positive) delta base layer rate ΔR_B indicates a decrease (increase) of the base layer bit rate for the same PSNR and enhancement layer bit rate.
The same principle applies to the delta enhancement layer rate ΔR_E.

Table 3.2b reports the influence of the TMO on the coding efficiency of Profile B. The values were averaged over the 17 HDR images. Surprisingly, the simple *gamma* TMO, which is very easy to inverse to predict the HDR image from the LDR image, reduces the PSNR of the reconstructed HDR image by 1.7dB to 6.2dB for the same base layer and enhancement layer rates. On the other hand, the *mantiuk06* TMO, which usually produces pleasant LDR images, allows reducing the bit rate of the enhancement layer by 28% to 46% when compared to other TMOs.

As it can be observed, the proposed model allows a more complete and detailed quantitative analysis when compared to the analysis reported by Pinheiro et al. (2014). The results reported in their study are more qualitative and their analysis was mostly performed on two dimensions only (distortion and enhancement layer bit rate), by fixing the quality parameter of the base layer. Additionally, the proposed model can be used for other applications than two-layer coding. For example, this model can be used for video plus depth or mutliview video plus depth coding, to find the optimal bit rate allocation between the texture and depth data, to maximize the quality of a synthesized viewpoint.

Note that the proposed model only considers one distortion, e.g., the distortion of the base or residual layer, or the distortion of a derived image/video sequence (see example above). To consider two different distortions, e.g., the base and enhancement layer distortions, a 4D model must be used. In this case, 20 or more quads are required, while most performance analysis are conducted with only 4×4 combinations of base and enhancement layer parameter values.

3.3 Extension for Calculation Based on Subjective Quality Scores

The coding efficiency of different compression algorithms can be adequately compared only by means of subjective tests, carried out according to common evaluation methodologies defined by experts. During the development phase of their compression standards, JPEG, MPEG, and VCEG have relied during past years on both objective and subjective evaluations to select and evaluate potential coding technologies, as well as for verification purposes. For example, subjective evaluations were conducted during the development of JPEG XR (De Simone et al., 2009b), MPEG-4 (Alpert et al., 1997), AVC (Baroncini and Quackenbush, 2012; Fenimore et al., 2004; Oelbaum et al., 2004), SVC (Baroncini and Quackenbush, 2012; Oelbaum et al., 2008), and HEVC (Baroncini and Quackenbush, 2012; De Simone et al., 2011; Weerakkody et al., 2014). Independent researchers have also conducted subjective evaluations, both during and after the development phase of compression standards, as a validation process or to evaluate the codecs in different scenarios. These evaluations have been conducted for both image and video compression.

To estimate a more realistic coding efficiency, subjective quality scores should be considered instead of PSNR measurements. Therefore, we propose a model to calculate the average coding efficiency based on MOSs gathered during subjective evaluations instead of PSNR

measurements. We call this approach subjective comparison of encoders based on fitted curves (SCENIC). To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the HVS, a logistic function is used to fit the R-D values. The average MOS and bit rate differences are computed between the fitted R-D curves similarly to the Bjøntegaard model. To consider the statistical property of subjective scores, the 95% CIs associated with the MOSs are considered to estimate corresponding confidence intervals on the calculated average MOS and bit rate differences. To provide meaningful measures, the R-D curves should ideally cover the full range of the rating scale. This recommendation is considered in the proposed model to estimate a confidence index on the calculated average MOS and bit rate differences.

3.3.1 Proposed Model

In this subsection, we propose a method for subjective comparison of encoders SCENIC. First, the function used to fit the R-D values is described. Then, the calculation of average MOS and bit rate differences between two fitted R-D curves is presented. Finally, the CIs and reliability index on the calculated average MOS and bit rate differences are presented. A MATLAB implementation of the proposed model can be downloaded from: `http://mmspg.epfl.ch/scenic`

Fitting Function

According to recommendation ITU-R BT.500-13 (2012), the relationship between MOS and the objective measure of picture distortion tends to have a sigmoid shape, provided that the natural limits of picture distortion extend far enough from the region in which the MOS varies rapidly. If the distortion parameter is measured in a physical unit, e.g., a time delay (ms), then a non-symmetrical function should be used to approximate this relationship (ITU-R BT.500-13, 2012). If the picture distortion is measured in a related unit, e.g., PSNR (dB), then a 4-parameter logistic function is commonly used (see Section 9.1). The 4-parameter logistic function (see Figure 3.3) is

$$y = f(x) = a + \frac{b-a}{1+\exp\left[-c(x-d)\right]} \tag{3.17}$$

where a, b, c, and d are the parameters of the fitting function.

As bit rate is not a direct measure of picture distortion, a non-symmetrical function should be used to map bit rate values to MOS, according to recommendation ITU-R BT.500-13 (2012). However, Gisle Bjøntegaard has observed that R-D values expressed in $\big(\log(\text{bit rate}), \text{PSNR}\big)$ do not deviate much from straight lines (Bjøntegaard, 2008), meaning that there is a somewhat linear relationship between log(bit rate) and PSNR. Therefore, based on this observation, and following the common practice to map PSNR values to MOS, we propose to use a logistic function to fit the R-D values expressed in $\big(\log(\text{bit rate}), \text{MOS}\big)$.

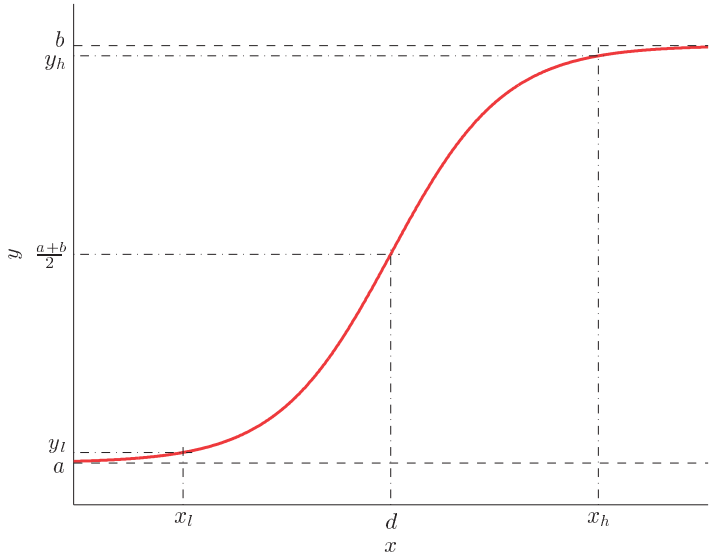


Figure 3.3: 4-parameter logistic function $y = f(x) = a + \frac{b-a}{1+\exp[-c(x-d)]}$.

Fitting a logistic function to a set of observed values is a nonlinear curve-fitting problem and can be expressed in least-squares sense. Several solutions have been proposed to solve this class of problem. However, the initial conditions may be critical to converge towards the optimal solution. Nevertheless, in most cases, constraints can be applied on the different parameters based on *a priori* knowledge to restrict the parameter search.

Most rating scales are divided into five categories with associated labels, such as (*Bad*; *Poor*; *Fair*; *Good*; and *Excellent*) or (*Very annoying*; *Annoying*; *Slightly annoying*; *Perceptible, but not annoying*; and *Imperceptible*) (see Section 2.4). The asymptotes of the relationship between MOS and bit rate, which are caused by the use of bounded rating scales and the saturation effects of the HVS, are typically associated with the two extreme categories of the rating scale. Moreover, the subjective scores should increase from the lower to the upper categories as the bit rate increases. Therefore, constraints are imposed on the logistic function such that the lower and upper asymptotes are associated with the lower and upper categories, respectively, and that the function is strictly increasing

$$u_{\min} \leq a \leq u_{\min} + \frac{1}{5}\Delta u \qquad u_{\max} - \frac{1}{5}\Delta u \leq b \leq u_{\max} \qquad c > 0 \tag{3.18}$$

where $\Delta u = u_{\max} - u_{\min}$, $u_{\min}$ and $u_{\max}$ are the boundaries of the rating scale, and $\frac{1}{5}\Delta u$ corresponds to the “length" of one category in a five categories scale.

Integration Bounds

Whereas the R-D curve based on PSNR measurements is unbounded, the R-D curve based on MOS is bounded due to the use of a bounded rating scale, the fact that many evaluation

methods consist in comparing the quality of a test stimulus against the quality of a reference stimulus, and the saturation effect of the HVS. Therefore, we think that it is not meaningful to compute average MOS or bit rate differences when both R-D curves have reached the saturation phase.

In statistics, it is common to consider only the values lying within the 95% CI. In the proposed model, we consider a similar approach by discarding the lower and upper parts of the fitted R-D curve and keeping only the values between y_l and y_h (see Figure 3.3), which covers 95% of the range spanned by the fitted R-D curve

$$y_l = a + 0.025(b-a) \qquad y_h = a + 0.975(b-a) \tag{3.19}$$

The x values corresponding to y_l and y_h are determined as

$$x_l = f^{-1}(y_l) \qquad x_h = f^{-1}(y_h) \tag{3.20}$$

where f^{-1} is the inverse function of the logistic function

$$x = f^{-1}(y) = g(y) = -\frac{1}{c}\ln\frac{b-y}{y-a} + d \tag{3.21}$$

Average MOS Difference

To approximate the R-D curve given by a set of N bit rate values $(R_1, ..., R_N)$ with corresponding MOSs $(D_1, ..., D_N)$, the R-D values are fitted in a least square sense using a logistic function with the constraints specified in Equation (3.18)

$$\hat{D}(r) = a + \frac{b-a}{1+\exp\left[-c(r-d)\right]} \tag{3.22}$$

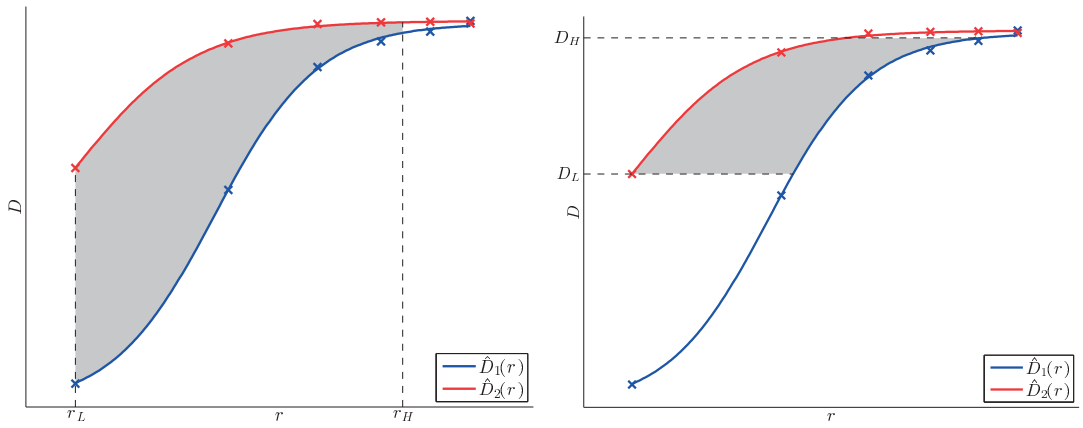
where $\hat{D}$ is the fitted distortion in MOS, r is the logarithm of the bit rate, and a, b, c, and d are the parameters of the fitting function.

Similarly to the Bjøntegaard model, the average MOS difference between two R-D curves is approximated by the difference between the integrals of the fitted R-D curves divided by the integration interval

$$\Delta D = \mathrm{E}\left[D_2 - D_1\right] \approx \frac{1}{r_H - r_L}\int_{r_L}^{r_H}\left[\hat{D}_2(r) - \hat{D}_1(r)\right]\mathrm{d}r \tag{3.23}$$

where ΔD is the delta MOS computed between the two fitted R-D curves $\hat{D}_1(r)$ and $\hat{D}_2(r)$, respectively, and the integration bounds, r_L and r_H, are

$$\begin{aligned} r_L &= \max\left\{\min\left(r_{1,1}, ..., r_{1,N_1}\right), \min\left(r_{2,1}, ..., r_{2,N_2}\right), \min\left(r_{1,l}, r_{2,l}\right)\right\} \\ r_H &= \min\left\{\max\left(r_{1,1}, ..., r_{1,N_1}\right), \max\left(r_{2,1}, ..., r_{2,N_2}\right), \max\left(r_{1,h}, r_{2,h}\right)\right\} \end{aligned} \tag{3.24}$$



(a) Computation of delta MOS.

(b) Computation of delta rate.

Figure 3.4: Integration bounds: the shaded area represents the integral of the difference of the two curves, evaluated between the lower and upper bounds.

where $r_{1,l}$ and $r_{1,h}$ ($r_{2,l}$ and $r_{2,h}$) are lower and upper rate bounds on $\hat{D}_1(r)$ ($\hat{D}_2(r)$) determined according to Equation (3.20).

To compute ΔD, the analytical expression of the integral of the logistic function is used

$$F(x) = \int f(x)\mathrm{d}x = \frac{b-a}{c} \ln\left\{1+\exp\left[-c(x-d)\right]\right\} + bx + (a-b)d + C \tag{3.25}$$

where C is an arbitrary constant.

Figure 3.4a illustrates the computation of the average MOS difference between two fitted R-D curves.

Average Bit Rate Difference

Instead of applying another fitting to express the (logarithm of the) bit rate as a function of the distortion, as in the Bjøntegaard model, the inverse function of Equation (3.22) is used

$$\hat{r}(D) = -\frac{1}{c} \ln \frac{b-D}{D-a} + d \tag{3.26}$$

where $\hat{r}$ is the fitted bit rate, D is the distortion in MOS, and a, b, c, and d are the parameters determined for Equation (3.22). Therefore, the logistic fitting is applied only once for a given set of R-D values.

Similarly to the Bjøntegaard model, the average bit rate difference between two R-D curves is

approximated as

$$\Delta R = \mathrm{E}\left[\frac{R_2 - R_1}{R_1}\right] \approx 10^{\frac{1}{D_H - D_L}\int_{D_L}^{D_H}[\hat{r}_2(D) - \hat{r}_1(D)]\mathrm{d}D} - 1 \tag{3.27}$$

where ΔR is the delta rate computed between the two fitted R-D curves $\hat{r}_1(r)$ and $\hat{r}_2(r)$, respectively, and the integration bounds, D_L and D_H, are

$$\begin{aligned} D_L &= \max\{\min(\hat{D}_{1,1},...,\hat{D}_{1,N_1}), \min(\hat{D}_{2,1},...,\hat{D}_{2,N_2}), \min(\hat{D}_{1,l}, \hat{D}_{2,l})\} \\ D_H &= \min\{\max(\hat{D}_{1,1},...,\hat{D}_{1,N_1}), \max(\hat{D}_{2,1},...,\hat{D}_{2,N_2}), \max(\hat{D}_{1,h}, \hat{D}_{2,h})\} \end{aligned} \tag{3.28}$$

where $\hat{D}_{1,l}$ and $\hat{D}_{1,h}$ ($\hat{D}_{2,l}$ and $\hat{D}_{2,h}$) are the lower and upper distortion bounds on $\hat{D}_1(r)$ ($\hat{D}_2(r)$) determined according to Equation (3.19).

To compute ΔR, the analytical expression of the integral of the inverse logistic function is used

$$G(y) = \int g(y)\mathrm{d}y = \frac{b-y}{c}\left[\ln(b-y) - 1\right] + \frac{y-a}{c}\left[\ln(y-a) - 1\right] + dy + C \tag{3.29}$$

where C is an arbitrary constant.

Figure 3.4b illustrates the computation of the average bit rate difference between two fitted R-D curves.

Confidence Interval

To consider the statistical property of a MOS, $\overline{u}_i$, the corresponding CI, $\left[\overline{u}_i - \delta_i, \overline{u}_i + \delta_i\right]$ (see Section 2.6.2), should be considered in the proposed model when computing the average MOS and bit rate differences. In recommendation ITU-R BT.500-13 (2012), it is proposed to consider three series of grades, constructed from the MOSs for each test condition and associated 95% CIs

i) minimum grade series $(\overline{u}_1 - \delta_1, ..., \overline{u}_N - \delta_N)$,
ii) mean grade series $(\overline{u}_1, ..., \overline{u}_N)$, and
iii) maximum grade series $(\overline{u}_1 + \delta_1, ..., \overline{u}_N + \delta_N)$.

According to this recommendation, the three grade series should be fitted independently.

Figure 3.5 depicts an example of MOSs and associated 95% CI. The fitting functions $\hat{D}^-(r)$, $\hat{D}(r)$, and $\hat{D}^+(r)$ (see Table 3.3) for the minimum, mean, and maximum grade series, respectively, are drawn on the same graph to provide an estimate of the 95% continuous confidence region, which can be used to determine a tolerance range. The space between $\hat{D}^+(r)$ and $\hat{D}^-(r)$ is not an exact 95% CI, but a mean estimate thereof (ITU-R BT.500-13, 2012).

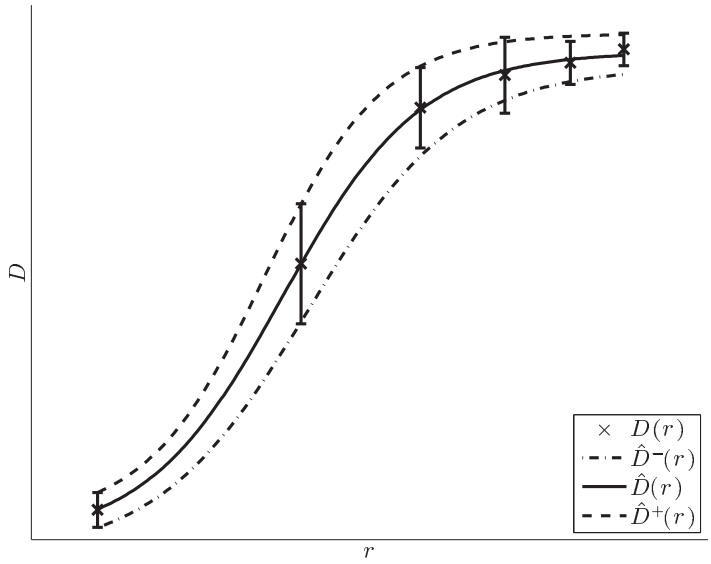


Figure 3.5: Different grade series: $\hat{D}^{-}(r)$, $\hat{D}(r)$, and $\hat{D}^{+}(r)$ are the fitting functions for the minimum, mean, and maximum grade series, respectively, constructed from the MOSs for each test condition and associated 95% CIs.

Table 3.3: Fitting functions for the different grade series.

Fitting functions	Fitting of	Values
$\hat{D}^{-}(r), \hat{r}^{-}(D)$	minimum grade series	$(\overline{u}_1 - \delta_1, ..., \overline{u}_N - \delta_N)$
$\hat{D}(r), \hat{r}(D)$	mean grade series	$(\overline{u}_1, ..., \overline{u}_N)$
$\hat{D}^{+}(r), \hat{r}^{+}(D)$	maximum grade series	$(\overline{u}_1 + \delta_1, ..., \overline{u}_N + \delta_N)$

The parameters a, b, and c of the logistic function are constrained as the subjective scores should increase from the lower to the upper categories as the bit rate increases. (see Equation (3.18)). These constraints should be modified when fitting the minimum and maximum grade series to consider the CIs. If we consider a typical R-D curve and rating scale divided into five categories, at the extreme parts of the curve, the CIs generally tends to become smaller, due to the intrinsic nature of bounded rating scales, but they may slightly span outside of the extreme categories. Therefore, for the fitting of the minimum (maximum) grade series, we decrease (increase) the lower (upper) bound on parameters a and b by half of the "length" of one category (see Table 3.4).

The average MOS and bit rate differences are computed from the mean grades series as described here above. The corresponding 95% CI is estimated using the minimum and maximum grade series to consider the CIs associated with the MOSs.

Table 3.4: Constraints for the different fitting functions.

Fitting functions	Constraints on parameter		
	a	b	c
$\hat{D}^-(r), \hat{r}^-(D)$	$u_{\min} - \frac{1}{10}\Delta u \le a \le u_{\min} + \frac{1}{5}\Delta u$	$u_{\max} - \frac{3}{10}\Delta u \le b \le u_{\max}$	$c > 0$
$\hat{D}(r), \hat{r}(D)$	$u_{\min} \le a \le u_{\min} + \frac{1}{5}\Delta u$	$u_{\max} - \frac{1}{5}\Delta u \le b \le u_{\max}$	$c > 0$
$\hat{D}^+(r), \hat{r}^+(D)$	$u_{\min} \le a \le u_{\min} + \frac{3}{10}\Delta u$	$u_{\max} - \frac{1}{5}\Delta u \le b \le u_{\max} + \frac{1}{10}\Delta u$	$c > 0$

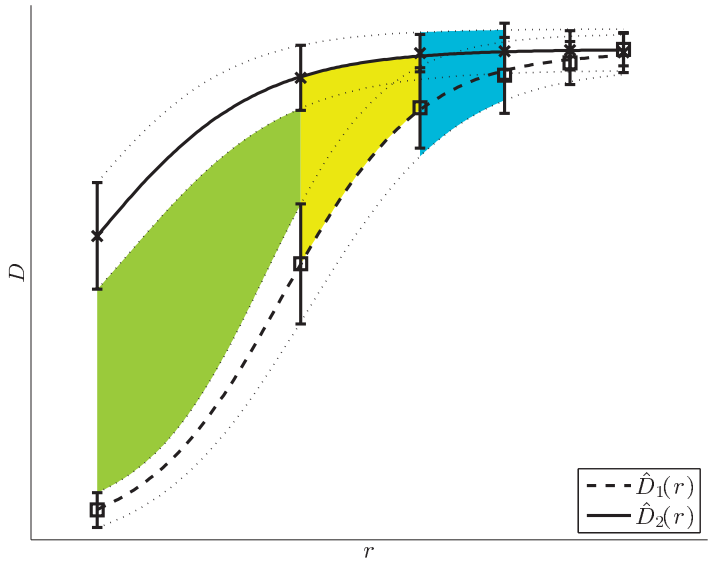


Figure 3.6: Confidence interval: the green, yellow, and blue areas illustrate the calculation of $\Delta D_{\min}$, ΔD, and $\Delta D_{\max}$, respectively. For illustration purpose, only part of the total area used in the calculation of each value is represented. The same principle applies for the calculation of $\Delta R_{\min}$, ΔR, and $\Delta R_{\max}$.

The average MOS difference, ΔD, and its corresponding estimated 95% CI $[\Delta D_{\min}, \Delta D_{\max}]$, are

$$\begin{aligned} \Delta D &= \phi\left(\hat{D}_1(r), \hat{D}_2(r), r_L, r_H\right) \\ \Delta D_{\min} &= \min\left\{\phi\left(\hat{D}_1^-(r), \hat{D}_2^+(r), r_L, r_H\right), \phi\left(\hat{D}_1^+(r), \hat{D}_2^-(r), r_L, r_H\right)\right\} \\ \Delta D_{\max} &= \max\left\{\phi\left(\hat{D}_1^-(r), \hat{D}_2^+(r), r_L, r_H\right), \phi\left(\hat{D}_1^+(r), \hat{D}_2^-(r), r_L, r_H\right)\right\} \end{aligned} \tag{3.30}$$

where r_L and r_H are the integration bounds computed from $(r_{1,1}, \ldots r_{1,N_1})$, $(r_{2,1}, \ldots r_{2,N_2})$, $\hat{r}_1(D)$, and $\hat{r}_2(D)$ according to Equation (3.24), and ϕ is a generic function to compute the average MOS difference between two fitted R-D curves $\hat{D}_1(r)$ and $\hat{D}_2(r)$, between r_L and r_H

$$\phi\left(\hat{D}_1(r), \hat{D}_2(r), r_L, r_H\right) = \frac{1}{r_H - r_L} \int_{r_L}^{r_H} \left[\hat{D}_2(r) - \hat{D}_1(r)\right] \mathrm{d}r \tag{3.31}$$

Figure 3.6 illustrates the calculation of $\Delta D_{\min}$, ΔD, and $\Delta D_{\max}$.

The average bit rate difference, ΔR, and its corresponding estimated 95% CI $[\Delta R_{\min}, \Delta R_{\max}]$, are

$$\begin{aligned} \Delta R &= \psi\left(\hat{r}_1(D), \hat{r}_2(D), D_L, D_H\right) \\ \Delta R_{\min} &= \min\left\{\psi\left(\hat{r}_1^-(D), \hat{r}_2^+(D), D_L, D_H\right), \psi\left(\hat{r}_1^+(D), \hat{r}_2^-(D), D_L, D_H\right)\right\} \\ \Delta R_{\max} &= \max\left\{\psi\left(\hat{r}_1^-(D), \hat{r}_2^+(D), D_L, D_H\right), \psi\left(\hat{r}_1^+(D), \hat{r}_2^-(D), D_L, D_H\right)\right\} \end{aligned} \tag{3.32}$$

where D_L and D_H are the integration bounds computed from $(D_{1,1},...D_{1,N_1})$, $(D_{2,1},...D_{2,N_2})$, $\hat{D}_1(r)$, and $\hat{D}_2(r)$ according to Equation (3.24), and ψ is a generic function to compute the average bit rate difference between two fitted R-D curves $\hat{r}_1(D)$ and $\hat{r}_2(D)$, between D_L and D_H

$$\psi\left(\hat{r}_1(D), \hat{r}_2(D), D_L, D_H\right) = 10^{\frac{1}{D_H - D_L}\int_{D_L}^{D_H}[\hat{r}_2(D) - \hat{r}_1(D)]\mathrm{d}D} - 1 \tag{3.33}$$

Confidence Index

To provide confident measures, the R-D curves should ideally cover the full range of the rating scale. In most quality evaluations, both objective and subjective, a predefined set of targeted bit rates is usually considered. In well-designed subjective tests, the lower bit rate is chosen such that at least one test stimulus (specific combination of content, codec, and bit rate) would have a quality corresponding to the lower category. However, care should be taken to avoid too low quality test stimuli. Therefore, it is possible that at the lower bit rate, one codec produces bad quality, whereas another codec produces fair or good quality, if there is a significant difference in terms of compression efficiency between the two codecs.

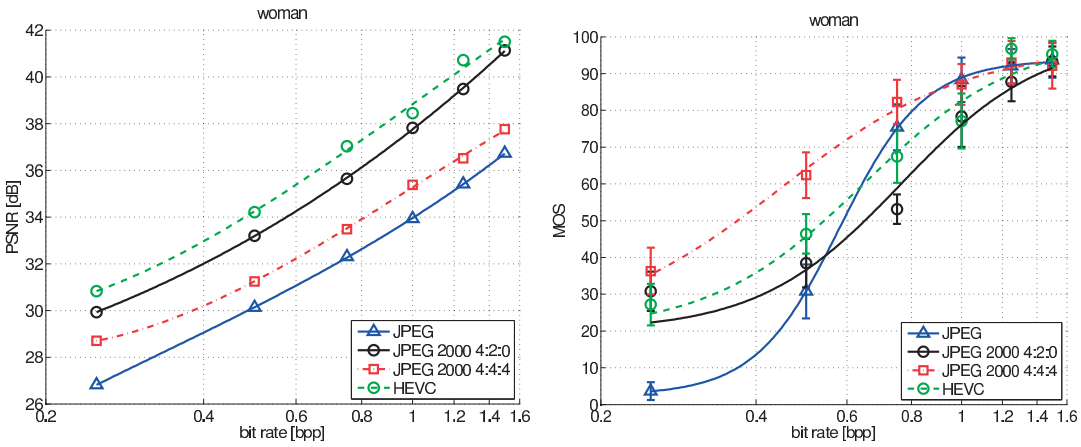
These considerations are incorporated in the proposed model to produce a confidence index on the calculated average MOS and bit rate differences. As it is impossible in most practical situations to cover the full range of the rating scale with both R-D curves for the above-mentioned reason, we assume that at least one of the two R-D curves should cover 80% of the rating scale to have a valid measure of the average MOS and bit rate differences. The range of the rating scale, Δu_1 and Δu_2, covered by the two R-D curves is

$$\begin{aligned} \Delta u_1 &= \max(D_{1,1},...,D_{1,N_1}) - \min(D_{1,1},...,D_{1,N_1}) \\ \Delta u_2 &= \max(D_{2,1},...,D_{2,N_2}) - \min(D_{2,1},...,D_{2,N_2}) \end{aligned} \tag{3.34}$$

We also consider the goodness of the fitting functions, measured in terms of the PCC

$$\begin{aligned} \rho_1 &= r\left((D_{1,1},...,D_{1,N_1}), (\hat{D}_{1,1},...,\hat{D}_{1,N_1})\right) \\ \rho_2 &= r\left((D_{2,1},...,D_{2,N_2}), (\hat{D}_{2,1},...,\hat{D}_{2,N_2})\right) \end{aligned} \tag{3.35}$$

where $r(\cdot)$ is the PCC.



(a) Bjøntegaard model. (b) Proposed model.

Figure 3.7: Rate-distortion curves for content *woman*.

The confidence index is computed as

$$\text{Confidence index} = \min\left\{1, \frac{\max(\Delta u_1, \Delta u_2)}{0.8(u_{\max} - u_{\min})}\rho_1\rho_2\right\} \tag{3.36}$$

where $u_{\min}$ and $u_{\max}$ are the boundaries of the rating scale.

3.3.2 Applications and Discussions

In this section, three case studies are presented where the Bjøntegaard and proposed models were used to calculate average coding efficiency. The aim of these examples is twofold. The first objective is to show that the Bjøntegaard model does not always provide an accurate measure of coding efficiency, whereas the proposed model should report more realistic coding efficiency. However, as there is no ground truth for the coding efficiency, it is impossible to quantify the performance of the two models, but rather to discuss when the two models do not agree. The second objective is to illustrate the usefulness of the CIs and confidence index provided by the proposed model.

Quality of High Resolution Images

In this case study, we used the results from the evaluation of HEVC image compression reported in Section 4.2. Tables 3.5 and 3.6 report the coding efficiency calculated for content *woman* using the Bjøntegaard and proposed models, respectively. Figure 3.7 shows the fitted R-D curves for content *woman*.

Table 3.5 reports an average bit rate difference for JPEG 2000 4:2:0 over JPEG 2000 4:4:4 of −31% based on the Bjøntegaard model. However, Table 3.6 reports an average bit rate difference

Table 3.5: Coding efficiency for content *woman*: Bjøntegaard model.

Encoding	Average bit rate difference relative to (%)				Average PSNR difference relative to (dB)			
	JPEG	J2000 4:2:0	J2000 4:4:4	HEVC	JPEG	J2000 4:2:0	J2000 4:4:4	HEVC
JPEG	-	+78	+26	+112	-	−3.4	−1.3	−4.4
JPEG 2000 4:2:0	−44	-	−31	+18	+3.4	-	+2.1	−1.0
JPEG 2000 4:4:4	−21	+44	-	+73	+1.3	−2.1	-	−3.1
HEVC	−53	−15	−42	-	+4.4	+1.0	+3.1	-

A negative (positive) bit rate difference indicates a decrease (increase) of bit rate for the same PSNR.
A negative (positive) PSNR difference indicates a decrease (increase) of PSNR for the same bit rate.

Table 3.6: Coding efficiency for content *woman*: Proposed model.

Encoding	Average bit rate difference relative to			
	JPEG	JPEG 2000 4:2:0	JPEG 2000 4:4:4	HEVC
(a) Delta rate				
JPEG	-	−9% [−25%,+6%] (100%)	+34% [+13%,+68%] (100%)	+5% [−14%,+26%] (100%)
J2000 4:2:0	+10% [−6%,+33%] (100%)	-	+61% [+24%,+109%] (77%)	+19% [−9%,+53%] (85%)
J2000 4:4:4	−25% [−40%,−12%] (100%)	−38% [−52%,−19%] (77%)	-	−26% [−45%,−3%] (86%)
HEVC	−5% [−21%,+17%] (100%)	−16% [−35%,+10%] (85%)	+36% [+3%,+81%] (86%)	-
	Average MOS difference relative to			
	JPEG	JPEG 2000 4:2:0	JPEG 2000 4:4:4	HEVC
(b) Delta MOS				
JPEG	-	−0.5 [−11.4,+11.2] (100%)	−18.7 [−30.0,−7.0] (100%)	−7.2 [−18.9,+3.8] (100%)
JPEG 2000 4:2:0	+0.5 [−11.2,+11.4] (100%)	-	−18.2 [−29.0,−5.5] (77%)	−6.6 [−17.8,+5.3] (85%)
JPEG 2000 4:4:4	+18.7 [+7.0,+30.0] (100%)	+18.2 [+5.5,+29.0] (77%)	-	+11.5 [−1.1,+23.0] (86%)
HEVC	+7.2 [−3.8,+18.9] (100%)	+6.6 [−5.3,+17.8] (85%)	−11.5 [−23.0,+1.1] (86%)	-

(a) A negative (positive) value indicates a decrease (increase) of bit rate for the same MOS. Reading: ΔR [$\Delta R_{\min}, \Delta R_{\max}$] (Confidence index).
(b) A negative (positive) value indicates a decrease (increase) of MOS for the same bit rate. Reading: ΔD [$\Delta D_{\min}, \Delta D_{\max}$] (Confidence index).

of +61% [+24%,+109%] based on the proposed model. Note that the 95% CI resulting from the proposed model does not contain the value calculated by the Bjøntegaard model. These results show that JPEG 2000 4:2:0 has better coding efficiency than JPEG 2000 4:4:4 according to the Bjøntegaard model, whereas the proposed model dictates the opposite. To understand why the two models lead to different conclusions, it is necessary to analyze the objective and subjective scores. According to PSNR measurements, JPEG 2000 4:2:0 performed always better than JPEG 2000 4:4:4 (see Figure 3.7a), whereas the subjective results dictate the opposite (see Figure 3.7b).

Visual weighting was disabled for JPEG 2000 4:2:0, whereas it was enabled for JPEG 2000 4:4:4. The lack of visual weighting creates distortions, particularly at lower bit rates, as reported during the development of JPEG 2000. This example shows that when PSNR fails to capture a specific distortion, the comparison of coding efficiency using the Bjøntegaard model may lead to wrong conclusion. In this case, the proposed model, which relies on subjective scores, should report more realistic estimation of coding efficiency.

Table 3.5 reports an average bit rate difference over JPEG of −44% and −53% for JPEG 2000 4:2:0 and HEVC, respectively, based on the Bjøntegaard model. However, Table 3.6 reports an average bit rate difference over JPEG of +10% [−6%,+33%] and −5% [−21%,+17%] for JPEG 2000 4:2:0 and HEVC, respectively, based on the proposed model. Note that the 95% CIs resulting from the proposed model do not contain the values calculated by the Bjøntegaard model. As it can be observed from Figure 3.7, HEVC outperformed JPEG by at least 3 dB on all bit rates, whereas JPEG was evaluated better than or equal to HEVC at 0.75bpp and above based on the subjective results. This example shows that the coding efficiency reported by the Bjøntegaard model may be over-estimated in some cases.

It is known that PSNR does not accurately reflect human perception of visual quality (Sheikh et al., 2006). As the Bjøntegaard model relies on PSNR measurements, it is not surprising that the coding efficiency calculated with this model may not accurately reflect the true coding efficiency in some cases. Using a different model relying on a perceptual metric that better correlates with perceived quality, e.g., structural similarity (SSIM), would probably result in more accurate estimation of coding efficiency.

Quality of UHD Video Sequences

In this case study, we used the results from the evaluation of HEVC video compression reported in Section 4.1. Table 3.7 report the coding efficiency for HEVC over AVC calculated on each test content using the Bjøntegaard and proposed models. Figure 3.8 shows the fitted R-D curves for content *Traffic*.

For content *Traffic*, subjects evaluated nine out of ten video sequences as *Imperceptible* (see Figure 3.8b). These results show that, at the selected bit rates, the R-D curves are mostly in the upper saturation phase. However, it is impossible to predict this behavior from the PSNR

Table 3.7: Average coding efficiency for HEVC over AVC.

(a) Bjøntegaard model.

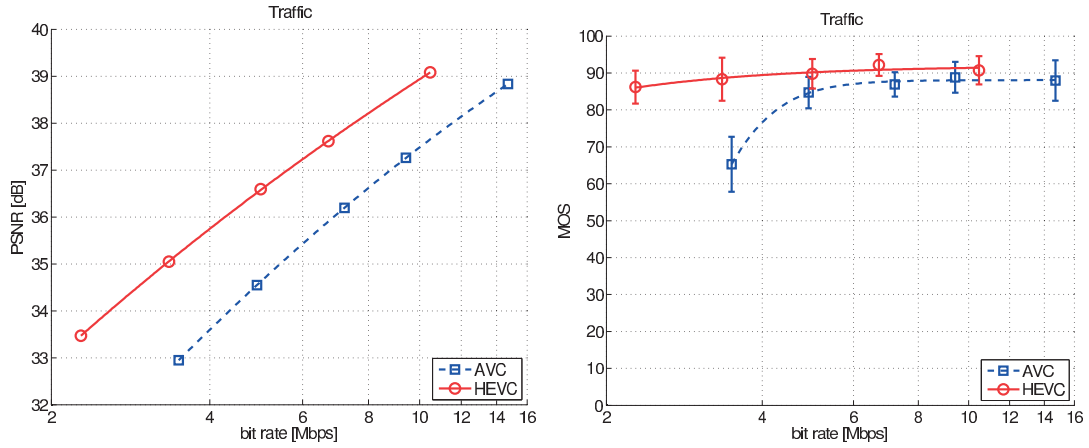
Content	Delta rate (%)	Delta PSNR (dB)
PeopleOnStreet	−27	+1.6
Traffic	−38	+1.8
Sintel2	−68	+4.4
Overall	−44	+2.6

A negative (positive) value indicates a decrease (increase) of bit rate for the same PSNR.
A negative (positive) value indicates a decrease (increase) of PSNR for the same bit rate.

(b) Proposed model.

Content	Delta rate ΔR $[\Delta R_{\text{min}}, \Delta R_{\text{max}}]$	Delta MOS ΔD $[\Delta D_{\text{min}}, \Delta D_{\text{max}}]$	Confidence index (%)
PeopleOnStreet	−53% [−69%,−27%]	+25.8 [+13.0,+38.4]	79
Traffic	−59% [-,−5%]	+10.8 [−2.2,+20.3]	28
Sintel2	−73% [-,−60%]	+40.7 [+28.9,+52.4]	62
Overall	−62% [-,−31%]	+25.8 [+13.2,+37.1]	56

A negative (positive) delta rate indicates a decrease (increase) of bit rate (MOS) for the same MOS (bit rate).
A negative (positive) delta MOS indicates a decrease (increase) of bit rate (MOS) for the same MOS (bit rate).

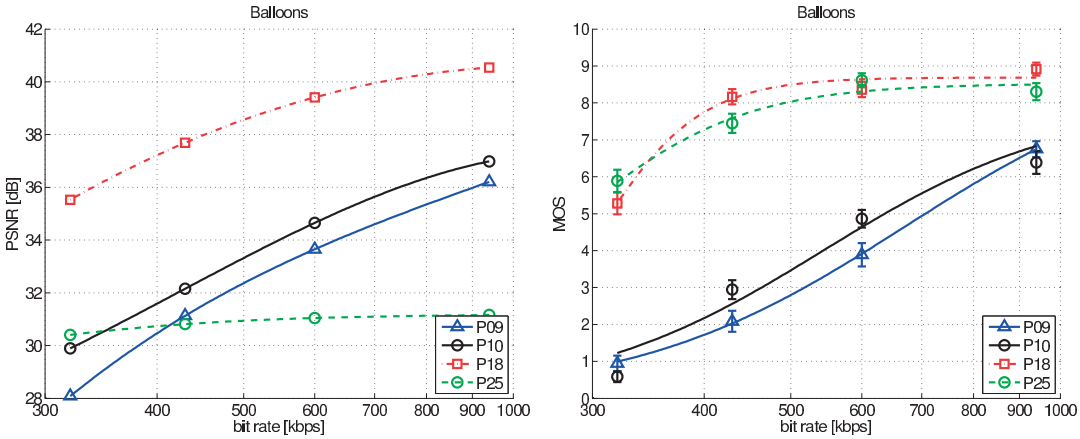


(a) Bjøntegaard model. (b) Proposed model.

Figure 3.8: Rate-distortion curves for content *Traffic*.

measurements as the two curves are continuously increasing and the PSNR values are below 40 dB, which is often considered as excellent quality.

For this particular content, the R-D values were mostly measured in the upper saturation phase, and not across the entire rating scale, as recommended. Therefore, the average PSNR/MOS and bit rate differences calculated using the two models are not representative of the true coding efficiency for this content. Nevertheless, for the proposed model, this problem is reflected in the low confidence index (28%) and wide CI reported in Table 3.7b. Note that the value for ΔR_{min} could not be determined as there was no overlap between the two R-D curves.



(a) Bjøntegaard model.

(b) Proposed model.

Figure 3.9: Rate-distortion curves for content *Balloons*.

However, the the Bjøntegaard model does not consider the saturation effect of the HVS and does not provide such indication regarding the confidence of the calculated coding efficiency.

Quality of 3D Video Sequences

In this case study, we used the same dataset as for the benchmarking of objective metrics on stereo pairs formed from two synthesized views reported in Section 10.2. Tables 3.8 and 3.9 report the coding efficiency calculated for content *Balloons* using the Bjøntegaard and proposed models, respectively. Figure 3.9 shows the fitted R-D curves for content *Balloons*.

The average bit rate reduction values calculated using the Bjøntegaard model (see Table 3.8) are in general similar to those calculated using the proposed model (see Table 3.9), except for the values related to proponent *P25*. To understand why the two models differ for this particular proponent, it is necessary to analyze the objective and subjective scores. As it can be observed from Figure 3.9, proponent *P25* obtained constant low PSNR values, whereas it obtained high subjective scores.

It is known that one proposal submitted in response to the CfP used a different view synthesis algorithm. As the data submitted by the proponents is anonymous, we cannot be certain that proponent *P25* used a different view synthesis algorithm. However, these results show that coding efficiency calculated based on PSNR measurements might not accurately reflect the true coding efficiency in the case of stereoscopic content formed from synthesized views, as PSNR is not accurate to assess perceived quality of synthesized views (see Section 10.2). Using a different model relying on a perceptual metric that better correlates with perceived quality of stereoscopic content, e.g., VIF or VQM, would probably result in more accurate estimation of coding efficiency.

Table 3.8: Coding efficiency for content *Balloons*: Bjøntegaard model.

Encoding	Average bit rate difference relative to (%)				Average PSNR difference relative to (dB)			
	P09	*P10*	*P18*	*P25*	*P09*	*P10*	*P18*	*P25*
P09	-	+16	+159	−11	-	−1.1	−5.9	+1.8
P10	−14	-	+123	−22	+1.1	-	−4.9	+2.9
P18	−61	−55	-	-	+5.9	+4.9	-	+7.8
P25	+12	+28	-	-	−1.8	−2.9	−7.8	-

A negative (positive) bit rate difference indicates a decrease (increase) of bit rate for the same PSNR.
A negative (positive) PSNR difference indicates a decrease (increase) of PSNR for the same bit rate.

Table 3.9: Coding efficiency for content *Balloons*: Proposed model.

Encoding	Average bit rate difference relative to			
	P09	*P10*	*P18*	*P25*
(a) Delta rate				
P09	-	+10% [+2%,+22%] (71%)	+146% [+131%,+164%] (72%)	+155% [+134%,+178%] (71%)
P10	−9% [−18%,−2%] (71%)	-	+131% [+111%,+149%] (71%)	+142% [+115%,+180%] (70%)
P18	−59% [−62%,−57%] (72%)	−57% [−60%,−53%] (71%)	-	−8% [−14%,+2%] (44%)
P25	−61% [−64%,−57%] (71%)	−59% [−64%,−53%] (70%)	+8% [−2%,+16%] (44%)	-
	Average MOS difference relative to			
	P09	*P10*	*P18*	*P25*
(b) Delta MOS				
P09	-	−0.5 [−1.1,−0.0] (71%)	−4.6 [−5.1,−4.1] (72%)	−4.3 [−4.8,−3.8] (71%)
P10	+0.5 [+0.0,+1.1] (71%)	-	−4.1 [−4.5,−3.6] (71%)	−3.8 [−4.2,−3.3] (70%)
P18	+4.6 [+4.1,+5.1] (72%)	+4.1 [+3.6,+4.5] (71%)	-	+0.3 [−0.1,+0.8] (44%)
P25	+4.3 [+3.8,+4.8] (71%)	+3.8 [+3.3,+4.2] (70%)	−0.3 [−0.8,+0.1] (44%)	-

(a) A negative (positive) value indicates a decrease (increase) of bit rate for the same MOS. Reading: ΔR [$\Delta R_{\text{min}}, \Delta R_{\text{max}}$] (Confidence index).
(b) A negative (positive) value indicates a decrease (increase) of MOS for the same bit rate. Reading: ΔD [$\Delta D_{\text{min}}, \Delta D_{\text{max}}$] (Confidence index).

3.4 Conclusion

This chapter described in details the Bjøntegaard model, which is commonly used to calculate the coding efficiency between different codecs. This model reports two values: average PSNR difference in dB for the same bit rate and the average bit rate difference in percent for the same PSNR. We proposed two extensions of the Bjøntegaard model. The first extension is designed for two-layer coding systems and aims at investigating the impact on quality of the interaction of the base and enhancement layers bit rates. The proposed model extends the Bjøntegaard model from R-D curve fitting to R^2-D surface fitting. It uses a cubic surface as fitting function and a more complex characterization of the domain formed by the data points to compute a more realistic estimate of the compression efficiency. We presented two applications of the proposed model to measure the compression efficiency of JPEG XT. The proposed model can also be used for other applications, e.g., to optimize the bit rate allocation between texture and depth in 3D video coding.

The second extension relies on subjective quality scores instead of PSNR measurements. To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the human visual system, a logistic function was used to fit the R-D values. The average MOS and bit rate differences were computed between the fitted R-D curves. To consider the statistical property of subjective scores, the 95% CIs associated with the MOSs were considered to estimate corresponding confidence intervals on the calculated average MOS and bit rate differences. We presented three case studies where the Bjøntegaard and proposed models were used to calculate average coding efficiency. Results showed that the Bjøntegaard model does not always report an accurate measure of the true coding efficiency as it relies on PSNR measurements, which does not accurately reflect human perception of visual quality. However, the proposed model, which relies on subjective scores, is expected to report more realistic estimation of coding efficiency. This model was also used in the performance analysis of image and video compression reported in the rest of the thesis.

4 Performance Analysis of Image and Video Compression

Every day, 350 million photos are shared on Facebook and 4 billion videos are viewed on YouTube, whereas 300 hours of video are uploaded every minute. For a typical 12 megapixels still image, captured with an iPhone or digital camera, each image needs 36MB of storage to be represented in uncompressed RAW format. Similarly, for a typical full HD video at 24 frames per second, which is typical for movie, each 1 second of video requires about 150MB to be stored in RAW format. Considering these characteristics, Facebook would require 12.6 peta bytes of additional storage capacity every day and YouTube's IT engineers would have to add about 40 4TB hard drive disks in their storage bay every minute. Without efficient image and video compression algorithms, it would not be feasible to transmit and store such a huge amount of multimedia content.

To compress an image or video sequence and reduce its file size, algorithms typically try to exploit correlation, e.g., spatial and temporal, in the data, for example to predict the current frame from previously encoded frames. Additionally, properties of the HVS are exploited to further reduce the amount of data, for example to adaptively quantize the data. Most coding schemes rely on the following processing. First, a non-linear mapping, referred to as gamma encoding (Poynton, 2012), is applied on the linear red, green, and blue (RGB) signal to optimize the usage of bits when encoding an image based on the observation that the HVS is more sensitive to changes in dark areas than bright areas. Then, the RGB image is converted to another color space, e.g., YCbCr, to decorrelate the redundant information in the red, green, and blue components. The new components are typically composed of a luma component, which is related to the luminance information after gamma expansion, and two chroma components, which are related to the chrominance information. Based on the observation that the HVS is more sensitive to loss of resolution in luminance than in chrominance, the chroma components are often downscaled from 4:4:4 (full horizontal and vertical resolutions) to 4:2:2 (half horizontal resolution, full vertical resolution) or 4:2:0 (half horizontal and vertical resolutions) sampling. The image is then decomposed into several blocks, e.g., 8×8 pixels, and each block is processed following a specific order. The image blocks can be transformed to another space, for example using the discrete cosine transform

(DCT) (Ahmed et al., 1974), which decomposes a block into a sum of different sinusoidal patterns, with different horizontal and vertical periods. As the HVS is more sensitive to low frequency than high frequency, an adaptive quantization of the DCT coefficients is performed to further reduce the amount of data. This block-based transform coding is referred to as intra-frame coding. Further prediction can be performed, for example to predict a block in the current frame from a neighboring and already coded block in the current frame, which is referred to as intra-frame prediction, or from another block in the previous frame, which is referred to as inter-frame prediction. These approaches try to exploit the spatial and temporal redundancy of the data to further reduce the amount of data that has to be transferred or stored. Finally, the encoded data is further reduced using entropy encoding (Wiegand and Schwarz, 2010). Note that this last step is lossless, whereas the most previous steps are lossy and can result in quality degradation. The reader is invited to have a look at the following text books for a more detailed and more comprehensive description of image and video compression schemes: (Bhaskaran and Konstantinides, 1997; Haskell et al., 1997; Mitchell, 1997; Netravali, 2013; Rabbani and Jones, 1991; H. R. Wu and Rao, 2005).

Several standardization bodies are at the roots of the still and moving pictures codecs used over the past 30 years. For example, the Joint Photographic Experts Group (JPEG) has published the following image compression standards: JPEG (Wallace, 1991), JPEG 2000 (Christopoulos et al., 2000), JEPG XR (Dufaux et al., 2009), JPEG XT (Artusi et al., 2015). The Moving Picture Experts Group (MPEG) and Video Coding Experts Group (VCEG) have published the following video compression standards: H.261, H.262/MPEG-2 Part 2 (Haskell et al., 1997), H.263, H.264/MPEG-4 Part 10 Advanced Video Coding (AVC) (Wiegand et al., 2003b) and its scalable (SVC) (Schwarz et al., 2007) and 3D (MVC, MVC+D, and 3D-AVC) (Y. Chen et al., 2014; Vetro et al., 2011) extensions, H.265/MPEG-H Part 2 High Efficiency Video Coding (HEVC) (Ohm et al., 2012) and its scalable (SHVC) (Boyce et al., 2016) and 3D (MV-HEVC) (Y. Chen and Vetro, 2014) extensions. Recently, Google was actively involved in the development of VP9 (Mukherjee et al., 2015a) and its successor VP10 (Mukherjee et al., 2015b), an open source alternative to AVC and HEVC that is used in YouTube.

The coding efficiency of different compression algorithms can be adequately compared only by means of subjective tests, carried out according to common evaluation methodologies defined by experts (see Chapter 2). During the development phase of their compression standards, JPEG, MPEG, and VCEG have relied during past years on both objective and subjective evaluations to select and evaluate potential coding technologies, as well as for verification purposes. For example, subjective evaluations were conducted during the development of JPEG XR (De Simone et al., 2009b), MPEG-4 (Alpert et al., 1997), AVC (Baroncini and Quackenbush, 2012; Fenimore et al., 2004; Oelbaum et al., 2004), SVC (Baroncini and Quackenbush, 2012; Oelbaum et al., 2008), and HEVC (Baroncini and Quackenbush, 2012; De Simone et al., 2011; Weerakkody et al., 2014). Independent researchers have also conducted subjective evaluations, both during and after the development phase of compression standards, as a validation process or to evaluate the codecs in different scenarios.

This chapter reports the results of performance analysis of HEVC for video (see Section 4.1) and image (see Section 4.2) compression. Additionally, Section 4.5 reports the performance analysis of potential coding technologies to further extends the capabilities of HEVC for HDR video compression. Sections 4.3 and 4.4 report the performance analysis of VP9 for video compression and JPEG XT for HDR image compression, respectively. Finally, Section 4.6 report the performance analysis of the MVC+D and 3D-AVC 3D video compression standards. The performance analysis performed on HDR and 3D video compression reported in this thesis were conducted in a collaboration with MPEG and reported to the standardization body. All subjective evaluations were conducted and analyzed following the guidelines described in Chapter 2 and coding efficiency was measured following the models described in Chapter 3.

4.1 Evaluation of HEVC Video Compression

The current trend in video consumption clearly shows that the already large quantity of video material distributed over broadcast channels, digital networks, and packaged media is going to increase in the coming years. As an effect of the growing popularity, the users' demand for increased resolution and higher quality is driving the efforts of the technological development. From this point of view, the evolution of video acquisition and display technologies is much faster than that of network capabilities. Thus, a clear need for a new video coding standard with higher efficiency when compared to the popular AVC codec (Wiegand et al., 2003a) was identified.

To develop the next-generation video coding standard, a group of video coding experts from VCEG and MPEG, called JCT-VC, has been created. The JCT-VC standardization effort is being referred to as HEVC. The new standard targets a wide variety of applications such as mobile TV, home cinema, and UHDTV. It aims at supporting next-generation acquisition and display devices featuring progressive scanned video with higher frame rates and resolutions (from WVGA to HDTV and UHDTV), as well as improved picture quality in terms of noise level, color gamut, and dynamic range. HEVC aims at a substantially improved coding efficiency compared to the AVC High Profile, i.e., reducing the bit rate requirements by half while keeping comparable image quality, but at the expense of increased computational complexity. Depending on the application scenario, a trade-off between computational complexity, compression ratio, robustness to errors, and processing delay should be supported.

A Joint CfP on Video Compression Technology (N11113) was issued by JCT-VC in January 2010. A total of 27 proposals were evaluated with respect to two AVC anchors in the largest subjective video quality testing effort ever conducted (Sullivan and Ohm, 2010). All proposals used a coding architecture conceptually similar to AVC, containing the following basic elements: (a) block-based coding, (b) variable block sizes, (c) block motion compensation, (d) fractional-pel motion vectors, (e) spatial intra prediction, (f) spatial transform of residual difference, (g) integer-based transform designs, (h) arithmetic or VLC-based entropy coding, and (i) in-loop filtering. However, the individual coding tools differed a lot between the indi-

vidual proposals. Key elements of some of the best proposals were combined to develop an initial Test Model, as a starting point for the definition of the new standard (Sullivan and Ohm, 2010). The initial Test Model was refined over the next JCT-VC meetings and, in January 2011, an official Test Model, named HEVC reference software (HM), was publicly released. The HM software integrates the latest developments that have been validated within the JCT-VC group and a new version is available at each JCT-VC meeting cycle.

The compression efficiency of different codecs can be reliably compared only by means of subjective tests, carried out according to common evaluation methodologies defined by experts. Therefore, the responses to the CfP were evaluated during a formal subjective test campaign (De Simone et al., 2011) and informal subjective tests were still carried out during the development of the standard to assess the improvements of the integrated coding tools (M22988; M23863). It was expected that HEVC could achieve double the compression efficiency of AVC, at the expense of a significant increase in computational complexity. In particular, it was expected that HEVC could achieve even better compression efficiency for resolutions beyond HDTV, especially due to increased prediction flexibility and a wider range of block sizes. However, until August 2012, no subjective evaluation had been performed on resolutions higher than HDTV, mostly because of hardware limitations and the lack of high quality uncompressed content. To address this problem, we performed the first subjective quality evaluation to benchmark the performance of HEVC and AVC on 4K/quad full high definition (QFHD) video content. This section reports the details and results of this performance analysis.

4.1.1 Dataset

At the time of this study, the availability of high quality 4K uncompressed video data free of use for research purpose was very limited. Only two contents were available to the JCT-VC group: *PeopleOnStreet* and *Traffic*. To cover a wider application scenario, synthetic content from the Sintel movie was included. Two synthetic scenes were included in the dataset: one for the test (*Sintel2*) and one for the training (*Sintel39*). The dataset was thus composed of four contents, one for the training and three for the test, with different visual characteristics, resolutions, and frame rates (see Table 4.1). Figure 4.1 shows the first frame of each content. Figure 4.2 shows the SI and TI indexes computed on the luminance component of each content (see Section 2.2). It can observed that sequences *Sintel2* and *Sintel39* have large TI values, whereas content *Traffic* shows a small TI index. Since the *Traffic* sequence is five seconds long only, it was decided to clip all contents to five seconds to maintain consistency during the test between the different contents. All test sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8-bit per sample.

The video sequences were compressed with AVC and HEVC using AVC reference software (JM) 18.3 and HM 6.1.1, respectively. The random access (RA) configuration was selected for this study since it gives better results than the low delay (LD) configuration. The group of pictures

(a) *PeopleOnStreet* (b) *Traffic*

(c) *Sintel2* (d) *Sintel39*

Figure 4.1: Sample frames of the individual contents considered in the subjective test.

Table 4.1: Dataset

Dataset	Video	Resolution	Framerate
Test	*PeopleOnStreet*	3840 × 2160	30
	Traffic	3840 × 2048	30
	Sintel2	3840 × 1744	24
Training	*Sintel39*	3840 × 1744	24

(GOP) size was set to 8 pictures and the Intra Period was set to 24 and 32 pictures for 24 and 30 fps contents, respectively. Hierarchical B-pictures were used, with a quantization parameter (QP) increase of 1 between each Temporal Level. The Coding Order was set to 0 8 4 2 1 3 6 5 7. The configuration parameters for AVC and HEVC were selected such that similarity was ensured between the two codecs to avoid penalization. For example, `BLevel0MoreRef` and `BIdenticalList` were set in the JM configuration file. More details on the configurations can be found in Table 4.2.

For each content and codec, five different bit rates were selected. Due to the different spatio-temporal characteristics of the contents and the presence of both natural and synthetic content, it was decided to select the targeted bit rates for each content separately. Since no Rate Control is implemented in HM 6.1.1, fixed QPs were used. Typical QPs for AVC are in the range of 25 to 37. First, a few sequences were compressed for each content using this range, keeping in mind the ~ 12.5% per QP rule (i.e., there is approximately a 12.5% bit rate reduction for every increase in QP), and trying to map the QPs of the HM to those of the JM. To be

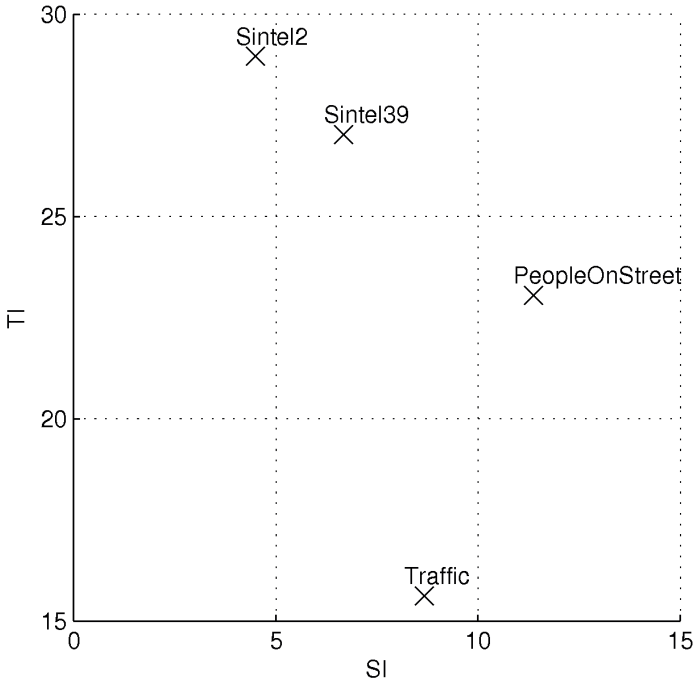


Figure 4.2: SI versus TI indexes of the selected contents.

Table 4.2: Selected encoder settings for AVC and HEVC.

Codec	AVC	HEVC
Encoder	JM 18.3	HM 6.1.1
Profile	High 5.1	Main
Reference Frames	4	4
R/D Optimization	On	On
Motion Estimation	EPZS	EPZS
Weighted Prediction	On	-
Search Range	128	64
GOP	8	8
Hierarchical Encoding	On	On
Temporal Levels	4	4
Intra Period	1s	1s
Deblocking	On	On
Rate Control	Off	-
8x8 Transform	On	-
Adaptive Loop Filter	-	Off
Coding Unit size / depth	-	64 / 4
Transform Unit size min / max	-	4 / 32

realistic, it was decided to set the upper bit rate limit to 20 Mbit/s. Then, an expert screening session was conducted to select the lower and upper bounds for each content separately, keeping in mind the standard QP range and targeting realistic bit rates, to try to cover the full quality scale. Finally, the targeting bit rates were refined and validated during a second expert screening session. The training material was selected during the last expert screening session

Table 4.3: Targeted bit rates (Mbit/s).

Content	Codec	R1	R2	R3	R4	R5
PeopleOnStreet	AVC	5.000	7.000	10.000	14.000	20.000
	HEVC	5.000	7.000	10.000	14.000	20.000
Traffic	AVC	3.500	5.000	7.000	10.000	14.000
	HEVC	2.500	3.500	5.000	7.000	10.000
Sintel2	AVC	1.200	1.600	2.000	2.500	3.500
	HEVC	0.768	1.200	1.600	2.000	2.500

Table 4.4: Actual bit rates (Mbit/s).

Content	Codec	R1	R2	R3	R4	R5
PeopleOnStreet	AVC	4.743	6.799	9.454	14.561	20.745
	HEVC	4.889	6.960	9.833	13.871	20.278
Traffic	AVC	3.490	4.914	7.208	9.429	14.717
	HEVC	2.277	3.346	4.997	6.720	10.474
Sintel2	AVC	1.205	1.571	1.935	2.389	3.455
	HEVC	0.705	1.204	1.616	1.903	2.674

Table 4.5: Quantization Parameters.

Content	Codec	R1	R2	R3	R4	R5
PeopleOnStreet	AVC	44	41	38	34	31
	HEVC	42	39	36	33	30
Traffic	AVC	37	34	31	29	26
	HEVC	38	35	32	30	27
Sintel2	AVC	35	32	30	28	25
	HEVC	32	28	26	25	23

to cover the full quality scale. For the three intermediate quality levels, examples of both AVC and HEVC degradations with similar strengths were selected. Tables 4.3 and 4.4 report the complete sets of targeted and actual bit rates, respectively. Table 4.5 report the QPs used to encode these sequences.

4.1.2 Methodology

Natural playback in native spatial and temporal resolutions of raw 4K/QFHD video sequences at 30 fps requires specific hardware. Particularly, reading and displaying in real time YUV 4:2:0 color subsampled QFHD (3840 × 2160 pixels) video sequences at 30 fps requires a data rate of 373.25 MB/s. Since the typical reading speed of current Hard Disk Drives (HDD) is below 160 MB/s, a hardware solution based on Solid State Drives (SSD) was adopted.

To display 4K/QFHD content, a 56” professional high-performance 4K/QFHD LCD reference

Figure 4.3: Experimental setup.

monitor Sony Trimaster SRM-L560 was used. The monitor consists of four full HD panels. The panels are driven by four display ports and mutually synchronized by the graphic board of the video server to prevent any tearing effect. This monitor can operate in tree different modes (4K/QFHD, Quad View, and 2K/HD Zoom), while only the first one is available when DVI inputs are used.

To assure the reproducibility of results by avoiding involuntary influence of external factors, the laboratory for subjective video quality assessment was set up according to Section 2.1. The monitor was calibrated using an EyeOne Display2 color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level. The room was equipped with a controlled lighting system that consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid grey. The illumination level measured on the screens was 20 lx and the ambient black level was 0.2 cd/m^2. The test area was controlled by an indoor video security system to keep track of all the test activities and of possible unexpected events, which could influence the test results. The experiments involved three subjects assessing the test material, seated in one row perpendicular to the center of the monitor, at a distance of about 3.5 times the height of the display.Figure 4.3 depicts the MMSPG test environment where assessments took place.

Test Method

Since the test sequences are only five seconds long and subjects were not used to watch UHDTV, the DSIS Variant II method with a continuous five-level impairment scale (see Section 2.4) was chosen to perform the subjective quality assessment experiments.

Test Planning

Since the evaluation task requires a lot of attention due to the short sequences duration, it was decided to split the test in sessions that are no longer than 15 min each, followed by a resting phase. Furthermore, to avoid a possible effect of the presentation order, the stimuli are randomized in a way that the same content is never shown consecutively. One DSIS Variant II presentation took about 46 s. We had to evaluate a total of 30 test sequences (2 codecs × 3 contents × 5 bit rates), thus it was decided to split the test in two sessions. We decided to include two dummy presentations and one reference vs reference pair at the beginning of the first session. The first test session contained 18 presentations (2 dummies + 1 ref vs. ref + 15 stimuli), corresponding to about 14 min. The second test session contained 15 presentations (15 stimuli), corresponding to about 11.5 min.

The test was planned over two days, with three time slots per day. Each time slot was attended by six subjects, which were split into two groups of three subjects each. While one group was evaluating one session in the test room, the other group was resting in a separate room. A total of 36 naïve people took part in the test campaign. 30% of the observers were female and the age of the subjects ranged from 20 to 61 years old, with a median of 25 years old. All participants were screened for correct visual acuity and color vision using Snellen charts and Ishihara charts, respectively.

The training of the subjects of each group was conducted before the first test session, as a 10 min training session, where oral instructions were provided to explain the task and a viewing session was performed to allow the subject to familiarize with the assessment procedure. The video sequences used as training samples had quality levels representative of the labels reported on the rating scales: the experimenter explained the meaning of each label reported on the scale and related them to the presented sample sequences.

To collect evaluation scores, subjects were provided with scoring sheets to enter their quality scores. The scores were then offline converted into electronic version. All the scores were converted by one operator and crosschecked by a second operator to identify and correct any eventual manual mistake.

Data Processing

To detect and remove subjects whose scores appear to deviate strongly from the other scores in a session, outlier detection was performed. The outlier detection was applied to the set of results obtained from the 36 subjects. The boxplot inspired outlier detection technique proposed by De Simone et al. (2011) (see Section 2.6.1) was used. In this study, no outlier subjects were detected. Then, the MOSs were computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% CIs, assuming a Student's t-distribution of the scores.

4.1.3 Results

Figure 4.4 shows the resulting PSNR and MOS/CI plots for the different contents. As it can be seen from the small CIs, the results are reliable and the variations between the subjects are rather small. The subjective results show that, especially for lower bit rates, the performance of HEVC exhibits a substantial quality improvement compared to AVC.

Traffic is relatively easy to encode since it has a small TI index. Therefore, bit rates as low as 5 Mbit/s and 2 Mbit/s for AVC and HEVC, respectively, are evaluated as transparent. *PeopleOnStreet* is more challenging since it has higher SI and TI indexes, but mostly because artifacts are more visible in the upper left corner due to higher sensitivity of the HVS in low intensity areas (Weber law). For this content, blockiness was perceived in AVC encoded sequences, whereas the content was smoothed out in HEVC encoded sequences, which is less annoying. For the synthetic content, HEVC exhibits a significant improvement over AVC and very low bit rates can be achieved due to the absence of camera noise in the original content. A bit rate as low as 1.2 Mbit/s is perceived as transparent with HEVC, whereas the same bit rate for AVC is evaluated as annoying.

To accurately analyze the performance of HEVC and evaluate whether the obtained results were significantly different from those obtained with AVC, a multiple comparison significance procedure has been applied to the data, for each combination of content and bit rate separately. To identify the test conditions that resulted in statistically different MOSs, a one-way ANOVA and multiple comparison tests were performed, considering as treatment the combination of codec and bit rate. Figure 4.5 shows the results comparing all the possible pairs of treatments, for each content separately. Comparing the two codecs at similar bit rates, HEVC outperforms AVC for four bit rates out of four for *Sintel2* (1.2, 1.6, 2, and 2.5 Mbit/s) and for four bit rates out of five for *PeopleOnStreet* (5, 7, 10, and 14 Mbit/s), whereas only for one bit rate out of four for *Traffic* (3.5 Mbit/s). For the remaining bit rates, the codecs show the same performance. A two-way ANOVA, considering the codec and the bit rate as two separate treatments, has also been performed, resulting in a significant codec effect and significant bit rate effect, but irrelevant interaction effect.

Table 4.6 reports the average coding efficiency for HEVC over AVC computed using the Bjøntegaard and SCENIC models (see Chapter 3). It can be noticed that BD-PSNR under estimates the actual bit rate reduction, especially for *PeopleOnStreet*. For this content, BD-PSNR under estimates the actual gain because PSNR does not fully capture the difference between AVC and HEVC artifacts. For *Sintel2*, the values are very similar since the relation between MOS and PSNR is almost linear for the considered bit rates. In the case of *Traffic*, the difference is due to the saturation effect in perceived quality, which is not captured by PSNR.

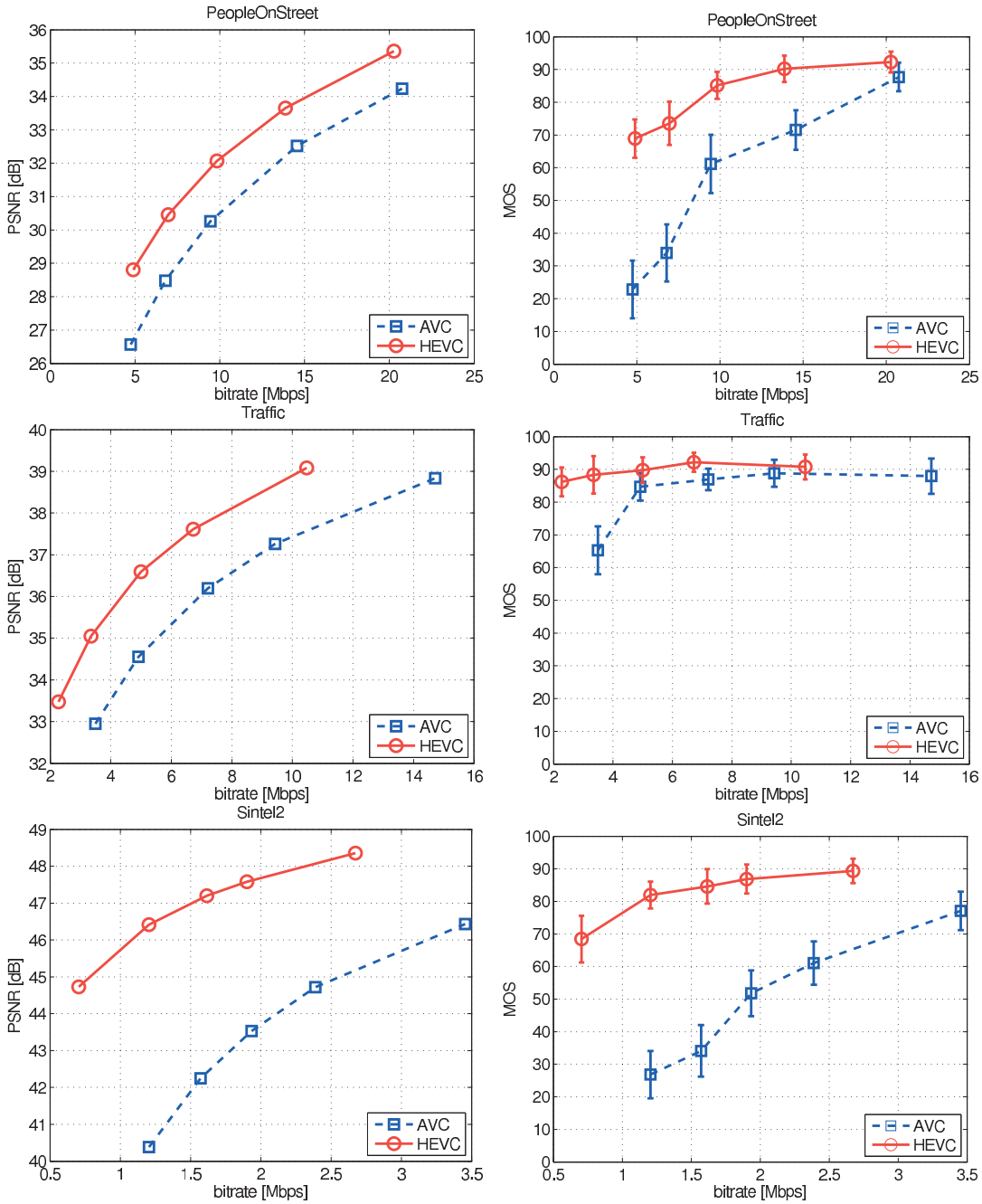


Figure 4.4: R-D curves: PSNR versus bit rate (left) and MOS versus bit rate (right).

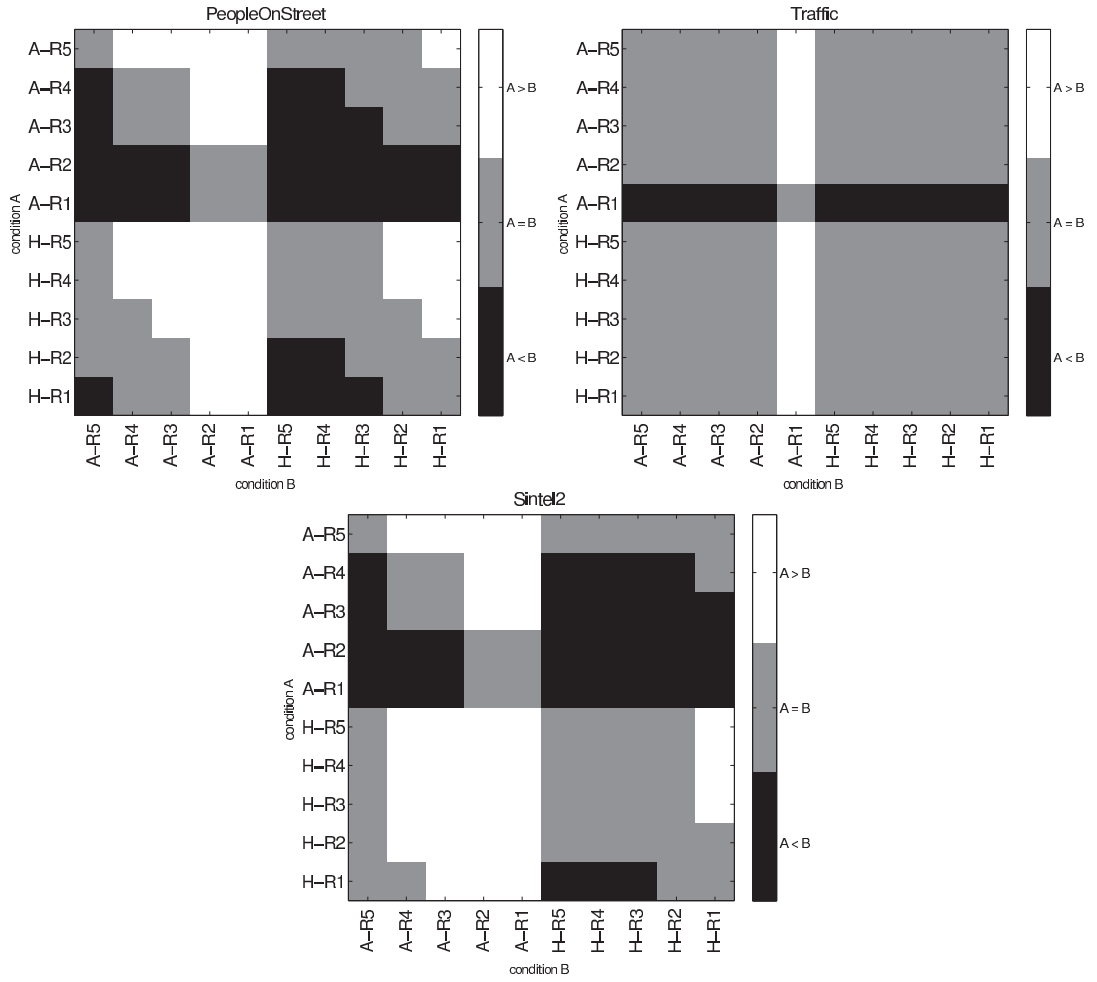


Figure 4.5: Results of the multiple comparison test for the different test conditions, i.e., combination of codec (A stands for AVC and H stands for HEVC) and bit rate (R1 to R5). In each plot, the color of each square shows the result of the significance test between the MOSs related to the two test conditions reported in the corresponding column and row. A white (black) square indicates that the MOS corresponding to condition A is statistically significantly better (worse) than the MOS corresponding to condition B, whereas a grey square indicates that there is no sufficient statistical evidence indicating differences between the two MOSs.

Table 4.6: Average coding efficiency for HEVC over AVC.

(a) Bjøntegaard model.

Content	Delta rate (%)	Delta PSNR (dB)
PeopleOnStreet	−27	+1.6
Traffic	−38	+1.8
Sintel2	−68	+4.4
Overall	−44	+2.6

A negative (positive) value indicates a decrease (increase) of bit rate for the same PSNR.
A negative (positive) value indicates a decrease (increase) of PSNR for the same bit rate.

(b) SCENIC model.

Content	Delta rate ΔR [$\Delta R_{\min}, \Delta R_{\max}$]	Delta MOS ΔD [$\Delta D_{\min}, \Delta D_{\max}$]	Confidence index (%)
PeopleOnStreet	−53% [−69%,−27%]	+25.8 [+13.0,+38.4]	79
Traffic	−59% [-,−5%]	+10.8 [−2.2,+20.3]	28
Sintel2	−73% [-,−60%]	+40.7 [+28.9,+52.4]	62
Overall	−62% [-,−31%]	+25.8 [+13.2,+37.1]	56

A negative (positive) delta rate indicates a decrease (increase) of bit rate (MOS) for the same MOS (bit rate).
A negative (positive) delta MOS indicates a decrease (increase) of bit rate (MOS) for the same MOS (bit rate).

4.2 Evaluation of HEVC Image Compression

As showed in the previous section, HEVC is demonstrating significant quality gains when compared to state of the art video codecs such as AVC. Such effectiveness in video compression suggests the potential efficiency of using HEVC intra coding for still images. In particular, when compared to previous standards, the following features of HEVC can contribute to improving coding efficiency for still images (Lainema et al., 2012): (a) quadtree-based coding structure following the HEVC block coding architecture, (b) angular prediction with 33 prediction directions, (c) planar prediction to generate smooth sample surfaces, (d) adaptive smoothing of the reference samples, (e) filtering of the prediction block boundary samples, (f) prediction mode dependent residual transform and coefficient scanning, and (g) intra mode coding based on contextual information. The coding efficiency of HEVC intra coding for still image compression was investigated in a few studies that compare still images compression standards with HEVC intra coding by using PSNR as an objective metric for visual quality (JCTVC-I0461; JCTVC-I0595). These objective evaluations demonstrated that HEVC can achieve a considerable gain even compared to the state of the art JPEG 2000 compression standard.

However, the PSNR metric, despite its popularity in visual quality evaluations, does not accurately reflect perceptual visual quality of the HVS (Sheikh et al., 2006). In addition, the lack of standardization in the field of objective quality assessment and the lack of extensive and commonly accepted comparisons of the different metrics make the PSNR-based assessments rather questionable. Therefore, to fully confirm the claim raised by objective evaluations on the effectiveness of HEVC intra coding for still images, a formal subjective evaluation is

(a) *p04* (b) *p14* (c) *p22* (d) *p30* (e) *p01*

(f) *p06* (g) *p10* (h) *bike* (i) *cafe* (j) *woman*

Figure 4.6: Dataset: training set (a-d) and testing set (e-j).

necessary. To address this problem, we performed both objective and subjective evaluations of HEVC intra coding for still image compression following the guidelines defined by the JPEG committee for the evaluation of JPEG XR (De Simone et al., 2009b). HEVC intra coding was compared to the existing JPEG and JPEG 2000 (both 4:2:0 and 4:4:4 configurations are used) standards using high resolution 24 bpp images. The compression efficiency was evaluated by means of PSNR objective metric, for comparison with previous work, and subjective tests, which were conducted in a specific testing environment and following formal evaluation methodology. This section reports the details and results of this performance analysis.

4.2.1 Dataset

The dataset from the JPEG XR evaluation (De Simone et al., 2009b) was used in this study. All the images had a resolution of 1280×1600 pixels and were available in RGB 4:4:4 uncompressed format. The whole image set was split into a training set of 4 images (referred to as *p04*, *p14*, *p22*, and *p30*) and a testing set of 6 images (referred to as *p01*, *p06*, *p10*, *bike*, *cafe*, and *woman*). Figure 4.6 provides an overview of the dataset. This set of images was coded using the 3 codecs and 4 different coding configurations described below. Similarly to the JPEG XR evaluation, the following bit rates were selected: 0.25, 0.50, 0.75, 1.00, 1.25 and 1.50 bpp. Thus, this resulted in a final test set of 144 coded images used for the subjective evaluation.

The JPEG compressed images were produced using the IJG implementation, version 6b. The images were coded in Baseline Profile and the target coding bit rates were controlled by varying the *quality factor* input parameter.

For JPEG 2000 coding, the Kakadu implementation version 6.0 was used. Two different configurations were considered. The first configuration uses chrominance subsampling, which requires external pre- and post-processing steps. Since the weighting tables in JPEG 2000 have been designed and optimized for 4:4:4 content, visual weighting was disabled in this configuration. The following parameters were used:

(i) pre-processing: RGB to YCbCr conversion and 4:4:4 to 4:2:0 downsampling,
(ii) 64 × 64 code block size, 1 layer, no precincts, 9 × 7 wavelets, and 5 decomposition levels,
(iii) no visual weighting, and
(iv) post-processing: 4:2:0 to 4:4:4 upsampling and YCbCr to RGB color conversion.

As visual weighting impacts the performance of the JPEG 2000 codec, a second configuration with visual weighting enabled was also included in the evaluations. The parameters in this second configuration were the same as before but the pre- and post-processing steps were discarded and the RGB 4:4:4 images were encoded directly without any subsampling. The *rate control* option was used to encode the images at the target coding bit rates.

For HEVC intra coding, the HM version 8.0rc2 was used. As for JPEG 2000, the images were converted from RGB 4:4:4 to YCbCr 4:2:0 prior to encoding and then back-converted to obtain the final decoded image. The images were coded in Main Intra Profile and the target bit rates were obtained by varying the QP.

4.2.2 Methodology

The experiments were conducted at the MMSPG quality test laboratory, which fulfills the recommendations for the subjective evaluation of visual data issued by ITU. The laboratory setup was intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors. The test area was controlled by an indoor video security system to keep track of all the test activities and possible unexpected events, which could affect the test results.

An Eizo CG301W LCD monitor with a native resolution of 2560 × 1600 pixels was used to display the test stimuli. The monitor was calibrated using an EyeOne Display2 color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level. The room was further equipped with a controlled lighting system that consists of neon lamps with 6500 K color temperature, whereas the color of all the background walls and curtains present in the test area were in mid grey. The illumination level measured on the screen was 15 lx and the ambient black level was 0.2 cd/m^2.

The experiment involved only one subject per display assessing the test materials. Subjects were seated in line with the center of the monitor, at a distance approximately equal to the height of the screen, but were encouraged to vary the viewing distance whenever needed, to inspect the high-resolution image shown on the screen.

Test Method

The subjective quality evaluations to compare the image compression algorithms described in Section 4.2.1 were conducted following the method proposed in (De Simone et al., 2009b). As an adaptation of the DSCQS method for video quality evaluation (see Section 2.4.4), the selected method implies that two images are displayed simultaneously by splitting the screen horizontally into two parts. One of the two images was always the reference (unimpaired) image. The other was the test image, which in this study was a compressed version of the reference. The position of the reference image on the screen was randomly selected at each visualization. Instead of judging the quality of both images, the subject was asked to detect the impaired image in the pair and rate its quality, using a continuous five-level quality scale.

Training Session

Before the test starts, oral instructions were provided to the subject to explain his/her task. Additionally, a training session was organized to allow subjects to familiarize with the assessment procedure and the graphical user interface. The contents shown in the training session were not used in the test session and the data gathered during the training were not included in the final test results. The four training contents, shown in Figure 4.6, were coded with the different codecs and bit rates described in Section 4.2.1. Five training samples were manually selected by expert viewers so that the quality of samples were representative of all categorical quality levels on the rating scale. The training materials were presented to subjects exactly as for the test materials, thus in side by side image pairs, where one of the two stimuli was always the unimpaired image.

Test Sessions

Since the total number of test samples was too large for a single test session, the overall experiment was split into 4 sessions of approximately 13 min each. After each session, each subject took a 5 min break before starting the next session. Each session included test materials corresponding to 3 contents (*p01*, *p06*, *p10* in sessions 1 and 3 and *bike*, *cafe*, *woman* in sessions 2 and 4), all the codecs under analysis, and only a subset of the bit rates, which were uniformly distributed across all the sessions.

Four dummy pairs, whose scores were not included in the results, were included at the beginning of each session to stabilize the subjects' ratings. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each subject, whereas the same content was never shown consecutively.

A total of 22 subjects, 6 female and 16 male, took part in the test, completing all the test sessions. All participants were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

Data Processing

To detect and remove subjects whose scores appear to deviate strongly from others in a session, outlier detection was performed. The outlier detection was applied to the set of results obtained from the 22 subjects and performed according to the guidelines described in Section 2.3.1 of Annex 2 of ITU-R BT.500-13 (2012). In this study, 2 outliers were detected in session 1 and 1 outlier was detected in session 2. Then, the MOSs were computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% CIs, assuming a Student's *t*-distribution of the scores.

4.2.3 Results

Figure 4.7 shows the PSNR based R-D performance for all compression algorithms and contents. It is clear that HEVC outperforms other codecs across the majority of contents and through most investigated bit rates. JPEG 2000 with 4:2:0 sampling configuration is the only competitive compression algorithm in comparison to HEVC, especially for content *p01*. The observed performance difference of JPEG 2000 4:2:0 and HEVC in terms of PSNR is between 0.0 - 3.0 dB for all tested bit rates and contents. Furthermore, the PSNR improvement of HEVC relative to JPEG 2000 4:4:4 and JPEG varies through all tested contents and bit rates between 0.7 - 4.9 dB and 1.1 - 8.6 dB, respectively.

Similar results for still image compression performance comparison based on objective metrics have been reported in two recent studies. Using the dataset containing, among others, the images described in Section 4.2.1, HEVC HM 6.0 encoder and reference software encoders for other standards, JCTVC-I0595 reports an average bit rate reduction of 43% and 22.6% for HEVC intra coding over JPEG and JPEG 2000 4:2:0, respectively. Additionally, JCTVC-I0461 reports an average bit rate reduction of 56% over JPEG. The BD-Rate values computed with the Bjøntegaard model (see Section 3.1) and reported in Table 4.7 are similar to those reported in above mentioned studies and confirm that, according to objective evaluations based on PSNR, a significant bit rate reduction can be achieved for HEVC intra coding over the JPEG standards.

Figure 4.8 shows the subjective R-D plots illustrating the MOS and CI values for each content. For each content, the MOS values span the entire range of quality levels. The only exception to this overall behavior is on content *cafe*, whose structure is sensitive to compression artifacts and therefore, even for the highest bit rate, the image quality is rated below 90.

An overall impression of the performance of the different codecs can be obtained when looking closely at the R-D plots in Figure 4.8. In general, all examined coding standards have the same or very similar performance at the highest bit rate. However, at lower bit rates, the performance of individual coding algorithms varies significantly depending on the content. Although HEVC outperforms (particularly at bit rates below 1.00 bpp) other coding algorithms for contents *bike*, *cafe*, and *p10*, its performance is quite comparable to both versions of JPEG 2000 for

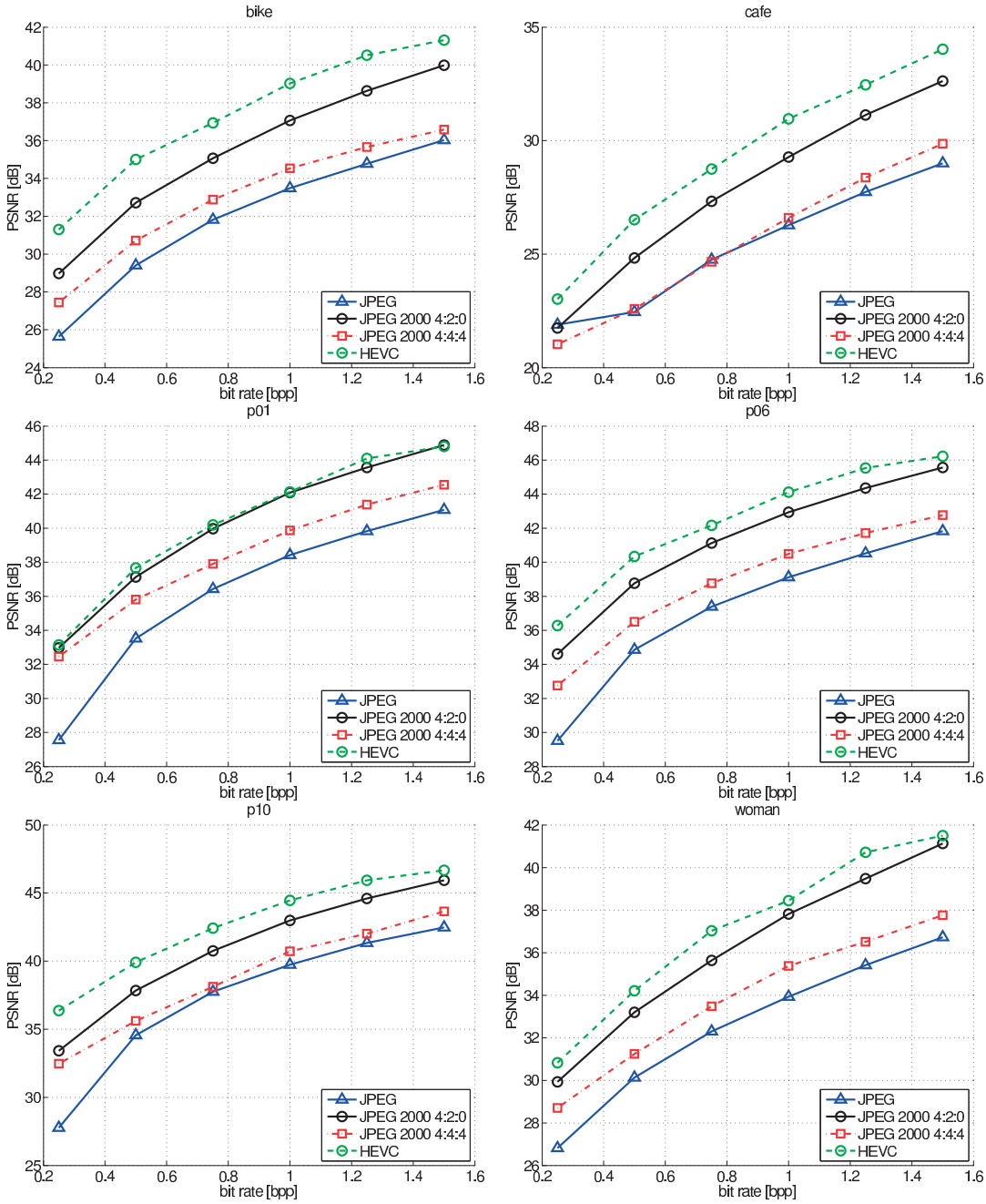


Figure 4.7: R-D performance.

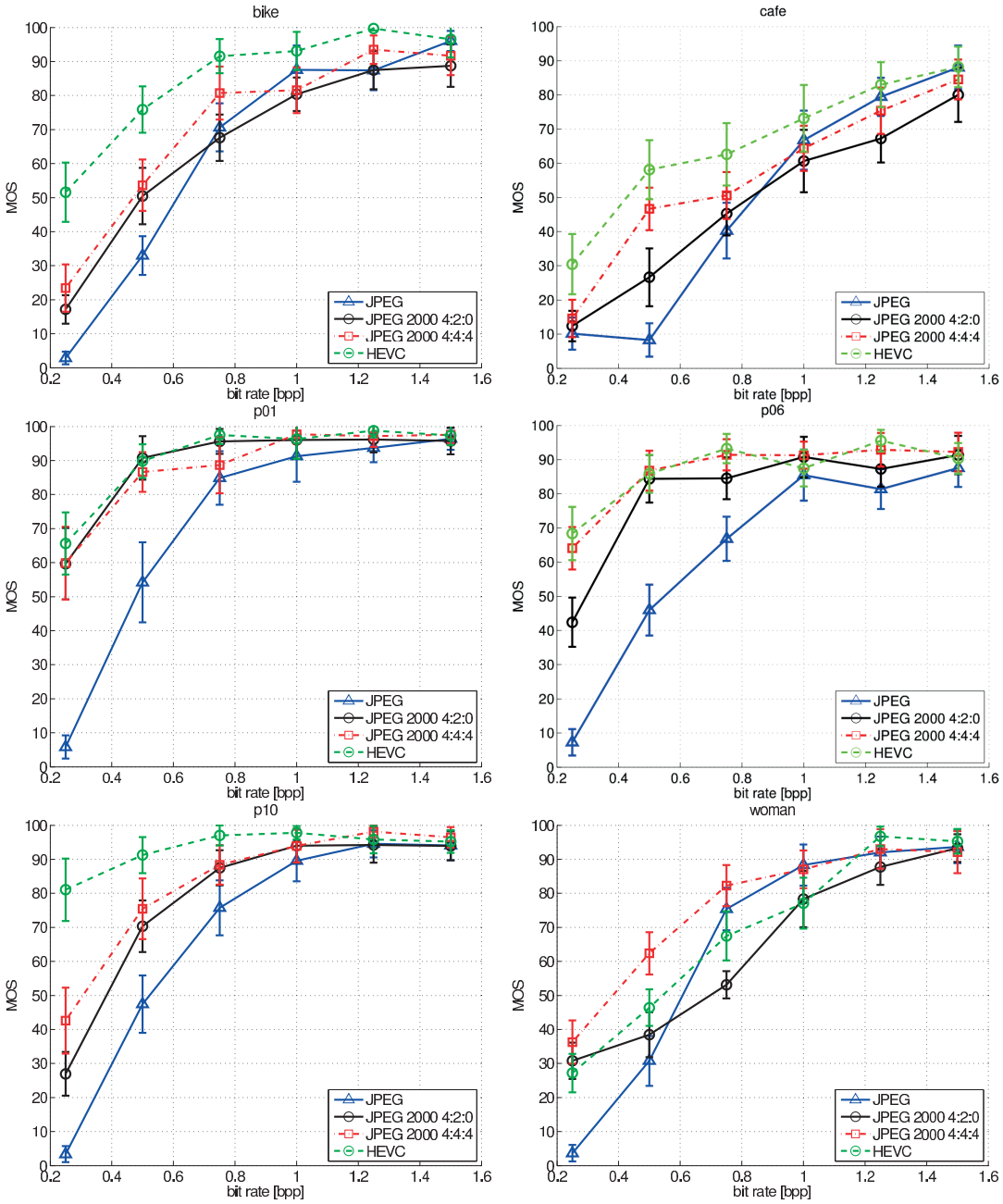


Figure 4.8: Mean opinion scores vs. bit rate for the different compression algorithm across the test images.

Table 4.7: Average coding efficiency: Bjøntegaard model.

Encoding	Average bit rate difference relative to (%)				Average PSNR difference relative to (dB)			
	JPEG	J2000 4:2:0	J2000 4:4:4	HEVC	JPEG	J2000 4:2:0	J2000 4:4:4	HEVC
JPEG	-	+71	+22	+119	-	−3.4	−1.3	−4.8
JPEG 2000 4:2:0	−41	-	−30	+25	+3.4	-	+2.1	−1.4
JPEG 2000 4:4:4	−17	+43	-	+84	+1.3	−2.1	-	−3.5
HEVC	−54	−19	−44	-	+4.8	+1.4	+3.5	-

A negative (positive) bit rate difference indicates a decrease (increase) of bit rate for the same PSNR.
A negative (positive) PSNR difference indicates a decrease (increase) of PSNR for the same bit rate.

Table 4.8: Average coding efficiency: SCENIC model.

Encoding	Average bit rate difference relative to			
	JPEG	JPEG 2000 4:2:0	JPEG 2000 4:4:4	HEVC
(a) Delta rate				
JPEG	-	+43% [+10%,+90%] (100%)	+63% [+23%,+128%] (100%)	+101% [+41%,+219%] (99%)
JPEG 2000 4:2:0	−23% [−40%,−3%] (100%)	-	+22% [−14%,+73%] (74%)	+57% [+4%,+120%] (74%)
JPEG 2000 4:4:4	−35% [−51%,−16%] (100%)	−15% [−41%,+26%] (74%)	-	+33% [−14%,+125%] (68%)
HEVC	−44% [−60%,−25%] (99%)	−31% [−51%,+3%] (74%)	−17% [−47%,+27%] (68%)	-
	Average MOS difference relative to			
	JPEG	JPEG 2000 4:2:0	JPEG 2000 4:4:4	HEVC
(b) Delta MOS				
JPEG	-	−14.5 [−26.4,−2.4] (100%)	−20.6 [−33.1,−7.9] (100%)	−27.8 [−39.0,−16.5] (99%)
JPEG 2000 4:2:0	+14.5 [+2.4,+26.4] (100%)	-	−7.1 [−20.3,+6.0] (74%)	−14.1 [−27.1,−1.0] (74%)
JPEG 2000 4:4:4	+20.6 [+7.9,+33.1] (100%)	+7.1 [−6.0,+20.3] (74%)	-	−6.3 [−18.8,+6.0] (68%)
HEVC	+27.8 [+16.5,+39.0] (99%)	+14.1 [+1.0,+27.1] (74%)	+6.3 [−6.0,+18.8] (68%)	-

(a) A negative (positive) value indicates a decrease (increase) of bit rate for the same MOS. Reading: ΔR [$\Delta R_{\min}, \Delta R_{\max}$] (Confidence index).
(b) A negative (positive) value indicates a decrease (increase) of MOS for the same bit rate. Reading: ΔD [$\Delta D_{\min}, \Delta D_{\max}$] (Confidence index).

Table 4.9: Results of the multiple comparison test expressed in terms of number of contents for which HEVC performs better, equal, or worse than the other codecs.

Condition			Bit rate (bpp)					
			0.25	0.50	0.75	1.00	1.25	1.50
HEVC	>	JPEG	6	6	5	1	3	0
HEVC	=	JPEG	0	0	1	4	3	6
HEVC	<	JPEG	0	0	0	1	0	0
HEVC	>	JPEG 2000 4:2:0	4	3	5	1	4	0
HEVC	=	JPEG 2000 4:2:0	2	3	1	5	2	6
HEVC	<	JPEG 2000 4:2:0	0	0	0	0	0	0
HEVC	>	JPEG 2000 4:4:4	3	3	4	1	1	0
HEVC	=	JPEG 2000 4:4:4	2	2	1	4	5	6
HEVC	<	JPEG 2000 4:4:4	1	1	1	1	0	0

contents *p01* and *p06*. Moreover, HEVC shows always better or equal performance than JPEG with the exception of content *woman*. Looking at the MOS results of the image *woman*, which consists in a woman's face portrait, one can see that HEVC is outperformed by JPEG and JPEG 2000 4:4:4. Whereas JPEG outperforms HEVC only at 0.80 bpp and 1.00 bpp, JPEG 2000 4:4:4 seems to be better for all bit rates below 1.00 bpp (1.00 bpp included). This might be explained by the specific banding artifacts introduced by HEVC at lower bit rates for this particular content. Such banding artifacts are subjectively more disturbing in comparison to the typical blurring effect introduced by JPEG 2000 4:4:4 coding.

Table 4.8 reports the average coding efficiency computed from the MOSs using the SCENIC model (see Section 3.3). The estimated bit rate saving based on MOS for HEVC relative to JPEG, JPEG 2000 4:2:0, and JPEG 2000 4:4:4 is about 44%, 31%, and 17%, respectively. The differences between the outcome of the Bjøntegaard and SCENIC models shows the importance of subjective tests to determine a more realistic estimation of the achievable bit rate reduction.

Interesting observations can be made by looking at the mutual comparison of both versions of JPEG 2000. Although JPEG 2000 4:2:0 performs always better than JPEG 2000 4:4:4 in terms of PSNR, the subjective results dictate the opposite. This might be explained by the fact that visual weighting was disabled for JPEG 2000 4:2:0 while it was enabled for the second color sampling configuration of JPEG 2000. The lack of the visual weighting creates strong distortions, especially on the skin texture at lower bit rates, as reported during development of JPEG 2000 standard, which is not captured by PSNR based metric.

Table 4.9 reports the results of the multiple comparison test, detecting the significant difference pairwise among individual codecs and comparing the performance of HEVC to all other codecs for all test conditions. These results confirm all the findings from the R-D plots. Although at the highest bit rate, all compression standards perform equally, at bit rates lower than 1.00 bpp, HEVC performs usually better, or at least equal, when compared to all other

standards, except for JPEG 2000 4:4:4 on content *woman*. JPEG 2000 4:4:4 is the second best performing compression algorithm while its performance is the same as for HEVC in 20 out of 36 cases. On the other hand, JPEG performs practically always worse than HEVC.

4.3 Evaluation of VP9 Video Compression

Recent dramatic increase in video consumption over IP-networks, with video data taking more than 75% of Internet traffic, prompted for the development of new video compression technologies that would be significantly more efficient than the existing video codecs, including the popular AVC. The development efforts led to the creation of two video codecs: HEVC and VP9 Mukherjee et al., 2015a. VP9 is an open source alternative to HEVC developed by Google and is positioned as a royalty-free, license-fee-free solution, with the main focus on supporting Internet-based video consumption.

The fact that VP9 was released at a similar time frame to HEVC and that it was announced as a superior alternative raised interest in the research and professional communities. It resulted in several studies comparing these two codecs to each other and to AVC. Most such studies relied on objective metrics to measure coding efficiency and resulted in conflicting conclusions depending on the study performed. In (Grois et al., 2013), the authors claim that VP9 is inferior to both AVC and HEVC and demonstrate that HEVC provides average bit rate savings of 43.3% compared to VP9. However, a different study by Mukherjee et al. (2013) comes to a different conclusion, with VP9 showing similar compression efficiency when compared to HEVC and a significantly higher compression efficiency when compared to AVC. Such conflicting conclusions are mainly caused by different usage scenarios assumed in the papers and by different encoding configurations used. The authors of (Grois et al., 2013) have further extended their study to a LD scenario (Grois et al., 2014), which is more suitable for real-time video applications, and by using PSNR measurements, and conclude that using HEVC results in average bit rate savings of 32.5% when compared to VP9.

In the above studies, authors relied only on PSNR as objective metric to compare compression efficiency of selected encoding schemes. However, human perception is subjective, and results of subjective assessments performed using standard quality evaluation methodologies is *a priori* a more reliable measure of compression efficiency. Therefore, a subjective evaluation of HEVC, VP9, and AVC codecs was performed by Rerabek and Ebrahimi (2014) to determine the actual perceived quality of compressed video content. The study assumed a broadcasting scenario using UHD video content in a standard test laboratory environment with controlled lighting conditions and a professional UHD reference monitor. According to the subjective evaluation results, HEVC outperformed VP9, showing on average a nearly 50% bit rate reduction for the same subjective quality.

However, no subjective quality evaluation had been performed to validate or refute the findings of Grois et al. (2014) on the LD configuration. To address this problem, we performed the first subjective quality evaluation to compare the compression efficiency between HEVC, VP9, and

AVC assuming a real-time Internet-based streaming scenario. In such a scenario, subjects receive a real-time streamed video content and watch it in a web browser in an uncontrolled environment. HD content is typical for current video consumption over the Internet and is compressed using parameters most suitable for Internet-based scenario. In our experiments, a total of 26 subjects took part in a crowdsourcing subjective assessment, evaluating 8 different video contents with resolutions ranging from 720p to 1080p, which were compressed to four different bit rates using HEVC, VP9, and AVC.

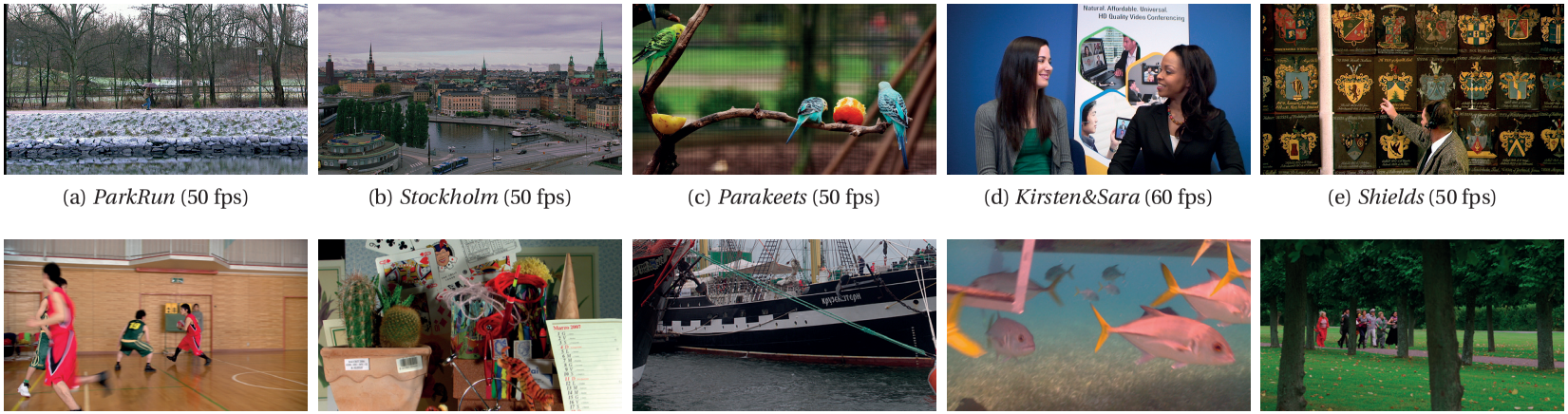
4.3.1 Dataset

Ten video sequences were used in the experiments, with different spatial and temporal characteristics, resolutions, and frame rates. Eight sequences were used for the subjective tests and two sequences were used for training. Figure 4.9 shows a representative frame sample of each video sequence. Each video sequence was ten seconds long and stored as raw, progressively scanned video file, with YCbCr 4:2:0 color sampling and 8 bits per sample. Furthermore, each video file was encoded with all three evaluated codecs at four bit rates. Since fixed QP configuration was used to control the quality of AVC, HEVC, and VP9 compressed bit streams, the sequences were first encoded at various QP values. Then, an expert screening session was conducted to select the lower and upper QP bounds for each content separately (including training), by targeting bit rates defined in (De Simone et al., 2011) and trying to cover the full quality scale for each content. Table 4.10 reports the final sets of targeted (R1' - R4') and actual (R1 - R4) bit rates, with corresponding QPs, for each codec.

Codecs Configuration

For HEVC, the HM reference software version 16.2 was selected, as it is a popular implementation. The latest version of the VP9 codec released by Google, i.e., release v1.3.0-4786-gbf44117, was selected and used in our experiments. Finally, the x264 library, release r2491, was used to evaluate the performance of AVC based coding scheme as it is fast, publicly available, and one of the most commonly used implementations of AVC. For each codec, the fixed quality parameter was set separately. Such setting allows fair mutual comparison of encoders as it removes all rate control adaptation between video frames. A more detailed description of the selected encoders, including their profiles and parameters configuration, is presented further in this section.

The latest versions of the HM reference software was used for encoding video sequences with HEVC. The LD configuration in default main profile with B frames was selected. LD configuration with B frames was selected since is achieves higher coding efficiency (because of bi-prediction), when comparing to low-delay configuration with P frames only. In this configuration, the first frame is encoded as an I frame and subsequent frames are encoded as B frames, while reordering of the B frames is not allowed, i.e., only the reference picture list 0, which references to past frames, is used. Therefore, this configuration introduces minimal coding delay and can be used for real-time application scenarios.



(a) *ParkRun* (50 fps) (b) *Stockholm* (50 fps) (c) *Parakeets* (50 fps) (d) *Kirsten&Sara* (60 fps) (e) *Shields* (50 fps)

(f) *Basketball* (50 fps) (g) *Cactus* (50 fps) (h) *Seedof* (60 fps) (i) *Underboat* (30 fps) (j) *ParkJoy* (50 fps)

Figure 4.9: Sample frames of individual contents used in the subjective evaluation campaign. Contents (a) - (e) and (f) - (j) have a resolution of 1280 × 720 and 1920 × 1080 pixels, respectively. Contents Shields (e) and ParkJoy (j) were used for training. Contents (c) - (d) and (f) - (j) were used in HEVC verification test, whereas the remaining contents were downloaded from Derf's dataset.

Table 4.10: Targeted (R') and actual (R) bit rates including the corresponding QP values for each codec.

Sequence	R1'	R1						R2'	R2						R3'	R3						R4'	R4					
		AVC		HEVC		VP9			AVC		HEVC		VP9			AVC		HEVC		VP9			AVC		HEVC		VP9	
	(kbps)	QP	(kbps)	QP	(kbps)	QP	(kbps)	(kbps)	QP	(kbps)	QP	(kbps)	QP	(kbps)	(kbps)	QP	(kbps)	QP	(kbps)	QP	(kbps)	(kbps)	QP	(kbps)	QP	(kbps)	QP	(kbps)
(a) 1280 × 720 *pixels*																												
ParkRun	384	50	374	45	376	62	773	512	48	530	43	603	61	990	850	46	818	41	941	60	1193	1500	43	1551	39	1420	58	1655
Stockholm	256	45	269	33	228	53	255	384	40	395	31	328	45	369	512	38	480	29	473	40	529	850	34	823	27	682	35	839
Parakeets	256	48	250	34	240	53	261	384	42	394	31	356	45	418	512	38	530	29	471	42	506	850	33	896	25	817	35	823
Kirsten&Sara	256	43	251	32	252	49	254	384	38	379	29	402	42	395	512	35	508	27	558	39	492	850	31	811	25	801	33	828
(b) 1920 × 1080 *pixels*																												
Basketball	2000	43	2014	33	2148	52	2117	3000	39	3020	31	2889	47	2961	4500	35	4722	30	3373	41	4540	7000	32	6800	26	6446	35	7320
Cactus	2000	41	1919	32	2059	49	2077	3000	37	2954	30	2818	43	3117	4500	33	4793	27	4585	39	4193	7000	30	7405	26	5519	33	6951
Seedof	2000	43	2119	37	2062	56	2336	3000	41	2865	35	2958	54	2897	4500	38	4817	33	4187	49	4508	7000	36	6858	30	7353	45	6489
Underboat	1000	40	1061	30	1000	43	1029	1600	36	1522	27	1692	39	1445	2500	32	2429	25	2510	33	2609	4000	28	4243	23	3997	29	3899

Table 4.11: Selected parameters and settings for the AVC, HEVC, and VP9 codecs.

Software	Parameters
HM	Default main LD profile with B frames. `IntraPeriod = -1` (only first frame encoded as I frame). List0 reference.
x264	`--profile high --tune psnr --ref 4 --direct auto --weightp 2 --level 5.1 --subme 8 --b-pyramid none --bframes 0 --b-adapt 0 --merange 24 --me tesa --no-fast-pskip --trellis 2 --min-keyint=9999 --keyint=9999 --pass 1 --slow-firstpass --fps <FR> --qp <QP> --psnr -v`
VP9	`--end-usage=3 --codec=vp9 --kf-max-dist=9999 --kf-min-dist=9999 --lag-in-frames=0 --good --cpu-used=0 --passes=1 --cq-level=<QP> -w <W> -h <H> --fps=<FR> --psnr -v -t 0`

For this work, the VP9 encoder and decoder were considered as a most recent implementation of the WebM Project. Due to the lack of official documentation and specifications for this encoder, the parameters were set based on recommendations received from the WebM Project lead developers. VP9 encoder allows to set the QP in two different ways. First approach (Grois et al., 2013; Grois et al., 2014) sets the `--min-q` and `--max-q` parameter to the same value. According to the comments of lead developers of VP9, such a setting apparently decreases the compression efficiency (Rerabek and Ebrahimi, 2014). Therefore, the available fixed quality mode `--end-usage=3`, which allows to vary the coding quality factor, was selected for VP9 encoding. Furthermore, the Intra Period parameters (`--kf-min-dist` and `--kf-max-dist`) were set to very large values to ensure that only the first frame is an I frame, which corresponds to LD configuration requirements for real-time scenarios considered in this paper. The selected configuration for VP9 allows comparative testing with AVC and HEVC.

Since the x264 implementation allows LD configuration only with P frames, it is only used as an orientation anchor to benchmark the other two next generation codecs. More detailed information about the configuration of all investigated encoders can be found in Table 4.11.

4.3.2 Methodology

The SS method with a five-grade quality scale (see Section 2.4.1) was chosen for evaluations. The subjects were asked to judge the overall quality of the evaluated video sequence. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each subject, while the same content was never shown consecutively.

To display the video sequences and collect individual scores, a slightly modified version of the QualityCrowd 2 framework (Keimel et al., 2012) was used. QualityCrowd 2 uses a simple scripting language that allows for the creation of test campaigns with high flexibility, e.g., specific pages for instructions, mixing different methodologies, etc. However, QualityCrowd 2 uses a fixed task order for each batch. To overcome this drawback, a plugin was developed to redirect each worker to a different batch, with a different display order for stimuli. Additionally, the VLC web plugin was used instead of the default Flash player, as it offers fullscreen playback. Fullscreen was automatically enforced for full HD video sequences.

All decoded video sequences were re-encoded with AVC, since transmitting uncompressed video data to remote workers is impractical and there is no browser video plugin capable of reliable real-time decoding and displaying for all evaluated codecs and resolutions, especially for HEVC and VP9 full HD content. The 720p contents at 50 fps were compressed at 20 Mbit/s, which is commonly considered as perceptually transparent quality for video broadcasting. For other frame rates and resolutions, the bit rate was set proportionally to their frame rate and resolution corresponding to the above mentioned bit rate. For example 1080p contents at 50 fps were encoded at 45 Mbit/s. A two-pass encoding was used and the deblocking filter was disabled to preserve the original blockiness artifacts when encoded at low bit rates. Expert viewing session was conducted prior to the main subjective assessment and the expert viewers evaluated the quality of this second encoding as visually lossless.

To mimic the realistic real-time application scenario, the subjective tests should ideally be conducted in form of crowdsourcing. Nevertheless, as it is relatively difficult to find online subjects equipped with a full HD monitor and because of the relatively large amount of transmitted video data, the experiments were conducted at EPFL in an uncontrolled lab room with desktop computers. Therefore, the workers' demographic was limited to university students participating on voluntary basis, and thus they were not remunerated for their effort. However, this approach helps to focus the subjective tests to compression part of the transmission chain only, as it limits the artifacts due to network transmission, transport protocol and playback settings.

A total of 26 subjects participated in the study. Each subject evaluated all test stimuli. Half of the subjects evaluated the 720p contents first, while the other half evaluated the 1080p contents first. To minimized visual fatigue effects, subjects took 10 min break between the two tasks.

Before the experiments, short written training instructions were provided to the subjects to explain their tasks. Additionally, three training samples, representative of *Excellent*, *Fair*, and *Bad* quality, were displayed to familiarize subjects with the assessment procedure. The training instructions and samples were presented using QualityCrowd 2.

To evaluate perceived quality, standard statistical indicators describing the score distribution across subjects for each test condition (combination of content, codec, and bit rate) were computed. First, outlier detection was applied to remove subjects whose scores deviated strongly from others. Assuming the reliability of subjects participating on voluntary basis, no crowdsourcing measures, such as honeypots, were used to detect the outliers. However, the outlier detection was performed according to the guidelines described in Section 2.3.1 of Annex 2 of ITU-R BT.500-13 (2012). In our experiments, none of the subjects was detected as an outlier for any of the test sessions. Then, the mean opinion score (MOS) and 95% confidence intervals (CI), assuming a Student's t-distribution of the scores, were computed for each test condition.

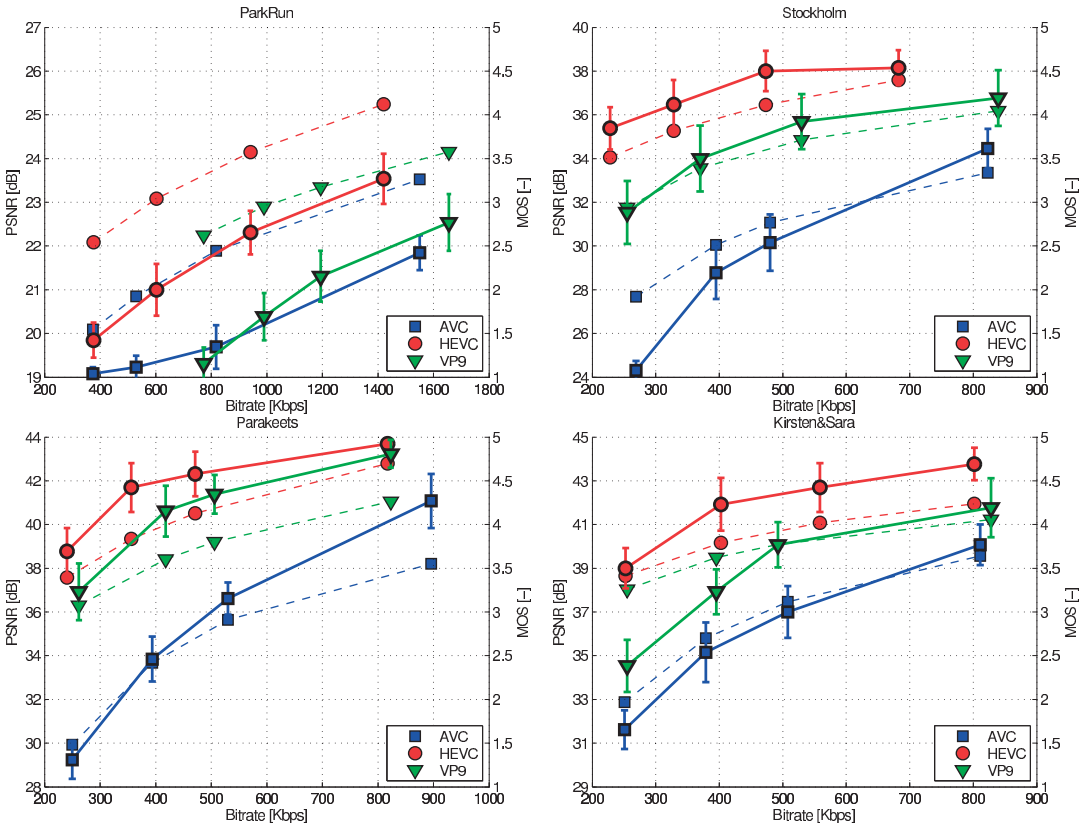


Figure 4.10: PSNR (dashed line) and MOS and CI (solid line): 720p contents.

4.3.3 Results

Figures 4.10 and 4.11 depict the R-D curves for the 720p and 1080p contents, respectively. The R-D curves based on PSNR measurements are plotted with dashed lines, whereas the subjective ratings, i.e., MOS and CI, are plotted with solid lines. Based on PSNR measurements, HEVC outperforms VP9 by 0.5 to 2 dB, while VP9 provides a gain ranging from 0.5 to 6 dB when compared to AVC. For all contents and bit rates, objective measurements show that HEVC outperforms both coding algorithms.

The subjective results show relatively small CIs, indicating a high reliability of the results and rather small variation across subjects. The ratings show similar trend to objective measurements: HEVC provides the best visual quality for a similar bit rate and largely outperforms AVC in most cases. Also, VP9 achieves better visual quality than AVC, except for contents *ParkRun* and *Seedof*, where CIs overlap significantly. However, in some cases (in particular, at high bit rates), HEVC and VP9 have similar ratings and there is no sufficient statistical evidence indicating differences in performance between these codecs at these bit rates. Finally, both HEVC and VP9 codecs can achieve *Good* to *Excellent* quality, i.e., MOS ≥ 4, at the highest bit rates used in our study, with the only exception of content *ParkRun*. Lower ratings for *ParkRun* content can be explained by the large values of spatial and temporal indices (see Section 2.2),

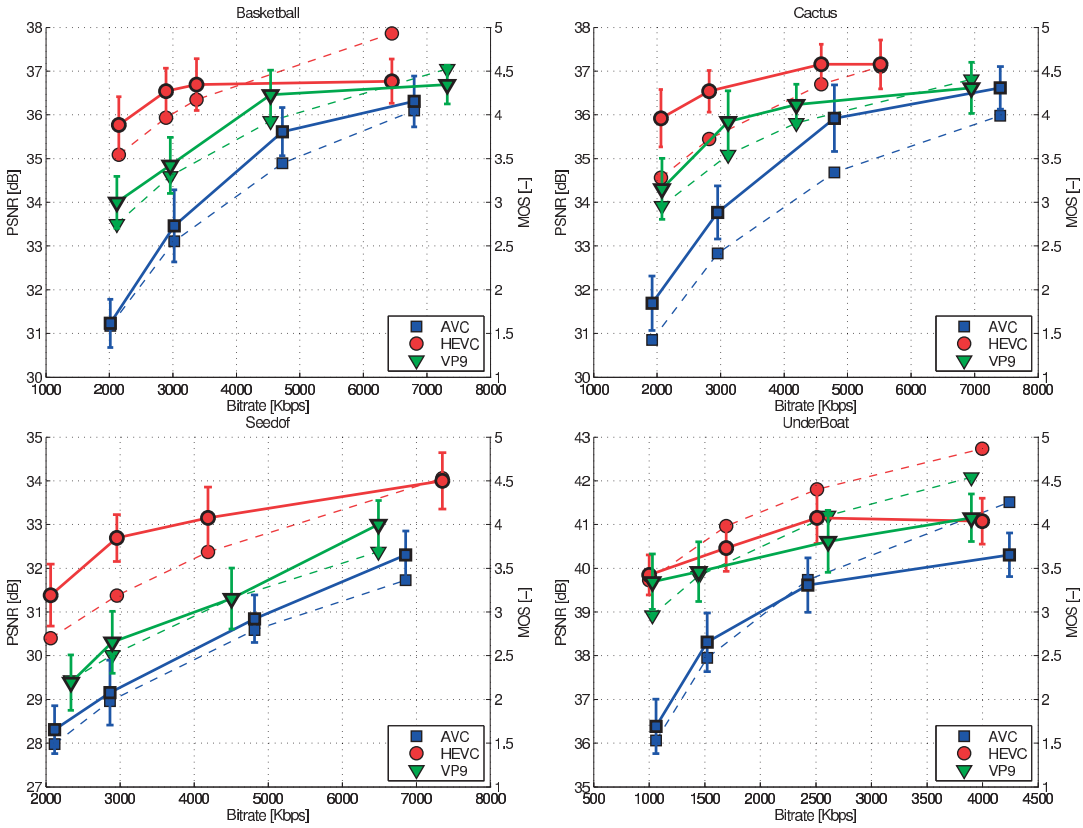


Figure 4.11: PSNR (dashed line) and MOS and CI (solid line): 1080p contents.

implying that this content contains areas with high level of details and a lot of motion, and thus it is very demanding in bit rate.

Figure 4.12 shows the results comparing all possible conditions for the 720p and 1080p contents. Comparing HEVC and AVC at similar bit rates, HEVC always provides statistically better visual quality when compared to AVC for contents *Stockholm*, *Parakeets*, *Kirsten&Kara*, and *Seedof*. For the other contents, there is not sufficient statistical evidence to show that HEVC outperforms AVC, especially at high bit rates. Looking at HEVC vs. VP9, HEVC is significantly better at the three lowest bit rates for contents *Seedof*, *Kirsten&Sara*, and *ParkRun*, whereas there are no statistical differences on contents *Parakeets* and *UnderBoat*. For the other contents, HEVC only outperforms VP9 at the lowest bit rates. Regarding the comparison between VP9 and AVC, VP9 outperforms AVC at the three lowest bit rates on the 720p contents, except for content *ParkRun* where there is no statistical difference. Similarly, VP9 shows better performance to AVC on the two lowest bit rates for the 1080p contents, with the exception of content *Seedof*, where there is not sufficient statistical evidence to show that VP9 outperforms AVC. Note that there is no case where AVC nor VP9 outperform HEVC, or when AVC outperforms VP9.

(a) 720p contents.

(b) 1080p contents.

Figure 4.12: Multiple comparison test results for all possible combinations of codecs (A stands for AVC, H stands for HEVC, and V stands for VP9 coding algorithm) and bit rates (R1 to R5). In each plot, the color of each square shows the result of the significance test between the MOSs related to the two test conditions reported in the corresponding column and row. A white (black) square indicates that the MOS corresponding to condition A is statistically significantly better (worse) than the MOS corresponding to condition B, whereas a grey square indicates that there is no sufficient statistical evidence indicating differences between the two MOSs.

Table 4.12: Comparison of investigated coding algorithms in terms of bit rate reduction for similar PSNR and MOS. Negative values indicate actual bit rate reduction. Note that the bit rate difference between HEVC and AVC on content Stockholm could not be computed as the R-D curves have no horizontal overlapping.

Content	HEVC vs. AVC		HEVC vs. VP9		VP9 vs. AVC	
	BD-Rate (%)	ΔR (%)	BD-Rate (%)	ΔR (%)	BD-Rate (%)	ΔR (%)
ParkRun	−54.8	−53.0	−44.0	−46.2	−18.3	−7.4
Stockholm			−46.1	−54.7	−55.9	−49.7
Parakeets	−69.1	−62.4	−32.1	−28.0	−55.5	−48.5
Kirsten&Sara	−60.1	−62.6	−20.8	−43.8	−52.8	−30.8
Basketball	−55.8	−59.3	−38.3	−45.6	−31.5	−28.3
Cactus	−54.3	−57.5	−23.6	−43.0	−42.9	−31.9
Seedof	−52.8	−61.8	−36.0	−51.1	−26.9	−21.9
Underboat	−54.2	−60.1	−27.6	−26.6	−39.2	−48.1
Average	−57.3	−59.5	−33.6	−42.4	−40.4	−33.3

Table 4.12 reports the average bit rate reduction BD-Rate and ΔR computed based on the Bjøntegaard and SCENIC models (see Chapter 3), respectively. Results based on objective measurements show that the average bit rate reduction of HEVC relative to AVC and VP9 is 57.3% and 33.6%, respectively. Although we used different encoders, different parameters (i.e., quality control parameters for VP9), and different PSNR metric, the results comparing HEVC to VP9 correspond to findings of (Grois et al., 2014), where authors claim 32.5% bit rate savings in favor of HEVC. In other studies (Grois et al., 2013; Mukherjee et al., 2013; Rerabek and Ebrahimi, 2014), authors used RA encoders configuration, and therefore mutual comparison of our results to those works is irrelevant.

On the other hand, results based on the subjective ratings indicate an average bit rate saving of 59.5% and 42.4% for HEVC when compared to AVC and VP9, respectively. Furthermore, the bit rate reduction achieved by VP9 relative to AVC is 40.4% and 33.3% based on estimated and perceived quality, respectively. These results show that the compression efficiency of HEVC over AVC predicted based on PSNR values is similar to the gain observed from subjective ratings. However, the performance of VP9 computed based on objective measurements seems to be overestimated, as the compression efficiency estimated from subjective ratings shows lower values. These results indicate that previous studies relying only on objective evaluations might have overestimated the performance of VP9.

4.4 Evaluation of JPEG XT HDR Image Compression

Despite a rapid increase of scientific activities and interests in HDR imaging, its adoption by industry is rather limited. One of the reasons is the lack of a widely accepted standard for HDR image coding that can be seamlessly integrated into existing products and applications. To resolve this problem, in 2012, the JPEG issued a CfP, which led to the initiation of JPEG XT, a JPEG backward compatible standard for HDR image compression. An important requirement

was the possibility for any legacy JPEG decoder to be able to recover a LDR version of the coded HDR image, resulting in a two-layer design of a base LDR and an extension codestream. Another important requirement was to impose both base and extension codestreams to use legacy JPEG compression tools in order to facilitate implementations. Compression efficiency was also considered as a third objective.

The JPEG XT standard defines a common codestream syntax and a common decoder architecture. To make practical implementations easier, the set of coding tools offered by the standard can be restricted to smaller subsets denoted as *Profiles*. Currently, the standard defines four profiles, referred to as profiles A, B, C, and D, of which Profile D is a very simple entry-level decoder that allows a 12 bit mode compatible to the 8 bit Huffman mode of JPEG while offering a precision similar to the 12 bit mode of legacy JPEG. Each profile offers a technical solution for coding HDR images considering additional requirements for different applications.

A few objective evaluations of JPEG XT have been performed (Pinheiro et al., 2014; Richter, 2013; Richter, 2014). However, only one subjective evaluation was performed by Mantel et al. (2014), but only for Profile C and only on six different images. To overcome the lack of subjective evaluations of JPEG XT, we performed an extensive subject quality assessment of the three main profiles, i.e., profiles A, B, and C. A subjective experiment was conducted with 24 naïve subjects to evaluate 20 different HDR images coded at 4 different bit rates and displayed on a SIM2 HDR47E S 4K monitor. This section reports the details and results of this performance analysis.

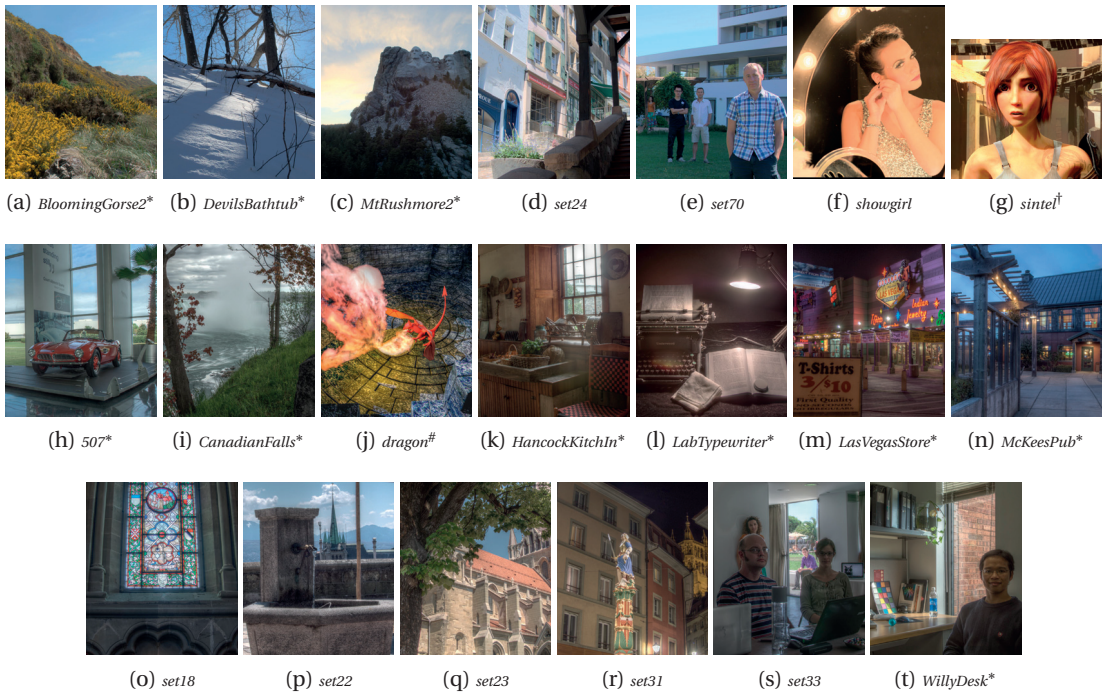
4.4.1 Dataset

The challenge of testing backward-compatible HDR compression is that the compression performance does not depend only on a single quality control parameter, but also on the quality settings for the base layer and on the choice of TMO, which produces this layer. To fully understand the implications of those parameters on perceptive viewing, a practical set of testing conditions was used in a subjective experiment (Section 4.4.2).

Image Selection

A set of 20 HDR images with resolutions varying from full HD (1920 × 1080) to larger than 4K (6032 × 4018) were selected (see Figure 4.13 for display-adapted versions). The dataset contains scenes with architecture, landscapes, and portraits. The original images were taken from other public datasets, including Fairchild, HdM-HDR-2014, and EPFL's HDR-Eye datasets. Then, the images were processed for subjective evaluation as follows.

Images were adjusted for a SIM2 HDR monitor. Images were first cropped and scaled by a factor of two with a bilinear filter to fit their size to 944 × 1080 for side-by-side subjective experiments (details in Section 4.4.2), and then tone-mapped using display-adaptive TMO (Mantiuk et al., 2008) to map the relative radiance representation of the images to an absolute



(a) *BloomingGorse2** (b) *DevilsBathtub** (c) *MtRushmore2** (d) *set24* (e) *set70* (f) *showgirl* (g) *sintel*†

(h) *507** (i) *CanadianFalls** (j) *dragon*# (k) *HancockKitchIn** (l) *LabTypewriter** (m) *LasVegasStore** (n) *McKeesPub**

(o) *set18* (p) *set22* (q) *set23* (r) *set31* (s) *set33* (t) *WillyDesk**

Figure 4.13: Display-adapted images of the dataset. The *reinhard02* TMO was used for images from (a) to (g) and the *mantiuk06* TMO was used for the remaining images. Copyrights: *2006-2007 Mark D. Fairchild, †Blender Foundation | www.sintel.org, under Creative Commons BY, #Mark Evans, under Creative Commons BY.

radiance and color space of SIM2 HDR monitor. The regions to crop were selected by expert viewers in such a way that cropped versions were representative of the quality and the dynamic range of original images. Downscaling together with cropping approach was selected as a compromise, so that a meaningful part of an image can be shown on the SIM2 HDR monitor. Figure 4.13 shows tone-mapped versions of images in the dataset and Table 4.13 presents different dynamic range and key characteristics (see Section 2.2) of these images.

Profiles Configuration

A common configuration for all tests in this paper has been chosen to ensure a fair comparison of profiles and to allow comparable evaluation results. For this purpose, the base layer always uses 4:2:0 chroma-subsampling, as it is traditionally employed in JPEG compression. To allow optimal quality, we decided to enforce 4:4:4, i.e., no chroma-subsampling, for the extension layer. All implementations enabled optimized Huffman coding, i.e., used a two-pass encoding to identify the optimal Huffman alphabet. Profile C in particular uses a 12-bit extension (8-bit legacy coding plus four refinement bits) for which no example Huffman table has been listed in the legacy JPEG; it should be noted, however, that the R-D curve of the 8-bit and 12-bit extension mode lie exactly on each other as quantization loss dominates, except that the 12-bit mode allows Profile C in particular to extend this curve towards higher bit rates and higher qualities, allowing scalable lossy to lossless coding.

Table 4.13: Characteristics of HDR images from the dataset.

	Dynamic range	Key
507	4.097	0.743
AirBellowsGap	4.311	0.768
BloomingGorse2	2.336	0.748
CanadianFalls	2.175	0.729
DevilsBathtub	2.886	0.621
dragon	4.386	0.766
HancockKitchenInside	4.263	0.697
LabTypewriter	4.316	0.733
LasVegasStore	4.131	0.636
McKeesPub	3.943	0.713
MtRushmore2	4.082	0.713
PaulBunyan	2.458	0.702
set18	4.376	0.724
set22	3.162	0.766
set23	3.359	0.764
set24	3.862	0.778
set31	4.118	0.678
set33	4.344	0.698
set70	3.441	0.735
showgirl	4.369	0.723
sintel	3.195	0.781
WillyDesk	4.284	0.777
min	2.175	0.621
max	4.386	0.781
mean	3.722	0.727
median	4.089	0.731

Despite these choices, we imposed no further restrictions or requirements on the encoder, though requested experts involved in their design to supply their recommendations for optimal coding performance. Like many other standards, JPEG XT itself does not specify the encoder and only imposes the requirement that it should create a syntactically correct codestream that describes the image with suitable precision.

Bit Rate Selection

Test images were created using the following procedure:

(i) Based on expert viewing on HDR monitor, for each of the 20 images, a tone-mapping algorithm was chosen out of 5 considered candidates (each TMO was applied with default parameters): a simple gamma-based algorithm, global logarithmic operator (Drago et al., 2003), global version of photographic operator *reinhard02* (Reinhard et al., 2002), operator optimized for encoding (Mai et al., 2011a) and local operator with strong contrast enhancement *mantiuk06* (Mantiuk et al., 2006a). For 7 images, *reinhard02* TMO was selected and for 13 images *mantiuk06* was selected as producing the best visual quality for these images.

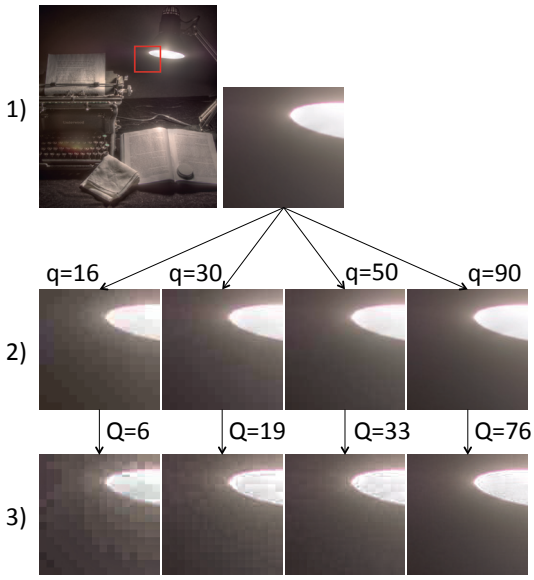


Figure 4.14: Illustration of the test images creation process for *LabTypewriter* and Profile A. 1) The TMO that produces the best visual quality is selected. 2) The tone-mapped image is encoded with JPEG at four different quality parameter (q) values such that they produce visual qualities corresponding to very annoying, annoying, slightly annoying, and imperceptible. 3) The HDR image is compressed with JPEG XT, using the base layer image and base layer quality parameter selected in 1) and 2), respectively. The quality parameter of the extension layer (Q) is set for each profile such that it produces the same bit rate as that of the base layer. For printed representation, the compressed HDR images were tone-mapped with *mantiuk06*.

(ii) Since JPEG XT images consist of a base and an extension layer, the overall bit rate has to be allocated to each of the layers. The bit rate allocation can be done differently and the strategy used can affect the performance of the profiles. To keep the overall number of samples small enough to allow subjective evaluation, for this study, we used the following allocation to generate codestreams.

(iii) We first fix for each image the bit rate of the base layer codestream. For the tone-mapped version of the image, the JPEG quality parameter was set to four different values such that they produce four different visual qualities based on the expert viewing: *very annoying*, *annoying*, *slightly annoying*, and *imperceptible* (see Figure 4.14).

(iv) The quality of the extension layer was then chosen for each profile in such a way that it would produce the same bit rate as that of the base layer. Such strategy resulted in a total of 12 (4 bit rates × 3 profiles) compressed versions for each HDR image (see Figure 4.14). Fixing the bit rate of the extension layer instead of its quality level ensured that profiles produced images with similar bit rates but potentially different perceptual qualities, which led to a fairer subjective evaluation of performance for each profile.

(v) A visual verification was then performed on SIM2 HDR monitor to confirm that 12 compressed versions of each HDR image cover the full quality scale from *very annoying* to *imperceptible*.

Figure 4.15: Three observers assessing a test image relative to a reference image shown on the SIM2 HDR monitor.

4.4.2 Methodology

Subjective evaluations were conducted at MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU. The laboratory setup ensures the reproducibility of subjective test results by avoiding unintended influence of external factors. In particular, the laboratory is equipped with a controlled lighting system with a 6500 K color temperature, a mid gray color is used for all background walls and curtains, and the ambient illumination did not directly reflect off of the monitor. During the experiment, the background luminance behind the monitor was set to 20 lx.

To display the test stimuli, a full HD 47" SIM2 HDR monitor with individually controlled LED backlight modulation, capable of displaying content with luminance values ranging from 0.001 to 4000 cd/m^2, was used. Prior to subjective tests, following a warm-up phase of an hour, a color calibration of the HDR display was performed using the software provided by SIM2. The red, green, and blue primaries were measured for white set to 1400 cd/m^2 level since the measurement probe (X-Rite i1Display Pro) is limited to a maximum value of 2000 cd/m^2.

In every session, three subjects assessed the displayed test images simultaneously, as illustrated in Figure 4.15. They were seated in an arc configuration, at a constant distance of 3.2 times the picture height (see Table 2.1).

Test Method

The DSIS Variant I method with a five-grade impairment scale (see Section 2.4.2) was selected, since this methodology is recommended for evaluating impairments and is typically used to evaluate compression algorithms. Two images were presented in side-by-side fashion to

reduce visual memory efforts by subjects. Due to the availability of only one full HD HDR monitor, each image was cropped and scaled to 944×1080 pixels with 32 pixels of black border separating the two images. One of the two images was always the reference (unimpaired) image. The other was the test image, which is a reconstructed version of the reference.

To reduce the effect of order of images on the screen, the participants were divided into two groups: the left image was always the reference image for the first group, whereas the right image was always the reference image for the second group. After the presentation of each pair of images, a six-second voting time followed. Subjects were asked to rate the impairments of the test images in relation to the reference image.

Test Design

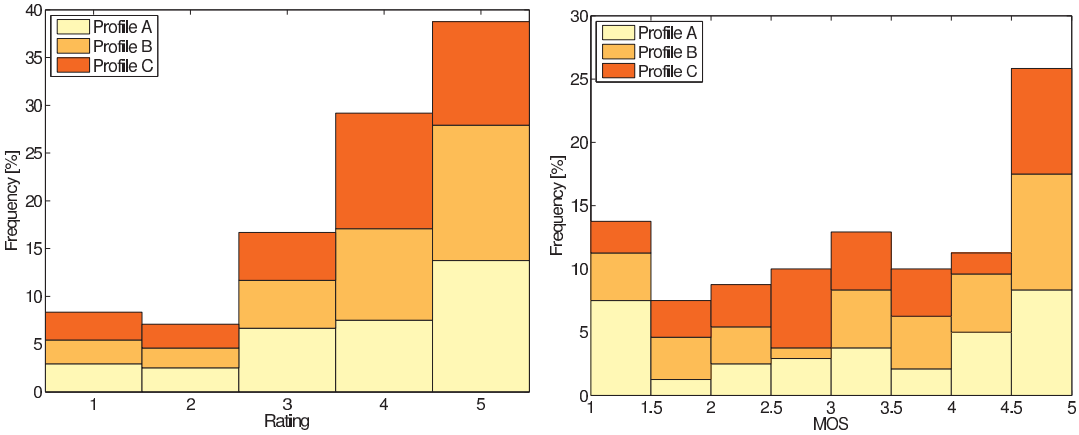
Before the experiment, a consent form was handed to subjects for signature and oral instructions were provided to explain their tasks. Additionally, a training session was organized allowing subjects to familiarize with the test procedure. For this purpose two images outside of the dataset were used. Five samples were manually selected by expert viewers for each image so that the quality of samples was representative of the rating scale.

Since the total number of test samples was too large for a single test session, the overall experiment was split into 3 sessions of approximately 16 min each. Between the sessions, subjects took a 15 min break. The test material was randomly distributed over the test sessions. To reduce contextual effects, the order of displayed stimuli was randomized applying different permutation for each group of subjects, whereas the same content was never shown consecutively.

A total of 24 naïve subjects (12 females and 12 males) took part in the experiments. Subjects were aged between 18 and 30 years old with an average of 22.1. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

Data Processing

The subjective scores were processed by first detecting and removing subjects whose scores deviated strongly from others. The outlier detection was applied to the set of results obtained from the 24 subjects and performed according to the guidelines described in Section 2.3.1 of Annex 2 of ITU-R BT.500-13 (2012). In this study, two outliers were detected. Then, the MOS was computed for each test stimulus as the mean across scores by valid subjects, as well as associated 95% CI, assuming a Student's t-distribution of the scores.



(a) Ratings distribution

(b) MOS values distribution

Figure 4.16: Scores distributions.

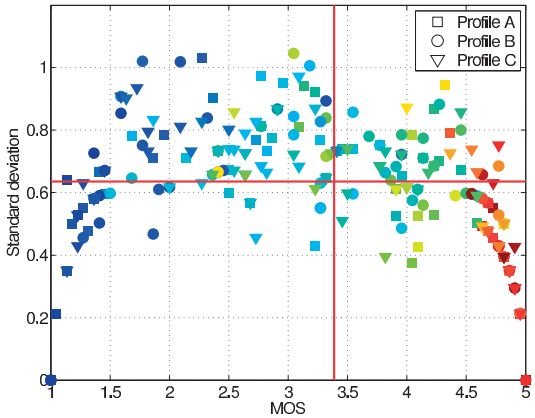


Figure 4.17: Standard deviation of subjective ratings versus MOS. The red lines represent the respective medians. Points are colored according to the bit rate of the corresponding compressed HDR image.

4.4.3 Results

Figures 4.16 to 4.18 show different characteristics of the obtained subjective scores. Figure 4.16a demonstrates that subjects' answers are well distributed within the rating scale and across profiles. As it can be observed in Figure 4.16b, MOS values reflect the subjects perception fairly with enough MOS samples for each meaningful value range. Figure 4.17 shows that subjective rating deviations do not exceed one rating point. Also, median value of the standard deviations is 0.62, which is about half of the rating scale step, and it leads to relatively small CIs, demonstrating that individual ratings are consistent across subjects. Median for the MOS values is about about 3.4, which is close to the middle of the rating scale with a slight skew towards the top of the scale. Figure 4.18 presents the distribution of MOS values for each evaluated content. It can be noted that, for most contents, MOS values cover almost the whole range from *very annoying* to *imperceptible*. While for some contents (e.g.,

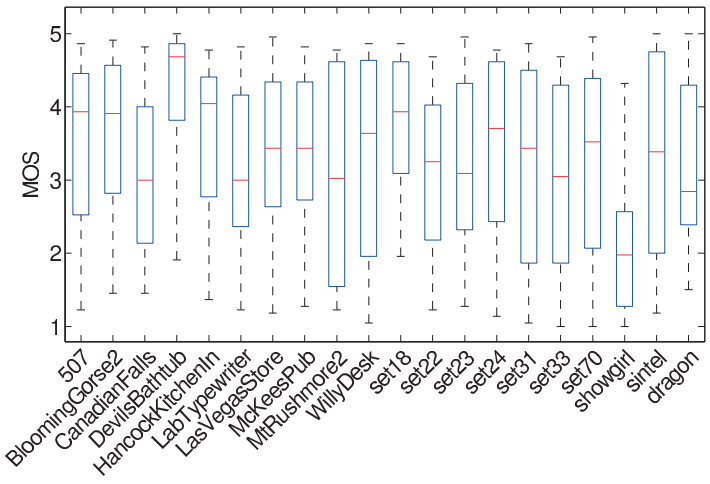


Figure 4.18: MOS distribution for each content. Whiskers are from minimum to maximum.

DevilsBathtub, *set18*, and *showgirl*) the MOS values are clustered nearer the extreme ends of the scale, Figure 4.18 shows that there are still enough of MOS values to cover the whole scale range. Such even distribution of MOS values means that the dataset is well-balanced overall, both in terms of quality distribution across the rating scale and across contents, which is a desirable feature for designing and benchmarking objective quality metrics.

Figure 4.19 shows the plots of MOSs and CIs at different bit rates for the three JPEG XT profiles. In most cases, there is not sufficient statistical evidence to indicate differences in performance between profiles. However, at the lowest bit rates, profiles B and C outperform Profile A on some contents. Likewise, for some contents, Profile C shows lower performance at medium bit rates. Nevertheless, at the highest bit rates, all three profiles reach transparent quality.

The results deviated strongly from the general trend for two contents: *MtRushmore2* and *showgirl*. For the first content, Profile B clearly outperforms the other two profiles. However, Profile B is outperformed by profiles A and C for the second content. For content *MtRushmore2*, many block coding artifacts can be observed for Profile A at the three lowest bit rates, as well as for Profile C at the lowest bit rate, which resulted in low quality scores. However, Profile B, as well as Profile C at medium bit rates, mostly exhibit color coding artifacts, and less block coding artifacts than Profile A, resulting in higher scores than Profile A. Regarding content *showgirl*, all profiles exhibit strong block and color coding artifacts at lower bit rates. Profile B shows block coding artifacts even at the highest bit rates, but mostly exhibits strange greenish and pinkish colors and some other color artifacts located near the top and bottom black borders. When encoded with Profile C, the image exhibits a one pixel wide red line near the transitions between the skin area and other areas, even at highest bit rate. Profile A encoded at the highest bit rate provides the best overall quality, but is not a perfect representation of the original image. This content is very challenging, because humans are very sensitive to artifacts in skin regions.

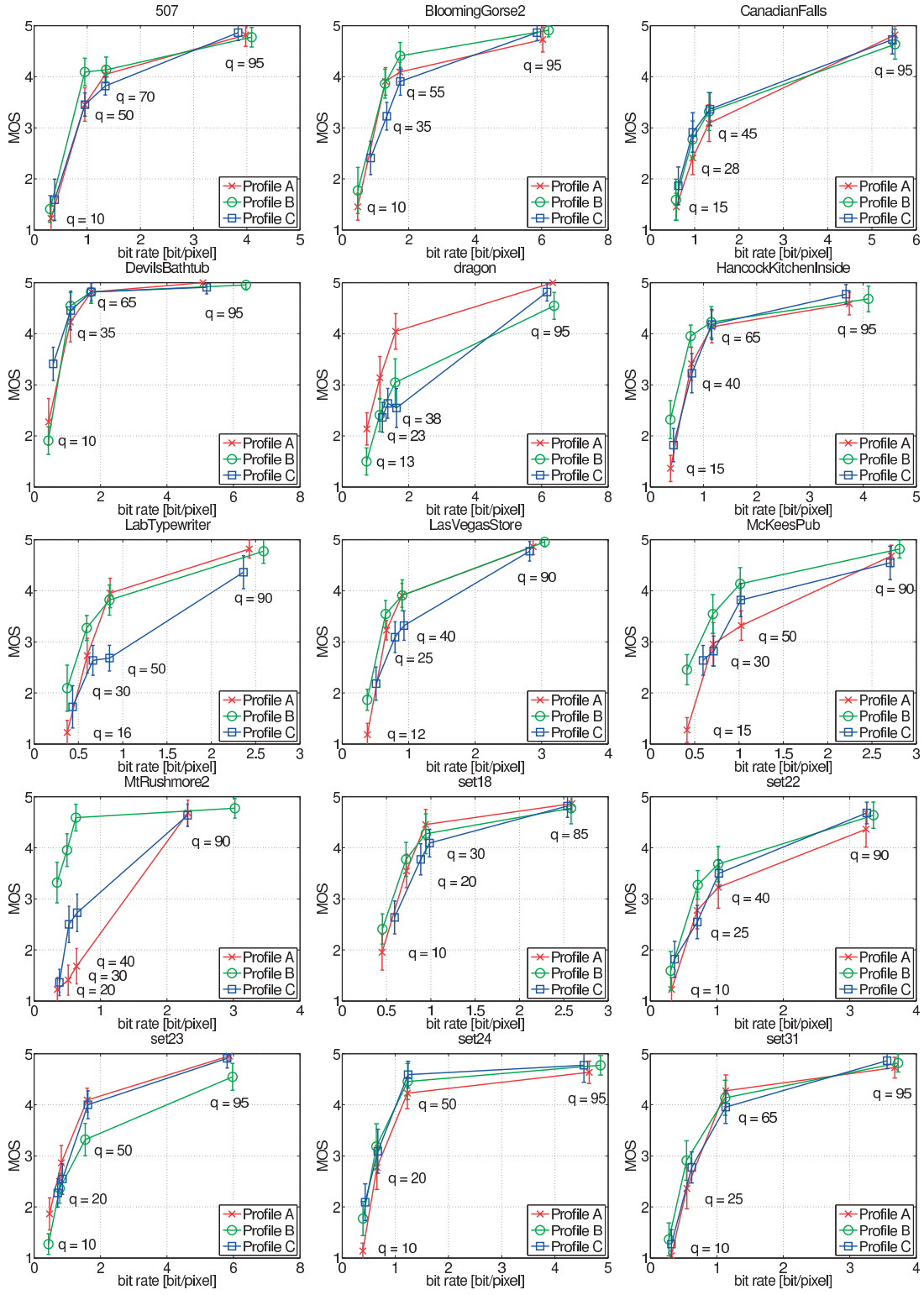


Figure 4.19: Plots of the MOS at different bit rates.

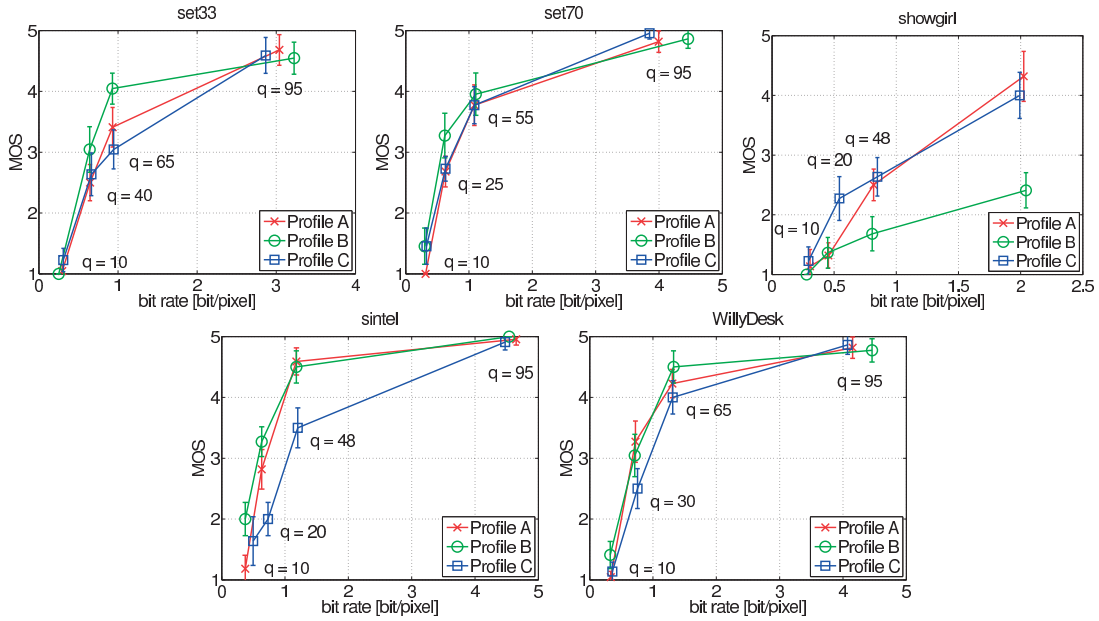


Figure 4.19: Plots of the MOS at different bit rates *(Continued)*.

Overall, we observed that Profile A exhibits a lot of block coding artifacts in flat areas, similar to JPEG, but usually preserves colors, except at very low bit rates. Profile B suffers from color bleeding on areas of uniform colors, but exhibits less block coding artifacts when compared to Profile A. In addition, Profile C performs better on flat uniform areas, but exhibits a checkerboard style color pattern on non-flat areas and introduces color noise near edges at low and medium bit rates, depending on content.

4.5 Towards HDR Extensions of HEVC

Since the completion of the first edition of the HEVC standard, several key extensions of its capabilities have been developed to address the needs of an even broader range of applications. Recognizing the rise of HDR applications and the lack of a corresponding video coding standard, MPEG released in February 2015 a CfE for HDR and WCG video coding (N15083). The purpose of this CfE was to explore whether the coding efficiency and/or the functionality of HEVC Main 10 and Scalable Main 10 profiles can be significantly improved for HDR and WCG content.

Potential evidence might include among others new video compression algorithms and coding tools, as well as new signal processing techniques, and different color spaces and transfer functions. The CfE addressed four different categories covering various applications, including backward compatibility with existing SDR decoders and/or displays, with either normative or non-normative changes to existing HEVC profiles. Note that non-normative changes are categorized as modifications that do not have impact on the decoding process, e.g., color sampling conversion. More particularly, the submission categories are defined as follows

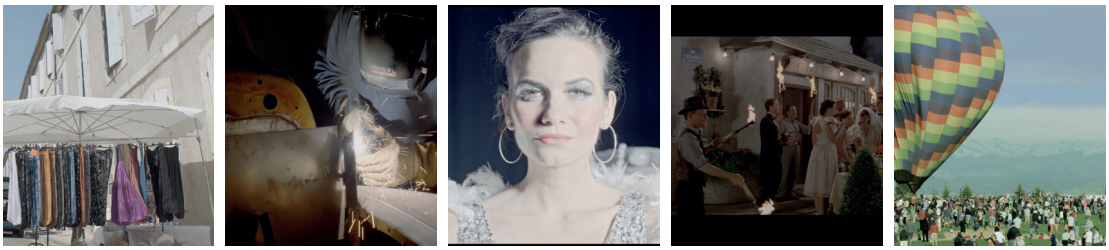
- Category 1: Single layer solution for HDR
- Category 2: Backward compatible solutions
  - 2a: Backward compatibility with legacy SDR decoders and displays, using an encoding system that has both HDR and SDR inputs
  - 2b: Technology Under Consideration for backward compatibility with legacy SDR decoders and displays, using an encoding system that has only an HDR input
  - 2c: Technology Under Consideration for backward compatibility with legacy SDR displays, but not SDR decoders, using an encoding system that has both HDR and SDR inputs
  - 2d: Technology Under Consideration for backward compatibility with legacy SDR displays, but not SDR decoders, using an encoding system that has only HDR input
- Category 3: Non-normative changes to the existing HEVC profiles
  - 3a: Main 10 Profile
  - 3b: Scalable Main 10 Profile

Each test condition, i.e., category, is described in more details within the CfE document (N15083). In the context of the CfE preparation for HDR/WCG video coding, HEVC Anchors of the selected content (M35480) were generated (M35852) using the official HM software with carefully selected bit rates as test points. These Anchors served as reference testing sequences as described in the CfE (N15083). Each proponent had provide the selected content encoded with a proprietary solution at the same bit rates as an attempt to improve compression efficiency of HEVC Main profiles.

In total, eight companies or aggregations of different companies and one university responded to the CfE and submitted responses to one or more of the different categories. Initially, responses to categories 1, 2b, 3a, and 3b were planned to be tested through formal subjective evaluations. However, based on the large number of responses, it was further agreed that only responses to categories 1 and 3a would be tested in the formal subjective evaluations. To benchmark the potential coding technologies submitted in response to the CfE, we performed a subjective quality evaluation to determine whether the proposed technologies could achieve better visual quality than the HEVC Anchor. The subjective tests were performed in the form of partial PC, where one video sequence of the pair was always the Anchor as a reference. Overall 48 naïve subjects participated in the subjective experiment, which leads to a total of 24 ratings per video stimuli. This section reports the details and results of this performance analysis.

4.5.1 Dataset

The dataset used for the subjective evaluation tests consists of five HD resolution HDR video sequences, namely, *Market3*, *AutoWelding*, *ShowGirl2*, *WarmNight*, and *BalloonFestival*. Figure 4.20 shows a typical frame example of each content. Each video sequence was cropped to 950×1080 pixels, so that the video sequences were presented side-by-side with a 20-pixels separating black border. Each video sequence was displayed at 24 fps, which is the native



(a) *Market3* (b) *AutoWelding* (c) *ShowGirl2* (d) *WarmNight* (e) *BalloonFestival*

Figure 4.20: Representative frames of the sequences used in the experiments. Tone-mapped versions are shown, since typical displays and printers are unable to reproduce higher dynamic range images.

Table 4.14: HDR test sequences used in the subjective evaluations.

Sequence	fps	window		frames		Anchor bit rates (kbit/s)			
						R4	R3	R2	R1
Market3	50	970	1919	0	239	1248	2311	4224	7913
AutoWelding	24	600	1549	162	401	454	778	1383	3157
ShowGirl2	25	350	1299	94	333	574	971	1652	3316
WarmNight	24	100	1049	36	275	462	780	1328	2441
BalloonFestival	24	0	949	0	239	1276	2156	3767	6644

frame rate of the display used in the experiments (see Section 4.5.2), and cut to 240 frames, which corresponds to 10 seconds. Note that the *Market3* sequence was played at a slower frame rate than the original content (50 fps). This solution was evaluated as visually more pleasant than playing every other frame, which created temporal distortions. The coordinates of the cropping window, selected frames, and bit rates are given in Table 4.14.

The data was stored in uncompressed 16 bit TIFF files, in 12 bit non-linearly quantized (using Dolby PQ EOTF) RGB signal representation, using the SDI data range (code values from 16 up to 4076) and Rec. 2020 RGB color space. The side-by-side video sequences were generated using the HDRMontage tool from the HDRTools package (M35471).

4.5.2 Methodology

The experiments were conducted at the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU. The test room is equipped with a controlled lighting system of a 6500 K color temperature. The color of all background walls and curtains in the room is mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective test results by avoiding unintended influence of external factors. In the experiments, the luminance of the background behind the monitor was about 20 cd/m^2. The ambient illumination did not directly reflect off of the display.

To display the test stimuli, a full HD (1920×1080 pixels) 42" Dolby Research HDR RGB backlight dual modulation display (aka Pulsar) was used. The monitor has the following specifications: full DCI P3 color gamut, 4000 cd/m^2 peak luminance, low black level (0.005 cd/m^2), 12-bits/color input with accurate and reliable reproduction of color and luminance. In every session, three subjects assessed the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of about 3.2 times the picture height (see Table 2.1).

Test Method

Two video sequences were presented simultaneously in side-by-side fashion. Since only one full HD 1920 × 1080 HDR monitor was available, each video was cropped to 950 × 1080 pixels with 20 pixels of black border separating the two sequences. One of the two video sequences was always the Anchor, with a randomized position on the screen (either on the left or on the right). The other video sequence was the Proponent to be evaluated, at the same (targeted) bit rate as the Anchor.

Subjects were asked to judge which video sequence in a pair ('left' or 'right') has the best overall quality, considering fidelity of details in textured areas and color rendition. The option 'same' was also included to avoid random preference selections.

Statistical Analysis

No outlier detection was performed on the raw scores, since there is no international recommendation or a commonly used outlier detection technique for PC results.

For each test condition, i.e., combination of content, algorithm, and bit rate, the winning frequency of the Anchor, w_{Ai}, winning frequency of the Proponent, w_{Pi}, and tie frequency, t_i, are computed from the obtained subjective ratings. Note that $w_{Ai} + w_{Pi} + t_i = N$, where N is the number of subjects. To compute the preference probability of selecting the proponent version over the Anchor, p_P, ties are considered as being half way between the two preference options.

To determine whether the visual quality difference between the Proponent and the Anchor is statistically significant, a statistical hypothesis test was performed. As ties are split equally between the two preference options, the data roughly follows a Bernoulli process $B(N, p)$, where N is the number of subjects and p is the probability of success in a Bernoulli trial and was set to 0.5, considering that, a priori, the Anchor and Proponent have the same chance of success. Figure 4.21 shows the CDF for Binomial distribution with $N = 24$ and $p = 0.5$. The CDF is used to determine the critical region for the statistical test.

To determine whether the proponent provides statistically significant results, a one-tailed binomial test was performed at 5% significance level with the following hypotheses

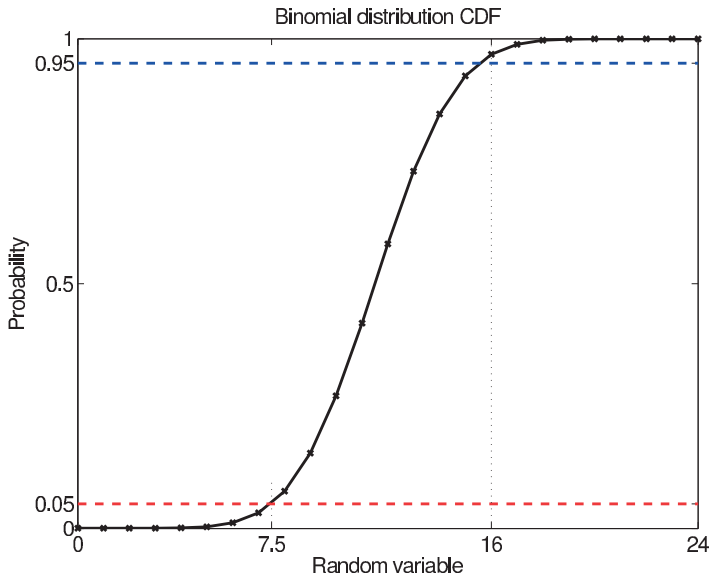


Figure 4.21: CDF for Binomial distribution with $N = 24$ and $p = 0.5$.

H0: Proponent is equal or worse than Anchor
H1: Proponent is better than Anchor

In this case, the critical region for the preference probability over Anchor, p_P, is $\left[\frac{16}{24}, 1\right]$, as the CDF for 16 or more successful trials is above 95% (see Figure 4.21, $B(16, 24, 0.5) = 0.9680$). Therefore, if there are 16 or more votes in favor of the Proponent, the null hypothesis can be rejected.

Similarly, to determine whether the Proponent provides statistically significantly lower visual quality than the Anchor, a one-tailed binomial test was performed at 5% significance level

H0: Proponent is equal or better than Anchor
H1: Proponent is worse than Anchor

In this case, the critical region for the preference probability over Anchor, p_P, is $\left[0, \frac{7.5}{24}\right]$, as the CDF for 7.5 or less successful trials is below 5% (see Figure 4.21, $B(8, 24, 0.5) = 0.0758$). Note that the Binomial distribution is not defined for non-integer values, and that extension is usually obtained using the floor function. Therefore, if there are 7.5 or less votes in favor of the proponent, the null hypothesis can be rejected.

Test Planning

Before the experiments, a consent form was handed to subjects for signature and oral instructions were provided to explain the evaluation task. A training session was organized to allow subjects to familiarize with the assessment procedure. The same contents were

used in the training session as in the test session to highlight the areas where distortions can be visible. Eleven training samples were manually selected by expert viewers. First, two samples, one of high quality and one of low quality, without any difference between left and right, were selected from the *AutoWelding* sequence. The purpose of these two examples was that subjects could get familiar with HDR content, as this content has both dark and bright luminance levels and fast luminance temporal changes, and see the extreme levels of quality observed in the test material. Then, one sample from *AutoWelding* with large visible difference was presented to illustrate the main differences that can be observed between the left and right video sequences, i.e., loss of texture/details and color artifacts. Finally, for each of the remaining contents, two samples were presented (one example with large difference and one example with small differences) in the following order: *Market3*, *BalloonFestival*, *ShowGirl2*, and *WarmNight*. The training materials were presented to subjects exactly as for the test materials, thus in side-by-side fashion.

The overall experiment was split into 6 test sessions. Each test session was composed of 30-31 basic test cells, corresponding to approximately 14 min each. To reduce contextual effects, the stimuli orders of display were randomized, whereas the same content was never shown consecutively. The test material was randomly distributed over the six test sessions.

Each subject took part to exactly three sessions. Three dummy pairs, whose scores were not included in the results, were included at the beginning of the first session to stabilize the subjects' ratings. Between the sessions, the subjects took a 14 min break.

A total of 48 naïve subjects (16 females and 32 males) took part in the experiments, leading to a total of 24 ratings per test sample. Subjects were between 18 and 49 years old with an average and median of 25.3 and 24 years of age, respectively. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

4.5.3 Results

Figure 4.22 reports the preference probability of selecting the Proponent version over the Anchor for each content separately. Category 1 submissions (P11, P12, P13, P14, and P22) are plotted with plain lines, whereas Category 3a submissions (P31, P32, P33, and P34) are plotted with dashed lines. Values on or above the horizontal upper dashed line provide statistically significant visual quality superior to the Anchor, while values on or below the horizontal lower dashed line provide statistically significant inferior visual quality when compared to the Anchor.

As it can be observed, there is evidence that potential coding technologies can do better than the Anchor in a statistically significant way, especially for contents *Market3* and *BalloonFestival*. For instance, on content *ShowGirl2*, Proponent P22 provides statistically significant superior visual quality when compared to the Anchor at rates R1 to R3. Improvements can also be observed for Proponents P11 and P12. Regarding content *WarmNight*, Proponents P32 and

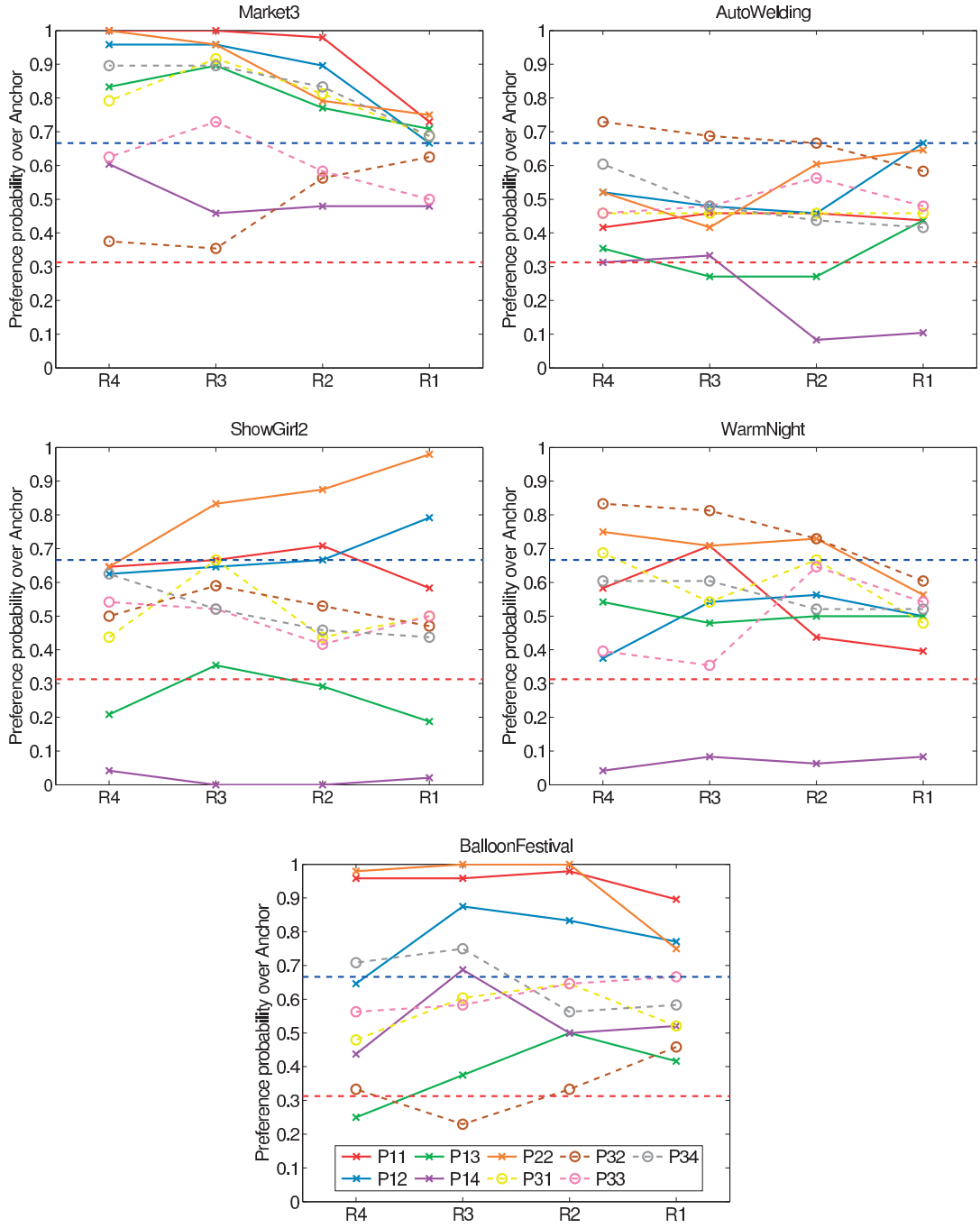


Figure 4.22: Preference probability of selecting the Proponent version over the Anchor.

P22 outperform the Anchor for rates R2 to R4. Proponents P31 and P11 also show gains for specific rate points. Finally, for content *AutoWelding*, Proponent P32 provides gain for rates R2 to R4, while Proponent P12 is at the limit for the rate R1.

In general, Proponent P32 seems to perform better on dark contents than on bright contents.

Regarding P14, wrong colors were observed throughout the test material, probably due to a wrong color transformation, as well as occasional green noise in the table scene on content *WarmNight*. Regarding the selection of contents, bright scenes are better to perceive color artifacts, especially in whitish parts, and loss of details and high frequencies, especially in textured areas. Sequences such as *ShowGirl2* and *Market3* are good for testing HDR compression. On the other hand, sequences with a wide dynamic range and strong luminance temporal changes, such as *AutoWelding* although good for demonstrating HDR, may not be necessarily best to assess HDR compression performance. Dark scenes are important too, as HDR is not only about high brightness, but it might be hard to see the improvements in these sequences, especially if the previous test sequence was bright, due to the adaptation time of the eye.

4.6 Cross-lab Evaluation of MVC+D and 3D-AVC 3D Video Compression

Consistent and imitable subjective measurement 3D video quality assessment is investigated for evaluating 3D service parameters and as an essential criterion towards the development of objective models. Quality assessment of 3D video is identified to range over numerous psychophysical extents, e.g., picture excellence, depth perception, and visual comfort, which may lead to higher level insights, e.g., visual experience and naturalness.

An important factor in subjective quality assessment experiments is the viewing conditions and it can be greatly influenced in the case of 3D video where the perception of depth is an additional factor when compared to 2D video. Furthermore, selecting and calibrating the display is very crucial in 3D video as it has a significant effect on the perceived brightness and overall quality, especially when more than one lab is involved in subjective evaluation. It is an interesting and challenging task to conduct the 3D video quality assessment in different labs and attempt to simulate the same conditions. It helps not only to measure the video quality accurately and precisely, but also gives us foundation to define the objective metrics for 3D video.

Perkis et al. (2012) performed cross-lab video quality assessment of 3D video to address various issues regarding certification of multimedia quality assessment. They evaluated two test scenarios, namely, a 2-view input configuration, on stereoscopic display, and a 3-view input configuration, on both auto-stereoscopic as well as stereoscopic display. However, in any single scenario, only two laboratories results were considered for cross validation.

Recently, Barkowsky et al. (2013) have studied cross-lab 3DTV quality assessment method with a main focus on defining the effect of different lab conditions like passive polarized displays, active shutter displays, viewing distance, number of parallel viewers, and voting device.

In November 2013, JCT-3V issued a test plan for 3D video subjective assessment (JCT3V-F1011) to evaluate the performance of two amendments of the AVC video coding standard, namely

Table 4.15: Multiview video plus depth contents used in the experiments.

Sequence	Resolution	Frame rate	Frames	QP settings	Depth resolution	Input views	*SS* stereo pair	*OS* stereo pair
Poznan Hall2	1920 × 1088	25 fps	0 – 199	26,31,36,41	Quarter	7 – 6 – 5	6.25 – 5.75	6 – 5.5
Poznan Street	1920 × 1088	25 fps	0 – 249	26,31,36,41	Quarter	5 – 4 – 3	4.25 – 3.75	4 – 3.5
Undo Dancer	1920 × 1088	25 fps	0 – 249	31,38,41,46	Full	1 – 5 – 9	4 – 6	5 – 7
GT Fly	1920 × 1088	25 fps	0 – 249	26,31,36,41	Quarter	9 – 5 – 1	6 – 4	7 – 5

1. MVC+D (Y. Chen et al., 2014): the main target of this extension is to enable 3D enhancements while maintaining MVC stereo compatibility.
2. 3D-AVC (Y. Chen and Vetro, 2014): the main aims for higher compression efficiency by jointly compressing texture and depth data.

To analyze and compare the performance of the proposed technologies, a formal subjective quality evaluation was carried out, and a set of test video sequences, encoded with the proposed technologies, was produced. Three laboratories took part in the evaluation campaign of this test material: at EPFL in Switzerland, UWS in Scotland, and FUB in Italy.

This section analyzes the results obtained from three subjective experiments on the aforementioned coding techniques using identical video content and following similar methodologies and instructions. Cross-laboratory analysis is performed to find out whether or not consistent results can be obtained. These analyses show that laboratories employing different displays and different subjects could still produce highly correlated results, as they follow similar guidelines to carry out the evaluations. This confirms that the participating laboratories have high correlation to conduct subjective evaluation.

4.6.1 Dataset

Four MVD sequences, with different visual characteristics, were used in the experiments (see Table 4.15). The encoded views used in the experiments were the same as those specified in the common test conditions (CTC) (JCT3V-E1100) of the 3DV Core Experiments conducted by JCT-3V. All sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bit per sample.

The sequences were compressed with MVC+D and 3D-AVC using 3D-ATM v9.0 (JCT3V-G1003) under the conditions defined in (JCT3V-E1100; JCT3V-F1011) (see Table 4.15). For each sequence, two stereo pair configurations were considered: a stereo pair formed from two synthesized views, referred to as *SS* in this paper, and a stereo pair formed from one original (decoded) view and one synthesized view, referred to as *OS* in this paper. For each codec, four rate points were considered. Additionally, a "reference" stereo pair was generated from the original data for each stereo pair configuration. Thus, this resulted in a set of $4 \times 2 \times (1 + 2 \times 4) = 72$ test stimuli. The synthesized views were generated using VSRS-1D-Fast v8.0 (JCT3V-G1005), under the conditions defined in (JCT3V-E1100; JCT3V-F1011) (see Table 4.15).

Table 4.16: Test environment.

	Lab1	Lab2	Lab3
Monitor	Hyundai S465D	LG 47LM660	Sony KDL-55X9005A
Size	46”	47”	55”
#Subjects (♂/♀)	22 (15/7)	24 (14/10)	18 (16/2)
Age (average)	20 – 31 (23.1)	18 – 28 (19.6)	20 – 31 (26.5)

Additionally, five training samples were generated using the *Poznan CarPark* sequence with similar conditions and manually selected by expert viewers so that the quality of samples were representative of all grades of the rating scale.

4.6.2 Methodology

In total, three laboratories conducted the subjective evaluation. All laboratories fulfill the recommendations for the subjective evaluation of visual data issued by ITU. Each test room is equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of the maximum screen luminance, whereas the color of all the background walls and curtains present in the test area are in mid grey.

The experiment involved up to three subjects assessing the test materials simultaneously. Subjects were seated in a row perpendicular to the center of the monitor, at a distance of about 3 times the picture height. All subjects were screened for correct visual acuity, color vision, and stereo vision using Snellen chart, Ishihara chart, and Randot test, respectively. The main differences between the laboratories were in terms of display characteristics and number of observers (see Table 4.16).

Test Method

The SS method with a five-grade numerical categorical scale (see Section 2.4.1) was chosen. The rating scale ranged from 1 to 5, with 1 indicating the lowest quality and 5 indicating the highest quality. After the presentation of each video sequence, a five-second voting time followed. Subjects were asked to rate the overall quality of the video sequence to be evaluated, and to express these judgments in terms of the wordings used to define the rating scale.

Test Design

Before the experiment, oral instructions were provided to the subjects to explain their tasks. Additionally, a training session using the five training samples was organized to allow subjects to familiarize with the assessment procedure. Since the total number of test samples was too large for a single test session, the overall experiment was split into two sessions of approximately ten minutes each. Between the sessions, the subjects took a ten minutes break. The test material was randomly distributed over the two test sessions.

Four dummy video sequences (one with high quality, one with low quality, and two of mid quality), whose scores were not included in the results, were included at the beginning of each test session to stabilize the subjects' ratings. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each group of subjects, whereas the same content was never shown consecutively.

4.6.3 Results

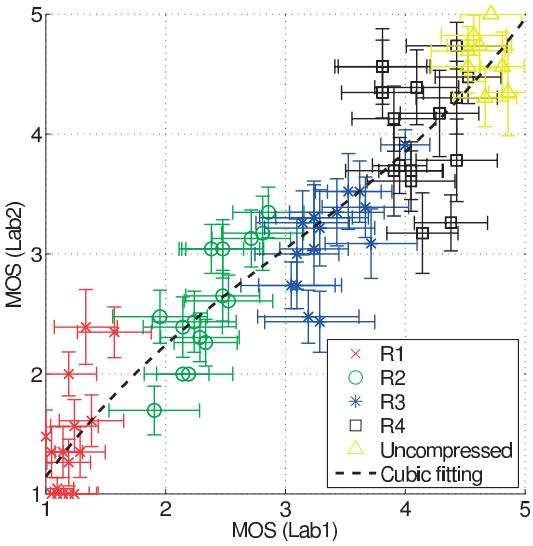
Figure 4.23 shows the scatter plots comparing the results of the different laboratories. The horizontal and vertical error bars represent the CI corresponding to the laboratories on the x- and y-axis, respectively. The data points are colored based on the different contents or rate points for better visualization. The cubic regressions fitted to each data set following the procedure described in Section 2.7.1 are represented to illustrate the trend of the data points.

Ideally, all points would be on a 45° line if the MOS values for each condition were the same between two laboratories. However, some points lie above the line, whereas others lie below. For example, subjects in Lab3 graded content *UndoDancer* at rate points R2, R3, and R4 lower than subjects in Lab1. Similarly, subjects in Lab3 graded content *Poznan Hall2* at rate points R2, R3, and R4 lower than subjects in Lab2. Nevertheless, no significant systematic offset can be observed between the MOS values of the different laboratories, which means that, in overall, subjects did not score more pessimistically nor more optimistically between the different laboratories.

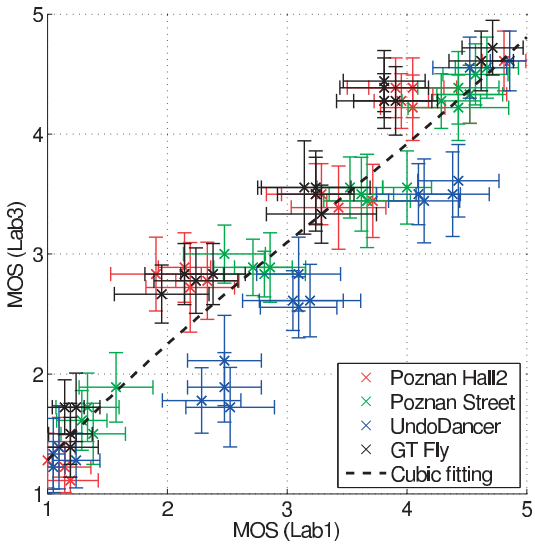
Regarding the comparison between Lab1 and Lab3, the cubic fitting is close to a straight line, but its slope is smaller than 45°. This indicates that subjects in Lab3 graded low quality stimuli higher than subjects in Lab1, whereas subjects in Lab1 graded high quality stimuli higher than subjects in Lab3. Regarding the comparison between Lab1 and Lab2 as well as between Lab2 and Lab3, the fitted cubic curves exhibit a sigmoid shape, which indicates non-linearity between the results of the different laboratories. For example, the ranges of grades associated with rate points R1 and R4 are wider in Lab2 than in Lab1. Nevertheless, the cubic regressions do not deviate much from a straight line.

As the mapping of MOS^{LabX} to MOS^{LabY} yields slightly different results when compared to mapping of MOS^{LabY} to MOS^{LabX}, both mappings are considered in the following subsections and results are reported for both cases. A value $v(i, j)$ on row i and column j is computed considering mapping of MOS^{Labi} to MOS^{Labj}.

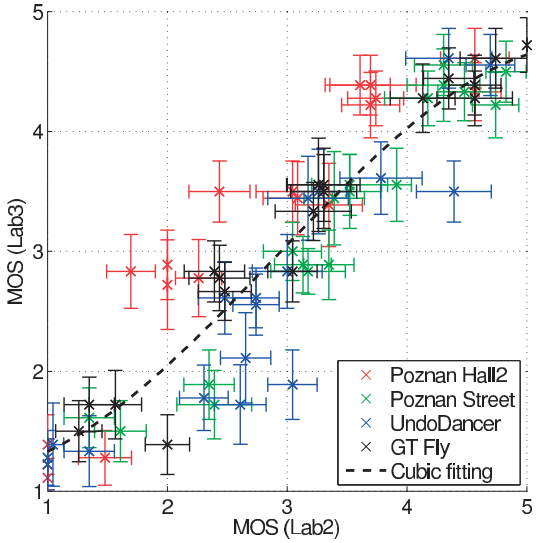
In the following part, to determine whether the difference between two sets of scores corresponding to the same stereo pair evaluated in two different laboratories is statistically significant, a multiple comparison test based on ANOVA was performed at a 5% significance level on the raw scores.



(a) Lab2 versus Lab1



(b) Lab3 versus Lab1



(c) Lab3 versus Lab2

Figure 4.23: Comparison of MOS values obtained in the different laboratories.

Statistical Evaluation Metrics

Table 4.17 reports the statistical evaluation metrics described in Section 2.7.2. Results show that there is a strong correlation between the different laboratories, as the correlation indexes are above 0.92 in all cases. The PCC, SROCC, and RMSE indexes are similar in all cases. However, the OR index shows a wider variation between the different cases. In particular, the OR values when mapping the results of Lab2 to Lab1 and Lab3 to Lab1 are above 27% and below 13%, respectively, whereas the average OR value is about 20% in the other cases. These results indicate that the correlation between Lab1 and Lab3 is the strongest.

Table 4.17: Statistical evaluation metrics.

	PCC			SROCC			RMSE			OR		(%)
	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3
Lab1	–	0.9461	0.9429	–	0.9393	0.9340	–	0.3962	0.4073	–	20.83	16.67
Lab2	0.9407	–	0.9321	0.9399	–	0.9356	0.3911	–	0.4177	27.78	–	19.44
Lab3	0.9430	0.9294	–	0.9340	0.9356	–	0.3737	0.4146	–	12.50	20.83	–

Table 4.18: Estimation errors.

	Correct estimation(%)			*Underestimation*(%)			*Overestimation* (%)		
	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3
Lab1	–	94.44	97.22	–	4.17	2.78	–	1.39	0.00
Lab2	93.06	–	93.06	2.78	–	4.17	4.17	–	2.78
Lab3	98.61	88.89	–	0.00	6.94	–	1.39	4.17	–

Table 4.19: Classification errors.

	Correct decision(%)			*False ranking* (%)			*False differentiation*(%)			*False tie*		(%)
	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3	Lab1	Lab2	Lab3
Lab1	–	82.20	79.50	–	0.00	0.00	–	7.63	7.00	–	10.17	13.81
Lab2	80.99	–	78.09	0.00	–	0.08	7.86	–	8.76	11.15	–	13.07
Lab3	79.03	78.48	–	0.00	0.20	–	10.95	10.45	–	10.02	10.88	–

Estimation Errors

Table 4.18 reports the estimation errors (see Section 2.7.3). Results again show that there is a strong correlation between the different laboratories; especially between Lab1 and Lab3 (*Correct estimation* above 97%). However, when mapping the results of Lab3 to those of Lab2, the *Correct estimation* is below 89%, whereas the *Underestimation* and *Overestimation* are above 4%.

Classification Errors

Table 4.19 reports the classifications errors (see Section 2.7.4). About 80% of all possible distinct combinations of two stereo pairs lead to the same conclusion in different laboratories. Moreover, *False ranking*, which is the most offensive error, almost never occurs. *False tie* occurs in more than 10% of the cases, but this is the least offensive error. Results for *False differentiation* are in overall lower between Lab1 and Lab2 than between Lab1 and Lab3, which indicates that the correlation between Lab1 and Lab2 is higher than between Lab1 and Lab3, as opposed to the results of the statistical evaluation metrics. However, the difference is not as big as for the statistical evaluation metrics.

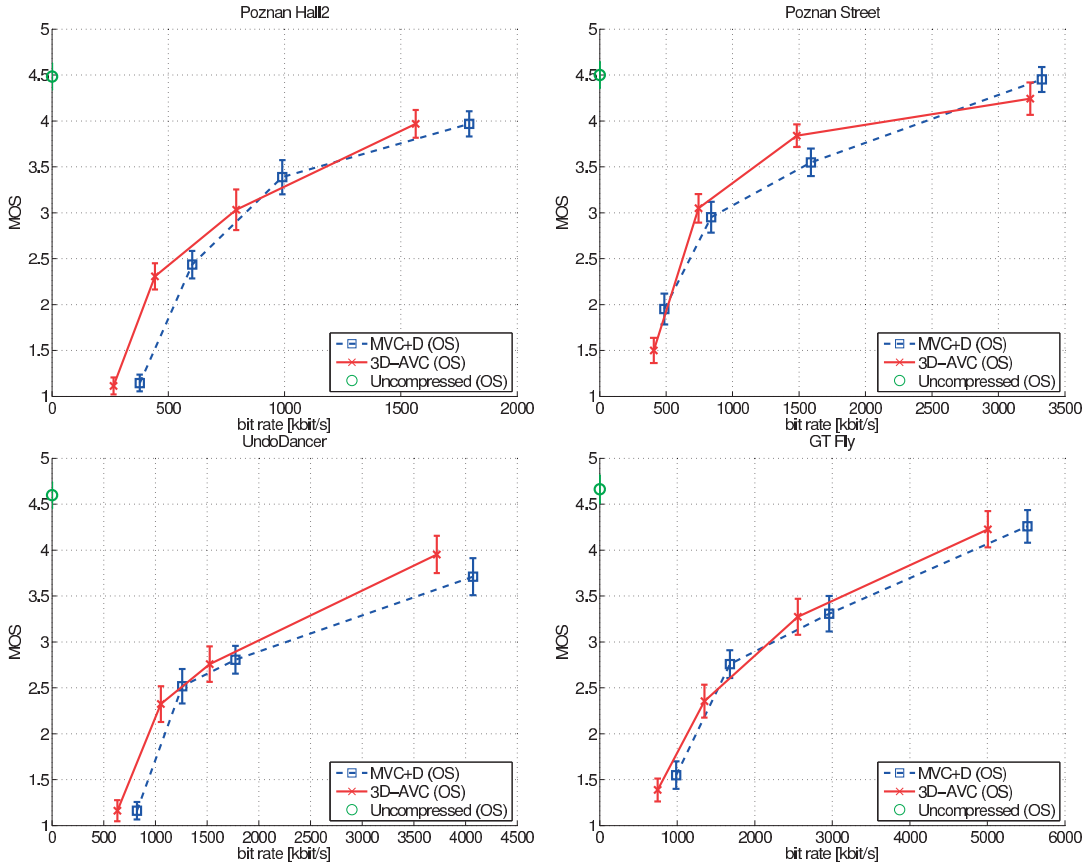


Figure 4.24: R-D curves: *OS* stereo pair.

Rate Distortion Curves

The previous results show a strong correlation between the different laboratories. To further determine whether the scores from the different laboratories can be merged, an ANOVA was performed at a 5% significance level on the raw scores. The main effect of laboratories was not significant. Therefore, the raw scores from the three laboratories were merged in the following analyses.

Figures 4.24 and 4.25 depict the R-D curves for the *SS* and *OS* stereo pairs, respectively. As it can be observed, 3D-AVC usually outperforms MVC+D, as most of the R-D curve of 3D-AVC lie above that of MVC+D. However, comparing the two codecs at specific QP settings show that the CIs overlap in most cases, which indicates that the difference between the two codecs is not significant in most cases.

Average Bit Rate Difference

Table 4.20 reports the average bit rate difference ΔR for 3D-AVC over MVC+D computed from the MOSs using the SCENIC model (see Section 3.3). For both stereo pair configurations,

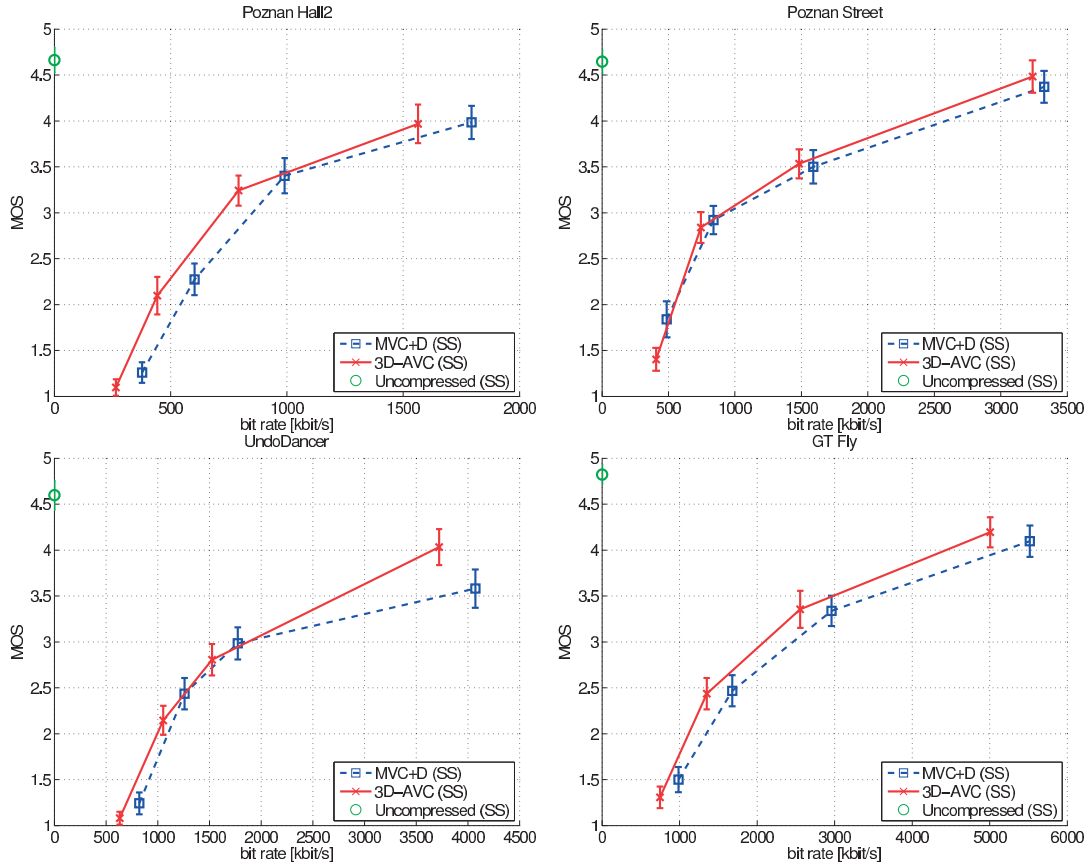


Figure 4.25: R-D curves: *SS* stereo pair.

Table 4.20: Bit rate differences for 3D-AVC over MVC+D.

(a) *SS* stereo pair

Sequence	ΔR	$[\Delta R_{\min}, \Delta R_{\max}]$	Confidence index (%)
Poznan Hall2	−17%	[−31%, −1%]	89
Poznan Street	−8%	[−25%, 17%]	93
UndoDancer	−14%	[−25%, 10%]	88
GT Fly	−16%	[−32%, 5%]	89
Average	−14%	[−28%, 8%]	90

(b) *OS* stereo pair

Sequence	ΔR	$[\Delta R_{\min}, \Delta R_{\max}]$	Confidence index (%)
Poznan Hall2	−13%	[−29%, 3%]	87
Poznan Street	−21%	[−35%, −3%]	84
UndoDancer	−15%	[−32%, 8%]	82
GT Fly	−6%	[−25%, 19%]	87
Average	−14%	[−30%, 7%]	85

results show that, in average, 3D-AVC offers 14% bit rate reduction when compared to MVC+D, which is lower than the 22.6% bit rate reduction measured based on objective results (JCT3V-F0094). However, from the CIs, it can be seen that the bit rate difference varies from −30% to +8%, which indicates that sometimes MVC+D is better than 3D-AVC, as it can observed from the R-D curves.

4.7 Conclusion

This chapter reported the performance of different coding formats for still image, video, HDR image, HDR video, and 3D video compression. Some of these coding formats, e.g., the HDR&WCG extension of HEVC, are still under development by international standardization organizations. However, other coding formats, e.g., HEVC, were recently standardized and are starting to appear in consumer applications, thanks to dedicated hardware implementation in new devices. These new coding formats provide a solution to the increasing amount of data due to higher resolution, faster frame rate, higher bit depth, higher dynamic range, wider color gamut, etc. In particular, results showed that HEVC can achieve more than 50% bit rate reduction when compared to AVC, which is the current standard used in video delivery, while providing the same visual quality. A significantly higher compression performance can be achieved on resolutions beyond HD, manly thanks to better flexibility, adaptability, and signaling. The upcoming HEVC video compression standard seems to be one of the key elements towards a wide deployment of 4K and 8K resolutions. Additionally, HEVC intra coding outperforms encoders for still images with an average bit rate reduction ranging from 17% (compared to JPEG 2000) up to 44% (compared to JPEG). These findings imply that both still images and moving pictures can be efficiently compressed by the same encoder, i.e., HEVC, and therefore specialized still image compression encoders may be becoming redundant, at least if judged by compression efficiency criteria only. Extensions of HEVC for higher bit depth and enhanced chroma sampling structures (Sullivan et al., 2013), screen content coding (J. Xu et al., 2016), 3D video coding (Y. Chen and Vetro, 2014), scalable video coding (Boyce et al., 2016), and HDR&WCG video coding make HEVC a perfect multi-purpose coding format.

All these performance analyses were mainly performed using subjective quality evaluations to provide a more realistic estimation of the true coding efficiency. Except for the study on VP9 video compression, which was targeting an Internet-based streaming scenario, all other evaluations were performed in a controlled test environment that fulfills the recommendations for the subjective evaluation of visual data issued by ITU to obtain repeatable results. All evaluations were performed following the guidelines provided by international recommendations to obtain reliable results. To illustrate the reliability and repeatability of the results, the study on 3D video compression was performed in three different laboratories in Europe. Even though the different laboratories used different displays and different subjects, analyses showed that the test laboratories could still produce highly correlated results, as they follow similar guidelines to carry out assessments.

5 Investigation of Alternative Evaluation Protocols

For more than 40 years, most subjective quality evaluations have been conducted on 2D LDR still images and video sequences. The first version of the ITU-R BT.500 recommendation on "Methodology for the subjective assessment of the quality of television pictures" was released in 1974. Since then, many technological revolutions have occurred in imaging and display technologies, e.g., the shifts from analog to digital, CRTs to LCDs, 2D to 3D, or LDR to HDR. Unfortunately, the guidelines and methodologies for subjective evaluations have not always been updated to reflect the requirements of new technologies. For example, in ITU-R BT.500-13 (2012), the maximum observation angle and preferred viewing distance are still based on the properties of CRT displays. Some other recommendations have been published to overcome some of these problems, e.g., ITU-R BT.2021 (2012) for "Subjective methods for the assessment of stereoscopic 3DTV systems" and ITU-R BT.2022 (2012) for "General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays". However, there is still a lack of evaluations guidelines, methodologies, and protocols for new applications, e.g., FTV, HDR, interactive technologies, or virtual reality.

For more than 40 years, most subjective quality evaluations have been conducted in laboratory environments. However, conducting subjective experiments is very time consuming and can be quite expensive, especially in countries such as Switzerland, even when hiring students. To reduce the costs of subjective evaluations and also to consider more practical environments, researchers are investigating crowdsourcing platforms, which allow employing workers online from around the world. The authors of (Hossfeld et al., 2014a) provide a comprehensive overview of crowdsourcing approaches for subjective evaluations of image and video content and (Hossfeld et al., 2014b) discusses and compares the corresponding existing implementation frameworks. Both works also discuss issues and limitations of crowdsourcing in the context of subjective evaluations. One of the constraints is the limited variety of display devices used by online workers. Due to this limitation, for example, a direct evaluation of 3D or HDR content is impossible, since 2D SDR displays are the most commonly used, especially for workers from Bangladesh, India, Pakistan, Philippines, Singapore, or Thailand, which are commonly found on crowdsourcing platforms such as Microworkers.com (Redi et al.,

2013). Therefore, it is necessary to use alternative representations of 3D and HDR content in crowdsourcing evaluations.

This chapter investigates alternative evaluation protocols for subjective quality assessment. In particular, to overcome the lack of standardize test methodologies for FTV scenarios, we propose a new subjective assessment protocol that consists in assessing the perceived image quality of FVV sequences corresponding to a smooth camera motion during a time freeze. Section 5.1 provides a detailed description of the proposed assessment protocol and investigate the assessment of the impact of depth compression on perceived image quality in a FTV scenario using the proposed protocol. This study considers depth maps compression only (and not color view compression, as in a classical scenario), as it has been shown that depth compression has a critical impact on the quality of synthesized views. Sections 5.2 and 5.3 investigate alternative representations of 3D and HDR content for crowdsourcing evaluations of MVD video and HDR image coding, respectively, on 2D LDR displays. The crowd-based evaluations results are compared to ground truth subjective scores obtained in lab-based evaluations to investigate the suitability of crowdsourcing evaluations for 3D and HDR content.

5.1 A Quality Assessment Protocol for Free-Viewpoint Video Sequences

With the growing interest for stereoscopic 3D imaging (Kubota et al., 2007), VCEG and MPEG have joined their efforts to develop new 3D video formats and coding standards. Among the numerous possible 3D scene representations is the MVD format (Smolic et al., 2007). This format consists of multiple texture views and associated depth maps acquired at different viewpoints of the represented scene. Although the history of stereoscopic video sequences dates back from the last century, the subjective quality assessment protocols that are essential to evaluate new 3D viewing systems are not standardized yet. This is very likely to be due to the complexity brought by 3D and the numerous possible 3D applications. The most popular applications are 3DTV and FTV. 3DTV provides a depth feeling thanks to an appropriate 3D display. FTV allows the user to interactively control the viewpoint of the scene.

Considering the demand for high-quality visual content, the success of 3D video applications is closely related to its ability to provide viewers with a high quality level of visual experience. While many efforts have been dedicated to visual quality assessment in the last twenty years, some issues still remain unsolved in the case of 3D video. The assessment of 3D contents arises different issues

i) Quality assessment of synthesized views: 3DTV and FTV are likely to require view synthesis, which is often performed via DIBR. This process can induce new types of artifacts. Since view synthesis is fundamental for both 3DTV and FTV, the quality assessment of synthesized views is crucial.

ii) Specific distortions in DIBR: artifacts in DIBR are mainly geometric distortions. These distortions are different from those commonly encountered in video compression and that are assessed by usual evaluation methods. Most video coding standards rely on discrete cosine transform, which results in specific artifacts (some of them are described by Yuen and H. R. Wu (1998)). These artifacts are often scattered over the whole image, whereas DIBR related artifacts are mostly located around the disoccluded areas. Thus, since most of the usual objective quality metrics were initially designed to address specific usual distortions, they may not be suitable to assess the quality of DIBR synthesized views (see Sections 10.1 and 10.2 and (Bosc et al., 2011a; Bosc et al., 2012b)).

iii) Use case and visualization scenario: the evaluation of DIBR systems is a difficult task because the type of evaluation differs depending on the context of use. Different factors are involved in the different 3D imaging applications. A major discriminatory factor is the stereopsis phenomenon (the fusion process of left and right images by the HVS), exploited by 3DTV systems. Psycho-physiological mechanisms are induced but they are not completely understood. A FTV application is not necessarily used in conjunction with a stereoscopic display, as FTV can be watched in a 2D context. Consequently, the quality assessment protocols differ as they address the quality of synthesized views in two different contexts (2D and stereoscopic visualization). It is obvious that stereoscopic impairments (such as cardboard effect, crosstalk, keystone, flickering depth, picket-fence, etc., as described by Meesters et al. (2004)), which occur in stereoscopic conditions, are not assessed in 2D conditions. Also, distortions detected in 2D conditions may not be perceptible in stereoscopic conditions.

iv) Assessment factors: Depending on the use case, except for the conventional image quality, new assessment factors can be considered such as comfort, naturalness, and depth perception (W. Chen et al., 2012).

v) Clear definition of assessment factors: even though a training session is usually performed before each subjective quality assessment test, subjects are generally non-expert. In addition, they may not be familiar with simulated stereoscopic viewing. Therefore, there is a risk of collecting erroneous results due to the novelty of the media display, which may not always be taken into account in these subjective quality assessment methodologies. The assessment factors need to be clearly defined to avoid confusion during the rating procedure.

vi) Need for no-reference metric: another limitation of usual objective metrics concerns the need for non-reference quality metrics. In particular use cases, such as FTV, references are unavailable because the generated viewpoint is virtual. In other words, there is no ground truth allowing a full comparison with the distorted view. Though, assessment tools are required to evaluate the quality of the synthesized views.

The ITU has recently released a new recommendation for the assessment of stereoscopic 3DTV systems: ITU-R BT.2021 (2012). This recommendation is mostly an extension for 3DTV of the well known recommendation ITU-R BT.500-13 (2012), which was established for 2D television. The recommendation includes a subset of four methods from ITU-R BT.500-13 (2012) (see

Table 5.1: MVD contents used in the experiment.

	Content	Resolution	Type	Encoded views	Frame no.
S1	*Balloons*	1024 × 768	Natural	1 – 5	1
S2	*Book Arrival*	1024 × 768	Natural	6 – 10	33
S3	*Undo Dancer*	1920 × 1080	Synthetic	1 – 9	250
S4	*GT Fly*	1920 × 1080	Synthetic	9 – 1	157
S5	*Kendo*	1024 × 768	Natural	1 – 5	1
S6	*Newspaper*	1024 × 768	Natural	2 – 6	1

Section 2.4), namely the SS, DSCQS, SC, and SSCQE methods. According to ITU-R BT.2021 (2012), picture quality, depth quality, and visual comfort of stereoscopic imaging technologies should be assessed. However, this recommendation does not address the specific issue of synthesized views. Therefore, subjective quality assessment of 3D contents represented in the video plus depth or MVD formats, and, as a consequence, of virtual synthesized views, has been conducted according to methods used for the assessment of conventional 2D contents. For example, Hewage et al. (2009) have used the DSCQS methodology to evaluate the quality of stereoscopic video sequences that were synthesized from video plus depth video sequences. Recently, the DSIS has been used to evaluate the responses of the MPEG CfP on 3D Video Coding Technology. The evaluations have been performed on both stereoscopic and multiview auto-stereoscopic displays. The displayed 3D contents were synthesized via DIBR from a limited number of input views represented in the MVD format.

In this section, we investigate the assessment of the impact of depth compression on perceived image quality in a FTV scenario. This study considers depth maps compression only (and not color view compression, as in a classical scenario), as it has been shown that depth compression has a critical impact on the quality of synthesized views. To overcome the lack of standardize test methodologies for FTV scenarios, we propose a new subjective assessment protocol that consists in assessing the perceived image quality of FVV sequences corresponding to a smooth camera motion during a time freeze, which were generated through DIBR from 3D content represented in the MVD format. This protocol is expected to enable the evaluation of different types of depth coding distortions. The proposed protocol is original because it proposes a novel approach to assess image quality of free-viewpoint content via synthesized frames in a *freezing time* scenario. In this section, we make a complementary use of simple and reliable statistical tools to analyze the subjective evaluation results. This section reports a detailed analysis of the results, which shows the importance of content selection when evaluating visual image quality of free-viewpoint data.

5.1.1 Dataset

Six MVD contents were used in this experiment: *Book Arrival*, *Newspaper*, *Kendo,* and *Balloons* are real scenes with estimated depth maps; and *GT Fly* and *Undo Dancer* are synthetic scenes with ground truth depth maps. Table 5.1 summarizes the characteristics of the contents and

encoded viewpoints. The contents are referred to as *S1* to *S6* (see Table 5.1). The contents and the key frames were selected for their availability and amount of depth.

The depth map compression algorithms under test are labeled from *C1* to *C7*, where *C1* to *C4* are state-of-the-art codecs

- *C1*: 3D-HEVC, 3D-HTM version 0.4, inter-view prediction and *View Synthesis Optimization* enabled,
- *C2*: MVC, JM version 18.4, inter-view prediction enabled,
- *C3*: HEVC, HM version 6.1,
- *C4*: JPEG 2000, Kakadu implementation,
- *C5*: based on (Gautier et al., 2012), a lossless-edge depth map coding based on optimized path and fast homogeneous diffusion,
- *C6*: based on (Pasteau et al., 2011), this algorithm exploits the correlation with color frames, and
- *C7*: Z-LAR-RP (Bosc, 2012), a region-based algorithm.

All compression algorithms were used in intra coding mode. In the case of 3D-HTM, colored views and their associated depth maps were provided as inputs to the encoder, but only the decoded depth maps were employed in our experiments. The stimuli were not classically selected relying on a list of bit rates to be evaluated. Instead, the stimuli were previously selected by expert viewers based on their visual quality. For each compression algorithm, the subjective visual quality of the views synthesized from decompressed depth data, at different bit rates, were first considered by the expert viewers. Then, for each compression algorithm, the expert viewers selected one stimulus corresponding to each of the following categories: *Good*, *Fair*, and *Poor*. Therefore, three QPs were selected for each depth map compression algorithm under test, according to the visual quality of the rendered views. For each compression algorithm, we refer to the highest, middle, and lowest bit rates evaluated as *R0*, *R1*, and *R2*, respectively. This choice was motivated by the need to cover a wide range of categories in the visual quality scale to properly assess each codec under test. Two additional methods were also included to increase the variety of distortions: low pass filtered depth maps (noted *F*) and depth maps with low-pass filtered applied on edges only (noted *FE*). Table 5.2 provides our observations regarding the specific distortions of each method, when using a coarse quantization. The first column indicates the effects of coarse compression on depth maps. The second column indicates the resulting effects on views synthesized from this decompressed depth data. The depth compression related artifacts mostly affect the strong depth gradients (object edges), which results in flickering around the object edges in the synthesized sequence.

Two different synthesis modes were considered in this study. The synthesis process was performed through the 3D-HTM renderer, also referred to as VSRS-1D-Fast, which is the view synthesis algorithm used by JCT-3V at the time of writing this paper. Two different modes for the synthesis process, referred to as *VS1* and *VS2* in the rest of the paper, were considered:

Table 5.2: Impact of coarse depth quantization on depth maps and synthesized views.

Method	Effects on depth maps	Effects on synthesized views
C1	scattered blocking effect	staircase effect on object edges
C2	blur	inaccurate edges
C3	blur	inaccurate edges
C4	blurred, ringing edges	deformed edges, crumbling edges
C5	blur, introduction of gradients	deformed objects
C6	blocking effect	blocking effect around edges
C7	smooth depth fading	reduced parallax
F	blur	deformed objects
FE	blurred edges	inaccurate edges

- *VS1*: *Blended Mode* disabled: all pixels visible in the closer reference view are copied to the virtual view, and only hole areas are filled from the farther reference view and
- *VS2*: *Blended Mode* enabled: a weighted blending based on the baseline distance is used for hole filling, such that pixels from the reference camera that are closer to the virtual view are assigned a higher weight.

5.1.2 Methodology

The proposed quality assessment protocol aims at highlighting the impact of depth coding only on the synthesized views in a FTV scenario. A specific case of use is considered to allow a reliable comprehension the impact of depth coding: smooth motion of camera when freezing time in a free-viewpoint application. In this subsection, the choices that motivated the design of this experimental protocol are presented. Then the subjective assessment conditions and analysis tools are discussed.

Proposed Experimental Protocol

In the absence of subjective methodologies specifically designed to assess the quality of 3D content, assessment protocols that have been developed and validated for 2D content are mostly used. The aim of this experiment is to evaluate the impact of depth compression on perceived image quality of free-viewpoint data. Depth maps compression only (and not color view compression, as in a classical scenario) is considered in this experiment as it has been shown that depth compression has a critical impact on the quality of synthesized views. We recall that stereopsis is not considered in this experiment. Considering the aim of this experiment, the design of a subjective quality assessment methodology should be based on consideration of reliability, accuracy, efficiency, and easiness of implementation of the available methodologies. Our experimental protocol relies on these concerns.

Regarding the construction of the assessed stimuli, we considered a scenario that involved the generation and the quality assessment of synthesized views in a FTV scenario. The assessment

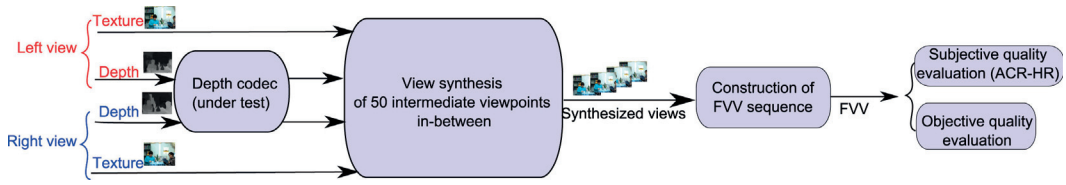


Figure 5.1: Overview of the experimental protocol.

protocol targets depth coding only and not color coding as in the "real" case of use because the goal is to provide an assessment protocol able to underline the impact of depth coding only on the synthesized views. Moreover, to allow a reliable comprehension of the studied phenomenon (impact of depth coding), a specific case of use is considered in the proposed methodology: smooth motion of camera when freezing time in a free-viewpoint application. Indeed, most of the free-viewpoint applications involve freezing time when moving from one viewpoint to another. This particular case is prone to meticulous observation by the user and distortions occurring at this stage may reduce the perceived QoE.

So, only the depth maps were encoded as for an example of evaluation of depth coding algorithms. The general scheme followed in this experiment is depicted in Figure 5.1. From a given MVD sequence, two different viewpoints at one time instant t (also referred to as key frames in the following) were considered. The associated depth maps were encoded through seven depth map codecs under test. From the decoded depth maps, fifty intermediate viewpoints (equally separated) were generated in-between the two considered viewpoints. A sequence of 100 frames (at 10 fps) was built from the 50 intermediate virtual frames to simulate a smooth camera motion from left to right and from right to left. This experimental protocol is expected to reveal the distortion specificity of each compression strategy. This leads to a specific case of use in free-viewpoint applications since it simulates a smooth motion of camera when freezing time.

Evaluation Method

Among the different standardized subjective quality assessment methods, the ACR-HR (see Section 2.4.1) has been widely used to assess 2D content and have also been used to assess content related to 3D video applications (Campisi et al., 2007; Kalva et al., 2007). This method is often chosen for its known reliability in the context of the evaluation of 2D media. Huynh-Thu et al. (2011a) have conducted a study to compare different methods and different rating scales. The tests were carried out in the context of HD video. The results showed that the ACR method produced reliable subjective results, even across different scales. In addition, this basis method is known for its easiness of implementation. Based on these previous findings, we selected the single-stimulus pattern presentation and ACR-HR method with a five-grade quality scale.

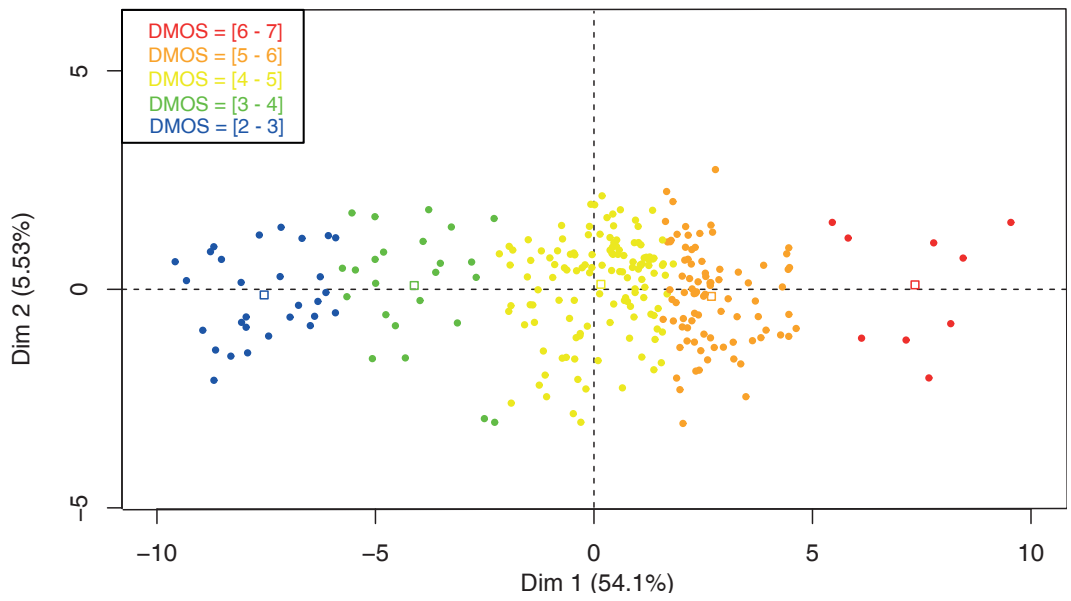


Figure 5.2: PCA plot with graphical emphasis on the DMOS values obtained by each stimulus.

Assessment Conditions

As specified above, the ACR-HR methodology was used to assess the FVV sequences. The combination of contents, view synthesis modes, depth map compression algorithms, and bit rates selected in this study (see Section 5.1.1) resulted in a total of 276 processed stimuli and 12 reference stimuli to be assessed. The subjective evaluations were conducted in an ITU conforming test environment. The stimuli were displayed on a Panasonic BT-3DL2550 screen (1920 × 1080p), and according to ITU-R BT.500-13 (2012). Twenty-seven naive observers participated in the subjective quality evaluation test into two 30-minutes sessions. All subjects underwent a screening to examine their visual acuity, color vision, and stereo vision. Four subjects were detected as outliers and all their scores were removed from the results.

5.1.3 Results

This subsection provides the results of statistical analyses of the obtained subjective scores. In the following figures, the results of the principal component analysis (PCA) over the obtained subjective scores are depicted. The distribution of scores of each observer has been normalized prior to the computation of the factor scores. In these figures, each point represents a stimulus. The points have different colors according to the view synthesis mode, content, depth map compression algorithm, and bit rate. In particular, in Figure 5.2, the points are colored depending on their DMOS values. As it can be observed in Figure 5.2, the stimuli obtaining the lowest DMOS (left part) are opposed to those obtaining the highest DMOS (right part). In the legend of Figure 5.2, the mentioned categories do not correspond to the classical five-grade categories of ACR-HR. They correspond to quintiles from the distribution of the obtained

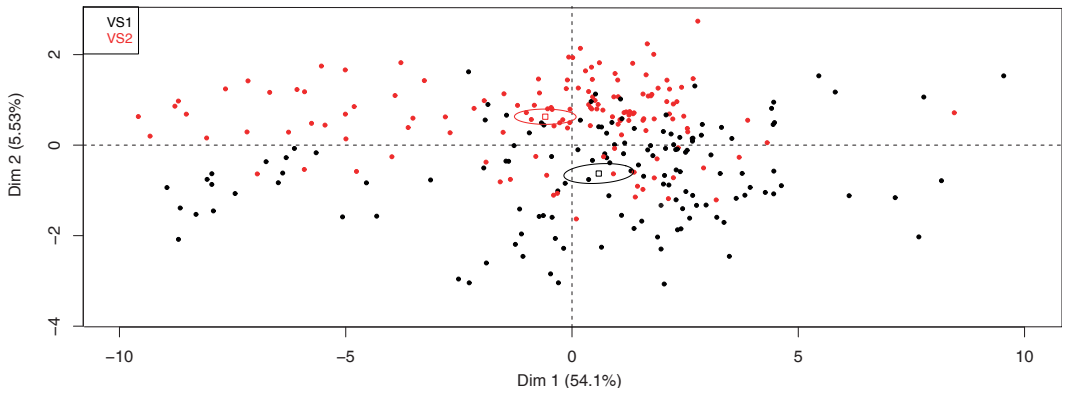


Figure 5.3: PCA plot with graphical emphasis on the synthesis modes.

DMOS values. So the blue points correspond to the quintile with the lowest DMOS values (ranged in $[2-3]$) and the red points correspond to the quintile with the largest DMOS values (ranged in $[6-7]$). The first two components of the PCA resumed 59.6% of the total inertia, i.e., the total variance of the dataset (the first axis resumed 54.1% of the total inertia while the second resumed 5.53% of the total inertia). The results of the ANOVA are given in the following figure to determine the influence of the different parameters, i.e., view synthesis mode, content, depth map compression algorithm, and bit rate. The results are discussed in the following parts.

View Synthesis Modes

Figure 5.3 shows the individual factor map from the PCA according to the view synthesis mode. It can be observed that although the ellipses centroids are close to each other, they are diametrically opposed. *VS1*'s ellipse centroid is located in the right part of the graph, which indicates that *VS1* generally obtained higher DMOS values and thus the perceived quality was higher. As the confidence ellipses do not overlap, the difference is statistically significant.

The ANOVA results showed that processed stimuli related to *VS1* obtained statistically better image quality scores when compared to *VS2*, with a p-value < 0.001, which confirms the PCA results. Since the depth and texture data was the same for the two view synthesis modes, this result shows that no blending (*VS1*) is better than blending (*VS2*) in terms of image quality.

To further investigate the difference between the two view synthesis algorithms, the MOS of the reference version of the stimulus, average MOS and DMOS of all processed stimulus for each content were computed. Figure 5.4 depicts the results for each content. For the hidden references (SRC), the view synthesis modes are statistically equivalent, except for content *S5*. As the depth maps used to generate the stimuli were not compressed in this case, the left and right depth maps are consistent and the artifacts are mostly due to the view synthesis process. We observed that the blending mode (*VS2*) seemed to smooth out the view synthesis artifacts, which is slightly visually more pleasant. On the other hand, when not blending was applied

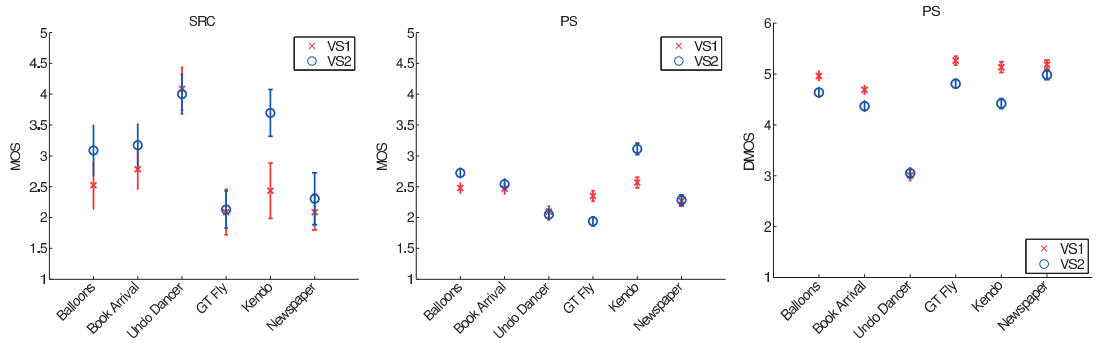


Figure 5.4: MOS of hidden references (SRC), average MOS of processed sequences (PS), and average DMOS of PS.

(*VS1*), some jitter was visible around the depth maps edges, which do not necessarily match the image edges (especially for content *S5*). This effect was visually less pleasant. Compression of the depth maps tends to remove the high frequencies and smooth out the sharp edges. As the left and right depth maps were compressed independently, some inconsistencies may arise. Therefore, when blending was applied (*VS2*), view synthesis errors resulted in large blurred areas, mostly around depth edges, which decrease the visual quality when compared to the SRC. On the other hand, when not blending was applied (*VS1*), the jitter was reduced as the depth edges were reduced, which results in similar visual quality when compared to the SRC. These observations explain why *VS1* usually outperformed *VS2*. These results show that the proposed assessment methodology allows the evaluation of different view synthesis algorithms.

Content Characteristics

Figure 5.5 shows the individual factor map from the PCA according to the content. A clear distinction between *S3* and the other contents of the dataset can be observed. The ellipse centroid is located in the left part of the graph, which indicates that *S3* generally obtained the lowest DMOS values. On the other hand, contents *S4* and *S6* obtained the highest DMOS.

Table 5.3 reports the results of the ANOVA. Contents *S1* and *S5* have equal means according to ANOVA. On the individual factor map (see Figure 5.5), the two confidence ellipses have almost collocated centroids. Contents *S4* and *S6* have equal means according to ANOVA and it can be also observed that their confidence ellipses cross each other on the PCA plot. However, content *S6* obtained statistically higher image quality scores than contents *S1* and *S5* according to ANOVA, whereas their confidence ellipses cross each others on the PCA plot.

To understand the influence of content characteristics on perceived quality, a set of content features was extracted from the original video sequences and compared to the average DMOS obtained on the different contents. Both 2D and 3D features were extracted from the original video sequences. For the 2D features, the well-known SI and TI (see Section 2.2) are often

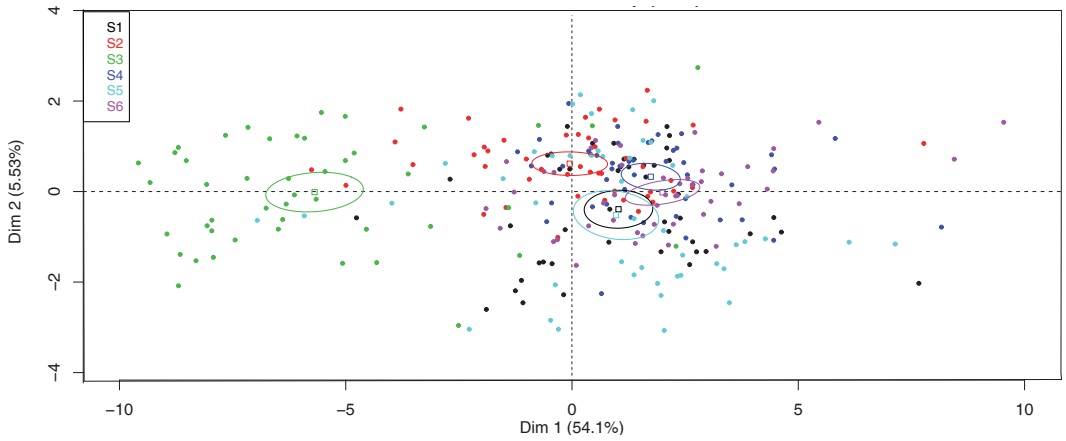


Figure 5.5: PCA plot with graphical emphasis on the contents.

Table 5.3: Results of the ANOVA test on the influence of the contents ($p < 0.001$).

	S1	*S2*	*S3*	*S4*	*S5*	*S6*
S1		↑	↑	↓	**o**	↓
S2	↓		↑	↓	↓	↓
S3	↓	↓		↓	↓	↓
S4	↑	↑	↑		↑	**o**
S5	**o**	↑	↑	↓		↓
S6	↑	↑	↑	**o**	↑	

Legend: ↑: superior, ↓: inferior, **o**: statistically equivalent. Reading: Line 1 is statistically superior to column 2.

used to characterize the amount of spatial detail of a picture and temporal changes of a video sequence, respectively. The 2D features were computed on the luminance component of each content. For the 3D features, Mittal et al. (2011) have proposed that 3D images have certain statistical properties that can be captured using simple statistical measures of the disparity distribution. Thus, the following 3D features were computed on the disparity map D of each content, according to (Mittal et al., 2011)

1) mean disparity $\mu = \mathrm{E}[D]$,
2) median disparity $med = median(D)$,
3) disparity standard deviation $\sigma = \sqrt{\mathrm{E}[(D-\mu)^2]}$,
4) kurtosis of disparity $\kappa = \frac{\mathrm{E}[(D-\mu)^4]}{(\mathrm{E}[(D-\mu)^2])^2}$,
5) skewness of disparity $skew = \frac{\mathrm{E}[(D-\mu)^3]}{(\mathrm{E}[(D-\mu)^2])^{(3/2)}}$,
6) mean differential disparity $\mu_d = \mathrm{E}[\delta D]$,
7) differential disparity standard deviation $\sigma_d = \sqrt{\mathrm{E}[(\delta D - \mu_d)^2]}$,
8) kurtosis of differential disparity $\kappa_d = \frac{\mathrm{E}[(\delta D-\mu_d)^4]}{(\mathrm{E}[(\delta D-\mu_d)^2])^2}$, and
9) skewness of differential disparity $skew_d = \frac{\mathrm{E}[(\delta D-\mu_d)^3]}{(\mathrm{E}[(\delta D-\mu_d)^2])^{(3/2)}}$

Table 5.4: Correlation between average DMOS and content characteristics.

SI	TI	DSI	DTI	μ	med	σ	κ	$skew$	μ_d	σ_d	κ_d	$skew_d$
-0.72	-0.26	0.12	0.19	0.04	-0.08	-0.27	0.13	0.72	-0.30	0.11	-0.72	-0.45

Figure 5.6: Texture and depth map of content *S3*.

where the differential disparity (δD) was computed using a Laplacian operator on the disparity map. The SI and TI features were also computed on the disparity map, and are referred to as DSI and DTI, respectively. For both 2D and 3D features, the value corresponding to the selected key frame was considered instead of the maximum or average value across the video sequence. A total of thirteen features, two 2D feature and eleven 3D features, were extracted for each content. Table 5.4 reports the correlation between the average DMOS for each content and the different features extracted. It can be observed that SI, κ_d, and $skew$ have the highest absolute correlation with the subjective scores. The results show that content with high texture information will be rated lower as the synthesis artifacts will be more visible. A negative skew in the distribution of the disparity values will have a negative impact on perceived quality, as most pixels will be shifted in the view synthesis process. Content *S3* is composed of two major transversal planes for the wall and floor (see Figure 5.6), which are highly textured, and lead to more visible synthesis errors than a flat depth plane with equal depth values. The negative correlation coefficient associated with κ_d seems to indicate that content with less depth discontinuities would be rated lower, which is counterintuitive as synthesis errors usually arise around depth discontinuities due to occlusion.

These observations show the importance of content selection choices when evaluating visual quality. Contents representing a typical range of content characteristics (texture, depth, etc.) should be used. To evaluate the performance of compression (view synthesis) algorithms, contents that are easier and harder to compress (render) should be used to cover the upper and lower range of the quality scale, respectively.

Depth Map Compression Algorithms

Figure 5.7 shows the individual factor map from the PCA according to the depth map compression algorithm. The location of the ellipses centroid can provide a ranking of the methods performances, based on the PCA. Many ellipses overlap each others, meaning that the image

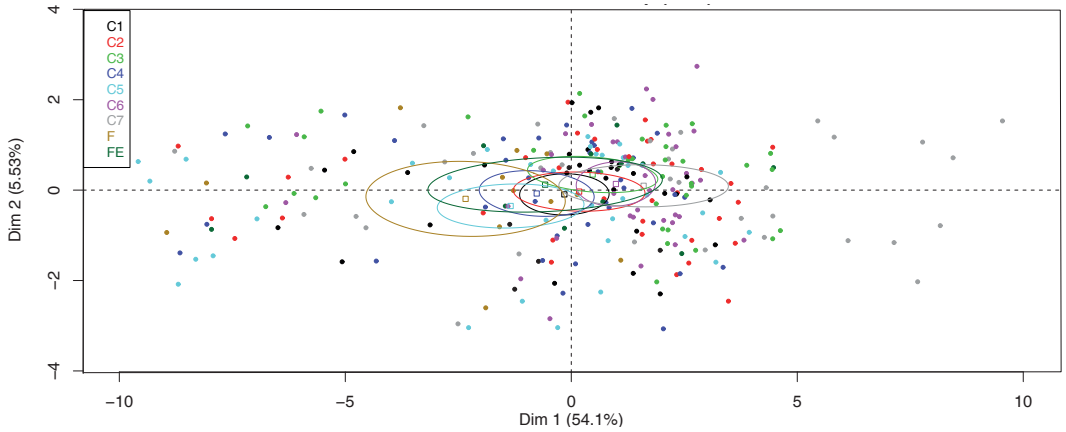


Figure 5.7: PCA plot with graphical emphasis on the compression algorithms.

quality is not statistically different in these cases, according to PCA. However, results show that *C7* and *C6* obtained the highest quality scores and are statistically better in terms of image quality than *F* and *C5*, which obtained the lowest DMOS. *C1*, *C2*, *C3*, *C4*, and *FE* obtained similar DMOS, according to PCA.

Table 5.5 reports the results of the ANOVA, which can be used to further analyze the impact of perceived quality due to the different codecs. According to ANOVA, *C7* obtained the best quality scores, followed by *C6*, whereas *F* obtained the lowest DMOS, followed by *C5*. These results are similar to those of the PCA. However, the ANOVA results indicate in general a statistical difference between the group formed by *C1*, *C2*, *C3*, *C4*, and *FE* and the group formed by *C7*, *C6*, *F*, and *C5*, which was not significant according to PCA. According to ANOVA, the depth maps encoded with *C1* produced similar image quality as those encoded with *C2* to *C4*. *C6* and *C7*, which are based on the same scheme basis, obtained similar image quality scores. An interesting remark lies in the fact that *FE* obtains equal means compared to three state-of-the-art methods (*C1*, *C2*, and *C4*) and to *C5*, which might suggest that the artifacts induced by *FE* may have been rated similarly to these methods. However, as the bit rates were not the same between the different compression algorithms (see Section 5.1.1), no conclusion can be drawn on the relative performance of the different compression algorithms.

Bit Rates

Figure 5.8 shows the individual factor map from the PCA according to the bit rates. Results show that the stimuli generated at *R0*, which corresponds to the highest bit rate for each compression algorithm, obtained statistically better quality scores than the stimuli generated at *R2*, which corresponds to the lowest bit rate for each compression algorithm. However, there is no statistical difference according to PCA between *R0* and *R1* as well as between *R1* and *R2*, as the respective confidence ellipses intersect each others.

Table 5.6 reports the results of the ANOVA, which can be used to further analyze the influence

Table 5.5: Results of the ANOVA test on the influence of the compression algorithms ($p < 0.001$).

	C1	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*F*	*FE*
C1		**o**	**o**	**o**	↑	↓	↓	↑	**o**
C2	**o**		**o**	↑	↑	↓	↓	↑	**o**
C3	**o**	**o**		↑	↑	**o**	↓	↑	↑
C4	**o**	↓	↓		**o**	↓	↓	↑	**o**
C5	↓	↓	↓	**o**		↓	↓	**o**	**o**
C6	↑	↑	**o**	↑	↑		**o**	↑	↑
C7	↑	↑	↑	↑	↑	**o**		↑	↑
F	↓	↓	↓	↓	**o**	↓	↓		↓
FE	**o**	**o**	↓	**o**	**o**	↓	↓	↑	

Legend: ↑: superior, ↓: inferior, **o**: statistically equivalent. Reading: Line 1 is statistically superior to column 5.

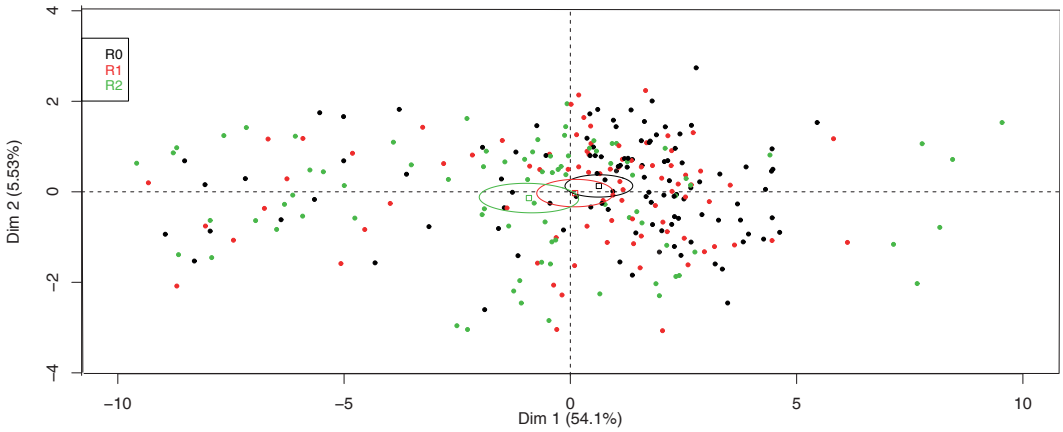


Figure 5.8: PCA plot with graphical emphasis on the bit rates.

of the different bit rates. According to ANOVA, *R0* lead to statistically better image quality than *R1* and *R1* lead to statistically better image quality than *R2*, which differs from the PCA results. From these results, it seems that ANOVA can distinguish more finely between different conditions.

A Specific Case: *C7*

Results showed that *VS1* lead to statistically better image quality than *VS2* (see the analysis of the impact of the view synthesis mode at the beginning of this subsection) with a p-value of 2.95×10^{-23}, which is clearly discriminative. This result was further analyzed by performing an ANOVA to determine the influence of the view synthesis mode on each depth map compression algorithm separately. Table 5.7 reports the results of this analysis. As it can be observed, the difference in p-value between *C7* and the other algorithms is higher than 25 orders of

Table 5.6: Results of the ANOVA test on the influence of the bit rates ($p < 0.001$).

	R0	*R1*	*R2*
R0		↑	↑
R1	↓		↑
R2	↓	↓	

Legend: ↑: superior, ↓: inferior, **o**: statistically equivalent. Reading: Line 1 is statistically superior to column 2.

Table 5.7: Influence of the view synthesis modes across codecs.

	Result	p-value	F-value
All	*VS1* > *VS2*	2.95×10^{-23}	99.47
C1	*VS1* = *VS2*	7.24×10^{-1}	0.12
C2	*VS1* > *VS2*	3.27×10^{-2}	4.58
C3	*VS1* > *VS2*	3.26×10^{-4}	13.02
C4	*VS1* > *VS2*	4.26×10^{-2}	4.12
C5	*VS1* > *VS2*	5.54×10^{-5}	16.43
C6	*VS1* = *VS2*	4.17×10^{-1}	0.66
C7	*VS1* > *VS2*	3.26×10^{-30}	141.41

magnitude.

As *C7* was reported to provide the best image quality when all processed stimuli were considered (see Section 5.1.3), a detailed analysis was performed to determine the compression algorithms leading to the best and worst image quality on each content, for both view synthesis modes together, as well as for each mode separately. Table 5.8 reports the results of this analysis. It can be observed that *C7* obtained the highest DMOS for contents *S1* and *S5* when no blending was applied (*VS1*), whereas it obtained the lowest DMOS on the same contents when blending was applied (*VS2*). Therefore, the blending mode seems to have a significant impact on perceived quality when *C7* is used to compress the depth maps.

To further investigate the influence of the view synthesis mode on the performance of *C7*, a analysis was performed to determine the influence of the bit rate for the different view synthesis modes. Table 5.9 reports the ANOVA results of this analysis. As it can be observed, when no blending was applied (*VS1*), the contents compressed at *R2* obtained significantly better ratings than contents compressed at *R1* and even *R0*, which is counterintuitive as higher compression introduces more visible artifacts. To explain this behavior, the test data corresponding to *C7* at *R2* was investigated by expert viewers. The processed stimuli showed high image quality but no perceived depth (no motion parallax), which was not evaluated during the subjective test. The compressed depth maps showed that for algorithm *C7*, at very low bit rate, the depth map tends to a smoothed, flattened, and uniform depth image. Therefore, no view synthesis is applied and the perceived quality is very high, but there is no motion parallax as one would experience in a normal FTV scenario.

Table 5.8: Compression algorithms obtaining the highest and lowest DMOS across the different contents.

	VS1 and *VS2*		*VS1* only		*VS2* only	
	Highest DMOS	Lowest DMOS	Highest DMOS	Lowest DMOS	Highest DMOS	Lowest DMOS
All	*C7,C6*	*F*	*C7*	*F*	*C6*	*F*
S1	*C3*	*F*	*C7*	*F*		*C7*
S2	*C6*	*F*	*C7*		*C6*	*F*
S3	*C6*	*C5*	*C6,C7*		*C6*	*C5*
S4	*C7*	*C5,F*	*C7*			*C5,F*
S5		*C5*	*C7*	*C5*		*C5,C7*
S6	*C7*	*F*	*C7*	*F*	*C7*	*F*

Table 5.9: Results of the ANOVA test on the influence of the bit rates for content *C7*.

	VS1 and *VS2*			*VS1* only			*VS2* only		
	R0	*R1*	*R2*	*R0*	*R1*	*R2*	*R0*	*R1*	*R2*
R0		↑	**o**		**o**	↓		↑	↑
R1	↓		**o**	**o**		↓	↓		↑
R2	**o**	**o**		↑	↑		↓	↓	

Legend: ↑: superior, ↓: inferior, **o**: statistically equivalent. Reading: Line 1 is statistically superior to column 2.

These observations show that the proposed experimental protocol (assessment of FVV generated from decompressed data) allows the evaluation of different compression algorithms and that particular behaviors can be highlighted by a detailed analysis of the subjective scores. Other assessment factors, such as depth perception, could be included in the experimental protocol to evaluate other aspects of FTV.

5.2 Crowd-based Evaluation of Multiview Video plus Depth Coding

One simple approach to representing 3D content on a 2D display is to play only one view of the 3D content. The intended depth perception would be lost with this approach, but it may be enough for the evaluation of compression, as many compression artifacts would still be visible even in one view. An alternative approach is to use the subjective assessment protocol proposed in the previous section. This protocol consists in evaluating a FVV (Smolic et al., 2009b) sequence corresponding to a smooth camera motion during a time freeze, which is generated from fifty intermediate views in-between the left and right views of the original content. The resulting effect is similar to the 'bullet time' visual effect used in such movies like "The Matrix". The intermediate views are generated through DIBR (Fehn, 2004b) using the depth maps and texture views. The FVV sequence can then be displayed on a regular 2D monitor and, more importantly, it still retains a depth perception without the aid of any

special glasses, thanks to the motion parallax, which is known to be a strong monocular depth cue (Rogers and Graham, 1979).

This section investigates the suitability of these two alternative representations for the quality assessment of MVD content on 2D displays. In particular, we use virtual view synthesized from the MVD content (referred to as 'mono' in the remaining of this section) and FVV sequence corresponding to a smooth camera motion during a time freeze (referred to as 'sweep' in the remaining of this section). Seven MVD sequences were encoded at different bit rates using the upcoming 3D-AVC video coding standard (see Section 4.6). A reference ground truth was obtained via a subjective evaluation of stereo pairs on a stereoscopic monitor in a laboratory environment. Then, two 'mono' and 'sweep' 2D representations were generated for each bit rate and evaluated in a crowdsourcing environment. To evaluate the suitability of crowd-based quality assessment of MVD coding, the results of the crowd-based evaluations were compared to the ground truth results of the lab-based evaluations. This section reports a detailed analysis of the results to determine if 3D content can be assessed in crowdsourcing experiments.

5.2.1 Dataset

Seven MVD sequences were used in the experiments, with different visual characteristics, resolutions, and frame rates (see Table 5.10). All sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bits per sample. The sequences were compressed with 3D-AVC using 3D-ATM v9.0 (JCT3V-G1003) under the conditions defined in (JCT3V-F1011). For each sequence, 5 stimuli were generated, 1 from the original data, and 4 from the decoded data, resulting in a total of 28 test stimuli. Five training samples were generated from the *Poznan CarPark* sequence, which was not used in the tests. Their quality was manually selected by expert viewers so that they represent all grades of the rating scale.

5.2.2 Methodology

To be consistent with the protocol proposed in Section 5.1, the SS method with a five-grade quality scale (see Section 2.4.1) was chosen. The subjects were asked to judge the overall quality of the evaluated video sequence.

Four dummy video sequences (one with high quality, one with low quality, and two of mid quality), whose scores were not included in the results, were included at the beginning of the test session to stabilize the subjects' ratings. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each subject, whereas the same content was never shown consecutively.

Table 5.10: Multiview video plus depth contents used in the experiments.

Sequence	Characteristics			Encoding			Lab			Mono			Sweep		
	Resolution	Frames	FPS	QP settings R4, R3, R2, R1	Depth resolution	Input views	Views	Frames	FPS	View	Frames	FPS	Views	Frame	FPS
Poznan Hall2	1920 × 1088	0-199	25	26, 31, 36, 41	Quarter	7-6-5	6.25-5.75	0-199	25	5.75	0-199	25	7:-0.02:5	199	15
Poznan Street	1920 × 1088	0-249	25	26, 31, 36, 41	Quarter	5-4-3	4.25-3.75	0-249	25	3.75	0-249	25	5:-0.02:3	249	15
Undo Dancer	1920 × 1088	0-249	25	31, 38, 41, 46	Full	1-5-9	4-6	0-249	25	6	0-249	25	1:0.08:9	249	15
GT Fly	1920 × 1088	0-249	25	26, 31, 36, 41	Quarter	9-5-1	6-4	0-249	25	4	0-249	25	9:-0.08:1	156	15
Kendo	1024 × 768	0-299	30	26, 31, 36, 41	Quarter	1-3-5	2.5-3.5	0-299	30	3.5	0-299	30	1:0.04:5	0	15
Balloons	1024 × 768	0-299	30	26, 31, 36, 41	Quarter	1-3-5	2.5-3.5	0-299	30	3.5	0-299	30	1:0.04:5	0	15
Newspaper	1024 × 768	0-299	30	26, 31, 36, 41	Quarter	2-4-6	3.5-4.5	0-299	30	4.5	0-299	30	2:0.04:6	0	15

Lab-based Evaluation

The stereo pairs were synthesized from the decoded data using VSRS-1D-Fast v8.0 (JCT3V-G1005), according to the parameters given in Table 5.10. The stereo pairs were displayed on a full HD 46" Hyundai S465D polarized stereoscopic monitor. The monitor was calibrated using an X-Rite i1Display Pro color calibration device according to the following profile: sRGB gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level. The test room was equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of maximum screen luminance.

The experiment involved up to three subjects assessing the test materials. Subjects were seated in a row perpendicular to the center of the monitor, at a distance of 3.2 times the picture height (see Table 2.1). A total of 22 naïve subjects took part in the experiment. All subjects were screened for correct visual acuity, color vision, and stereo vision using Snellen chart, Ishihara chart, and Randot test, respectively.

Before the experiment, oral instructions were provided to the subjects to explain their tasks. Additionally, a training session using the five training samples was organized to allow subjects to familiarize with the assessment procedure.

Crowd-based Evaluation

Since no video player is capable of decoding 3D-AVC bit streams and synthesizing virtual views in real time, the video sequences were generated offline. The video sequences were synthesized from the decoded data using VSRS-1D-Fast v8.0 (JCT3V-G1005), according to the parameters given in Table 5.10. For the 'mono' representation, the right view of the stereo pair was used. For the 'sweep' representation, the FVV sequences were generated from a stack of 100 frames (at 15 fps), which was built from 50 intermediate views in-between the left and right views of the original content. One key frame, which maximizes the amount of depth, was selected as the freeze point for each content.

The sequences were encoded with AVC High Profile, since transmitting uncompressed video data to remote workers is impractical, especially for full HD content. Original full HD sequences of 25 fps were compressed at 20 Mbit/s, which is commonly considered as perceptually transparent quality for video broadcasting. For other sequences, the bit rate was set proportionally. A two-pass encoding was used and the deblocking filter was disabled to preserve the original blockiness due to 3D-AVC at low bit rates. Expert viewers evaluated the quality of the compressed sequences as visually lossless. The full HD sequences were cropped to 1856 × 1016 pixels such that workers were able to see the whole video in the web browser on a WUXGA (1920 × 1200) monitor. To display the video sequences and collect individual scores, the QualityCrowd 2 framework (Keimel et al., 2012) was used.

The experiments were conducted at EPFL in an uncontrolled computer lab, as it is relatively difficult to find workers equipped with a full HD monitor and because of the relatively large

amount of transmitted video data (up to 670 MB). Therefore, the workers' demographic was limited to EPFL students. Also, no financial compensation was provided to the workers. Each worker evaluated all test stimuli. The same set of workers took part in the 'mono' (20 subjects) and 'sweep' (21 subjects) experiments. However, half of the subjects took part in the 'mono' experiment first, while the other half started with the 'sweep' experiment. To minimized memory effects, subjects took a break between the two experiments.

Before the experiments, short written instructions were provided to the workers to explain their tasks. Additionally, three training samples, representative of *Excellent*, *Fair*, and *Bad* quality, were displayed to familiarize workers with the assessment procedure.

Data Processing

The subjective scores were processed by first detecting and removing subjects whose scores deviated strongly from others (for each experiment independently). The boxplot inspired outlier detection technique proposed by De Simone et al. (2011) (see Section 2.6.1) was used. In this study, no outlier subjects were detected. Then, the MOSs were computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% CIs, assuming a Student's t-distribution of the scores.

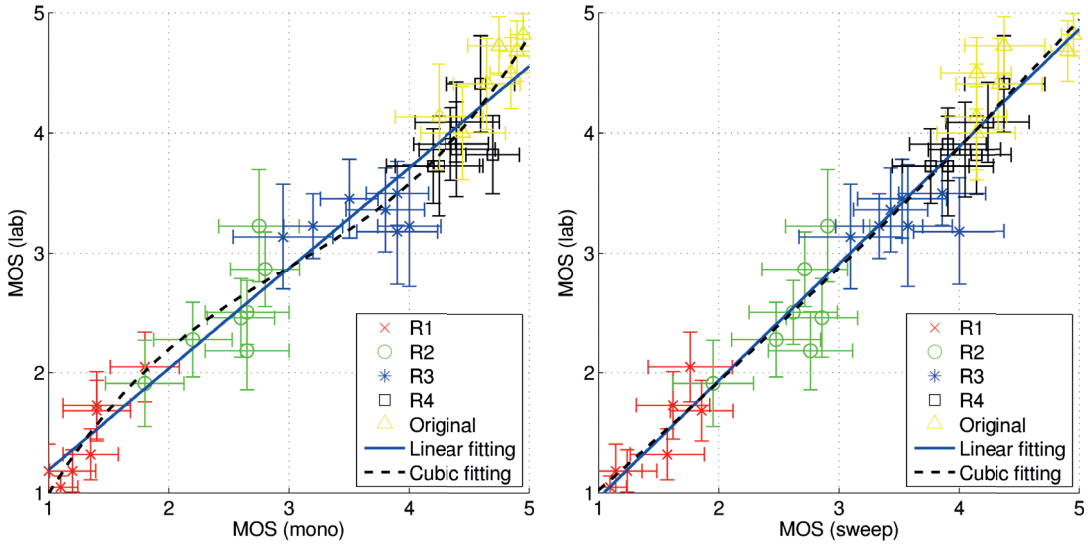
The results from the different evaluations were compared following the procedure described in Section 2.7. First, a regression was fitted to each $[MOS^{crowd}, MOS^{lab}]$ data set, using linear and cubic fitting, with the constraint that the function is monotonic on the interval of observed values. Then, the statistical evaluation metrics were computed between the two groups of MOS values. Finally, the estimation errors and classification errors were computed using a two-tailed t-test at a 95% confidence level to determine whether two distributions of subjective scores are different or not.

5.2.3 Results

Figures 5.9a and 5.9b show the results obtained for 'mono' and 'sweep' experiments respectively, with x-axis corresponding to the crowdsourcing data and y-axis to the data from the lab experiment. The horizontal and vertical error bars are the CIs of the respective experiments. To illustrate the trends of the data points, linear and cubic regressions fitted to each data set are also shown.

Ideally, all points would be on a 45° line if the MOS values for each condition were the same between two experiments. While, in the figures, some points lie above the line and others lie below, no significant systematic offset can be observed among MOS values of the compared experiments. It means that, overall, crowdsourcing workers scored closely to the lab experiment.

In case of 'mono' vs. lab, the slope of the linear regression is a little smaller than 45° (see



(a) Lab versus mono.

(b) Lab versus sweep.

Figure 5.9: Comparison of MOS values obtained in the different experiments.

Figure 5.9(a)), indicating that workers scored more pessimistically on lower quality stimuli, which is probably due to using AVC for encoding of the test stimuli instead of showing the original uncompressed data. More optimistic scores for high quality stimuli might be due to the uncalibrated monitors and uncontrolled lighting conditions. In Figure 5.9(b), the linear regression is very similar to a 45° line. The cubic regression for 'sweep' vs. lab is very close to a straight line, which means the relationship between 'sweep' and lab is more linear than between 'mono' and lab.

Table 5.11 reports the statistical evaluation metrics. Results show that there is a very strong correlation between crowd-based and lab-based evaluations, as the correlation coefficients are above 0.97, which is similar to the correlation between different laboratories conducting the same experiment on stereoscopic monitors (Barkowsky et al., 2013). The PCC, RMSE, and OR values are slightly better for 'sweep' than 'mono' when no fitting or linear fitting are considered. However, there are no statistically significant differences between the statistical evaluation metrics computed for 'mono' and 'sweep'.

Regardless of the fitting applied to the data sets, both crowd-based evaluations were able to correctly estimate the results of the lab-based evaluation with a *Correct estimation* of 100%, whereas the *Underestimation* and *Overestimation* were always null.

Table 5.12 reports the classifications errors. More than 84% of all possible distinct pairs of decoded 3D data lead to the same conclusion in crowd-based evaluations when compared to the lab-based evaluation. Moreover, *False Ranking* never occurs. Results for *False differentiation* show a slight advantage for 'sweep', but differences are not significant.

Displaying one view of the 3D content allows judging spatial and temporal impairments,

Table 5.11: Statistical evaluation metrics.

Fitting	Mono				Sweep			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
None	0.9750	0.9753	0.3697	2.86%	0.9761	0.9717	0.2629	0%
Linear	0.9750	0.9753	0.2495	2.86%	0.9761	0.9717	0.2440	0%
Cubic	0.9798	0.9753	0.2243	0%	0.9764	0.9717	0.2422	0%

Table 5.12: Classification errors.

Fitting	Mono				Sweep			
	Correct decision	*False ranking*	*False differentiation*	*False tie*	*Correct decision*	*False ranking*	*False differentiation*	*False tie*
None	86.05%	0.00%	6.89%	7.06%	87.90%	0.00%	4.54%	7.56%
Linear	84.54%	0.00%	5.71%	9.75%	87.90%	0.00%	4.54%	7.56%
Cubic	87.06%	0.00%	5.88%	7.06%	88.24%	0.00%	4.37%	7.39%

whereas depth impairments are difficult to evaluate. Nevertheless, some depth impairments may be visible when considering a virtual view that is synthesized from video and depth data. The FFV sequence is better to judge depth impairments, but temporal impairments cannot be evaluated. However, the selection of the key frame may impact the perceived quality, as the strength of the impairments typically varies in time.

In our experiments, 2D impairments were mostly visible in the test material, even though depth maps were also compressed, and the strength of the spatial impairments was similar across time. Therefore, it is reasonable to have high correlation with ground truth results in 'mono' and 'sweep'. However, if the test material mostly contains depth impairments, the 'sweep' methodology is expected to be more suitable.

5.3 Crowdsourcing Evaluation of HDR Image Compression

One simple approach for representing HDR content on an LDR display is to tone map the HDR image to reduce its dynamic range. As the mapping from a large set of values to a smaller set of values is not unique, the tone mapped image might look quite different depending on the TMO and its parameters. For example, the tone mapped image can have more emphasis on the darker or brighter areas, which might change the visibility of artifacts when compared to the HDR image.

This section investigates whether crowdsourcing approach combined with preprocessing by a TMO is suitable for evaluation of compressed HDR images. For that purpose, five HDR images were encoded with JPEG XT Profile A (see Section 4.4) at four different bit rates and evaluated using PC method (see Section 2.4.3), which was selected for its high accuracy and reliability in constructing a scale of perceptual preferences. Eleven TMOs were used to convert the compressed HDR images to the corresponding tone-mapped LDR versions. In addition, JPEG LDR versions were also used in the evaluations, which were manually produced from original

Table 5.13: HDR images information.

Image	Resolution	Dynamic range (dB)	Encoding parameters (q, Q)
BloomingGorse2	4288 × 2848	42	(11,12), (20,13), (32,15), (62,15)
CanadianFalls	4312 × 2868	41	(16,29), (30,30), (65,30), (80,33)
McKeesPub	4258 × 2829	60	(5,64), (15,91), (48,88), (83,91)
MtRushmore2	4312 × 2868	50	(5,20), (24,82), (67,80), (89,78)
WillyDesk	4288 × 2848	70	(5,63), (15,79), (57,90), (85,91)

HDR images by JPEG ad hoc group. A reference ground truth was obtained via a subjective evaluation of the compressed HDR images and their manually produced JPEG LDR versions on a Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar') in a laboratory environment. To evaluate the suitability of crowd-based quality assessment of HDR coding, the results of the crowd-based evaluations were compared to the ground truth results of the lab-based evaluations. An additional subjective evaluation was conducted using semantic differentiators to better understand the characteristics of the different TMOs. This section reports a detailed analysis of the results, which shows the importance of the TMO selection when evaluating HDR content on SDR displays.

5.3.1 Dataset

Five HDR images from the Fairchild dataset, with different dynamic ranges and representing different scenes, were used in the experiments (see Figure 5.10 and Table 5.13 for details). Originally, these images were selected by JPEG Committee for the verification tests of JPEG XT standard. JPEG Committee also provided LDR versions of these images that were manually tone-mapped using Adobe Photoshop from the original HDR images.

To prepare images for subjective experiments, both HDR and LDR versions were first downscaled by a factor of two with bicubic interpolation. The resulted images were then compressed using JPEG XT Profile A to four different bit rate values, ranging from a minimum of 0.3 bpp to a maximum of 2.2 bpp for different images. The bit rate values were selected for each content separately (see Table 5.13 for details) in such a way that there is a noticeable visual difference between images with different bit rates when they are displayed on an HDR monitor.

Compressed images were then cropped to 950 × 1080 pixels regions for side-by-side lab- and crowd-based subjective experiments (see Section 5.3.2 for details). The regions to crop were selected by expert viewers in such a way that cropped versions are representative of the quality and the dynamic range of the original images. Red rectangles in Figure 5.10 show the corresponding cropped regions. Downscaling together with cropping approach was selected as a compromise, so that a meaningful part of an image can be shown on the HDR monitor.

Eleven TMOs were used in the crowdsourcing subjective evaluation (see Table 5.14). These algorithms were selected based on their popularity in the research community and their visual

(a) *BloomingGorse2* (b) *CanadianFalls* (c) *McKeesPub*

(d) *MtRushmore2* (e) *WillyDesk*

Figure 5.10: HDR images used in the experiments.

Table 5.14: TMOs used in the subjective evaluation.

TMO	Description	Type
Ashikhmin (2002)	A tone mapping algorithm for high contrast images	Local
Chiu et al. (1993)	Spatially nonuniform scaling functions for high contrast images	Local
Drago et al. (2003)	Adaptive logarithmic mapping for displaying high contrast scenes	Global
Durand and Dorsey (2002)	Fast bilateral filtering for the display of HDR images	Local
Fattal et al. (2002)	Gradient domain high dynamic range compression	Local
Lischinski et al. (2006)	Photographic tone reproduction for digital images	Local
Linear	Simple linear scaling	Global
Logarithmic	Simple logarithmic scaling	Global
Reinhard et al. (2002)	Photographic tone reproduction for digital images	Local
Tumblin and Rushmeier (1993)	Two methods for display of high contrast images	Global
WardHistAdj (Ward, 1994)	A contrast-based scale factor for luminance display	Global

characteristics. The HDR Toolbox was used to apply the TMOs, as this toolbox implements many different operators. The TMOs were applied as follows. Each HDR original image was first compressed to four bit rates with JPEG XT producing four compressed HDR versions for each content. Each compressed HDR image was decoded and eleven TMOs were applied, producing eleven LDR images. These LDR images were compressed with high quality JPEG to be used in crowdsourcing evaluations.

5.3.2 Methodology

The PC method with a ternary scale (see Section 2.4.3) was selected to evaluate the quality of the different images, as this methodology provides a high accuracy and reliability in constructing a scale of perceptual preferences. Moreover, this methodology is very natural for test

subjects to understand and use, which makes it perfectly suitable for crowdsourcing evaluations. The image pairs were presented in side-by-side fashion to minimize visual working memory limitations.

In the lab-based evaluation (see below), one full HD (1920 × 1080 pixels) HDR monitor was used to display the images and a 20 pixels black border separated the pair of images. For the crowd-based evaluation, the images were downscaled by a factor two, using bicubic interpolation, so that workers were able to see the image pair in a web browser set for XGA (1024 × 768 pixels) monitor, which is one of the most common display resolutions.

Lab-based Evaluation

Lab-based experiments were conducted at the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU. The test room is equipped with a controlled lighting system with a 6500 K color temperature, while the color of all the background walls and curtains present in the test area are mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors.

To display the test stimuli, a full HD (1920×1080 pixels) 42" Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar') was used. The monitor has the following specifications: full DCI P3 color gamut, 4000 cd/m^2 peak luminance, low black level (0.005 cd/m^2), 12 bits/color input with accurate and reliable reproduction of color and luminance. In the experiments, the luminance of the background behind the monitor was about 20 cd/m^2. The ambient illumination did not directly reflect off of the display.

In every session, three subjects were assessing the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of about 3 times the picture height (see Table 2.1).

Before the experiment, a consent form was handed to subjects for signature and oral instructions were provided to explain their tasks. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively. A training session was organized using additional contents to allow subjects to familiarize with the assessment procedure.

For each of the five contents, all the possible combinations of the four bit rates were considered. The HDR image decoded with JPEG XT, as well as the base layer decoded with JPEG, were evaluated, leading to a total of $5 \times 2 \times \binom{4}{2} = 60$ comparisons. To reduce contextual effects, the stimuli orders of display were randomized applying different permutations for each group of subjects and special care was taken for the same content not to be shown consecutively.

A total of 18 naïve subjects (11 females and 7 males) took part in the evaluation. They were between 20 and 34 years old with an average of 25.3 years of age.

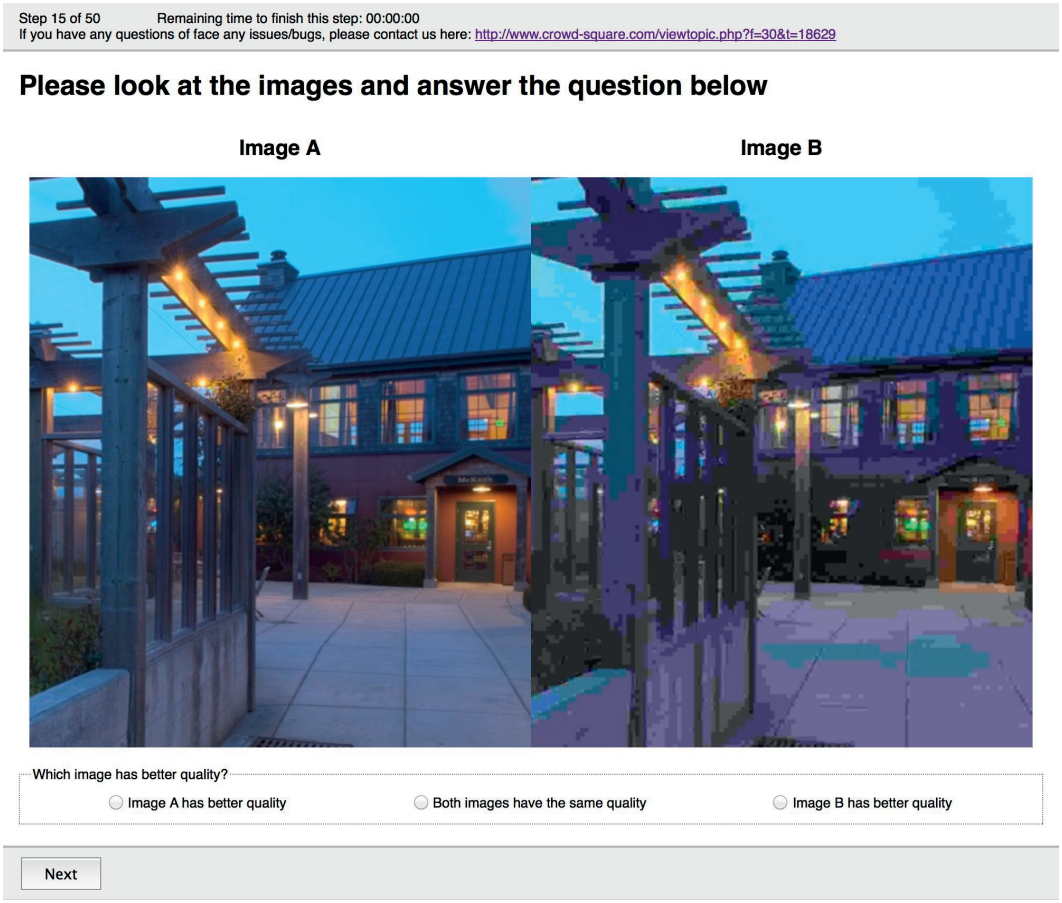


Figure 5.11: Evaluation of HDR image compression: screenshot of the crowdsourcing interface.

Crowd-based Evaluation

The tone mapped images were compressed with JPEG at quality 95, which produces images that have visually lossless quality and file sizes still suitable for transmission to remote crowdsourcing workers. Figure 5.11 shows a screenshot of the crowdsourcing interface for JPEG XT compression experiment and Figure 5.12 shows a screenshot of the TMO characteristics evaluation. In both crowdsourcing experiments, a slightly modified version of the QualityCrowd 2 framework (Keimel et al., 2012) was used.

Evaluation of JPEG XT Compression Before the experiments, short written instructions were provided to the workers to explain their tasks. Additionally, three training samples, with a different content, were displayed to familiarize workers with the assessment procedure. The training instructions and samples were presented using QualityCrowd 2.

For each of the 5 contents, all the possible combinations of the 4 bit rates were considered. In total, for each of 5 contents, 11 versions corresponding to 11 TMOs (see Table 5.14) and the base layer decoded with JPEG were evaluated, leading to $5 \times 12 \times \binom{4}{2} = 360$ comparisons. As

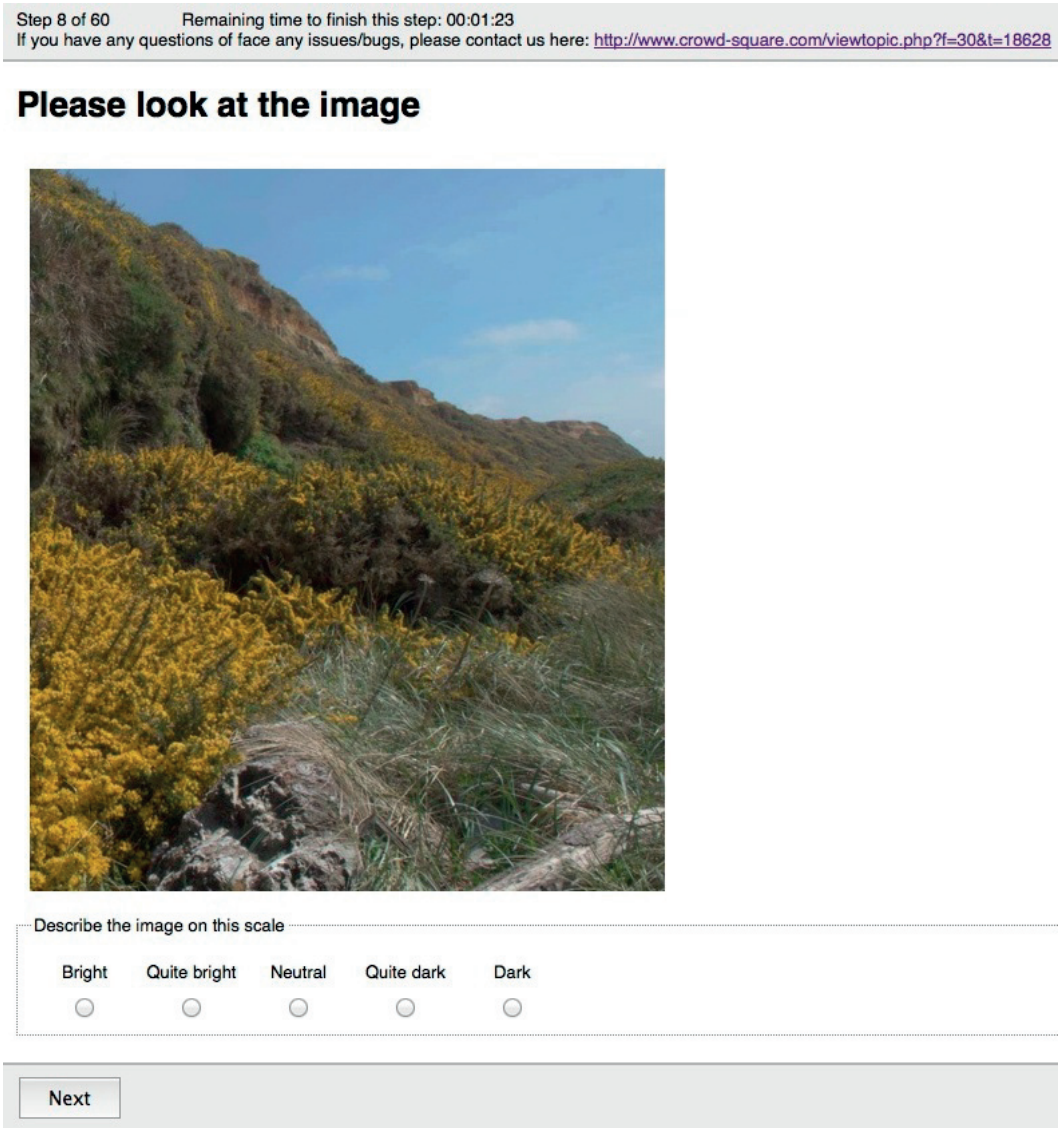


Figure 5.12: Evaluation of TMO characteristics: screenshot of the crowdsourcing interface.

the number of pairs to be evaluated is very high, it is impossible for one worker to evaluate all pairs. Therefore, the pairs were randomly split into 9 batches of 40 pairs each. To reduce contextual effects, the stimuli orders of display were randomized and special care was taken for the same content not to be shown consecutively. Each worker was allowed to take only one batch. Therefore, the evaluation required a relatively large amount of workers. Subjects were recruited through the Microworkers crowdsourcing platform. Only countries where English is a dominant language were chosen, with either more than 50% of population or more than 10 million of people speaking English, according to Wikipedia. Workers received a compensation of $0.50 for completing the 40 evaluation tasks. To detect unreliable workers, two honeypots were inserted around the 1st and 3rd quarter of the evaluation task. The honeypots were related to the content of the image pair seen on the previous evaluation task.

Evaluation of TMO Characteristics To evaluate the characteristics of the different TMOs, the semantic differential method was used considering the following bipolar adjective pairs

1) *Dark - Bright*
2) *Cold - Hot*
3) *Smooth - Sharp*
4) *Dirty - Clean*
5) *Dusty - Vivid*
6) *Static - Dynamic*
7) *Bad - Good*
8) *Unrealistic - Realistic*
9) *Unpleasant - Pleasant*
10) *Ugly - Beautiful*

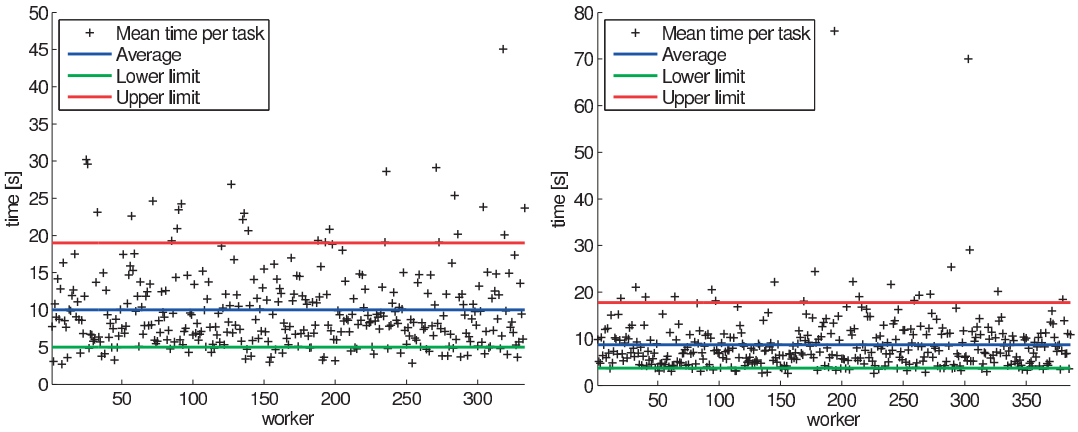
These adjective pairs were chosen based on the results from previous studies using semantic differential on image characteristics (Iwanami et al., 2009; Kimoto and Kato, 2014; Seetzen et al., 2003) and the relevance of adjective pairs in the context of tone mapped images.

The crowdsourcing experiment to evaluate TMO characteristics was conducted in a similar way to JPEG XT compression experiment, except that instead of pairs of images, a single image per page was displayed to a worker. This resulted in $5 \times 12 = 60$ stimuli (five contents, eleven TMOs, and the base layer). The stimuli were randomly split into 12 batches of 5 stimuli each, one for each content, and the 10 bipolar adjective pairs were considered for each stimuli. Workers received a compensation of $0.50 for completing the 50 evaluation tasks.

5.3.3 Results

Unlike lab-based subjective experiment where all subjects can be observed by operators and its test environment also can be controlled, the major shortcoming of the crowdsourcing-based subjective evaluation is the inability to supervise participants behavior and to restrict their test conditions. When using crowdsourcing for evaluation, there is a risk of including untrusted data into the analysis due to wrong test conditions or unreliable behavior of some workers who try to submit low quality work to reduce their effort while maximizing their compensation (Hossfeld et al., 2014a). For this reason, unreliable workers detection is an inevitable process in crowdsourcing evaluations. To identify a worker as 'trustworthy', the following factors were considered in our experiments

i) Mean observation time per question,
ii) Honeypots, i.e., trap questions related to the content of the image seen in the previous question (Hossfeld et al., 2014a), and
iii) Post-screening according to the guidelines described in Section 2.3.1 of Annex 2 of ITU-R BT.500-13 (2012) (see Section 2.6.1) (only for the evaluation of TMO characteristics).



(a) HDR image compression.

(b) TMO evaluation.

Figure 5.13: Mean time for one question for each worker.

The observation time per question is measured as the time from when the question is displayed until the time the answer is given by the worker. The mean observation time can be calculated using this data. If the mean observation time per question is too short or too long when compared to the average of all workers, it can be deduced that the worker did not take the test seriously or was distracted during the task. Figure 5.13 depicts the mean response time per question for each worker for both evaluations. The mean observation time is between 9 and 10 s. As it can be observed, some workers demonstrate large values when compared to the corresponding mean, especially for the TMO evaluation campaign. To filter out unreliable workers, we set a lower and upper limit at −5 and +9 s from the mean respectively.

Table 5.15 reports the number of valid workers per batch for both evaluation campaigns. The results are reported for the individual outlier detection techniques, as well as for the combination of the different techniques. As it can be observed, the results are quite different from one batch to another. For example, 50 workers were necessary, considering all outlier detection techniques, to obtain 16 valid sets of scores for 2nd batch of the TMO evaluation campaign, whereas 22 workers out of 33 produced reliable results in the 3rd batch. In general, it can be observed that one to two third of the workers were considered unreliable according to the specified criteria.

For the JPEG XT image compression campaign, we targeted 18 reliable workers per batch, to reach the same number of subjects as in the lab-based evaluation (see Section 5.3.2). For the TMO characteristics evaluation campaign, we targeted 16 reliable workers per batch, which is the number of subjects commonly considered in subjective evaluations. In both cases, the most restrictive outlier detection was considered, since it includes mean time per question, two honeypots, and recommendation ITU-R BT.500-13 (2012). Originally, we ran 30 workers per batch. However, additional slots were opened to reach the required minimum number of valid workers. Therefore, in total, 375 and 434 workers took part in the JPEG XT image compression and TMO characteristic evaluation campaigns respectively. In the final

Table 5.15: Number of reliable workers per batch depending on the considered outlier detection technique.

(a) HDR image compression.

Outlier detection	Batch								
	1	2	3	4	5	6	7	8	9
None	42	48	41	43	43	33	44	41	40
1st honeypot	29	32	32	27	27	24	23	29	27
2nd honeypot	28	31	25	34	29	25	30	28	28
Time	29	40	32	32	33	26	35	29	33
1st honeypot + time	26	29	28	22	24	21	22	25	24
2nd honeypot + time	23	28	21	27	29	22	28	22	25
2 honeypots	25	25	24	26	21	23	22	24	22
2 honeypots + time	22	22	20	21	21	20	21	20	20

(b) TMO evaluation.

Outlier detection	Batch											
	1	2	3	4	5	6	7	8	9	10	11	12
None	35	50	33	35	34	31	43	33	32	32	37	39
1st honeypot	28	23	27	25	24	26	21	23	27	24	28	22
2nd honeypot	26	34	28	23	25	21	34	27	22	23	24	30
Rec. ITU-R BT.500	30	44	29	29	29	28	34	29	29	29	29	37
Time	28	36	23	27	24	25	28	25	23	23	26	32
1st honeypot + Rec. ITU-R BT.500	28	23	27	25	24	25	21	23	27	24	28	22
2nd honeypot + Rec. ITU-R BT.500	26	34	28	23	25	21	34	27	22	23	24	30
1st honeypot + time	26	17	22	23	21	22	16	21	23	20	25	19
2nd honeypot + time	24	28	22	21	21	17	28	23	18	17	21	26
1st honeypot + Rec. ITU-R BT.500 + time	26	17	22	23	21	21	16	21	23	20	25	19
2nd honeypot + Rec. ITU-R BT.500 + time	24	28	22	21	21	17	28	23	18	17	21	26
2 honeypots	24	21	27	22	21	21	21	21	21	20	23	19
2 honeypots + Rec. ITU-R BT.500	24	21	27	22	21	21	21	21	21	20	23	19
2 honeypots + time	22	16	22	20	18	17	16	19	18	16	20	17
2 honeypots + Rec. ITU-R BT.500 + time	22	16	22	20	18	17	16	19	18	16	20	17

results, only the first 18 and 16 workers were considered to be valid for the JPEG XT and TMO campaigns, respectively.

Evaluation of TMO Characteristics

Figure 5.14 shows the semantic differential profiles of the different TMOs considered in the experiment (see Table 5.14). Different patterns can be observed from this figure. The *Dark-Bright* pair seems to have the most diversity of results, whereas the other pairs have somehow similar results across the different TMOs, except for *Chiu* TMO. This TMO was usually rated lower than the other TMOs. In particular, results show that *Chiu* produces less pleasant and realistic images, that are not as good, clean, and beautiful than the other TMOs. Therefore, it can be concluded that this TMO does not produce good tone mapped images, which may affect

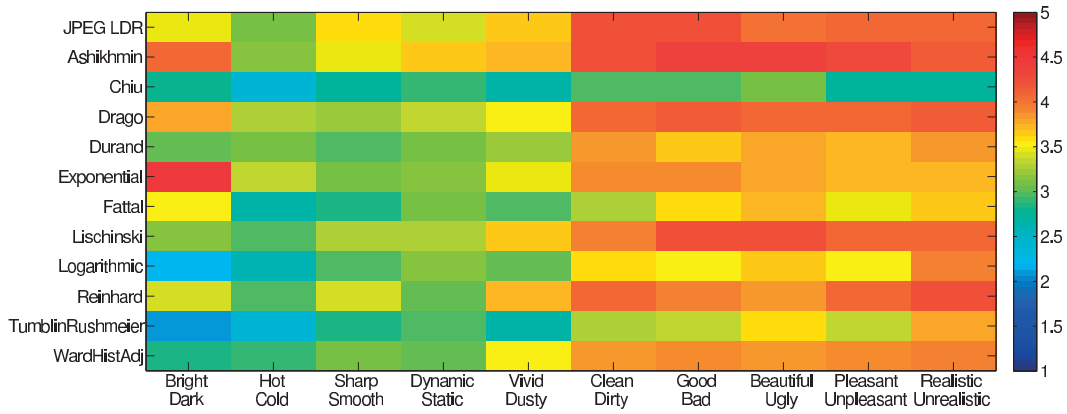


Figure 5.14: Semantic differentiator.

the evaluation of HDR compression later. On the other hand, *Ashikimin*, *Drago*, *Exponential*, *Lischinski*, and *Reinhard* seem to produce overall satisfactory tone mapped images, with *Fattal* and *WardHistAdj* being the next best subset of TMOs, since the corresponding characteristics tend to be towards the top of the scale.

The pairs *Dirty-Clean*, *Bad-Good*, *Ugly-Beautiful*, *Unpleasant-Pleasant*, and *Unrealistic-Realistic* seem to be quite correlated and most scores are above 3.5 (except for *Chiu*). The pairs *Smooth-Sharp*, *Static-Dynamic*, and *Dusty-Vivid* are also quite correlated, but with lower scores (around 3). The pair *Cold-Hot* appears to correlate more with the pair *Dark-Bright* than the other pairs, but the values are lower and less spread. Intuitively, *Cold-Hot* can be related to *Dark-Bright*, but in the context of images, the second one is more obvious and easier to understand. This probably explains why the scores are more spread over the scale for *Dark-Bright*, whereas most values for *Cold-Hot* are around the neutral value (3).

Figure 5.15 depicts the results of the PCA applied on the semantic differential profiles. The first component seems to be mostly related to the quality of the tone mapped image, although the contribution of the different related pairs is lower than 0.4 in all cases. The second dimension is mostly related to the brightness of the tone mapped image, as the contribution of the pairs *Dark-Bright* and *Cold-Hot* for this dimension is above 0.6. However, it appears that the brightness is also related to image quality. Table 5.16 reports the contribution of the different bipolar adjective pairs to the two principal components. The principal components explain 72.43% of the variance observed in the data. However, most of the contribution comes from the first component, which explains 62.03% of the variance.

Evaluation of HDR Image Compression

To determine whether the PC results from the crowdsourcing evaluation are significantly different from the results of the ground truth lab-based evaluation, the Barnard's test (see Section 2.6.4) was used. This test was used to compare the results of a pair of images evaluated

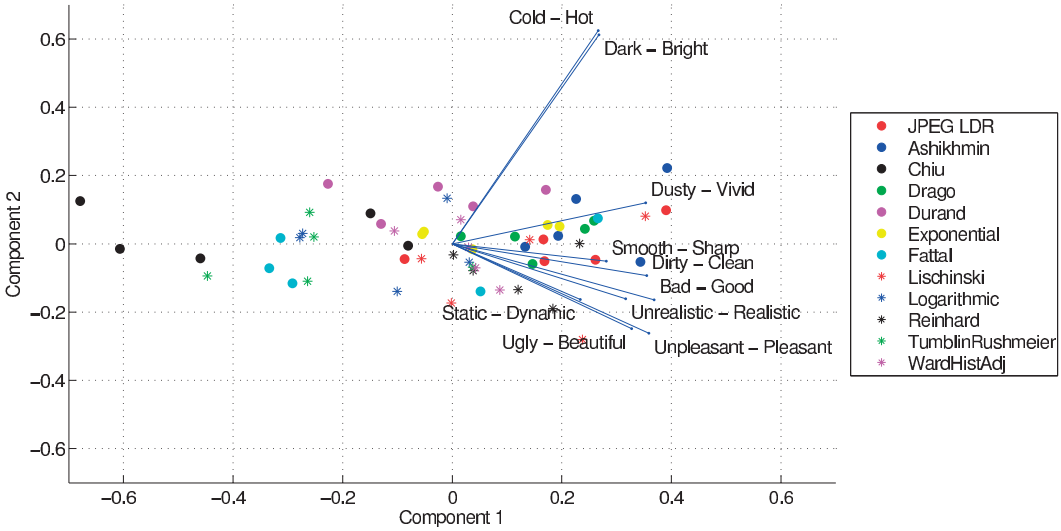


Figure 5.15: PCA of the semantic differential profiles.

Table 5.16: Two factor analysis.

Adjective pairs	First factor	Second factor
Bad - Good	0.3689	−0.1642
Unpleasant - Pleasant	0.3591	−0.2624
Dirty - Clean	0.3550	−0.0934
Dusty - Vivid	0.3532	0.1200
Ugly - Beautiful	0.3275	−0.2487
Unrealistic - Realistic	0.3167	−0.1608
Smooth - Sharp	0.2813	−0.0510
Dark - Bright	0.2682	0.6117
Cold - Hot	0.2663	0.6245
Static - Dynamic	0.2339	−0.1628
Contribution ratio	62.03%	10.40%
Cumulative contribution ratio	62.03%	72.43%

in the crowdsourcing evaluation to the ground truth results corresponding to the same pair. The test was repeated for all possible pairs, i.e., six pairs per content and per TMO.

Table 5.17 reports the number of significantly different pairs between crowd-based and lab-based evaluation of JPEG LDR. In this case, no tone mapping was applied to produce the images for the crowdsourcing evaluation and the same images were shown in both environments. Therefore, these results can be used to determine whether crowdsourcing produces reliable results when compared to the ground truth lab results. As it can be observed, the number of significantly different pairs is 0.6 on average for all contents or 3 out of 30 pairs, which means that 10% of the pairs obtained different results. The difference, though not very significant, is probably due to the influence of the different viewing conditions, which cannot be controlled in crowdsourcing evaluations, or the unreliability of workers, even though we

Table 5.17: Number of significantly different pairs between crowd-based and lab-based evaluation of JPEG LDR.

BloomingGorse2	*CanadianFalls*	*McKeesPub*	*MtRushmore2*	*WillyDesk*	Average
1	1	0	0	1	0.6

Table 5.18: Number of significantly different pairs between JPEG LDR and HDR.

Evluation	Content					Average
	BloomingGorse2	*CanadianFalls*	*McKeesPub*	*MtRushmore2*	*WillyDesk*	
Laboratory	1	0	0	0	0	0.2
Crowdsourcing	0	2	0	0	0	0.4

used several mechanisms to detect outliers and potential cheaters.

Table 5.18 reports the number of significantly different pairs between JPEG LDR and JPEG XT for both lab- and crowd-based evaluations. Results show that, in the laboratory, only one pair was significantly different. This indicates that, for the compression parameters considered in this experiment, the relative quality difference between two LDR base layers was similar to that of the corresponding HDR images reconstructed with an additional enhancement layer. In the crowdsourcing evaluation, only two pairs were significantly different, but the differences occurred for a different content. This might indicate that content characteristics might influence the results in crowd-based evaluations. However, we have too few pairs per content to determine whether content characteristics have a statistically significant influence.

Table 5.19 reports the number of significantly different pairs between the tone mapped images evaluated in the crowdsourcing experiment and the HDR images evaluated in the laboratory. As it can be observed, most of the differences occur for contents *BloomingGorse2* and *CanadianFalls*, which have a limited dynamic range (see Table 5.13). Since the TMOs were essentially designed to handle images with higher dynamic range, they may not produce good tone mapped images when the input image has a limited dynamic range, which might influence the perception of artifacts. For example, if noise is present in a dark area, its intensity might be reduced by the compression of the dynamic range performed by the TMO. However, if the dynamic range is rather limited, this compression will be less and the noise might be more visible.

Results show that *Chiu* resulted in 11 significantly different pairs out of 30 pairs, which means that this TMO is not suitable to assess the performance of HDR compression. As it was observed previously, this TMO produces worse tone mapped images, that were less pleasant and realistic, than the other TMOs. *Ashikhmin*, *Fattal*, *Lischinski*, and *WardHistAdj* are the only TMOs that resulted in less than 6 significantly different pairs. These TMOs could be considered to assess the performance of HDR compression on LDR monitors. However, as the number of contents, bit rates, and subjects considered in this study are rather limited, it is impossible to draw general conclusion and recommend one particular TMO.

Table 5.19: Number of significantly different pairs between TMO and HDR.

TMO	Content					Average
	BloomingGorse2	*CanadianFalls*	*McKeesPub*	*MtRushmore2*	*WillyDesk*	
Ashikhmin	3	2	0	0	0	1.0
Chiu	4	3	2	1	1	2.2
Drago	3	2	1	1	0	1.4
Durand	1	3	0	1	1	1.2
Exponential	3	3	0	1	0	1.4
Fattal	0	3	0	0	0	0.6
Lischinski	0	2	0	0	1	0.6
Logarithmic	2	1	1	1	4	1.8
Reinhard	2	2	0	2	1	1.4
TumblinRushmeier	2	3	1	2	2	2.0
WardHistAdj	1	0	0	1	1	0.6
Average	1.7500	2.1667	0.4167	0.8333	0.9167	1.2167

5.4 Conclusion

This chapter investigated alternative evaluation protocols for subjective quality assessment. We proposed an experimental protocol to evaluate the impact of depth compression on perceived image quality in a FTV scenario. A specific use case was considered to allow a reliable comprehension of the impact of depth coding: a smooth camera motion during a time freeze. The analyses of the resulting subjective scores revealed that the proposed experimental protocol allows the evaluation of different compression and view synthesis algorithms. The use of statistical tools (ANOVA and PCA) to analyze the subjective scores showed particular behaviors such as the influence of different view synthesis modes on the performance of a specific compression algorithm. These results showed the originality and the effectiveness of the proposed assessment protocol as well as the importance of subjective quality assessment. This methodology can be considered to evaluate the performances of various depth compression algorithms and can be extended to the assessment of classical MVD compression schemes.

Crowdsourcing is becoming a popular cost effective alternative to lab-based evaluations for subjective quality assessment. However, crowd-based evaluations are constrained by the limited availability of display devices used by typical online workers, which makes the evaluation of 3D or HDR content a challenging task. We investigated two possible approaches to crowd-based quality assessment of MVD content on 2D displays: by using a virtual view and by using a FVV, which corresponds to a smooth camera motion during a time freeze. We conducted the corresponding crowdsourcing experiments using seven MVD sequences encoded at different bit rates. The crowdsourcing results showed high correlation with ground truth results obtained in a subjective evaluation performed on a stereoscopic monitor in a laboratory environment. No statistically significant differences between the two approaches were found.

In this chapter, we also investigated the feasibility of using LDR versions of original HDR content obtained with TMOs in crowdsourcing evaluations. We conducted two crowdsourcing

experiments. In the first experiment, we evaluated five HDR images encoded at four bit rates. To find best suitable TMO, we created eleven tone-mapped versions of these five HDR images by using eleven different TMOs. The crowdsourcing results were compared to a reference ground truth obtained via a subjective assessment of the same HDR images on a HDR monitor in a laboratory environment. The second crowdsourcing evaluation used semantic differentiators to better understand the characteristics of the different TMOs. The crowdsourcing evaluations showed that some TMOs are more suitable for evaluation of HDR image compression.

6 Evaluation of Immersive Video Technologies

Over the last decades, several technological revolutions have impacted the television industry, such as the shifts from black & white to color and standard to high definition. Nevertheless, to provide a better experience and a better picture quality, considerable improvements may still be achieved along several axes, including resolution, frame rate, contrast, brightness, and color. Immersive video technologies, e.g., UHD, HFR, HDR & WCG, and 3D, improve these characteristics and aim at providing better, more realistic, and emotionally stronger experiences.

An important question however is how significantly these technologies impact the viewers' QoE? To investigate this problem, a variety of video sequences should be recorded in UHD HFR HDR WCG 3D format (all at the same time) and displayed on a UHD HFR HDR WCG 3D monitor in all possible combinations while evaluating viewers' QoE. Such investigation could be used to measure the added value of the different immersive video technologies, as well as the added value induced by their interactions. Unfortunately, even if the capture would be possible with a very high-end multi-cameras setup, there is currently no display that can render all these video technologies at the same time, especially if considering autostereoscopic 3D. Therefore, we can only investigate the added value of each technology independently. Unfortunately, this approach is not sufficient to model QoE, as the QoE is most likely be much more complex than a simple addition of the individual benefits offered by each technology.

This chapter investigates some of these immersive video technologies and their impact on viewers' QoE. In particular, Section 6.1 investigates the impact of 3D rendering on sensation of reality. Different video sequences were presented in 2D and 3D modes, with low and high quality levels, to investigate the impact of these factors on perceived overall quality, perceived depth quantity, content preference, and sensation of reality. Section 6.2 investigates the impact of higher dynamic range on viewers' preference. Different video sequences were presented at 100, 400, 1000, and 4000 nit peak luminance levels. Two subjective experiments with different test methods and designs were conducted to measure the impact of peak luminance on viewers' preference.

6.1 Sensation of Reality in 3DTV

An important part of our impressions and understanding about our surroundings are based on sight. Thus, our perception of the world is mainly three-dimensional. An efficient representation of real scenes should therefore provide a three-dimensional feeling to enhance sensation of reality through multimedia devices. The importance of sensation of reality has been recognized in the field of games and virtual reality (Stoakley et al., 1995; von der Pütten et al., 2012), through user-system interactions. Also, recent advances in imaging and displays have enabled implementation of more immersive multimedia environments, offering improved sensation of reality to users (Hayward et al., 2004; Kulkarni et al., 2012).

As a result, immersive multimedia, which allows users to experience enhanced immersion and involvement in comparison to traditional multimedia, is receiving a rapidly increasing amount of attention. It has strong impact on users' emotion, sense of presence, and degree of engagement, which can eventually be used to provide users more satisfactory media experience (Sanchez-Vives and Slater, 2005; Slater, 2009; Slater and Wilbur, 1997).

For instance, 3D image and video technologies are gaining ground in multimedia applications since they incorporate depth perception, leading to more realistic scenes, and consequently to emotionally stronger experiences. However, in order for the experience to be as realistic as possible, the quality of the rendering should be as good as possible. Thus, it is important to understand the QoE perceived by users from various multimedia rendering schemes to design and optimize human-centric immersive multimedia systems.

This chapter attempts to investigate immersive video presentation experience via explicit subjective rating analysis for 2D and 3D multimedia contents. Various QoE-related aspects are investigated and compared. In particular, depth perception, sensation of reality, content preference, and perceived quality are investigated with respect to how they influence each other. This section reports a detailed analysis of the results to investigate how QoE is perceived.

6.1.1 Dataset

At the time of this study, the availability of high quality stereoscopic content of sufficient duration to induce immersiveness was almost inexistent. In our experiments, we used video clips recorded during the Montreux Jazz music festival (MJF) by NVP3D, a professional 3D video production company. The dataset was composed of eight video contents: one for the training and seven for the tests. All contents were recorded with two RED SCARLET-X cameras mounted on a Genus Hurricane Rig. All video sequences were recorded in REDCODE RAW (R3D) format, DCI 4K resolution (4096 × 2160 pixels), at 25 fps, and had a duration of about one minute long. Stereo audio was recorded in PCM format, sampled at 48 kHz, 24 bits. Table 6.1 describes the contents and their characteristics. The recorded video sequences were cropped and downsampled to Full HD resolution (1920 × 1080 pixels) and then compressed with AVC. Two different QPs were selected: QP=2 for high quality (HQ) and QP=35 for low quality (LQ).

Table 6.1: Characteristics of the contents used in our experiments.

Content	Description and characteristics
Training	Rock band playing at the Auditorium Stravinski. Dark. Bright spots. Shot from the back of the auditorium.
Jazz	Jazz band playing at the Funky Claude's Lounge at the Opening Party. Wide shot.
Rock	Rock band playing at the Auditorium Stravinski. Dark. Bright spots. Shot from the back of the auditorium.
Stage	MJF general manager on stage introducing the next artist. Very dark. In French. Wide shot.
Speech1	MJF general manager giving a speech at the Opening Party. In French. Mid shot.
Speech2	Speech at the Opening Party. In French. Mid shot.
Outdoor	Crowd walking on the street near the lake. Lot of depth. Wide shot.
Interview	Interview of Quincy Jones. Medium close up.

For each content, four different versions were considered: 2D HQ, 3D HQ, 2D LQ, and 3D LQ, leading to a total of 28 video sequences, 14 of which in 2D and 14 in 3D.

6.1.2 Methodology

The experiments were conducted at the MMSPG quality test laboratory, which fulfills the recommendations for the subjective evaluation of visual data issued by ITU. The laboratory setup was controlled to ensure the reproducibility of results by avoiding involuntary influence of external factors. The test room was equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of the maximum screen luminance.

To display the video stimuli, a HD 46" Hyundai S465D stereoscopic monitor with passive 3D glasses were used. The monitor has a 60 Hz refresh rate and relies on a line-interleaved display and circular polarizing filters to separate the left- and right-eye images. For the audio playback, the PSI A14-M professional studio full range speakers were used.

The experiment involved only one subject at a time assessing the test material. The participants were seated at a distance of 3.2 times the picture height (see Section 2.1), corresponding to roughly 1.8 m from the stereoscopic monitor.

Test Method

The SS method with a 9-point rating scale (see Section 2.4.1) was chosen for the evaluations. Subjects were asked to evaluate the video sequences in terms of four different aspects: perceived overall quality, perceived depth quantity, content preference, and sensation of reality. The 9-point rating scale ranged from 1 to 9, with 1 representing the lowest value, and 9 the highest value of each aspect. In particular, the two extremes (1 and 9) correspond to “low” and

"high" for perceived overall quality and content preference, "no presence" and "very strong presence" for sensation of reality, and "no depth" and "a lot of depth" for perceived depth quantity.

Training Session

Before the test starts, oral instructions were provided to the subject to explain his/her task. Additionally, a training session was organized to allow subjects to familiarize with the assessment procedure. The content shown in the training session was selected by experts to include 2D and 3D examples of various quality levels.

Test Sessions

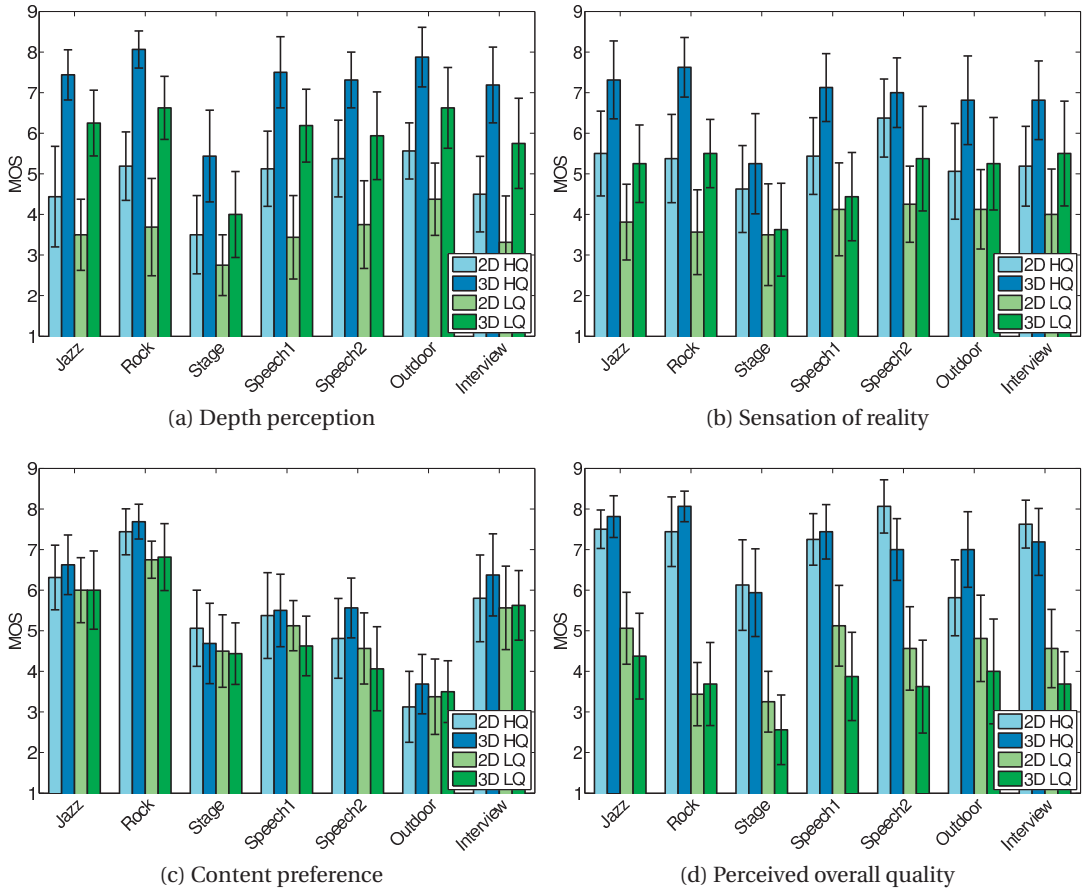
Experiments were conducted in three sessions. To avoid subjects' fatigue, a fifteen-minute break was provided between consecutive sessions. Nine video sequences were presented in the first and second sessions, and ten in the last one, leading to a total of 28 video sequences, and thus, to a total of 28 trials. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each subject, whereas the same content was never shown consecutively.

The 2D and 3D video sequences were mixed, such that subjects could not predict the rendering mode and to reduce any a priori that could influence subjects' ratings. Therefore, all video sequences were viewed with 3D glasses. Watching 2D video content while wearing 3D glasses reduces the horizontal resolution by a factor two due to characteristics of the monitor used in the experiments , which can reduce perceived quality. However, the loss of vertical resolution in passive 3D display is very low and, in our results, no statistical difference was found between 2D and 3D modes on the perceived overall quality (see Section 6.1.3).

A total of 16 subjects (5 females, 11 males) took part in our experiments. They were between 19 and 30 years old with an average of 23.8 years of age. All subjects were screened for correct visual acuity, color vision, and stereo vision using the Snellen chart, Ishihara chart, and Randot test, respectively.

Data Processing

To detect and remove subjects whose scores appear to deviate strongly from others in a session, outlier detection was performed. The outlier detection was applied to the set of results obtained from the 16 subjects and performed according to the guidelines described in Section 2.3.1 of Annex 2 of ITU-R BT.500-13 (2012) (see Section 2.6.1). During the training session, examples of the lowest and highest quality levels were shown to guide subjects to bound their own perceived overall quality ratings in a similar way. Since quality was the only factor in which subjects could be trained, the outlier detection was performed only on the



(a) Depth perception

(b) Sensation of reality

(c) Content preference

(d) Perceived overall quality

Figure 6.1: MOSs and CIs for each of the perceptual factors.

perceived overall quality ratings. No outliers were detected, thus, for the subjective ratings analysis all 16 subjects were included. Then, the MOSs were computed for each test stimulus as the mean across the rates of the valid subjects, as well as associated 95% CIs, assuming a normal distribution of the scores.

6.1.3 Results

Figure 6.1 shows the resulting MOS and CI for the four perceptual factors. As it can be observed, for a given quality level, perceived depth and sensation of reality are both higher for 3D when compared to 2D stimuli. Similarly, high quality sequences generally obtained higher ratings for perceived depth quantity, sensation of reality, and perceived overall quality when compared to their corresponding low quality versions. However, the difference in terms of perceived depth and sensation of reality between 3D LQ stimuli and 2D HQ stimuli is not significant as the CIs considerably overlap in all contents. This observation shows that depth cues in 3D stimuli are effective for depth perception only if a certain level of visual quality is reached. As content *Stage* is very dark, the perceived 3D effect was not very strong and the perceived depth and

sensation of reality were rated relatively low.

To investigate quantitatively whether the objective factors, such as the rendering mode (2D vs. 3D), actual quality level (LQ vs. HQ), and content have a significant influence on the perceptual factors (perceived depth, sensation of reality, content preference and perceived overall quality), an ANOVA analysis was performed on the subjective ratings for each case. In particular, the null hypothesis was that the rendering mode, quality level, and content do not influence neither of the perceptual factors.

The null hypothesis was rejected for the cases of perceived depth and sensation of reality for all three objective factors, $p < 0.001$, indicating that the effects of the rendering mode, actual quality level, and content on perceived depth quantity and on sensation of reality were significant. Regarding the effects of the objective factors on content preference and on perceived overall quality, only the actual content and the actual quality level influenced these perceptual factors significantly, $p < 0.001$. Two sequences (*Jazz* and *Rock*) out of seven are from music concert and contain a musical audio track, whereas the other five sequences are quite general. As the interview of Quincy Jones, who is a famous musician, got similar ratings for content preference when compared to the *Jazz* sequence, we believe that the presence of a musical audio track was not the only factor influencing content preference. Although the rendering mode itself did not influence neither content preference nor perceived overall quality, the interactions between rendering mode and quality level, as well as the interactions between actual content and quality level influence significantly, $p < 0.05$, perceived overall quality. For the rest of the cases, interactions among the objective factors did not influence any other perceptual factor. The findings confirmed our expectations.

To understand the impact of the perceptual factors, such as sensation of reality, content preference, perceived overall quality, and perceived depth quantity on each other, the correlation between the MOS for all four factors was measured using the PCC. Table 6.2 reports the correlation coefficients. The results show that there is a strong correlation between perceived depth quantity and sensation of reality ($\rho > 0.88$). Also, there is a strong correlation between sensation of reality and perceived overall quality ($\rho > 0.73$). However, the correlation between perceived overall quality and perceived depth quantity is relatively low ($\rho = 0.42$), but statistically different from zero, $p = 0.03$. Since the correlation between sensation of reality and perceived depth quantity, as well as between sensation of reality and perceived overall quality, is strong, it is rational that the correlation between perceived overall quality and perceived depth quantity is also different from zero, due to the transitivity property. On the other hand, the correlation between perceived depth quantity and content preference ($\rho < 0.16$) is very weak. Thus, apparently content *per se* impacts on depth perception, but content preference does not. Additionally, depth perception is significantly influenced by the presentation mode, as binocular depth cues are quite powerful, while this factor has no significant effect on content preference, which also explains the weak correlation between content preference and perceived depth. The correlation between sensation of reality and content preference is very low ($\rho < 0.3$) and not statistically different from zero, $p = 0.12$. Again, the low correlation

Table 6.2: PCC between the ratings of different perceptual aspects.

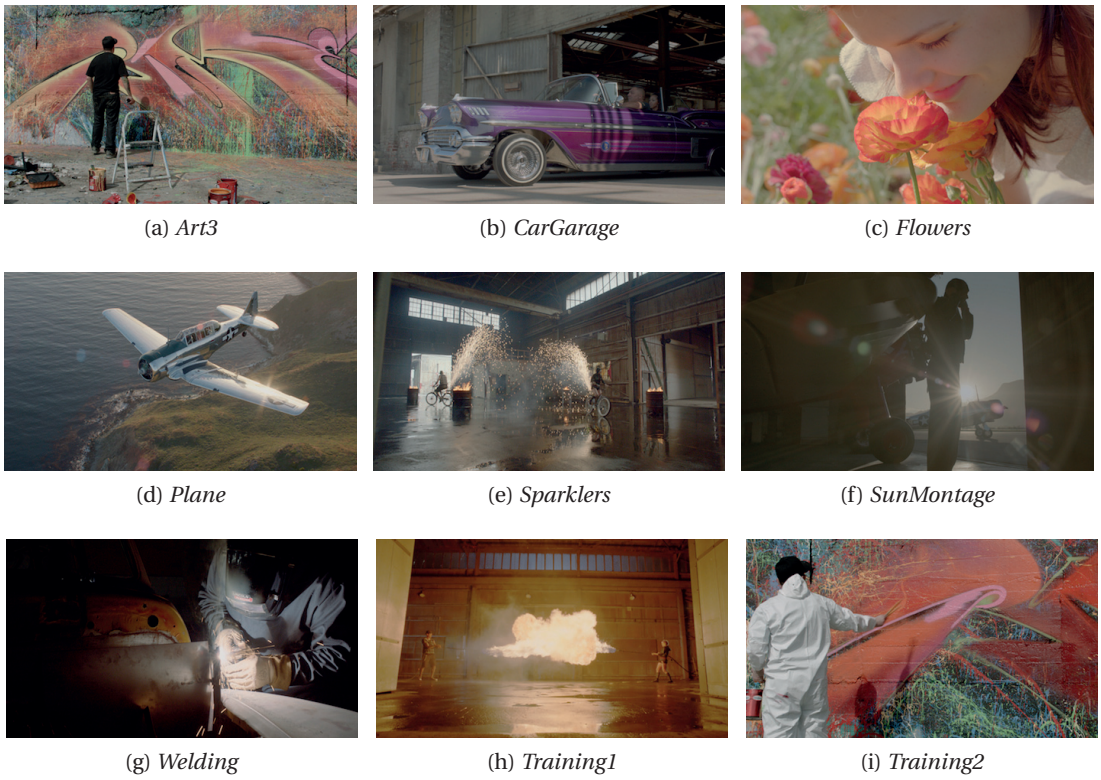
	Content preference	Sensation of reality	Depth quantity
Overall quality	0.3392	0.7308	0.4172
Content preference		0.3017	0.1527
Sensation of reality			0.8835

between sensation of reality and content preference can be explained by the fact that the rendering mode has a significant impact on first perceptual factor, but not on the former one.

6.2 Evaluation of Higher Dynamic Range Video

Until now, most of the research in HDR imaging focused on the ability to capture, store, and display HDR content using conventional imaging technology. A significant amount of effort has been spent on designing algorithms, coined TMOs, for accurate reproduction of HDR content on conventional displays with limited dynamic range (Drago et al., 2003; Kuang et al., 2007b; Mantiuk et al., 2008; Reinhard et al., 2002). Several formats and compression techniques were also developed for HDR, mostly focusing on backward compatibility with popular coding formats such as JPEG (Korshunov and Ebrahimi, 2013; J. Liu et al., 2013; Ward and Simmons, 2006) and MPEG (Mantiuk et al., 2006b; Mantiuk et al., 2007), including the upcoming JPEG XT standard for HDR image compression (Artusi et al., 2015). The TMOs and coding formats were studied and compared using different evaluation methodologies (subjective and objective) to determine the most suitable for different usage scenarios and displays (Annighöfer et al., 2010; Čadík et al., 2008; M. Chen et al., 2006; Korshunov and Ebrahimi, 2012; Kuang et al., 2007a; Mai et al., 2011b; Yoshida et al., 2005). However, since high quality HDR displays did not exist, the direct effect of HDR video technology on viewing experience was little studied, except for a few works (Akyüz et al., 2007; Ledda et al., 2005; Rempel et al., 2009) that relied on the limited first generation of HDR displays, such as BrightSide DR37-P monitor, or the work from Mantel et al. (2015b), which investigated only peak luminance levels up to 490 cd/m^2.

The recent breakthrough in the HDR capturing and displaying technologies allowed to bridge this gap and, hence, this section evaluates the added value of higher dynamic range to user preference using high quality HDR video sequences and the latest professional HDR monitor. For this purpose, several HDR video sequences were displayed at four different peak luminance levels, including the maximal supported level of 4000 nit, in a side-by-side fashion on a professional reference HDR monitor, 'Pulsar' prototype developed by Dolby, renown for the accurate and reliable reproduction of color and luminance. The black level was held constant, so the luminance dynamic range was solely determined by the maximum luminance. The tested luminance levels reflect four levels of dynamic range that are typical for current and



(a) *Art3* (b) *CarGarage* (c) *Flowers*

(d) *Plane* (e) *Sparklers* (f) *SunMontage*

(g) *Welding* (h) *Training1* (i) *Training2*

Figure 6.2: Representative frames of the sequences used in the experiments ((a)-(g)) (one additional sequence from "Star Trek: Into Darkness" is not included here due to Copyright) and training session ((h)-(i)). The 100 nit versions are shown, since typical displays and printers are unable to reproduce higher dynamic range images.

future consumer scenarios, given today's current displaying technology and latest advances in HDR displays. Two different evaluation methodologies were selected and compared in terms of accuracy and reliability in constructing a scale of perceptual preferences. This section reports a detailed analysis of the results to determine the subjective preference among different luminance levels at which the high dynamic range video sequences were displayed.

6.2.1 Dataset

Eight video sequences representing various levels of dynamic range and with different visual characteristics (see Figure 6.2) were used in the experiments. Two additional sequences were used during a training session. Each video sequence was about 20 s long. All video clips, except one from "Star Trek: Into Darkness", were shot by professional film directors specifically for this experiment.

For each video content, four dynamic range levels were selected to represent several key use cases, as opposed to using uniform perceptual distances

i) 4000 nit version, which was manually graded by professional colorists from the originally captured video: this value was determined by the HDR monitor used in the experiments (see Section 6.2.2) and the availability of professionally color graded content for this luminance range,
ii) 1000 nit version, which was tone-mapped from manually graded 4000 nit version: this value represents some very high end consumer TVs,
iii) 400 nit version, which was tone-mapped from manually graded 4000 nit version: this value is a good representation of the maximum luminance level of current high-quality consumer TVs, and
iv) 100 nit version, which was tone-mapped from manually graded 4000 nit version: this value is a commonly used maximum luminance level for reference monitors in a production environment.

For tone-mapping, an automated proprietary algorithm was used. This algorithm was designed to preserve overall appearance to the input version and is not intended for image enhancement or to bias importance to specific image regions, as often occurs in a human-guided color grading. The original video sequences were uncompressed, with 12 bits per color, in a domain that has characteristics of gamma and log nonlinearities, as suited for HDR (Miller et al., 2013). The combination of high bit-depth and uncompressed video is intended to remove secondary issues of dynamic range effects on needed bit-depth and compression algorithm parameters, since the study's aim was to isolate the question of dynamic range.

6.2.2 Methodology

The experiments were conducted at the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU. The test room is equipped with a controlled lighting system of a 6500 K color temperature. The color of all background walls and curtains in the room is mid grey. The laboratory setup is intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors.

To display the test stimuli, a full HD (1920×1080 pixels) 42" Dolby Research HDR RGB backlight dual modulation display (aka 'Pulsar') was used. The monitor has the following specifications: full DCI P3 color gamut, 4000 cd/m^2 peak luminance, low black level (0.005 cd/m^2), 12 bits/color input with accurate and reliable reproduction of color and luminance. In the experiments, the luminance of the background behind the monitor was about 20 cd/m^2. The ambient illumination did not directly reflect off of the display.

In every session, three subjects assessed the displayed test video content simultaneously. They were seated in one row perpendicular to the center of the monitor, at a distance of about 3 times the picture height (see Table 2.1).

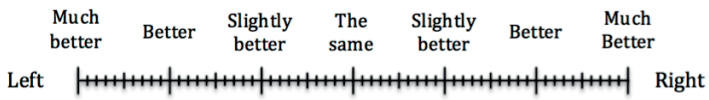


Figure 6.3: SC scoring scale.

Test Method

The video sequences were presented in pairs in side-by-side fashion to minimize visual working memory limitations. Since only one full HD 1920 × 1080 HDR monitor was available, each video was cropped to 950 × 1080 pixels with 20 pixels of black border separating the two sequences.

Subjects were asked to rate the overall quality of pairs of displayed video sequences. To select a score, subjects were instructed to consider such video characteristics as color rendition, quality of the reproduction of skin tones, details of shadows in the scene, contrast and the details of highlights, presentation of light sources appearing in the scene, etc. They were also asked to consider visual discomfort.

Two different subjective tests were performed using two different evaluation methods: full PC and SC with hidden reference.

Full Pair Comparison In this evaluation, subjects were asked to judge which video sequence in a pair ('left' or 'right') is preferred. The option 'same' was also included to avoid random preference selections. For each of the 8 contents, all the possible combinations of the 4 grades (100, 400, 1000, and 4000 nit) were considered, as well as an extra pair corresponding to 4000 nit vs 4000 nit, leading to a total of $8 \times \left[\binom{4}{2} + 1\right] = 56$ comparisons. The comparison of identical video content is useful for side analysis of observer performance and display symmetry.

Stimulus Comparison with Hidden Reference Instead of considering all possible pairs, as it was done in the full PC, the 4000 nit version was treated as a hidden reference. Therefore, only pairs with the 4000 nit version were evaluated. Again, an extra pair corresponding to 4000 nit vs 4000 nit was included for reliability checks. The hidden reference was shown in every pair, with a randomized position on the screen (either on the left or on the right). The adjectival categorical judgment method with a horizontal continuous scale (see Section 2.4.3) was used in the evaluation to provide a finer comparison of the two conditions. Figure 6.3 depicts the scoring scale of SC method.

Similarly to full PC, in SC evaluation, subjects were asked to judge which video sequence in a pair ('left' or 'right') is preferred, however, the option 'same' was not included and subjects were instructed to randomly select one option when both sequences appear equal. Basically, this forced choice (FC) preference is a binary scale that directly identifies which condition is preferred.

Table 6.3: Details of the two experiments.

	Full PC	SC
Number of sessions	2	1
Session length	15 minutes	17 minutes
Break length	10 minutes	N/A
Number of subjects (♂/♀)	21 (10/11)	20 (10/10)
Age range (average)	18 – 33 (25.8)	18 – 31 (24.1)

Test Planning

Before the experiments, a consent form was handed to subjects for signature and oral instructions were provided to explain the evaluation task. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts respectively. A training session was organized using the selected video sequences (see Figure 6.2 for the screenshots) to allow subjects to familiarize with the assessment procedure.

In terms of session planning, the main differences between the two experiments using two methods are related to the different number of the evaluated stimuli and observers as summarized in Table 6.3. To reduce contextual effects, the order in which stimuli were displayed on the screen was randomized differently for each different group of subjects with the same video content never shown consecutively. In the full PC experiment, the test material was also randomly distributed over two test sessions.

Data Processing

A typical way to analyze and compare subjective evaluations is to compute MOSs. However, from both subjective experiments, only preference scale scores from the SC method can be used to compute MOS values (as the mean across the rates of the valid subjects) and corresponding 95% CI directly. Indeed, the FC and PC methods do not provide such scores directly, and hence an estimation of MOS values needs to be computed instead. The Thurstone Case V model and our proposed technique for estimating CIs (see Section 2.6.3) were used in this paper for computation of these values.

Next parts describe in details the outlier detection and the methods for estimation of MOS values from the results of FC and PC methods. Also, four incomplete full PC designs were identified by taking incomplete subsets of scores from full PC results. These four incomplete designs include pairs with one grade only (referred to as 'pairs with x nit only'). Analysis of these designs is useful for comparisons to single anchor methodologies, e.g., DSIS and DSCQS (see Section 2.4). Additionally, one incomplete design was analyzed considering pairs with consecutive grades, i.e., 100 vs. 400, 400 vs. 1000, and 1000 vs. 4000. As the result, incomplete designs include the evaluation results for 3 pairs of video content only, as opposed to 6 pairs forming the full PC design.

Outlier Detection No outlier detection was performed for the scores from the full PC evaluation, since there is no international recommendation or a commonly used outlier detection technique for PC results.

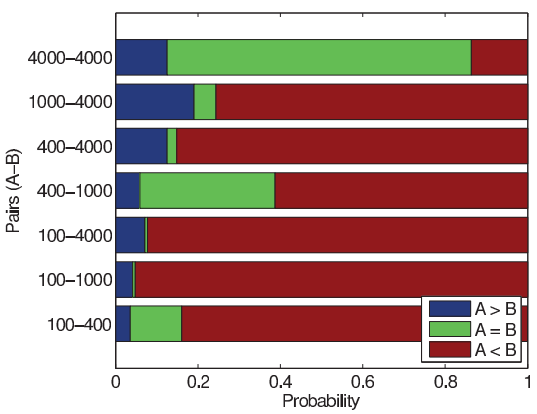
However, the subjective results of the SC experiment were processed by first detecting and removing subjects whose preference scale scores appeared to deviate strongly from others. The outlier detection was performed according to the guidelines described in Section 2.3.1 of Annex 2 of ITU-R BT.500-13 (2012) (see Section 2.6.1). As the result, one outlier was detected and the corresponding scores were removed from the subjective results.

Full Pair Comparison Before estimating MOS values for PC results, the winning frequency and the tie frequency are computed from the obtained subjective ratings for each pair of stimuli. This can be done individually for each test video content or jointly over all contents. Then, the Thurstone Case V model was used to estimate MOSs and associated CIs considering Laplace smoothing to help regularize the estimates.

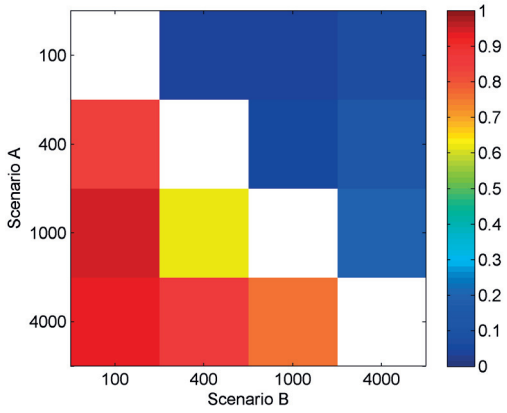
Forced Choice and Incomplete Pair Comparison Designs For the FC method or for incomplete PC designs, a different analysis has to be applied to consider the missing results for the pairs that were not evaluated. Morrissey (1955) and Gulliksen (1956) have formulated an incomplete matrix solution of Thurstone's Law for the estimation of the quality scores from a subset of PC data. The incomplete matrix solution is formulated as the least squares solution to a system of equations using only the valid data entries, i.e., missing data and pairs with 0 or 1 proportions are ignored. For a more detailed description of this model, please refer to (Tsukida and Gupta, 2011).

Similar to the full PC, the preference matrix is constructed from the winning frequencies. The matrix is incomplete, since it has no entries for the missing scores. The Morrisey-Gulliksen incomplete matrix solution is then used to convert the ratings from the binary/ternary scale to continuous-scale quality score values, based on Thurstone's case V model. When considering a subset of the full PC experiment, ties were equally distributed. Again, Laplace smoothing was used to help regularize the estimates. However, the CIs were not estimated for the incomplete matrix solutions, since there is no analysis method to estimate CIs for an incomplete matrix. The approach to estimate CIs for full PC might not work for incomplete matrix, since it might introduce uncertainty and make reliable estimation of CIs hard.

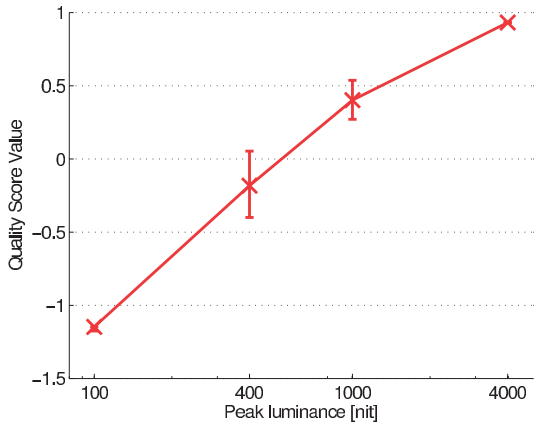
Correlation Between Different Designs and Scales To be compliant with the standard procedure for comparing estimated MOS values from different experiments (see Section 2.7), a regression was fitted to each $\left[MOS^{ExpX}, MOS^{ExpY}\right]$ data set using cubic fitting. The PCC and SROCC were computed to estimate linearity and monotonicity, respectively (see Section 2.7.2). Since estimated MOSs span different range values for different experiments, the computation of RMSE is not possible and was not done. Also, since there was no estimation of CIs for the incomplete PC designs, the OR was not computed.



(a) Preference probabilities



(b) Preference probability of choosing "Scenario A" over "Scenario B"



(c) Quality scores

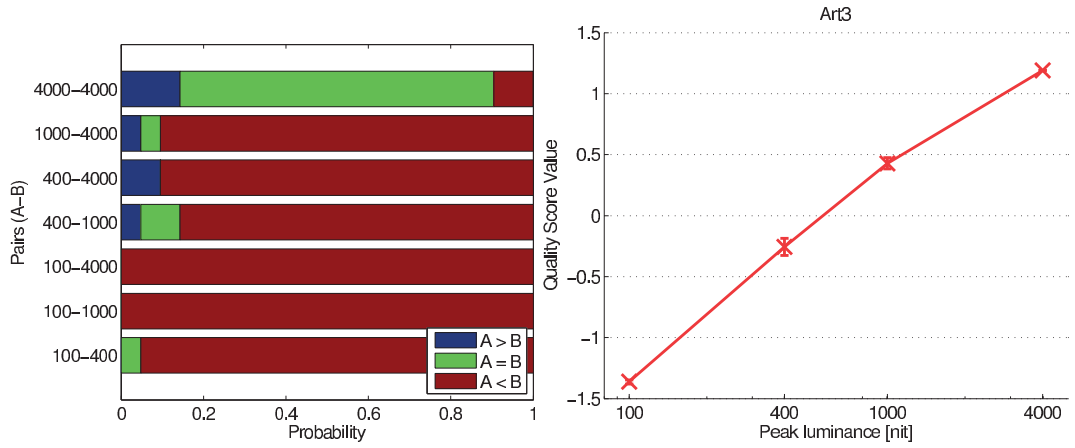
Figure 6.4: Full PC: Overall results.

6.2.3 Results

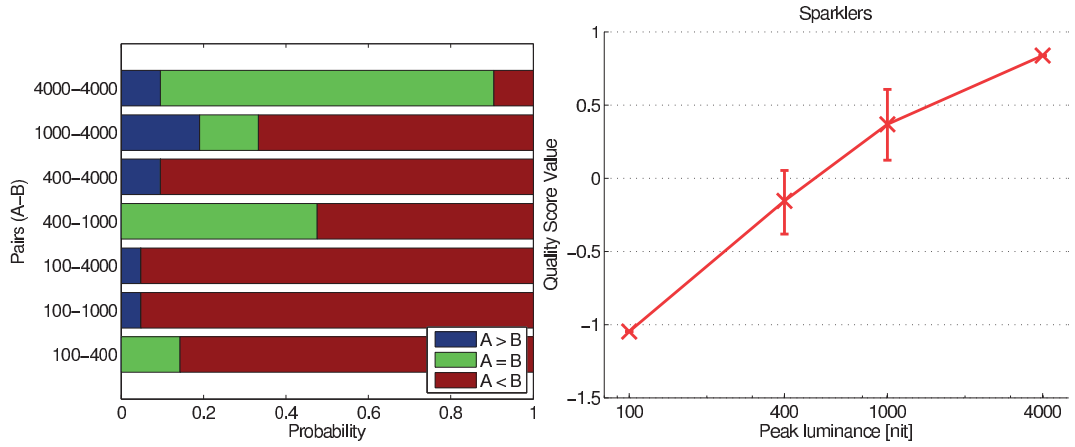
This subsection reports the results of the subjective experiments. First, the results of the full PC experiments are reported and analyzed. Then, the results of the SC with hidden reference are reported and analyzed. Finally, a comparison between the different methods and designs is made.

Full Pair Comparison

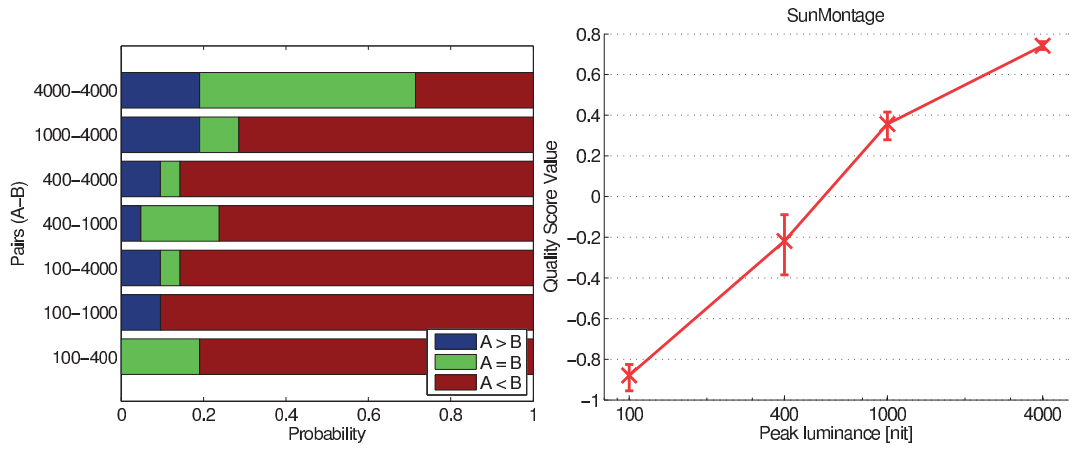
Figure 6.4 shows the preference probabilities, preference matrix, and estimated MOS values computed over all video contents for the full PC experiment. Figure 6.5 illustrates the preference probabilities and estimated MOS values computed for selected individual contents. All figures of MOS values demonstrate that the score value increases with the increase in peak



(a) Results for content *Art3*

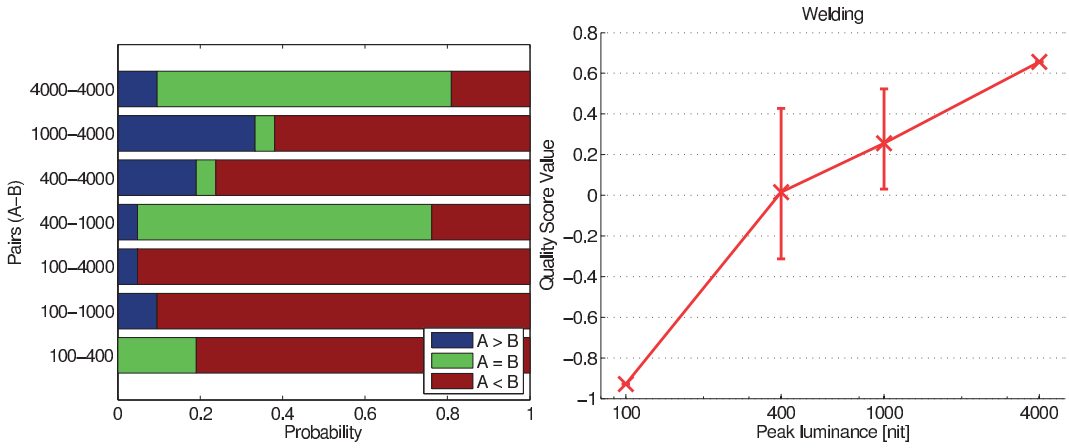


(b) Results for content *Sparklers*



(c) Results for content *SunMontage*

Figure 6.5: Full PC: Examples of results for individual contents.



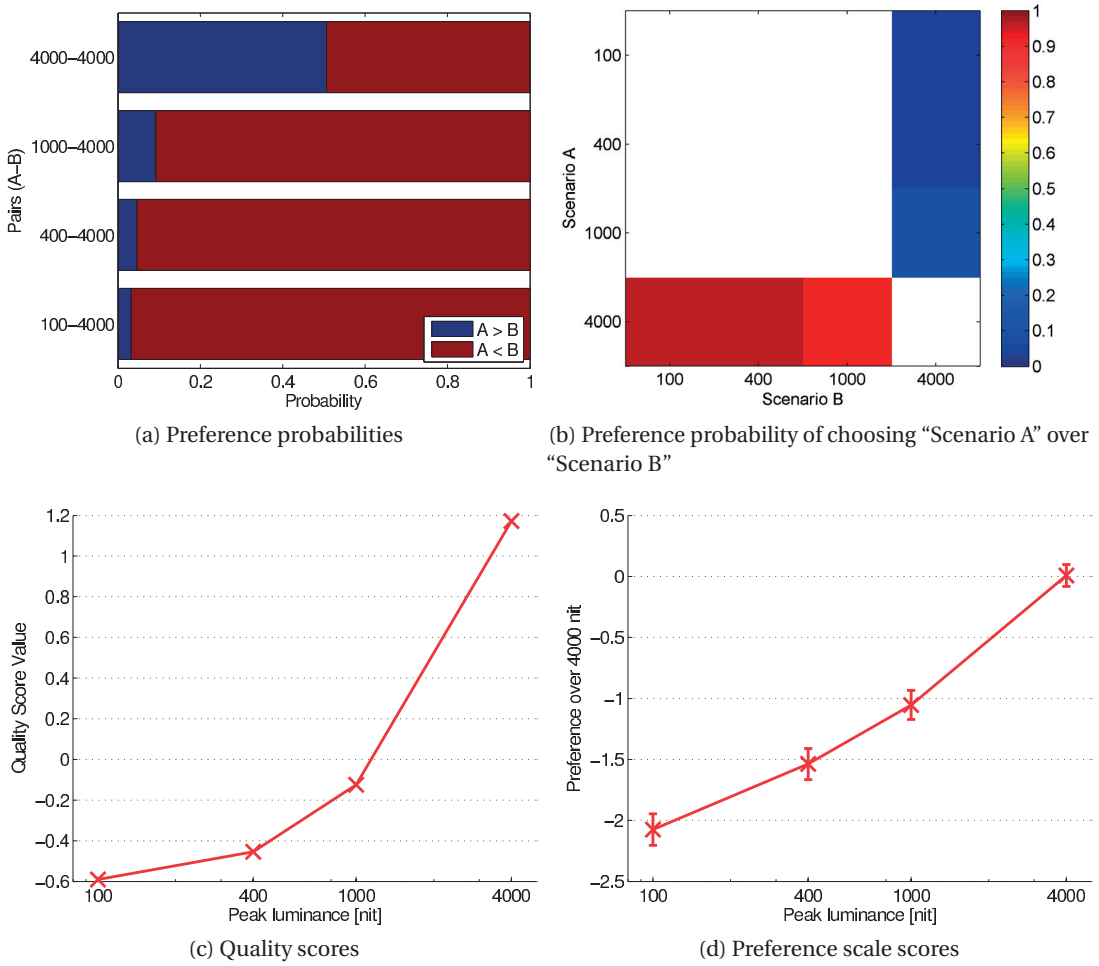
(a) Results for content *Welding*

Figure 6.5: Full PC: Examples of results for individual contents *(Continued)*.

luminance. The quality score values tend to increase linearly (with log luminance), although they exhibit a slight concave shape for some contents, which indicates the existence of a saturation level. In most cases, the difference between individual grades is significant, as the CIs do not overlap, except for content *Welding*, where about 70% of the subjects had no preference between the 400 and 1000 nit grades. For this content, 100 nit peak luminance was likely insufficient to represent the high contrast in the original (graded) image between the strong brightness created by the welding sparks and the dark garage, whereas the improvements in the 4000 nit peak luminance version were most likely related to the very strong luminance of the welding sparks.

In most comparisons, higher peak luminance was largely preferred and most ties occurred in the 400 nit versus 1000 nit pair. For example, for content *Sparklers*, 11 subjects preferred the 1000 nit grade, whereas the remaining 10 subjects did not express any preference between the two grades. As most ties occurred in the 400 nit versus 1000 nit pair, most of the uncertainty lies between those two grades, which is represented by generally wider CIs associated with the 400 and 1000 nit grades. Note that the 400 and 1000 nit levels are, in fact, closer together on a log luminance scale. The log luminance scale is a likely candidate for a perceptually uniform scale based on Weber's law. So having closer ratings, as well as more ties for these levels, is entirely expected based on a Weber's law model.

As stated previously, the dynamic range was explored by fixing the black level, and increasing the peak luminance level. Thus, there is some conflation of dynamic range and brightness. Trying to isolate brightness alone, with constant dynamic range, would elevate the black level, and this is already known to cause reductions in preference (Seetzen et al., 2006). On the other hand, trying to isolate dynamic range alone, without changes in brightness, means that the black level must change. It is already known that black level improvements lower than 0.005 cd/m^2 do not lead to preference improvements, except for specific imagery and viewed



(a) Preference probabilities

(b) Preference probability of choosing "Scenario A" over "Scenario B"

(c) Quality scores

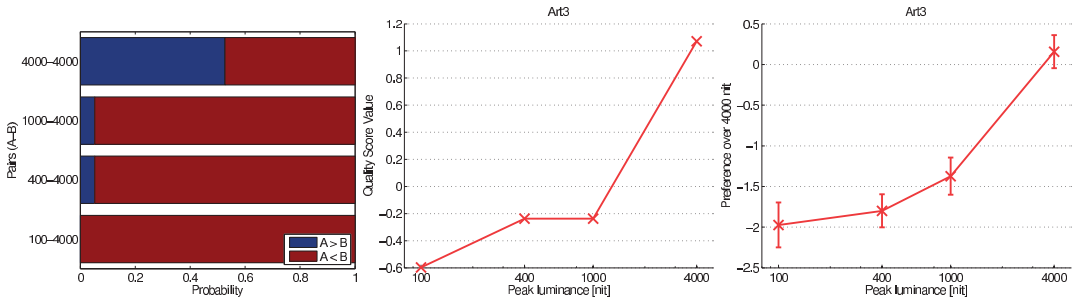
(d) Preference scale scores

Figure 6.6: SC with hidden reference: Overall results.

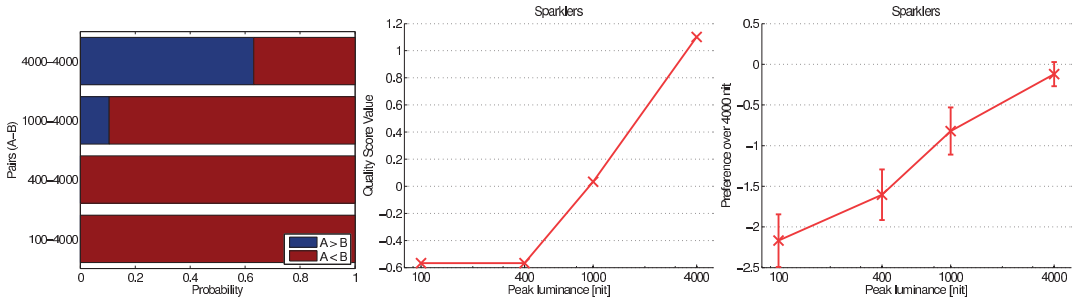
in total dark environments. The method chosen in this study of fixing the black level and changing the dynamic range via changes in brightness is most relevant to current display technology and viewing conditions.

Stimulus Comparison with Hidden Reference

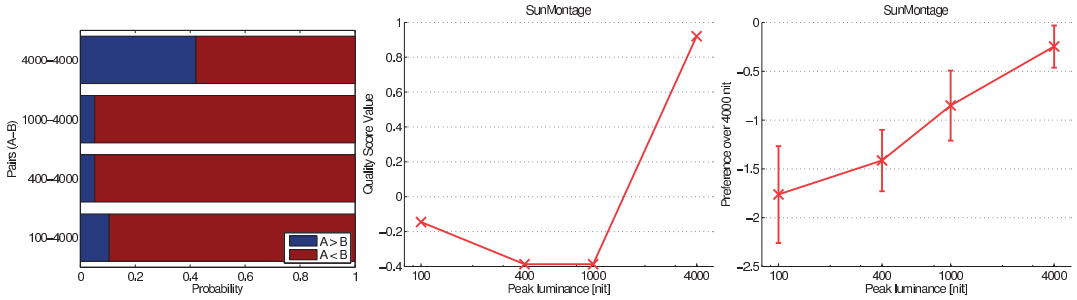
Figure 6.6 depicts the preference probabilities, preference matrix, estimated MOS values, and preference scale scores computed over all contents for the SC with hidden reference experiment. Figure 6.7 illustrates the preference probabilities, estimated MOS values, and preference scale scores computed for some individual contents. Regarding the plots for the preference scale scores, values −3, −2, and −1 on the y-axis indicate that the hidden reference, i.e., 4000 nit grade, was judged as 'much better', 'better', and 'slightly better', respectively, than the tested peak luminance level on the x-axis, whereas 0 indicates no preference.



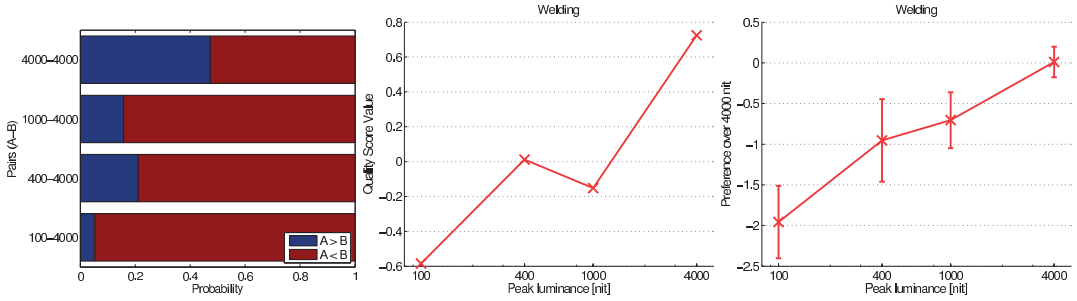
(a) Results for content *Art3*



(b) Results for content *Sparklers*



(c) Results for content *SunMontage*



(d) Results for content *Welding*

Figure 6.7: SC with hidden reference: Examples of results for individual contents.

Again, all figures of MOS values demonstrate that the score value increases with the increase in peak luminance. The overall results for the preference scale tend to increase linearly (with log luminance), although they exhibit a slight convex shape. However, when considering all contents, the quality score values tend to increase exponentially (with log luminance). These results indicate that preference would increase linearly as peak luminance increases, which is contradictory with the findings of the full PC experiment. Specific properties of the HVS, e.g., the Hunt effect (Hunt, 1952), may explain this behavior in some particular cases. However, for extremely bright monitors (well above 4000 cd/m^2), it is expected that visual discomfort might severely impact the overall QoE.

When considering the overall results for the preference scale, the difference between individual grades is significant, as the CIs do not overlap. However, when considering individual results, the difference between two consecutive peak luminance levels is not significant, as the CIs overlap in most cases, except between 1000 and 4000 nit grades. In general, the CIs are wider when compared to the results for the full PC experiment. However, the estimation of CIs is different between the two experiments. Therefore, one should not conclude that the PC methodology necessarily produces more precise results, although it is easier for the subjects to indicate their preference on a ternary scale than on a continuous scale, as it is difficult to have a clear, unambiguous, and commonly agreed definition of the different levels of the rating scale.

In general, results obtained for the FC method are comparable to those of the corresponding pairs in the full PC experiment. However, the quality scores values estimated from the incomplete design are quite different from those estimated from the full design, as most relations between the different grades were not evaluated. Because of the hidden reference in all pairs, it is impossible to estimate relative scores between the other grades from the binary scale. The preference scale might better represent the relative differences between the other grades, but the hidden reference could act as an upper anchor and influence the score difference between the other grades.

Comparison Between Different Designs and Scales

As stated previously, the quality score values tend to increase exponentially (with log luminance) for the FC method (see Figure 6.6c). However, the scores from the preference scale tend to increase linearly (with log luminance), although they exhibit a slight convex shape (see Figure 6.6d). Figure 6.8 depicts the difference between the results obtained over all contents for different designs estimated from the full PC. When considering pairs with 4000 nit only, the quality score values exhibit a convex shape, whereas they exhibit a concave shape when considering pairs with 100 nit only. However, the quality score values tend to increase linearly (with log luminance) when considering all pairs, which is somehow a combination of the different trends observed when considering one specific grade as a hidden reference.

These results suggest that complex mechanisms of the HVS are involved when comparing

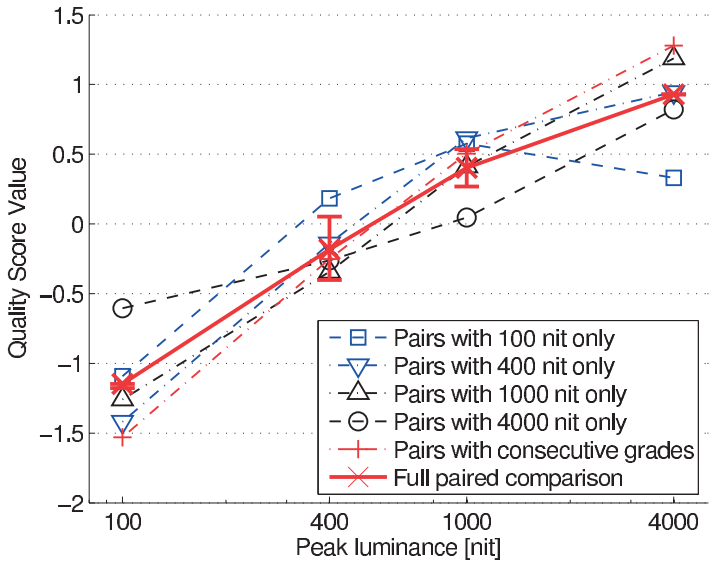


Figure 6.8: Full PC: Comparison between different designs.

different brightness levels and that deeper analysis is required to understand what are the significant factors impacting subjects' preference. However, to consider all these factors and their interactions, a full PC design reveals more information than incomplete designs in the estimation of quality score values. Nevertheless, the full PC methodology requires more time as more pairs need to be assessed as the number of test conditions increases. To overcome this drawback, carefully designed limited set of pairs can be considered, e.g., using pairs with consecutive grades.

To further investigate the correlation between different designs and scales, the PCC and SROCC were computed according to Section 6.2.2. As the mapping of MOS^{ExpX} to MOS^{ExpY} yields slightly different results when compared to mapping of MOS^{ExpY} to MOS^{ExpX}, both mappings are considered and results are reported for both cases. In the following, a value $v(i,j)$ on row i and column j is computed considering mapping of MOS^{Expi} to MOS^{Expj}.

Table 6.4 reports the PCC and SROCC statistical evaluation metrics. The table demonstrates that there is a strong correlation between results of the full PC experiment and results estimated considering pairs with 1000 nit only and pairs with consecutive grades, as the correlation indexes are above 0.98 in both cases. These results show that considering only pairs with consecutive grades could be an alternative to the full PC. On the other hand, considering incomplete designs with pairs of 100 or 4000 nit only have the lowest correlation with the full design. Considering only pairs with 4000 nit, the two experiment produced quite similar results according to the PCC (0.91-0.94). However, the SROCC is slightly lower (0.8). The difference is probably due to individual preferences, as the pool of subjects was different between the two experiments, and the non-linear process involved in the Thurstone Case V model to convert the preference scores to continuous quality scores.

Table 6.4: Statistical evaluation metrics.

(a) PCC

		Full PC						SC w/ hidden ref	
		Full	100 nit ref	400 nit ref	1000 nit ref	4000 nit ref	Consecutive	4000 nit ref	Pref scale
Full PC	Full	-	0.9383	0.9852	0.9870	0.9324	0.9917	0.8987	0.9428
	100 nit ref	0.8592	-	0.8912	0.7573	0.5800	0.8338	0.4329	0.6606
	400 nit ref	0.9779	0.9303	-	0.9469	0.8345	0.9718	0.7913	0.8841
	1000 nit ref	0.9823	0.8902	0.9598	-	0.9214	0.9811	0.8924	0.9418
	4000 nit ref	0.8867	0.6611	0.8065	0.8916	-	0.8616	0.9431	0.9328
	Consecutive	0.9928	0.9106	0.9876	0.9891	0.9099	-	0.8651	0.9280
FC	4000 nit ref	0.7951	0.5255	0.6951	0.8418	0.9154	0.7761	-	0.9075
	Pref scale	0.9341	0.8100	0.9117	0.9359	0.9404	0.9214	0.9539	-

(b) SROCC

		Full PC						SC w/ hidden ref	
		Full	100 nit ref	400 nit ref	1000 nit ref	4000 nit ref	Consecutive	4000 nit ref	Pref scale
Full PC	Full	-	0.6658	0.9692	0.9930	0.8831	0.9883	0.7980	0.9186
	100 nit ref	0.6673	-	0.6716	0.6601	0.4374	0.6617	0.4399	0.5597
	400 nit ref	0.9692	0.6716	-	0.9593	0.8189	0.9692	0.7177	0.8699
	1000 nit ref	0.9930	0.6601	0.9593	-	0.8721	0.9846	0.8050	0.9146
	4000 nit ref	0.8831	0.4361	0.8216	0.8721	-	0.8622	0.8015	0.8795
	Consecutive	0.9883	0.6586	0.9692	0.9846	0.8622	-	0.7527	0.9014
FC	4000 nit ref	0.7980	0.4382	0.7270	0.8050	0.8015	0.7597	-	0.8258
	Pref scale	0.9186	0.5561	0.8699	0.9146	0.8768	0.9014	0.8303	-

6.3 Conclusion

This chapter investigated the impact of 3D and HDR on viewers' QoE. In particular, we conducted an experiment during which the participants were experiencing 2D and 3D multimedia contents of various quality levels. The subjects provided their self-assessed ratings after each video, in which they were asked to rate various aspects that may influence QoE, namely, perceived overall quality, perceived depth, content preference, and sensation of reality. The subjective ratings analysis revealed that the effects of the rendering mode, actual quality level, and content on perceived depth and on sensation of reality were significant. It also revealed that there is a strong correlation between perceived depth and sensation of reality, as well as between sensation of reality and perceived overall quality. Finally, for a given quality level perceived depth and sensation of reality are both higher for 3D when compared to 2D stimuli. Similarly, high quality sequences generally obtained higher ratings for perceived depth quantity, sensation of reality, and perceived overall quality when compared to their corresponding low quality versions. However, the difference in terms of perceived depth and sensation of reality between 3D low quality stimuli and 2D high quality stimuli was not significant.

In this chapter, we also investigated the added value of higher dynamic range to user preference using stimulus comparison with hidden reference and full pair comparison methods. Subjective tests were conducted to evaluate the preference between video sequences at four

different peak luminance levels (ranging from 100 cd/m^2 to 4000 cd/m^2), which were displayed side-by-side on a professional reference HDR monitor. The analysis of the subjective results demonstrated that the increase in maximum luminance level at which higher dynamic range video is displayed is preferred by average viewers, with a steady increase in preference as the maximum luminance increases. The results showed a significant increase in the perceptual experience when viewing HDR content at 4000 cd/m^2 peak luminance compared to the current standards in TV and cinema.

7 Visual Attention in Immersive Video Technologies

Visual attention is a widely studied topic and its practical applications include gaze-adaptive image and video compression (Z. Chen et al., 2010; Itti, 2004), objective image quality metrics (Redi et al., 2009), image retargeting (D. Wang et al., 2011), and image retrieval (Vu et al., 2003). It even reaches beyond computing, proving useful in areas such as attention-based advertising, art, and cinema. To take advantage of visual attention information in practical applications, *salient regions* in images, i.e., regions that attract most of the attention, are either detected using an eye tracking device or predicted using computational models of visual attention. One of the first computational model was proposed by Itti et al. (1998), which uses image features such as luminance intensity, color, and orientation to construct a *saliency map*, i.e., a map predicting visual attention of a corresponding visual scene. The practical usefulness of computational models fueled the research for many years, resulting in many visual attention models, creation of different evaluation datasets with ground truth eye tracking data, and various evaluation methodologies and metrics.

Although a significant number of public image and video datasets for visual attention exist (Winkler and Subramanian, 2013), there are very few eye tracking datasets for immersive technologies. To the best of our knowledge, for 3D content, there are only two datasets for stereoscopic images (Lang et al., 2012; K. Wang et al., 2013) and two for stereoscopic video sequences (Fang et al., 2014; Hanhart and Ebrahimi, 2014c). Regarding HDR content, there is only one dataset for HDR images by Narwaria et al. (2014) and two for HDR video sequences (Dong et al., 2014; Narwaria et al., 2014). No dataset with eye tracking data is available for UHD content. However, without this subjective data, it is hard to understand what is the impact of immersive technologies on visual attention and whether it is significant for practical applications.

Since immersive technologies have the ability to provide more details and depth, as well as better color, contrast, and brightness, at the expense of higher data rate when compared to current technology, understanding human attention patterns and viewing strategies for immersive image and video content is important for developing efficient data compression algorithms, as well as accurate objective quality metrics and computational models of visual

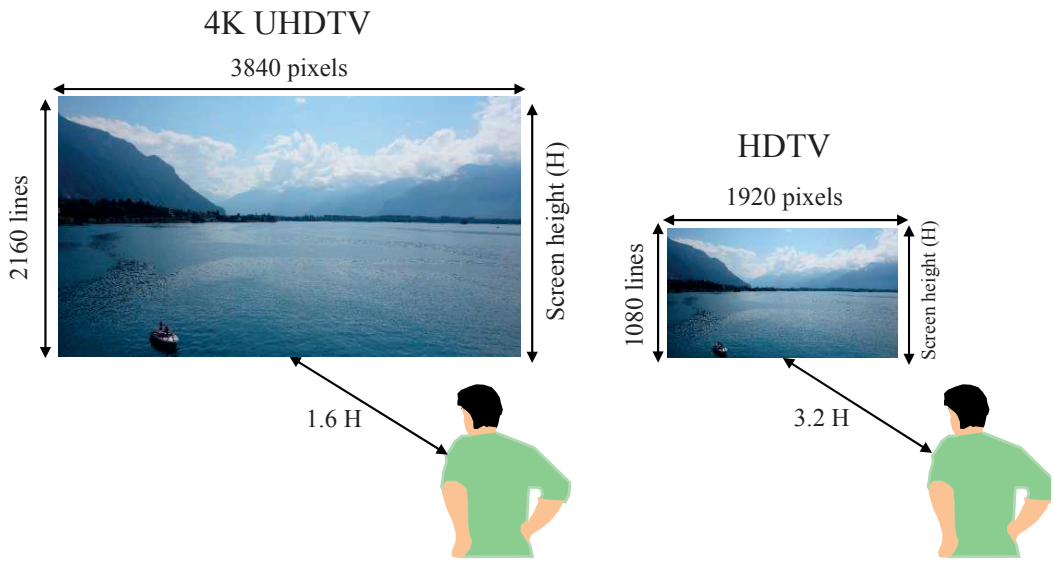


Figure 7.1: Recommended viewing conditions for 4K UHD and HD.

attention. The knowledge of visual attention for immersive image and video technologies can also help electronics manufactures to create better acquisition and display devices and content creators, such as photographers, movie and TV makers, to create images and video sequences with higher appeal value.

This chapter investigates the impact of UHD and HDR on visual attention in Sections 7.1 and 7.2, respectively. Eye tracking experiments were conducted to collect eye movements data for UHD images and their corresponding HD versions, as well as for HDR images and their corresponding LDR versions. FDMs were computed from the eye movements following the procedures described in Section 2.8.1. The similarity between the FDMs of the different technologies was measured and analyzed following the procedures described in Section 2.8.2 to investigate the impact of immersive technologies on visual attention.

7.1 Impact of Ultra High Definition on Visual Attention

UHD has the ability to provide more details, which enhances sense of presence and provides better viewing experience (Ito, 2010; Masaoka et al., 2006). The increased resolution of UHDTV typically leads to larger display sizes and, hence, for the full enjoyment of UHD content, ITU recommends certain viewing conditions (ITU-R BT.2022, 2012), as illustrated in Figure 7.1. The figure demonstrates the difference in viewing conditions between HD and UHD, suggesting that there might be also large differences in viewing strategies and in visual attention patterns of people watching HD and UHD TVs.

Although a significant number of public image and video datasets for visual attention exist (Winkler and Subramanian, 2013), no dataset with eye tracking data is available for UHD

content. However, without this subjective data, it is hard to understand what is the impact of UHD on visual attention and whether it is significant for practical applications. To answer this question, this section investigates the impact of UHD content on visual attention. An eye tracking experiment involving 20 naïve subjects was conducted to collect eye movements for a dataset of 45 UHD 4K images and their resized HD versions. FDM computed from the eye tracking data for UHD and HD resolutions are compared using three metrics to understand if there is a difference in visual attention between UHD and HD resolutions. This section reports the details and results of this eye tracking experiment.

7.1.1 Dataset

Since there is no publicly available standard dataset, at least to our knowledge, of UHD resolution images suitable for visual attention modeling, we constructed such dataset. For the dataset, we used still images with native resolution higher than UHD acquired by some of the latest digital cameras, including Sony DSC-RX100 II, Sony NEX-5N, FUJIFILM XF1, Olympus E-PL2, and RED SCARLET-X. Additionally, some high resolution painting images were obtained from the Europeana internet portal. A total of 45 images were selected to cover a wide variety of content, e.g., natural scenes (both indoor and outdoor), humans, ships, animals, music gigs, historical scenes, etc. For the dataset, all images were cropped to 3840 × 2160 pixels for UHD resolution and then downsampled to 1920 × 1080 pixels for HD resolution using Lanczos resampling. Figure 7.2 shows the images used in the experiments.

7.1.2 Methodology

The eye tracking experiments were conducted at the MMSPG quality test laboratory, which fulfills the recommendations for the subjective evaluation of visual data issued by ITU. The test room was equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of the maximum screen luminance, whereas the color of all the background walls and curtains present in the test area were in mid grey. The test room was separated in two by a curtain to isolate the subject and equipment from the test operators, which were present during the test session to supervise the recording of the eye tracking data. The laboratory setup was intended to ensure the reproducibility of the results and to avoid unintended influence of external factors.

Test stimuli were displayed on a professional high-performance 4K/QFHD 56” LCD reference monitor Sony Trimaster SRM-L560. A Smart Eye Pro 5.8 remote eye tracking system was employed to determine the gaze position on the screen of the left and right eyes independently. The system was equipped with three Sony HR-50 cameras at a frame rate of 60 fps and two infrared flashes, which enabled us to measure the gaze position with under 0.5 visual degrees error, while an accurate gaze output was available for at least ±45 degrees of head rotation. All measurements from the eye tracker were recorded on a separate computer.

Figure 7.2: Images used in the experiments.

The experiment involved one subject per test session. The subject was seated in line with the center of the monitor at the distance of 3.2 and 1.6 times the picture height for HD and UHD contents, respectively, as suggested as optimal viewing distance (OVD) (see Table 2.1), which corresponds to roughly 1.1 m from the monitor in both cases. The eye tracking system was placed at 0.7 m from the monitor such that the face was well captured by the cameras. Figure 7.3 depicts the conditions of the experiments.

At the beginning of the test, the aperture and focus settings of the eye tracker cameras were adjusted for optimal conditions and a full camera calibration was performed to maximize the accuracy of the measurements. For each subject, a personal profile was created by recording several head poses and gaze calibrations using four calibration points close to the screen corners and one at the center of the screen. To ensure the accuracy of the eye tracking data,

Figure 7.3: Experimental setup.

Table 7.1: Arrangement of test sessions for HD and UHD resolutions.

	Group #1	Group #2
First session	UHD resolution	HD resolution
Second session	HD resolution	UHD resolution

subjects were instructed to hold their head still while watching the images, and test operators made sure that all features were correctly detected by at least two out of three cameras during the experiment.

Test Methodology

The experiment was separated into two different sessions to avoid inter-resolution comparison: one sessions was dedicated to UHD resolution only and another session to HD resolution only. To reduce the influence of potential memory effects on visual attention from viewing the same contents twice, the participants were divided into two groups of ten subjects each: the first group watched the images in UHD resolution first and then in HD resolution, whereas the reverse order was considered for the second group. Table 7.1 depicts the arrangement of the test sessions. To reduce contextual effects, the stimuli orders of display were randomized by applying different permutation for each subject. To reduce fatigue effects, each subject took a 15 min break between the two sessions.

According to Engelke et al. (2013), the FDM is almost saturated at about four seconds presentation time. However, since the images used in our experiments were about four times larger than the ones used in (Engelke et al., 2013), it is possible that the subjects are not able to watch all salient regions in the image if the presentation time is too short. Therefore, each image was shown for 15 s in our experiments. Additionally, a two seconds mid-grey background was

displayed prior to the presentation of each test stimuli to reset subject's attention. With this timing, each session was approximately 15 min long.

Since the purpose of these experiments was to investigate the difference in visual attention and viewing strategies for HD and UHD resolutions, subjects were instructed to watch the images in a free-viewing scenario. Additionally, a training session was organized to allow subjects to familiarize with the procedure. The training materials were presented to subjects exactly as for the test materials.

To understand the influence of the memory effect on the subjective data, the following categories of FDMs were analyzed separately

1) UHD-First: Group #1, first session (10 subjects),
2) HD-First: Group #2, first session (10 subjects),
3) UHD-Second: Group #2, second session (10 subjects); they watched UHD contents after watching the same images with HD resolution, followed by a 15 min resting phase,
4) HD-Second: Group #1, second session (10 subjects); they watched HD contents after watching the same images with UHD resolution, followed by a 15 min resting phase,
5) UHD-All: all 20 subjects, and
6) HD-All: all 20 subjects.

Participants

A total of 20 naïve subjects (7 females and 13 males) took part in the experiments. Subjects were between 18 and 28 years old with an average of 23.8 years of age. Before the experiment, a consent form was handed to subjects for signature. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts respectively.

Data Analysis

First, the FDMs were computed following the procedure described in Section 2.8.1. Note that the eye tracking system used in our experiments (see above) automatically discriminates between saccades and fixations based on the gaze velocity information. More specifically, during a time frame, all gaze points associated with gaze velocity below a fixation threshold are classified as fixation points, whereas saccades are detected when the gaze velocity lies above the fixation threshold. Blinks are also detected automatically by the eye tracking system based on the distance between the two eyelids of each eye. Then, the FDMs were compared following the procedures described in Section 2.8.2.

(a) Presented image (b) FDM of HD resolution (HD-all) (c) FDM of UHD resolution (UHD-all)

Figure 7.4: Examples of FDMs for a presentation time of 15 s.

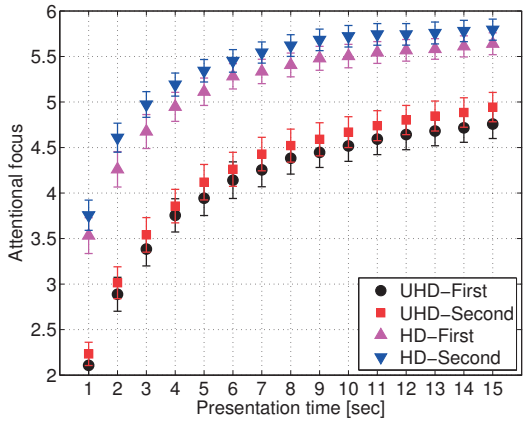


Figure 7.5: Attentional focus of FDMs with CIs.

7.1.3 Results

Figure 7.4 shows examples of FDMs for HD and UHD resolutions computed from the eye-tracking data (across all subject groups). It can be noted from the figure that FDM of UHD resolution is more scattered and more 'focused' compared to FDM of HD resolution. In both cases, subjects look at various objects in the images. However, subjects watched specific objects in UHD images with higher intensity but browsed HD images in a more 'relaxed' way. In the following parts, the FDMs for UHD and HD contents are compared using the following metrics: attentional focus, similarity score, and KLD.

Attentional Focus

Figure 7.5 shows the attentional focus computed separately for categories of FDMs versus varying presentation time. From the figure, it can be noted that, at each presentation time, the attentional focus of UHD resolution has lower value, which means that UHD has lower entropy or higher focus when compared to HD resolution, regardless of the presentation order. A possible explanation is that the higher level of details in UHD images make subjects' attention more focused and concentrated compared to HD images.

Also, attentional focus saturates faster for HD resolution than for UHD resolution, since UHD

Table 7.2: p-value computed for attentional focus (t = 15 s).

	HD-Second	UHD-First	UHD-Second
HD-First	0.074	< 0.001	< 0.001
HD-Second		< 0.001	< 0.001
UHD-First			0.11

resolution images are four times bigger. To estimate the presentation time at which the FDMs are saturated, the attentional focus values were fitted using the response curve of a first order lag system according to the equation

$$f(t) = a\left[1 - \exp\left(-\frac{t}{\tau}\right)\right] + b, \tag{7.1}$$

where t is the presentation time (how long the image was viewed by the subjects), a and b are the amplitude and the offset of the resulted attentional focus curve, and τ is a constant representing the time at which the attentional focus reaches 63.2% of its maximum value.

Considering that the saturation (95% of the maximum value) is achieved at 3τ, the FDMs are saturated after about 10.67 s for HD and after about 13.02 s for UHD. It means that 10 s is not enough to get a stable FDM for UHD resolution and that a longer presentation time is required.

Figure 7.5 shows that there is no influence of presentation order on the attentional focus and there is a difference between UHD and HD resolutions, but it does not show if these findings are statistically significant. To answer this question, an ANOVA was performed on attentional focus results at the presentation time equal to 15 s. ANOVA analysis was done for different pairs of FDMs with results shown in Table 7.2. The table shows that attentional focus is statistically significantly different for HD and UHD resolutions, while the presentation order of HD or UHD content, i.e., the order in which a subject viewed content, does not affect attentional focus in a statistically significant way. It means that even though each image was presented twice to the subjects, the influence of potential memory effects does not significantly impact attentional focus, indicating that the results from both groups of subjects can be combined.

Similarity Score

While attentional focus only measures one FDM, similarity score compares two different FDMs. The similarity scores were computed for all meaningful pairs of FDMs: HD-First vs. UHD-First, HD-Second vs. UHD-Second, HD-First vs. HD-Second, and UHD-First vs. UHD-Second; and the corresponding scores are shown in Figure 7.6 for all presentation times varying from 1 to 15 s. The figure demonstrates that HD-First (HD images were viewed before UHD) is more similar to HD-Second (UHD images were viewed before HD) than UHD-First and UHD-Second FDMs. This high similarity between FDMs for HD can also be noticed visually, for instance by comparing FMD of HD-First in Figure 7.7b of a sample image in Figure 7.7a with

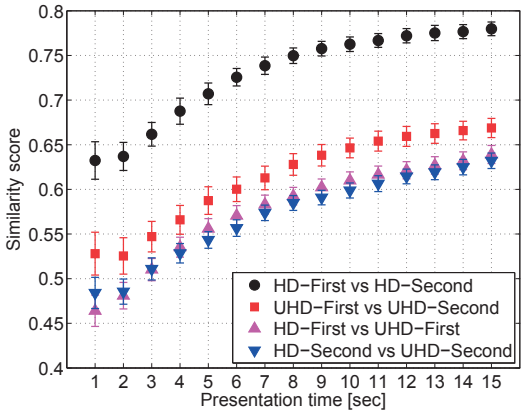


Figure 7.6: Similarity score of FDM pairs with CIs.

(a) Presented image

(b) FDM of HD-First

(d) FDM of UHD-First

(c) FDM of HD-Second

(e) FDM of UHD-Second

Figure 7.7: Examples of FDMs for different resolutions and viewing orders.

FDM of HD-Second in Figure 7.7c. In turn, the FDMs of UHD-First, shown in Figure 7.7d, and UHD-Second, shown in Figure 7.7e, are quite different visually too.

This observation with the fact that other two pairs, HD-First vs. UHD-First and HD-Second vs. UHD-Second have almost the same similarity as UHD-First vs. UHD-Second, indicate that the fixation patterns for UHD resolution have higher diversity compared to HD resolution, i.e., different subjects look at UHD images in many different ways compared to a more unified way of viewing HD images. It also means that the presentation order does not influence the similarity score.

To analyze the statistical significance of the similarity score results, an ANOVA analysis was performed comparing similarity scores between different pairs of FDMs. Table 7.3 reports the p-values. The table demonstrates that all results show statistically significant difference, except when comparing HD-First vs. UHD-First with HD-Second vs. UHD-Second. This analysis

Table 7.3: p-value computed for similarity score (t = 15 s).

	UHD-1st vs. UHD-2nd	HD-1st vs. UHD-1st	HD-2nd vs. UHD-2nd
HD-1st vs. HD-2nd	< 0.001	< 0.001	< 0.001
UHD-1st vs. UHD-2nd		< 0.001	< 0.001
HD-1st vs. UHD-1st			0.30

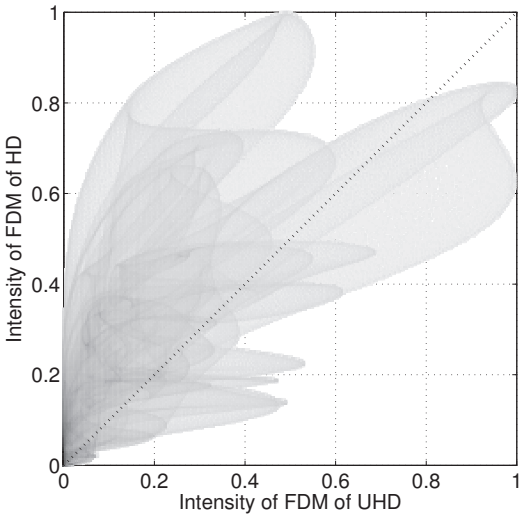


Figure 7.8: Scatter-like plot of the conjoint intensity values between the two FDMs of Figure 7.4.

confirms the observation that there is a significant difference between FDMs of HD and UHD resolutions but the presentation order, in other words, memory effect, has no influence on the results, confirming the conclusions formulated based on attentional focus.

To better understand the dissimilarity between HD and UHD resolutions, scatter-like plot of the conjoint intensity values between two FDMs can be used (Engelke et al., 2013). Figure 7.8 shows such plot for the FDMs given in Figures 7.4b and 7.4c. In this plot, highly correlated FDM values lie closer to the main diagonal (dashed line). As it can be observed, there are several structural dissimilarities, especially for highly fixated points, which are due to the difference in the number of peaks and their respective positions in the actual FDMs.

Also, similarly to attentional focus, estimated τ time constant of similarity scores for HD (τ = 4.2 s) is lower than for UHD resolution (τ = 5.1 s).

Kullback-Leibler Divergence

KLD metric was computed in the same way as similarity scores metric and results are shown in Figure 7.9. As it can be observed, KLD values are clearly saturated after 3 s for both HD and UHD. Since KLD measures the dissimilarity of the two histograms, this metric does not

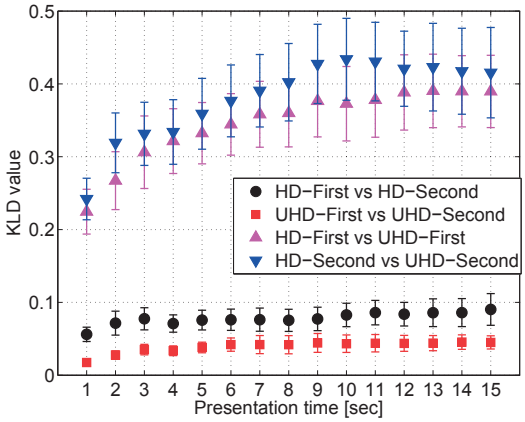


Figure 7.9: KLD of FDM pairs with CIs.

Table 7.4: p-value computed for KLD (t = 15 s).

	UHD-1st vs. UHD-2nd	HD-1st vs. UHD-1st	HD-2nd vs. UHD-2nd
HD-1st vs. HD-2nd	< 0.001	< 0.001	< 0.001
UHD-1st vs. UHD-2nd		< 0.001	< 0.001
HD-1st vs. UHD-1st			0.53

consider the spatial distribution but only evaluates the difference in the number of points of attention and their intensities. Therefore, this metric shows the difference between the viewing strategy of the subjects. Results in Figure 7.9 show very small KLD values when comparing FDMs of the same resolution, which suggests that the strategy to browse the images does not change much across subjects for a specific resolution. The fact that KLD values for UHD resolution (UHD-First vs. UHD-Second pair) are the lowest suggests that subjects are focusing on a fewer attentive regions in UHD compared to HD, probably, due to the higher resolution and higher level of details in UHD images. It can also be noted from the figure that KLD values for HD vs. UHD FDM pairs are much higher than for the same resolution pairs, which means that viewing strategies for HD and UHD resolutions are different.

To investigate the statistical significance of KLD results, an ANOVA analysis was performed in the same way as for similarity score metric. Table 7.4 reports the p-values of the ANOVA analysis for KLD metric computed on different pairs of FDMs. From the table, it is clear that only HD-First vs. UHD-First is not significantly different compared to HD-Second vs. UHD-Second, which, similarly to the earlier observations, means that influence of memory effects is insignificant, but HD is significantly different from UHD.

7.2 Visual Attention in LDR and HDR Images

HDR imaging is able to capture a wide range of luminance values, closer to what the human eye can perceive. Since luminance contrast significantly affects visual attention (Einhäuser and König, 2003), HDR content may lead to different human visual attention patterns compared to LDR content. Despite the recent advances in HDR imaging, there are few reports about the effect of HDR on human visual attention. The first study was reported by Narwaria et al. (2014), who investigated the impact of TMOs on visual attention. They compared the FDMs corresponding to HDR content to that of different tone-mapped LDR versions. They found that TMOs modify visual attention in a significant manner for images and substantially in case of video sequences. Nasiopoulos et al. (2014) investigated the differences in visual fixations between HDR video sequences and their LDR versions produced using Reinhard TMO. Even though subjects reported a clear preference for the HDR display, they found no significant differences in human fixations between HDR and LDR. However, they used rather simple statistical measurements, e.g., number of fixations and fixations duration in area of interest.

Previous studies only considered tone-mapped versions, which were generated from the HDR content using different TMOs, but did not consider a LDR version corresponding to auto-exposure settings, as it would be produced by a DSLR camera. To fill this gap, this section investigates the difference in human visual attention between a HDR image generated from multiple exposure pictures and a single exposure LDR image of the same scene. An eye tracking experiment was conducted with 20 naïve subjects to collect eye movements data for 46 HDR images together with their LDR versions. From the raw eye movements data, the FDMs were computed following the procedure described in Section 2.8.1. To analyze the difference between salient regions in HDR and LDR images, the FDMs were compared using the similarity score metric (see Section 2.8.2). This section reports the details and results of this eye tracking experiment.

7.2.1 Dataset

Although there are several publicly available HDR image dataset, most of them contain only the resulted HDR images without providing the original bracketed LDR images. A few datasets that include original LDR images contain also color artifacts caused by image fusion, visible camera noise, or blurring artifacts caused by moving objects such as cars, moving trees, or walking people. For focus of attention experiments, to obtain practically useful results, a large variety of content is also desirable.

Therefore, in addition to a few selected images from the existing datasets (several images from EMPA HDR Image Database and a few frames from 'Tears of Steel' short film, we have built a new public HDR dataset, called EPFL HDR-Eye dataset, by combining nine bracketed images acquired with several cameras, including Sony DSC-RX100 II, Sony NEX-5N, and Sony α6000, with different exposures settings (-2.7, -2, -1.3, -0.7, 0, 0.7, 1.3, 2, 2.7 EV). Several images (obtained with Nikon D70 camera) from PEViD-HDR dataset (Korshunov et al., 2014) that

Table 7.5: Dynamic range of the scenes in the dataset

Dynamic range (dB)	Number of scenes
<48	13
48-60	7
60-72	8
72-84	11
>84	7

shows different people under different lighting conditions were also used.

To avoid ghost artifacts in the fused HDR images due to camera shake and moving items, the cameras were placed on a tripod and special care was taken to avoid moving objects appearing in the pictures during the shooting. The open source Pictureaunt 3.2 software was used for linearizing the bracketed exposures using the inverse of the camera response, and combining them into a single radiance map. For the better picture quality of fused images, ghost removal and image alignment provided by the software were used.

The resulted dataset contains 46 images that cover a wide variety of content, e.g., natural scenes (both indoor and outdoor), humans, stained glass, sculptures, historical buildings, etc. Table 7.5 provides dynamic ranges of the scenes in the dataset.

Brightness Adjustment

To reflect the real luminance of actual scenes, HDR images need to be reproduced with physically correct values using measured data, as suggested by Akyüz et al. (2007). However, most of the selected HDR pictures do not have this data, and the HDR monitor used in the test is not capable of generating more than 4000 cd/m^2. This peak luminance is not sufficient to display some of the bright scenes. Therefore, to make all HDR pictures look visually acceptable on the HDR monitor, the brightness of the HDR images was adjusted in accordance with the following equation (Krawczyk et al., 2007)

$$\log R_{new} = \log R - f(L) + c \tag{7.2}$$

with

$$f(L) = 0.28 \cdot L^{[9]} + 0.37 \cdot L^{[42]} + 0.35 \cdot L^{[100]} \tag{7.3}$$

where R and R_{new} are the original linear and adjusted luminance values, L represents the original logarithm luminance values, $L^{[p]}$ denotes the p-th percentile of the original logarithm luminance values, and c is a target logarithm luminance on a display. This approach can be interpreted more intuitively as the logarithm luminance of the original image being scaled in a way that matches the target logarithm luminance of the display. According to the literature, to estimate the best preferred brightness, the reference logarithm luminance of the original

image $f(L)$ has to be computed based on the relative distribution of low, high, and mid-tones of the images, as shown in Equation (7.3), and 60% of the white luminance of the display is used as a target luminance c. We used 2000 cd/m^2 as white luminance, since this value was used for the color calibration of the monitor

To display LDR contents with the HDR monitor, the LDR images were converted into radiance map representation. The images were first linearized with a typical gamma curve ($\gamma = 2.2$), then the pixel values were adjusted proportionally so that theoretical maximum pixel values of LDR image can match the peak luminance of common LDR monitor. ITU-R BT.2022 (2012) specifies optimal peak luminance between 70 and 250 cd/m^2 in general viewing condition. We chose 120 cd/m^2 as the peak luminance since it is the default value in most monitor calibration software. Assuming that the LDR images taken with middle exposure setting of 0 EV are the most common LDR images, we used middle-exposed LDR images in radiance format in the eye tracking experiments.

7.2.2 Methodology

The eye tracking experiments were conducted at the MMSPG test laboratory, which fulfills the recommendations for subjective evaluation of visual data issued by ITU. The laboratory is equipped with a controlled lighting system with a 6500 K color temperature, a mid gray color is used for all background walls and curtains, and the ambient illumination did not directly reflect off of the monitor. During the experiment, the background luminance behind the monitor was set to 20 lx. The test room was separated in two by a curtain to isolate the subject and equipment from the test operators, which were present during the test session to supervise the recording of the eye tracking data. The laboratory setup was intended to ensure the reproducibility of the results and to avoid unintended influence of external factors.

To display the test stimuli, a full HD 47" SIM2 HDR monitor with individually controlled LED backlight modulation, capable of displaying content with luminance values ranging from 0.001 to 4000 cd/m^2, was used. Prior to subjective tests, following a warm-up phase of an hour, a color calibration of the HDR display was performed using the software provided by SIM2. The red, green, and blue primaries were measured for white set to 2000 cd/m^2, which corresponded to the maximum value of the measurement probe (X-Rite i1Display Pro).

A Smart Eye Pro 5.8 remote eye tracking system was used to record the eye movements. The system was equipped with three Sony HR-50 cameras at a frame rate of 60 fps and two infrared flashes, which enabled us to measure the gaze position with under 0.5 visual degrees error, while an accurate gaze output was available for at least ±45 degrees of head rotation. All measurements were recorded on a separate computer.

The experiment involved one subject per test session. The subjects were seated in line with the center of the monitor, at a distance of 3.2 times the picture height (see Table 2.1), corresponding to roughly 1.9 m meters from the monitor. The eye tracking system was placed at 0.7 m from

Figure 7.10: Experimental setup.

the monitor such that the face was well captured by the cameras. Figure 7.10 depicts the conditions of the experiments.

At the beginning of the test, the aperture and focus settings of the eye tracker cameras were adjusted for optimal conditions and a full camera calibration was performed to maximize the accuracy of the measurements. For each subject, a personal profile was created by recording several head poses and gaze calibrations using four calibration points close to the screen corners and one at the center of the screen. To ensure the accuracy of the eye tracking data, subjects were instructed to hold their head still while watching the images, and test operators made sure that all features were correctly detected by at least two out of three cameras during the experiment.

Test Methodology

Each image was shown for 12 s in the experiments. Additionally, a two seconds mid-grey background was displayed prior to the presentation of each test stimuli to reset subject's attention. Each subject participated in two sessions of 13 minutes each with a 15 minutes break in between. All 46 contents were viewed by each subject in one session, and both HDR and LDR contents were displayed in the same session in a random order. Also, for half of the tested images, their HDR versions were displayed in the first session followed by the corresponding LDR versions in the second session. And for the other half of the images, the order was reversed: LDR versions were shown during the first session and HDR during the second. This approach was used to reduce the influence of potential memory effects on visual attention from viewing the same content twice. To reduce contextual effects, the stimuli orders of display were randomized applying different permutation for each subject.

Since the purpose of these experiments was to investigate the difference in visual attention

between HDR and LDR, subjects were instructed to watch the images in a free-viewing scenario. Additionally, a training session was organized to allow subjects to familiarize with the procedure. The training materials were presented to subjects exactly as for the test materials.

Participants

A total of 20 naïve subjects took part in the experiments. Subjects were aged between 18 and 56 years old with an average of 25.3 years of age. Before the experiment, a consent form was handed to subjects for signature. All subjects were screened for correct visual acuity and color vision using Snellen and Ishihara charts, respectively.

Data Analysis

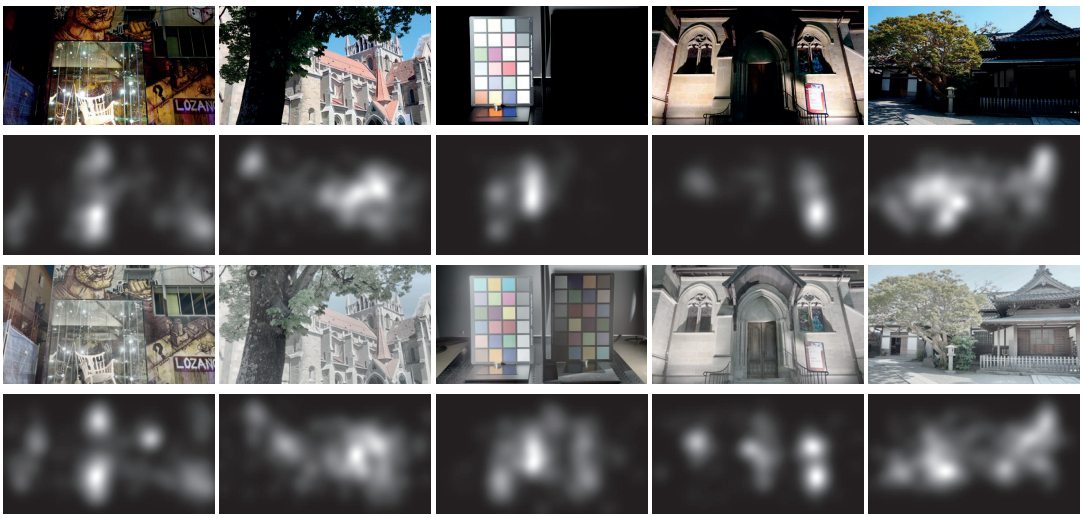
First, the FDMs were computed following the procedure described in Section 2.8.1. Note that the eye tracking system used in our experiments (see above) automatically discriminates between saccades and fixations based on the gaze velocity information. More specifically, during a time frame, all gaze points associated with gaze velocity below a fixation threshold are classified as fixation points, whereas saccades are detected when the gaze velocity lies above the fixation threshold. Blinks are also detected automatically by the eye tracking system based on the distance between the two eyelids of each eye. Then, the FDMs were compared using the similarity score metric (see Section 2.8.2).

7.2.3 Results

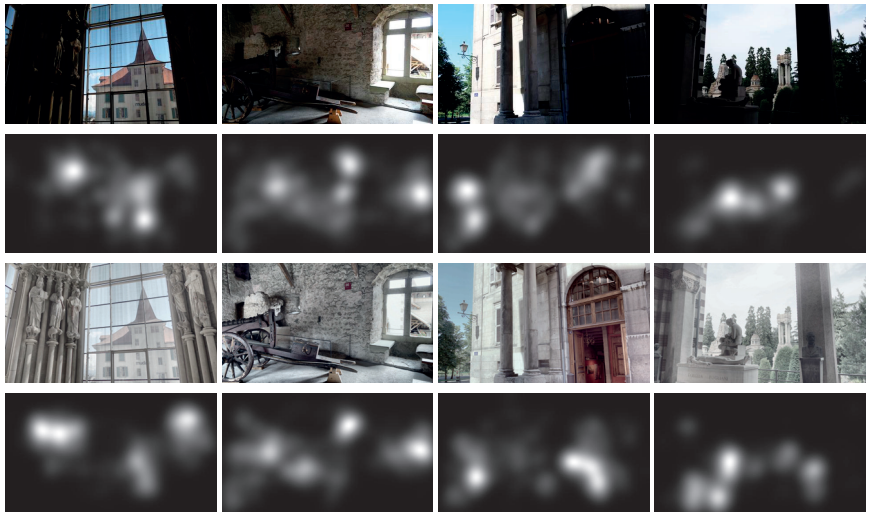
Figure 7.11 shows the LDR image, LDR FMD, tone mapped HDR image, and HDR FDM for contents exhibiting significant differences between LDR and HDR. In these examples, different FDM patterns can be observed, depending on scene characteristics. For example, it can be noted that viewers looked at more objects in some HDR images, e.g., the color chart in the dark part of content *C09* or the inscription below the statue on content *C40*.

While results show that viewers tend to look more at the bright objects in LDR images, details in the dark regions become more visible in HDR, resulting in the increased visual attention in these areas. This effect can be observed for content *C10* where viewers looked more attentively at the entrance door of the cathedral. Also, in some contents, focus of attention can shift from the bright areas of the LDR image to details in the darker areas of the HDR image. For example, in content *C16*, the attention was mostly focused on the building visible through the window in the LDR image, whereas the viewers mostly looked at the details of the statues located in the darker parts on the right and left side of the HDR image.

For some contents, the HDR FDM is mostly a modulated version of the LDR FDM, i.e., viewers looked at the same objects in both cases but with a different intensity. On the other hand, some contents did not show any significant difference between LDR and HDR FDMs. In particular,



(a) *C05* (b) *C08* (c) *C09* (d) *C10* (e) *C11*



(f) *C16* (g) *C22* (h) *C32* (i) *C40*

Figure 7.11: Examples showing significant visual differences between FDMs for HDR images and FDMs for LDR versions. First row: LDR version, second row: FDM of LDR, third row: tone-mapped HDR image, fourth row: FDM of HDR.

scenes containing human faces do not show any difference, as humans are very sensitive to human faces and are able to detect silhouettes easily, even in the dark regions.

Based on these observations, three clusters were manually created: (i) scenes that induce a change in visual attention pattern, (ii) scenes that induce a change in fixation intensity, and (iii) scenes that induce similar visual attention between LDR and HDR. Table 7.6 reports mean similarity score and its deviation computed on the images from these three clusters. From the table, it can be noted that the similarity score is lower when a change in the visual attention pattern or fixation intensity is observed in the FDMs. However, the difference between similarity scores for different clusters is not very large, which also indicates that

Table 7.6: Average similarity score between the FDMs of LDR and HDR.

Change in	Nb scenes	Similarity score	
		mean	std
Visual attention pattern	9	0.6742	0.1101
Fixation intensity	14	0.7447	0.0359
No change	23	0.7720	0.0379

Table 7.7: *p*-value between each cluster of similarity score.

	Fixation intensity	No change
Visual pattern	0.0358	0.0007
Fixation intensity		0.0372

similarity metric may not be the most suitable metric (note that FDMs in Figure 7.11) are visually different for LDR and HDR versions) to measure the changes in FDM that are caused by HDR.

To determine whether the difference between the three clusters is statistically significantly, an ANOVA was performed on the similarity scores. As shown in Table 7.7, the computed *p*-values indicate that the similarity scores are significantly different between the three clusters, in particular, between the scenes corresponding to visual attention pattern cluster and the scenes from the 'no change' cluster with LDR and HDR having similar FDMs. These findings show that, for some contents, HDR imaging impacts visual attention significantly, but it is not clear whether existing measurement tools can adequately measure this impact.

7.3 Conclusion

This chapter investigated the impact of some immersive technologies on visual attention. In particular, we studied the influence of UHD resolution on human visual attention. We conducted subjective eye tracking experiments with both HD and UHD resolution images covering a wide variety of scenes. We then computed the FDMs for HD and UHD images and compared them using three different statistical evaluation metrics: attentional focus, similarity score, and KLD. The assessment results demonstrated that (i) UHD resolution images can grab the focus of attention more than HD images, (ii) humans tend to look at a few attentive regions in the images with more intent when viewing UHD, and (iii) viewing strategy is different for HD and UHD.

In this chapter, we also investigated the impact of HDR imaging on human visual attention. For this purpose, a public HDR image dataset with images of wide variety of natural scenes was created. The dataset also contains original bracketed LDR images and FDMs from the eye tracking experiment. The eye tracking test demonstrated that FDMs of HDR images for some scenes are significantly different from the FDMs of the corresponding LDR versions. Three

clusters of HDR images were then identified: (i) with FDMs having different visual attention pattern compared to FDMs of LDR versions, (ii) with FDMs showing different distribution of fixation intensities when compared to FDMs of LDR versions, and (iii) with FDMs that are similar to FDMs of LDR images. The similarity metric demonstrated that these clusters are dissimilar in statistically significant way. However, the similarity scores for clusters (i) and (ii) are not as small compared to cluster (iii) as it was expected, which means the metric did not capture the difference between FDMs adequately. Therefore, the impact of HDR on human visual attention is scene-dependent and it is hard to measure it using existing statistical evaluation metrics.

Predicting Quality of Experience Part II

8 Objective Quality Metrics

Subjective visual experiments are time consuming, expensive, and not always feasible (for example, in the context of real-time streaming). Therefore, objective quality metrics are needed to predict perceived visual quality. The result of execution of a particular objective metric is an objective quality rating (OQR), which is expected to be the estimation of the MOS (see Section 2.6.2) corresponding to an image or video sequence.

Algorithm design usually follows either a bottom-up or a top-down approach, though some algorithms also use a hybrid design and combine the two approaches. The most common approach consists in studying the visual pathways in the HVS, from the eyes to the brain, through the lateral geniculate nucleus and visual cortex. All relevant components and their characteristics, e.g., light adaptation, contrast sensitivity, color perception, and spatial and temporal masking, are then modeled as basic blocks. The basic blocks are assembled to form several subsystems and different levels of subsystems are combined to build a computational system that mimics the HVS for a specific task. This is the bottom-up approach, also referred to as the psychophysical approach, and is similar to building complex objects from simple LEGO bricks.

The other approach considers the HVS as a black box and tries to formulates an overview of the system, e.g., the input-output relationship, and makes some high-level assumptions on its mechanisms, without detailing any first-level subsystems. The subsystems are also treated as black boxes and can be further refined in greater detail, using several additional subsystem levels, until the entire specification is reduced to basic elements. Hypotheses about the different HVS functionalities are made and implemented, but may be different from what the HVS really does. This is the top-down approach, also referred to as the engineering approach, and is similar to reverse engineering a system to gain insight into its compositional subsystems. Top-down models primarily extract and analyze features, e.g., structural elements, contours, colors, and so on. The top-down approach may provide much simpler solutions than the bottom-up approach.

Depending on the amount of information required about the source reference image or video

sequence, objective quality metrics can be classified into three categories

(i) FR metrics, which compare the test image or video sequence with its source reference. The majority of objective metrics fall into this category. However, it is impossible to implement such metrics in practical situations where the source reference is not available.
(ii) NR metrics, which do not use any information about the source reference. Therefore, these metrics can be used anywhere. Nevertheless, they are more complicated owing to the difficulty of distinguishing between distortions and actual content.
(iii) Reduced-reference (RR) metrics, which have access to a number of features from the source reference, extract the same features from the test image or video sequence, and compare them. Thus, this category lies in between the two extremes.

Different media types have different properties that need to be considered in the quality assessment, for example to assess spatial, color, temporal, or depth distortions. Thus, specific objective quality metrics have been designed for specific tasks, e.g., image, video, HDR, or 3D quality assessment. However, image quality metrics can be adapted to assess the quality of video sequences or stereoscopic images by pooling the individual scores across frames and views, respectively, or to assess HDR content by considering nonlinearities of the HVS. Conversely, some video quality metrics can also be computed on still images.

This chapter describes some of the most common image and video quality metrics. A detailed description of the HDR quality metrics released so far is also provided. Finally, the last part gives an overview of the different metrics proposed for 3D quality assessment. Some of the objective quality metrics described in this chapter were benchmarked in the performance evaluations reported in the rest of the thesis.

8.1 Image Quality Metrics

A large number of objective quality metrics has been proposed over the years for the purpose of image quality assessment. From simple difference measures working on a pixel based level to more complex algorithms that aim reproducing different characteristics of the HVS, a wide variety of metrics has been developed. Most metrics are computed on the luma component only, but other metrics are computed in different color spaces and take color information into account. Simple pixel based metrics are very fast to compute, whereas there are more complex algorithms that require more processing time and power. Additionally, a few metrics also rely on machine learning to better predict perceived quality and use a ground truth dataset to training the algorithm. Specific metrics have been also developed for particular applications, e.g., quality assessment of discrete cosine transform based image compression schemes (Watson, 1993), or specific visual artifacts, e.g., blurriness and blockiness.

The goal of this section is not to provide a complete review of all existing image quality metrics. For this purpose, the reader is invited to have a look at the following text books and overview

papers for a more detailed and more comprehensive description of image quality metrics: Engelke and Zepernick (2007); W. Lin and Kuo (2011); Pedersen and Hardeberg (2009); Z. Wang and Bovik (2006); Winkler (2005); H. R. Wu and Rao (2005). The goal is rather to list the most common metrics, as well as those used in the rest of the thesis, to classify them based on their characteristics, and to briefly described these characteristics.

8.1.1 Full-Reference Metrics

The following FR metrics can be divided into different categories: difference measures and statistical-oriented metrics, structural similarity measures, visual information measures, information weighted metrics, HVS inspired metrics, and objective color difference measures (studied in the vision science).

(i) Difference measures and statistical-oriented metrics
These metrics are based on pixel color differences and provide a measure of the difference between the reference image and the distorted image
- mean squared error (MSE): mean squared error,
- PSNR: peak signal-to-noise ratio, and
- SNR: signal-to-noise ratio.

(ii) Structural similarity measures
These metrics model the quality based on pixel statistics to model the luminance (using the mean), the contrast (variance), and the structure (cross-correlation) (Laparra et al., 2010).
- UQI: universal quality index (Z. Wang and Bovik, 2002),
- SSIM: structural similarity index (Z. Wang et al., 2004),
- MS-SSIM: multiscale SSIM index (Z. Wang et al., 2003),
- M-SVD: measure - singular value decomposition (Shnayderman et al., 2006), and
- QILV: quality index on local variance (Aja-Fernandéz et al., 2006).

The MS-SSIM index is a multiscale extension of SSIM, which has a higher correlation with perceived quality when compared to SSIM. It is a perceptual metric based on the content features extraction and abstraction. This quality metric considers that the HVS uses the structural information from a scene (Z. Wang et al., 2004). The structure of objects in the scene can be represented by their attributes, which are independent of both contrast and average luminance. Hence, the changes in the structural information from the reference and distorted images can be perceived as a measure of distortion. The MS-SSIM algorithm calculates multiple SSIM values at multiple image scales. By running the algorithm at different scales, the quality of the image is evaluated for different viewing distances. MS-SSIM also puts less emphasis on the luminance component when compared to contrast and structure components (Z. Wang et al., 2003).

(iii) Visual information measures
These metrics aim at measuring the image information by modeling the psycho-visual features of the HVS or by measuring the information fidelity. Then, the models are

applied to the reference and distorted images, resulting in a measure of the difference between them

- IFC: image fidelity criterion (Sheikh et al., 2005),
- VIF: visual information fidelity (Sheikh and Bovik, 2006),
- VIFp: VIF pixel-based version (Sheikh and Bovik, 2006), and
- FSIM: feature similarity index (L. Zhang et al., 2011).

The VIF criterion analyses the natural scene statistics, using an image degradation model and the HVS model. This FR metric is based on the quantification of the Shannon information present in both the reference and the distorted images. VIFP is derived from the VIF criterion.

FSIM is a perceptual metric that results from SSIM. FSIM adds the comparison of low-level feature sets between the reference and the distorted images (L. Zhang et al., 2011). Hence, FSIM analyzes the high phase congruency extracting highly informative features and the gradient magnitude, to encode the contrast information. This analysis is complementary and reflects different aspects of the HVS in assessing the local quality of an image.

(iv) Information weighted metrics

The metrics in this category are based on the modeling of relative local importance of the image information. As not all regions of the image have the same importance in the perception of distortion, the image differences computed by any metrics have allocated local weights resulting in a more perceptual measure of quality

- IW-MSE: information content weighting MSE (Z. Wang and Q. Li, 2011),
- IW-PSNR: information content weighting PSNR (Z. Wang and Q. Li, 2011), and
- IW-SSIM: information content weighting SSIM (Z. Wang and Q. Li, 2011).

(v) HVS inspired metrics

These metrics try to model empirically the human perception of images from natural scenes

- JND_st: just noticeable distortion (X. Yang et al., 2005),
- WSNR: weighted SNR (Mannos and Sakrison, 1974; Mitsa and Varkur, 1993), and
- DN: divisive normalization (Laparra et al., 2010).

(vi) Objective color difference measures

The color difference metrics were developed because the CIE1976 color difference (CIE1986) magnitude in different regions of the color space did not appear correlated with perceived colors. These metrics were designed to compensate the nonlinearities of the HVS present on the CIE1976 model

- CIE1976 (CIE1986),
- CIE94 (CIE1995),
- CMC (Clarke et al., 1984), and
- CIEDE2000 (Luo et al., 2001).

The CIEDE2000 metric is a color difference measure that not only includes weighting factors for lightness, chroma, and hue, but also factors to handle the relationship between chroma and hue. The CIEDE2000 computation is not reliable in all color spaces.

However, in this case it can be used because the tested images are represented in the CIELAB color space that allows a precise computation.

8.1.2 No-Reference Metrics

The following NR metrics are based on the analysis of a set of well-known sharpness measures

- JND: just noticeable distortion (X. Yang et al., 2005),
- variance (Erasmus and Smith, 1982),
- laplacian (Batten, 2000),
- gradient (Batten, 2000),
- frequency threshold metric (Murthy and Karam, 2010),
- HP Metric (Shaked and Tastl, 2005),
- marziliano blurring metric (Murthy and Karam, 2010),
- kurtosis based metric (N. Zhang et al., 2003),
- kurtosis of wavelet coefficients (Ferzli et al., 2005),
- auto correlation (Batten, 2000), and
- Riemannian tensor based metric (Ferzli and Karam, 2007).

8.2 Video Quality Metrics

The quality assessment of video sequences is a more complex problem than that of still images because of the temporal dimension and increased amount of data. Thus, a common approach consists of performing a frame-by-frame analysis using a standard image quality metrics and pooling the individual scores to compute an overall quality score. The simplest and yet most common pooling strategy consists of computing the average quality score across all frames. However, this simplistic strategy tends to smooth out temporal quality fluctuations, which can severely impact the overall quality perception. To overcome this problem, several pooling strategies have been proposed and investigated (Rimac-Drlje et al., 2009; Rohaly et al., 1999). In applications where quality can fluctuate largely along the temporal domain, e.g., dynamic adaptive streaming (Seufert et al., 2013), a proper pooling strategy is necessary. Nevertheless, in many cases, the average pooling is sufficient, especially when considering short video sequences, e.g., 10 s long, which are widely used in subjective visual experiments.

More complex video quality algorithms have been proposed to take into account the objects' motion and perform some sort of temporal alignment before assessing quality across frames, e.g., the TetraVQM metric (Barkowsky et al., 2009). Other strategies based on optical flow or simply temporal slicing have been proposed. The reader is invited to have a look at the following text books and overview papers for a more detailed and more comprehensive description of video quality metrics: (Chikkerur et al., 2011; Engelke and Zepernick, 2007; W. Lin and Kuo, 2011; Winkler, 2005; H. R. Wu and Rao, 2005).

The most common video quality metrics, which are also considered as the state-of-the-art, are

- VQM: video quality metric (Pinson and Wolf, 2004) and
- MOVIE: motion-based video integrity evaluation (Seshadrinathan and Bovik, 2010)

The VQM metric was developed by the Institute for Telecommunication Services (ITS), the research and engineering branch of the National Telecommunications and Information Administration (NTIA). The metric provides different models, e.g., Television Model, General Model, and Video Conferencing Model, for different applications. To better predict video quality, the different models have different calibration options prior to feature extraction. The VQM metric configured using the General Model was the only metric that provided a linear correlation coefficient above 0.9 on the VQEG FRTV Phase II database (VQEG, 2003). Thus, the NTIA VQM General Model was adopted by the American National Standards Institute (ANSI) as a national standard in 2003 (ANSI-T1.801.03, 2003) and as ITU recommendations in 2004 (ITU-T J.144, 2004; ITU-T BT.1683, 2006).

The General Model considers the following distortions: blur, block, jerky/unnatural motion, noise in luminance and chrominance channels, and error blocks (e.g., transmission errors). Blur is measured using an information filter (SI13), which is a perceptually significant edge impairment filter. The shift of horizontal and vertical edges with respect to diagonal orientation due to strong blurring is also measured using the SI13 filter to detect jerky/unnatural motion. Additionally, the shift of edges from the diagonal to horizontal and vertical orientations due to tiling or blocking artifacts is also measured using the SI13 filter. To estimate color impairments, the chroma distribution is computed over blocks of 8×8 pixels. These features are also used to estimate localized color impairments, e.g., transmission errors. The model also considers enhancements, e.g., edge sharpening, that might results in quality improvements. The amount of spatial detail is used to modulate temporal impairments and also combined with contrast information to compute temporal distortion. Finally, the overall quality score is computed using a linear combination of all impairments scores.

The MOVIE metric models the middle temporal visual area of the visual cortex, which is critical for the perception of video quality, using separable Gabor filter banks. This filter bank is applied on both the reference and distorted video sequences and the squared difference of their coefficients is used to capture spatial distortions. The local energy content is used as a masking coefficient to normalize spatial distortions. A Gaussian filter is also applied on the reference and distorted video sequences and their squared difference is used to capture low frequency distortions. The Gabor and Gaussian error measures are pooled to give the spatial error measure. The ratio of standard deviation to mean of the spatial error is computed at each frame. Then, the results are averaged across all frames to generate the spatial MOVIE index. The optical flow fields of the reference video and Gabor coefficients are used to estimate temporal distortions. The temporal distortions are pooled at a frame level and averaged across all frames to generate the temporal MOVIE index. The product of the spatial and temporal indexes gives the final MOVIE index.

8.3 HDR Quality Metrics

To overcome the lack of HDR objective metrics, LDR metrics, e.g., PSNR, were also used to evaluate HDR quality, especially in early HDR studies. However, LDR metrics are designed for gamma encoded images, typically having luminance values in the range 0.1-100 cd/m^2, while HDR images have linear values and are meant to capture a much wider range of luminance. Originally, gamma encoding was developed to compensate for the characteristics of CRT displays, but it also takes advantage of the non-linearity in the HVS to optimize quantization when encoding an image (Poynton, 2012). Under common illumination conditions, the HVS is more sensitive to relative differences between darker than brighter tones. According to Weber's law, the HVS sensitivity approximately follows a logarithm function for light luminance values (Shevell, 2003). Therefore, in several studies, LDR metrics have been computed in the log domain to predict HDR quality. However, at the darkest levels, the HVS sensitivity is closer to a square-root behavior, according to Rose–DeVries law (De Vries, 1943; Rose, 1948).

To extend the range of LDR metrics and to consider the sensitivity of the HVS, Aydın et al. (2008) have proposed the PU encoding. Other transforms that map absolute luminance values to perceptual codewords, e.g., the Dolby PQ (Miller et al., 2013), can be used instead of the PU encoding. Another approach to apply LDR metrics on HDR images was proposed by Munkberg et al. (2006). This technique consists in tone-mapping the HDR image to several LDR images with different exposure ranges and to take the average objective score computed on each exposure. However, this approach is more time consuming and requires more computational power, proportionally to the number of exposures.

The only true HDR quality metrics proposed so far are

- HDR-VDP: high dynamic range visible difference predictor (Mantiuk et al., 2005; Mantiuk et al., 2011; Narwaria et al., 2015a),
- DRIM: dynamic range independent metric (Aydin et al., 2008), and
- HDR-VQM: an objective quality measure for high dynamic range video (Narwaria et al., 2015b).

Note that only HDR-VDP and HDR-VQM have a publicly available implementation.

The original HDR-VDP metric (Mantiuk et al., 2005) was the first metric designed for HDR content. It is an extension of the VDP model (Daly, 1992) that considers a light-adaptive contrast sensitivity function (CSF), which is necessary for HDR content as the ranges of light adaptation can vary substantially. The metric was further extended (Mantiuk et al., 2011) with different features, including a specific model of the point spread function (PSF) of the optics of the eye, as human optical lens flare can be very strong in high contrast HDR content. The front-end amplitude non-linearity is based on integration of the Weber-Fechner law. HDR-VDP is a calibrated metric and takes into account the angular resolution. The metric uses a multi-scale decomposition. A neural noise block is defined to calculate per-pixel probabilities maps of visibility and the predicted quality metric.

The DRIM metric (Aydin et al., 2008) detects visible changes in the image structure that belongs to the following categories: loss of visible contrast, amplification of invisible contrast, and reversal of visible contrast. To detect changes in contrast, the HDR-VDP detection model is used and re-calibrated using a set of basic visual stimuli, e.g., Gabor patches, with ground truth detection thresholds. Similarly to HDR-VDP, the metric splits the perceptually normalized response into several bands of different orientation and spatial bandwidth. Finally, the results are pooled over the different bands to generate a distortion map that will predict the three types of visible distortions (one per color channel). Nevertheless, the distortion map is difficult to interpret and there is no pooling of its values into a single quality score that can be used to predict perceived quality.

HDR-VQM was designed for quality assessment of HDR video content. The metric is computed in the PU space and relies on a multi-scale and multi-orientations analysis, similarly to HDR-VDP, based on a subband decomposition using log-Gabor filters to estimate the subband errors. The subband errors are pooled over non-overlapping spatio-temporal tubes to account for short-term memory effects. Further spatial and long-term temporal poolings are performed to compute the overall quality score. In the case of still images, only spatial pooling is performed.

8.4 3D Quality Metrics

Quality assessment of monoscopic images and video sequences has been widely investigated and several 2D metrics have been proposed over the years. Hence, early studies on 3D quality assessment were focused on assessing the performance of 2D metrics considering several possible ways of combining the scores from the left and right views of a stereo pair. An early study from Campisi et al. (2007) compared the *average* approach against the *main eye* approach and the *visual acuity* approach. The results have shown no performance improvement over the *average* approach, which is commonly used. Hewage et al. (2009) investigated objective quality assessment of 3D video sequences represented in the video plus depth (2D+Z) format. They established that the quality of the color video was more correlated with perceived quality than the average quality of the rendered left and right views. To consider asymmetric distortions, You et al. (2010) benchmarked eleven 2D metrics on stereo pairs where only the right view was degraded. An evaluation of the impact of coding artifacts on stereoscopic 3D video quality with 2D objective metrics is reported in (K. Wang et al., 2013). The authors established that the 2D stereoscopic video quality seems to be a dominant factor in the overall 3D quality estimation, while Bosc et al. (2012b) have recently shown that most state-of-the-art 2D metrics fail at predicting perceived quality of synthesized views.

To consider depth quality, a few studies have proposed to compare the original and degraded disparity maps corresponding to the original and degraded stereo pairs, respectively. Benoit et al. (2008) have computed the correlation coefficient between the original and degraded disparity maps for the depth quality score and the average quality score of the left and right views for the image quality score. Similar or slightly lower performance is obtained when

considering both image and depth quality as opposed to considering only the image quality. A similar study was conducted by You et al. (2010) considering several metrics for the image and depth quality. It was found that an appropriate combination of the image and depth quality performs better than using only the image quality or depth quality alone. To further improve the correlation with perceived quality, image and depth quality scores can also be combined at pixel level (Benoit et al., 2008; X. Wang et al., 2011; You et al., 2010).

Further investigations have been conducted to incorporate directly the depth information in the estimation of image quality. For example, a technique based on 3D discrete cosine transform, contrast sensitivity function, and luminance masking is proposed in (Jin et al., 2011). The authors reported a significant gain of the proposed metric over state-of-the-art 2D metrics. Ryu et al. (2012) presented a stereo-version of SSIM based on binocular quality perception. Three components, luminance, contrast, and structural similarities, are combined into a quality index using a binocular quality perception model.

Different metrics have been proposed based on the so-called cyclopean image, which is a single mental image obtained from the fusion of the two images received from the two eyes. The earliest work on the subject was reported by Boev et al. (2006), who produced monoscopic and stereoscopic quality scores based on the cyclopean image and perceptual depth map, respectively. Later, Maalouf and Larabi (2011) used a metric based on a wavelet transform and contrast sensitivity function to measure the quality of the cyclopean image and a coherence measure for the perceptual depth map. A hybrid approach that performs a weighted sum of the qualities of the individual views, the cyclopean image, and the depth maps is proposed in (Banitalebi-Dehkordi et al., 2012). The metric proposed by M.-J. Chen et al. (2013a) relies on the cyclopean image and a binocular rivalry model to account for asymmetric distortions. Battisti et al. (2015b) have proposed an algorithm that considers the qualities of the cyclopean image and disparity maps, as well as binocular rivalry. Several quality metrics are used to generate different features and a linear regression is performed to compute the overall quality score.

Another set of metric relies on a binocular energy model, which provides a good description of the first stages of cortical binocular processing (Hibbard, 2008). The metric proposed by Bensalma and Larabi (2013) aims at reproducing the binocular signal generated by simple and complex cells to estimate the associated binocular energy. The simple cells in the visual cortex were modeled using spatial-frequency transforms, e.g., discrete wavelet transform, and their output is integrated by complex cells to perform binocular fusion. This process was modeled by finding for each simple cell of the dominant eye the corresponding cell in the other eye that maximizes the binocular energy. An improved version of this metric was proposed by Perera et al. (2014) to discriminate between orientations and sizes at the complex cells level and to include the binocular suppression theory. The metric proposed by Shao et al. (2014) relies on binocular energy and binocular masking. The binocular energy is computed for the original and distorted stereo pairs based on Gabor filter responses and the disparity maps. Their difference is modulated considering binocular masking to form the overall quality score.

To account for different quality levels between the left and right views, binocular rivalry was incorporated in stereoscopic quality metrics. For example, J. Yang et al. (2015) have built a model based on the combination of the binocular summation and differentiation channels to compute the overall quality score. The metric proposed by Shao et al. (2013) performs a left–right consistency check to classify the image into non-corresponding, binocular fusion, and binocular suppression regions. Then, each region is evaluated independently by considering its binocular perception property and results are pooled into an overall score. H. Lin and L. Wu (2014) proposed a model based on gain-Control theory and binocular frequency integration behaviors to account for asymmetric distortions. In particular, a difference-of-Gaussian decomposition is performed for each view and a standard quality metric is computed for each component. Finally, the scores are linearly combined using the gain-Control theory to form the overall quality value.

Machine learning is also becoming more and more popular in the design of objective metrics. In the case of 3D quality assessment, De Silva et al. (2013) have proposed a FR metric that estimates structural distortions, asymmetric blur, and spatial/temporal (depth) information to estimate the quality of a stereo pair. A deep learning approach was used to predict 3D quality in (Mocanu et al., 2014). The authors have modeled the left and right eye images separately using two layers of visible neurons, which are superimposed by a layer of hidden neurons, using three way multiplicative interactions. Khan Md et al. (2015) have proposed an algorithm that introduces natural stereo scene statistics. The metric relies on a generalized Gaussian distribution to fit the luminance wavelet coefficients and disparity information to improve the prediction performance.

While the metrics described here above are FR metrics, a few RR metrics have also been proposed. Maalouf and Larabi (2011) proposed an algorithm based on the discrepancy in the disparity maps and the perceptual difference between the reference and the distorted cyclopean images. The metric proposed by Zhou et al. (2014) relies on image structure variation in the stereoscopic pair, which is estimated using the relation of the horizontal and vertical components of the gradient vectors between the two views, and the stereoscopic perception quality variation, estimated using the disparity map. disparity map of the stereoscopic image, is used to reflect the stereoscopic perception quality variation Qi et al. (2015) have proposed to compute the entropy of each view to represent monocular cue and their mutual information to represent binocular cue. The difference of the monocular and binocular cues between the original and distorted image pairs is used to construct a perceptual loss vector and a support vector regression is used to estimate the quality score.

Several NR stereoscopic quality metrics were also proposed. For example, Sazzad et al. (2010) proposed an algorithm based on segmented local features of artifacts and disparity, whereas Silva et al. (2015) relied on a combination of a blockiness metric, a disparity metric, and a motion measure. The metric proposed by Ryu and Sohn (2014) estimates local blurriness, blockiness, and visual saliency and then combines them into an overall quality score using a binocular quality perception model. M.-J. Chen et al. (2013b) have proposed to extract

both 2D and 3D features, including natural scene statistics, and relied on a support vector machine model to predict 3D quality. Recently, Shao et al. (2015b) have proposed a blind quality assessment algorithm based on binocular feature combination. To encode monocular features, the difference of Gaussian filter and a singular value decomposition are used to learn a dictionary. A binocular feature combination of the left and right image coefficients is performed using a support vector regression to compute the quality score. A FR metric based on a similar idea was also proposed by the same authors (Shao et al., 2015a).

In the scenario of FTV, the quality assessment of synthesized viewpoints is needed. Except for computer-generated imagery, there is no pristine reference of the synthesized viewpoint. Kilner et al. (2009) proposed full and NR metrics to assess the quality of 3D objects in FVV production for human performance capture and sports production. However, the proposed metrics are not suitable for content represented in the MVD format. A NR metric was proposed by Oh et al. (2010) to assess the quality of synthesized views for FTV and 3DTV applications. Spatial and temporal consistencies are measured through spatial and temporal noise caused by the view synthesis process, respectively. Bosc et al. (2011b) proposed to analyze the contours shifts in the synthesized view or to compute the SSIM score of the disoccluded areas. The second idea was further extended by Jung et al. (2015). Later on, Bosc et al. (2012a) proposed a RR metric that estimates the edged-based structural distortion between the original and synthesized views. However, the proposed metric does not take the color consistency into account. A FR metric was also proposed by Conze et al. (2012) and relies on textures, diversity of gradient orientations, and presence of high contrast to detect artifacts in synthesized views. Finally, Battisti et al. (2015a) proposed a FR metric that compares features of wavelet subbands of the original and synthesized views. To ensure shifting-resilience, a registration step is included before the comparison.

8.5 Conclusion

This chapter described some of the most common image and video quality metrics. A detailed review of 3D and HDR quality metrics was also reported. We explained how standard quality metrics for still images can be extended for video and 3D quality assessment by considering the multiple frames and views, respectively. Finally, we explained how LDR quality metrics can be extended to HDR quality assessment by considering specific properties of the HVS. Some of these objective quality metrics were also used in the performance evaluation reported in the rest of the thesis.

9 Procedures for Statistical Evaluation of Objective Quality Metrics

The most important step in the development process of an objective quality metric is its verification with regards to subjective data. Indeed, it is essential to verify that the model can reliably and accurately predict perceived visual quality. This verification process is essential to assess the performance of an objective quality metric and to determine its scope of validity. For this purpose, ground truth subjective quality scores obtained via subjective visual quality experiments are used to evaluate a set of performance indexes. Different databases should ideally be used to evaluate the metric's performance in different conditions. The outcome of the performance indexes will be used to determine when the metric fails and when it should be used. For example, it is known that PSNR is quite reliable to tune the performance of a particular codec on a specific content (Huynh-Thu and Ghanbari, 2008; Huynh-Thu and Ghanbari, 2012; Korhonen and You, 2012), but that it fails at predicting visual quality when different contents and distortions are considered (Z. Wang et al., 2004).

The verification process is also essential to compare the performance of different objective quality metrics. For example, this process is used to determine which metric performs the best or to show that a proposed metric outperforms the state-of-the-art. To compare the performance of two objective metrics, their performance index values are compared using statistical tests.

The results of subjective visual experiments, i.e., MOSs and CIs, are considered as ground truth to evaluate how well an objective quality metric estimates perceived quality. The result of execution of a particular objective metric is an OQR, which is expected to be the estimation of the MOS (see Section 2.6.2) corresponding to an image or video sequence. The relationship between the OQR and MOS values does not need to be linear, as human perception of quality is nonlinear and saturates at the extremes of the quality range. Moreover, rating scales are also bounded, whereas the output value of some quality metrics, e.g., PSNR, is not bounded. A regression is fitted to the [OQR,MOS] data set to map the objective scores to the subjective ratings. The mapping process can also be seen as an estimation of the relationship between the model's prediction and the subjective ratings. The mapped OQR values, MOS_p, are a prediction of the MOS values.

Different performance indexes are commonly used or recommended when evaluating the performance of objective quality metrics. A few international recommendations, tutorials, and scientific papers have been published on the topic of benchmarking of objective quality metrics (Brill et al., 2004; ITU-T Tutorial, 2004; ITU-T J.149, 2004; ITU-T P.1401, 2012; Korhonen et al., 2012). Recommendation ITU-T P.1401 (2012) suggests considering the accuracy, consistency, and linearity of the OQR estimation of MOS using the RMSE, OR, and PCC, respectively. On the other hand, ITU-T Tutorial (2004) recommends considering accuracy, consistency, and monotonicity using the RMSE, OR, and SROCC, respectively. Monotonicity is a very important property, as an increase (decrease) in OQR values is expected to correspond to an increase (decrease) in visual quality. Therefore, we believe that all four performance indexes should be considered.

Statistical tests should be applied to determine whether the difference between two performance index values is statistically significant. Recommendation ITU-T P.1401 (2012) describes procedures to test the significance of RMSE, OR, and SROCC values. However, most scientific publications usually do not conduct any statistical analysis to check whether the performance difference between two objective metrics is significant.

This chapter describes in details the procedures to map objective to subjective scores, compute performance indexes to estimate the performance of the objective metric, and test for significant differences in performance between two objective metrics. The different procedures described in this chapter were used in the performance evaluations reported in the rest of the thesis.

9.1 Mapping Objective Values to Subjective Data

A regression is fitted to the [OQR,MOS] data set to map the objective scores to the subjective ratings. Note that different objective metrics typically have different range of values, so the mapping to a common scale also facilitates the comparison of different models. To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the HVS, a non-linear mapping function is typically used. The parameters of the mapping function are determined in a least-square sense, as to minimize the RMSE, but under the constraint that the function should be monotonic over the interval of observed OQR values, such that the rank-order is not changed. As the nature of the nonlinearities are not well known beforehand, several mapping functions will be considered and the one that results in the lowest RMSE should be used for that objective metric. The following mapping functions are commonly used (including the simple linear function)

i) linear function (ITU-T P.1401, 2012)

$$MOS_p = aOQR + b \tag{9.1}$$

ii) cubic polynomial function (ITU-T Tutorial, 2004; ITU-T P.1401, 2012)

$$MOS_p = aOQR^3 + bOQR^2 + cOQR + d \tag{9.2}$$

iii) 4-parameter logistic function (ITU-T J.149, 2004; ITU-T P.1401, 2012)

$$MOS_p = a + \frac{b}{1 + \exp\left[-c\left(OQR - d\right)\right]} \tag{9.3}$$

iv) 5-parameter logistic function (ITU-T Tutorial, 2004; ITU-T J.149, 2004)

$$MOS_p = a + \frac{b}{1 + c\left(OQR - d\right)^e} \tag{9.4}$$

where a, b, c, d, and e are the parameters of the fitting functions and are constraint such that the function is monotonic on the interval of observed quality values.

9.2 Performance Indexes

The following properties of the OQR estimation of MOS should be considered: linearity, monotonicity, accuracy, and consistency. For this aim, four different performance indexes are computed between the ground truth and predicted subjective scores. In particular, the PCC and SROCC are computed between MOS and MOS_p to estimate linearity and monotonicity, respectively. Accuracy and consistency are estimated using the RMSE and OR, respectively. Note that none of these performance indexes takes into account the subjective uncertainty, i.e., the CI associated with the MOS.

9.2.1 Pearson Correlation Coefficient

The Pearson correlation coefficient (PCC) is defined in Section 2.7.2. The PCC is computed to estimate the linearity between MOS and MOS_p. An absolute PCC closer to 1 means that the mapped metric's predictions are more linear. Note that the PCC is related to the average difference error (lower average difference error leads to higher correlation) and can thus be also considered as an indirect estimator of consistency.

9.2.2 Spearman Rank Order Correlation Coefficient

The Spearman's rank correlation coefficient (SROCC) is defined in Section 2.7.2. Note that the SROCC is a non-parametric measure and does not make any assumption regarding the form of the relationship (linear, polynomial, etc.) between the data. Thus, the SROCC value is independent of the mapping function if the function is constraint to be strictly monotonic over the interval of observed OQR values. Some mapping functions, e.g., the logistic function, are theoretically strictly monotonic, but the SROCC value might be different when compared

to another mapping function, e.g., the linear mapping, because of the numerical precision, especially near the lower and upper horizontal asymptotes. Therefore, the SROCC should be computed between MOS and OQR directly, without considering any mapping.

The SROCC is computed to estimate the monotonicity between MOS and OQR. An absolute SROCC closer to 1 means that the metric's predictions are more monotonic.

9.2.3 Root Mean Square Error

The root-mean-square error (RMSE) of the absolute prediction error computed between MOS and MOS_p is defined as

$$\text{RMSE} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(MOS_i - MOS_{pi}\right)^2} \tag{9.5}$$

where M is the total number of points. Note that the division by $M-1$ corresponds to the unbiased estimator for the RMSE.

The RMSE is computed to estimate the accuracy between MOS and MOS_p. A smaller RMSE means that the metric's predictions are more accurate.

9.2.4 Outlier Ratio

The OR represents the ratio of the number of outlier-points divided by the total number of points

$$\text{OR} = \frac{\text{total number of outliers}}{M} \tag{9.6}$$

where M is the total number of points and an outlier is defined as a point i for which the error exceeds the 95% CI of the MOS value

$$\left|MOS_i - MOS_{pi}\right| > \delta_i \tag{9.7}$$

where δ_i is related to the 95% CIs (see Section 2.6.2) corresponding to MOS_i.

The OR is computed to estimate the consistency between the two groups of MOS values. A smaller outlier fraction means that the metric's predictions are more consistent.

9.3 Statistical Significance Evaluation

To determine whether the performance difference between two objective quality metrics is statistically significant, statistical tests are performed on their performance index values. This section describes the procedures suggested in recommendation ITU-T P.1401 (2012) to test the

significance of correlation coefficients, RMSE, and OR values. Note that the same statistical test is performed to determine whether the difference between two PCC or SROCC values is statistically significant, as these two correlation coefficients have similar statistical properties.

9.3.1 Significance of the Difference between the Correlation Coefficients

Based on the assumption that MOS and MOS_p follow a bivariate normal distribution, the Fisher transformation of the correlation coefficient (CC) (PCC or SROCC), $F(CC)$, approximately follows a normal distribution with mean

$$z = F(CC) = \frac{1}{2}\ln\frac{1+CC}{1-CC} \tag{9.8}$$

and standard deviation

$$\sigma_z = \sqrt{\frac{1}{M-3}} \tag{9.9}$$

To determine whether the difference between two correlation coefficient values corresponding to two different objective metrics is statistically significant, a two-sample statistical test is performed. The null hypothesis under test is that there is no significant difference between correlation coefficients, against the alternative hypothesis that the difference is significant, although not specifying better or worse

$$\begin{aligned} H_0&: \quad CC_1 = CC_2 \\ H_1&: \quad CC_1 \neq CC_2 \end{aligned}$$

The observed value z_{obs} is computed from the observations for each comparison

$$z_{obs} = \frac{z_1 - z_2 - \mu_{z_1-z_2}}{\sigma_{z_1-z_2}} \tag{9.10}$$

where

$$\mu_{z_1-z_2} = 0 \tag{9.11}$$

due to the null hypothesis and

$$\sigma_{z_1-z_2} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2} \tag{9.12}$$

If the observed value z_{obs} is inside the critical region determined by the 95% two-tailed z-value, then the null hypothesis is rejected at a 5% significance level.

If the sample size M is lower than 30 samples, then the z-value should be replaced by a t-value corresponding to a two-tailed Student's t-distribution with $M-1$ degrees of freedom.

9.3.2 Significance of the Difference between the Root Mean Square Errors

Based on the assumption that MOS and MOS_p follow a normal distribution, the root mean square error follows approximately a chi-squared distribution with $M-d$ degrees of freedom, where d is the degrees of freedom of the fitting function, which is equal to the number of parameters of the fitting function minus one.

To determine whether the difference between two RMSE values corresponding to two different objective metrics is statistically significant, a two-sample statistical test is performed. The null hypothesis under test is that there is no difference between RMSE values, against the alternative hypothesis that the difference is significant, although not specifying better or worse

$$H_0: \quad RMSE_1 = RMSE_2$$
$$H_1: \quad RMSE_1 \neq RMSE_2$$

The statistic defined in Equation (9.13) follows a F-distribution with M_1 and M_2 degrees of freedom

$$F_{obs} = \frac{RMSE_1^2}{RMSE_2^2} \tag{9.13}$$

The observed value F_{obs} is computed from the observations for each comparison. If the observed value F_{obs} is inside the critical region determined by the 95% two-tailed F-value with $M_1 - d$ and $M_2 - d$ degrees of freedom, then the null hypothesis is rejected at a 5% significance level.

9.3.3 Significance of the Difference between the Outlier Ratios

The OR follows a binomial distribution with mean

$$p = OR \tag{9.14}$$

and standard deviation

$$\sigma_p = \sqrt{\frac{p(1-p)}{M}} \tag{9.15}$$

To determine whether the difference between two OR values corresponding to two different objective metrics is statistically significant, a two-sample statistical test is performed. The null hypothesis under test is that there is no significant difference between ORs, against the

alternative hypothesis that the difference is significant, although not specifying better or worse

$$\begin{aligned} H_0&: \quad OR_1 = OR_2 \\ H_1&: \quad OR_1 \neq OR_2 \end{aligned}$$

If the sample size is large ($M \geq 30$), then the distribution of differences of proportions from two binomially distributed populations can be approximated by a normal distribution.

The observed value z_{obs} is computed from the observations for each comparison

$$z_{obs} = \frac{p_1 - p_2 - \mu_{p_1 - p_2}}{\sigma_{p_1 - p_2}} \tag{9.16}$$

where

$$\mu_{p_1 - p_2} = 0 \tag{9.17}$$

and

$$\sigma_{p_1 - p_2} = \sqrt{p(1-p)\frac{2}{M}} \qquad p = \frac{p_1 + p_2}{2} \tag{9.18}$$

because the null hypothesis in this case considers that there is no difference between the population parameters p_1 and p_2.

If the observed value z_{obs} is inside the critical region determined by the 95% two-tailed z-value, then the null hypothesis is rejected at a 5% significance level.

If the sample size M is lower than 30 samples, then the z-value should be replaced by a t-value corresponding to a two-tailed Student's t-distribution with $M-1$ degrees of freedom.

9.4 Resolving Power

An important question when relying on an objective metric to assess quality is which relative difference in objective scores leads to a visible difference in perceived quality? For example, does a 0.1 dB increase in PSNR lead to a visible improvement in perceived quality? This problem refers to the resolving power of the objective metric.

In recommendation ITU-T J.149 (2004), the resolving power of an objective metric is defined as the difference in the metric values, ΔOM, above which the conditional subjective score distributions have means that are statistically different at a certain degree of confidence, typically 95%. The algorithm uses a one-tailed Z-test to determine the probability p that, given a pair of objective scores, the greater score corresponds to the greater true underlying MOS. Note that more appropriate statistical tests, e.g., Student's t-test or ANOVA, could be used for this purpose (see Section 2.6.4).

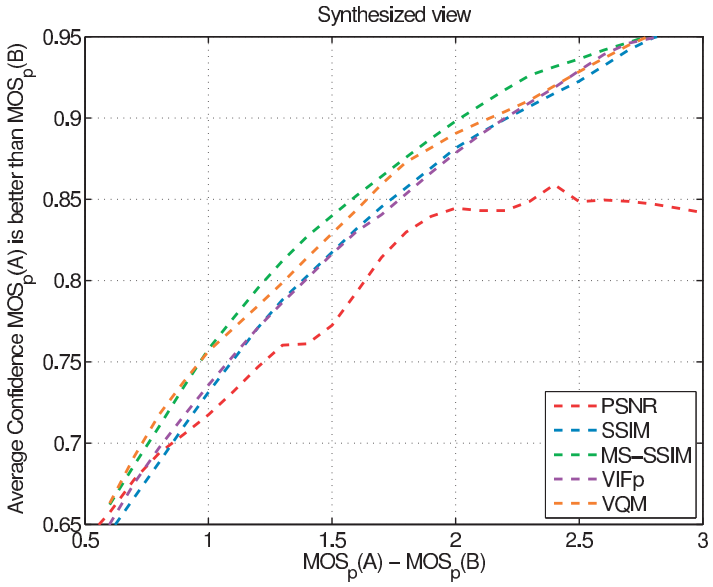


Figure 9.1: Sample plot of confidence versus common-scale ΔMOS_p score.

The resolving power can be computed either in the native scale of the objective metric or in a transformed scale, common to different metrics. This common scale allows an easier comparison between metrics that have different ranges. A common scale can be obtained by fitting the objective scores to the subjective data.

By plotting the probability p as a function of ΔOM, one can determine the resolving power of each metric corresponding to a specific degree of confidence, e.g., 68%, 75%, 90%, or 95% (see Figure 9.1). By stacking curves corresponding to different metrics on the same graph, one can determine which metric reaches the highest degree of confidence for a fixed resolving power, or which metric has the highest discriminability for a fixed degree of confidence.

9.5 Classification Errors

When comparing two pairs of images or video sequences, the comparison outcome can be classified into three categories: better, same, or worse. Another important question related to the resolving power is whether the objective metric leads to the same conclusion as the subjective evaluation. For example, if there is a 0.5 dB increase in PSNR, does the visual quality improves, deteriorates, or remains stable? If the outcome of the objective metric does not match that of the subjective evaluation, then a classification error is made by the metric.

In recommendation ITU-T J.149 (2004), it is suggested to compute the classification errors to evaluate the performance of an objective metric. A classification error is performed when the objective metric and subjective evaluation lead to different conclusions on a pair of images or video sequences, A and B, for example. Three types of error can happen

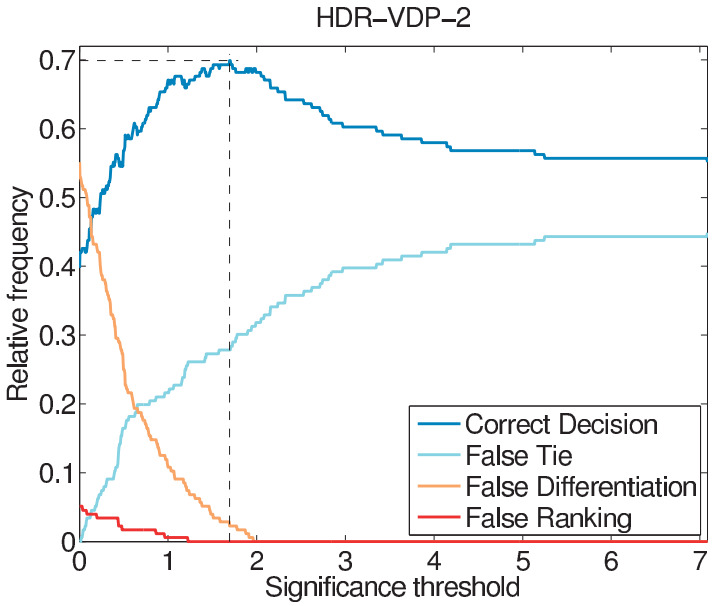


Figure 9.2: Sample plot of frequencies of classification error.

i) *False Tie*, the least offensive error, which occurs when the subjective evaluation says that A and B are different, whereas the objective scores say that they are identical,
ii) *False Differentiation*, which occurs when the subjective evaluation says that A and B are identical, whereas the objective scores say that they are different,
iii) *False Ranking*, the most offensive error, which occurs when the subjective evaluation says that A (B) is better than B (A), whereas the objective scores say the opposite.

To determine whether two groups of subjective scores are statistically different, a simple one-tailed Z-test is used in the MATLAB code provided in Appendix II of ITU-T J.149 (2004). However, there are more appropriate statistical tests for this purpose, especially when multiple comparisons are performed (see Section 2.6.4). The percentage of *Correct Decision*, *False Tie*, *False Differentiation*, and *False Ranking* are then recorded from all possible distinct pairs as a function of the difference in the metric values, ΔOM.

As ΔOM increases, more pairs of data points are considered as equivalent by the objective metric. This reduces the occurrences of *False Differentiations* and *False Rankings*, but increases the occurrence of *False Ties*. On the other hand, as ΔOM tends towards 0, the occurrence of *False Tie* will tend towards 0, while the occurrence of *False Differentiation* will tend towards the proportion of pairs of data points where there was not enough evidence to show a statistical difference in the subjective evaluation.

The relative frequencies are plotted as a function of the significance threshold ΔOM (see Figure 9.2). Ideally, the occurrence of *Correct Decision* should be maximized and the occurrence of *False Ranking* should be minimized when the ΔOM tends towards 0. The occurrences of *False Differentiations* and *False Rankings* should decrease as fast as possible as ΔOM in-

creases. Based on this, different graphs corresponding to different metrics can be compared to determine the best metric for the application under analysis.

To challenge the objective metrics and test critical situation, Ciaramello and Reibman (2011) have proposed to construct pairs with potential *False Tie* and *False Ranking*. This technique is inspired from software testing and can provide additional understanding on the weaknesses of the metrics, but it should not replace the standard benchmarking.

9.6 Conclusion

This chapter provided a detailed description of the different procedures available to benchmark objective quality metrics. From the mapping of objective scores to the statistical tools used for testing significant differences in performance between two objective metrics, we reviewed the guidelines suggested by the relevant international recommendations, as well as some common practice. These procedures were used to benchmark and analyze the performance of objective quality metrics in the rest of the thesis.

10 Performance Evaluation of Objective Quality Metrics

It is essential to evaluate the performance of an objective quality metric in predicting perceived visual quality and to determine its scope of validity. For this purpose, ground truth subjective quality scores obtained via subjective visual quality experiments are used to evaluate the performance of objective metrics following different procedures (see Chapter 9). For new applications, e.g., FTV, or types of content, e.g., 3D and HDR, it is also fundamental to determine the performance of existing metrics that are widely used.

Several databases of distorted images and video sequences with associated ground truth MOS values are publicly available. For image quality assessment, the following databases are commonly used to benchmark objective quality metrics: IRCCyN/IVC Image Database, LIVE Image Database (Sheikh et al., 2006), TID2008 (Ponomarenko et al., 2009), and TID2013 (Ponomarenko et al., 2015). In the case of video quality assessment, several public databases have also been made available, e.g., EPFL/PoliMI Video Database (De Simone et al., 2009a), IRCCyN/IVC 1080i Database (Péchard et al., 2008), LIVE Video Database (Seshadrinathan et al., 2010), and VQEG HDTV Database (VQEG, 2003). Several public image and video databases for quality assessment have been analyzed by Winkler (2012) and a list of publicly available databases is frequently updated on his personal website (Stefan Winkler's website). Additionally, the Qualinet Databases provide an exhaustive list of public image and video datasets with over 200 registered datasets.

This chapter reports the results of performance evaluation of several objective metrics in different scenarios. Sections 10.1 and 10.2 investigate the performance of state-of-the-art 2D metrics for quality assessment of stereo pairs formed from decoded and synthesized views and from two synthesized views, respectively. For this purpose, we used the ground truth subjective ratings collected during the evaluations of the CfP on 3D Video Coding Technology issued by MPEG (N12036). Section 10.3 reports the performance of state-of-the-art 2D metrics for quality assessment of FVV considering as ground truth the subjective scores collected in the evaluation reported in Section 5.1. Section 10.4 reports the results of an extensive benchmarking of state-of-the-art LDR and HDR quality metrics using the results of the subjective evaluation reported in Section 4.4. Finally, Section 10.5 evaluates the

effectiveness of different LDR and HDR quality metrics for HDR video quality assessment considering the results of the subjective evaluation reported in Section 4.5 as ground truth. All performance evaluations were performed following the procedures described in Chapter 9.

10.1 Benchmarking of Objective Metrics on Asymmetric Stereo Pairs formed from Decoded and Synthesized Views

Despite the efforts of the scientific community in recent years, 3D video quality assessment is still an open challenge and there are no objective metrics that are widely recognized as reliable predictors of human 3D quality perception. The assessment of 3D quality is particularly challenging for mismatched or asymmetric stereoscopic videos, which have different strength and/or types of degradation between the left and right views. In general, the perceived quality of an asymmetric stereo pair is closer to the average quality of the two views. However, Stelmach et al. (2000) have shown that, depending on the type of degradation and the difference of quality between the individual views, the 3D quality can be closer to the highest quality. Therefore, specific properties of the HVS, such as binocular suppression, i.e., the masking of low-frequency content in one view by the sharp visual content in the other view, should be taken into account when building models that objectively quantify the 3D quality of a stereo pair.

In March 2011, a CfP on 3D Video Coding Technology was issued by MPEG (N12036). The main objective is to support high-quality multiview autostereoscopic displays through generation of many high-quality views from a limited number of input views. For this application, a 3-view configuration is assumed, as illustrated in Figure 10.1. In this configuration, the decoded data, i.e., texture views and corresponding depth maps, is used to synthesize a set of virtual views at selected positions. The 3-view configuration was evaluated both on multiview autostereoscopic and stereoscopic displays. In the latter case, the displayed stereo pair is formed from two synthesized views and both views contain compression and view synthesis artifacts. A second objective is to allow advanced processing of stereoscopic content to cope with varying display types and sizes, as well as different viewing preferences. For this application, a 2-view configuration is assumed, as illustrated in Figure 10.2. In this configuration, the decoded data, i.e., texture views and corresponding depth maps, is used to synthesize a virtual view at a selected position. The stereo pair displayed on the stereoscopic monitor consists of the decoded left or right view and the synthesized view. Due to the artifacts introduced by the view synthesis algorithm and the compression of the depth maps, it is expected that the individual quality of the virtual view is lower than that of the decoded view. Thus, the displayed stereo pair is considered as asymmetric, as one view contains only compression artifacts whereas the other view contains both compression and view synthesis artifacts.

Bosc et al. have shown that traditional objective metrics have a very low correlation with perceived quality when used for objective quality assessment of synthesized views (Bosc et al.,

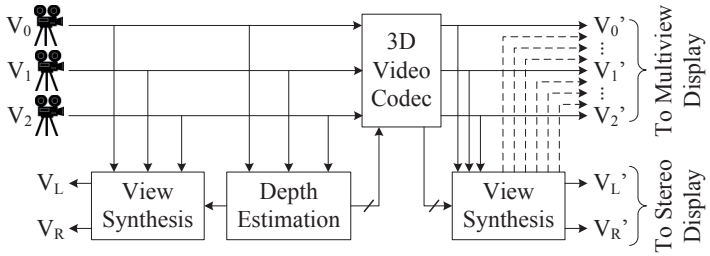


Figure 10.1: Stereoscopic and autostereoscopic output with 3-view configuration.

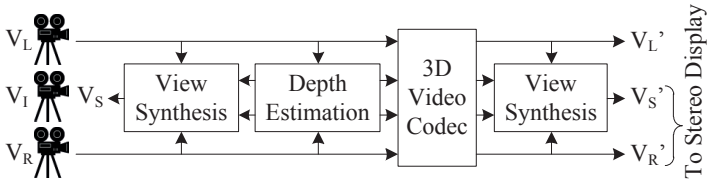


Figure 10.2: Advanced stereoscopic processing with 2-view configuration.

2011a; Bosc et al., 2012b). Therefore, for a stereo pair formed from a decoded view and a synthesized view, it is unclear whether objective metrics correlate well with perceived quality and which views should be taken into account: the decoded view, the synthesized view, or both views? To answer to this question, we measured the performance of state-of-the-art 2D quality metrics, including perceptual based metrics, in predicting quality of asymmetric stereo pairs formed from a decoded view and a synthesized view. The subjective results of the MPEG 3DV evaluations were used as ground truth to benchmark a set of objective metrics. This section reports the details and results of this performance evaluation.

10.1.1 Methodology

In this study, the performance of the following objective metrics (see Sections 8.1 and 8.2) is assessed

1) PSNR,
2) SSIM,
3) MS-SSIM,
4) VSNR: visual signal-to-noise ratio (D. Chandler and S. Hemami, 2007),
5) VIFp: VIF pixel domain version,
6) WSNR: weighted signal-to-noise ratio (Damera-Venkata et al., 2000),
7) PSNR-HVS (Egiazarian et al., 2006),
8) PSNR-HVS-M (Ponomarenko et al., 2007), and
9) VQM (NTIA General Model, no calibration).

All above objective metrics, except for VQM, were computed on the luma component of each frame and the resulting values were averaged across the frames to produce a global index for

the entire video sequence.

Most of the objective metrics, except for WSNR, VSNR, and VQM, were computed using our Video Quality Measurement Tool (VQMT). WSNR was computed using MeTriX MuX Visual Quality Assessment Package. VSNR was obtained from its developer website (VSNR). VQM was obtained from the Institute for Telecommunication Sciences (ITS) website (VQM).

In the 2-view configuration, as considered in the MPEG CfP, a pair of cameras is used to produce the input views at the encoder side. At the decoder side, the displayed stereo pair is formed from the decoded right view and a synthesized view, located in-between the input views, as depicted in Figure 10.2. The baseline (inter-camera distance) of the displayed stereo pair is roughly equal to half of the baseline of the input stereo pair. For one Class A content and all Class C contents, the location of the synthesized view matched the location of a real view, called intermediate view, available in the original data (but not used by the encoder). Thus, five different objective video quality models are considered

i) Quality of the decoded view, calculated between the decoded view and the original view,
ii) Quality of the intermediate view, calculated between the synthesized view at the decoder side and the intermediate view from the original data (when available),
iii) Quality of the synthesized view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side,
iv) Average quality of the decoded view and the intermediate view, computed as the mean value of i) and ii), and
v) Average quality of the decoded view and the synthesized view, computed as the mean value of i) and iii).

The objective metrics were benchmarked following the procedure described in Chapter 9. First, a linear function was used to map the objective scores to the subjective ratings. Then, the performance indexes were computed between the predicted MOS values and ground truth MOSs. Finally, the resolving power and classification errors were computed for some of the best metrics. Regarding the resolving power, a one-tailed Z-test was used to determine the probability that the greater score corresponds to the greater true underlying subjective MOS. To compute the classification errors, the same statistical test was used at a 95% confidence level to determine whether two distributions of subjective scores are different or not. Since some metrics, e.g., PSNR, are very content dependent, the resolving power and classification errors were computed on each content separately and the results were averaged across the different contents.

Dataset

The test material used in the MPEG CfP is composed of eight different contents encoded at four target bit rates. The contents are divided in two classes: Class A, with a spatial resolution of 1920 × 1088 pixels and a temporal resolution of 25 fps, and Class C, with 1024 × 768 pixels at

Table 10.1: Input views and displayed stereo pairs.

Sequence	Class	2-view conf.		3-view configuration		
		Input views	Stereo pair	Input views	Fixed stereo pair	Random stereo pair
Poznan Hall2	A	7-6	6.5-6	7-6-5	6.125-5.875	-
Poznan Street		4-3	3.5-3	5-4-3	4.125-3.875	-
Undo Dancer		2-5	3-5	1-5-9	4.5-5.5	-
GT Fly		5-2	4-2	9-5-1	5.5-4.5	-
Kendo	C	3-5	4-5	1-3-5	2.75-3.25	2.25-2.75
Balloons		3-5	4-5	1-3-5	2.75-3.25	4.375-4.875
Lovebird1		6-8	7-8	4-6-8	5.75-6.25	4.0833-4.5833
Newspaper		4-6	5-6	2-4-6	3.75-4.25	4.3333-4.8333

30 fps. All contents are 10 s long. All test sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8-bit per sample. A total of 22 coding algorithms, submitted by the proponents, and 2 anchors were evaluated at 4 different bit rates for each sequence.

In the 2-view test scenario, the stereo pair consisted of one of the two input views reconstructed at the decoder side and a synthesized view (see Table 10.1). However, for the 3-view scenario, the displayed stereo pair is formed from two synthesized views. More specifically, two different stereo pairs were evaluated: one referred to as fixed stereo pair, which is centered on the central decoded view, and one referred to as random stereo pair, which is located in-between two decoded views. The reader can refer to the 3DV CfP (N12036) for more details.

The evaluation was performed using a 46" Hyundai S465D polarized stereoscopic monitor with a native resolution of 1920 × 1080 pixels. Eighteen naive viewers evaluated the quality of each test sequence. The viewers were seated at a distance of about four times the height of the active part of the display. The laboratory setup was controlled to produce reliable and repeatable results. The test room was equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of maximum screen luminance.

The DSIS method with an 11-grade numerical categorical scale (see Section 2.4.2) was used. A basic test session including 24 test pairs, 3 dummy stimuli pairs, and 1 reference versus reference pair, was designed. Thus, the test material resulted in a total of sixteen sessions for each of the two classes of data. In each session, the stimulus pairs were presented in random orders, but never with the same video content in consecutive pairs.

All subjects taking part in the evaluations underwent a screening to examine their visual acuity, color vision, and stereo vision using Snellen chart, Ishihara chart, and Randot test, respectively. Before each test session, written instructions and a short explanation by a test operator were provided to the subjects. Also, a training session was run to show the graphical user interface, the rating sheets, and examples of processed video sequences. These training sequences were

produced using two contents not included in the test material, with coding conditions similar to those used to produce the actual test material.

The subjective results were processed by first detecting and removing subjects whose scores appeared to deviate strongly from other scores in each test session. Then, the MOSs were computed for each test sequence as the mean across the rates of the valid subjects. Regarding the results for the 2-view configuration, the results were computed on a total of 18 naïve viewers coming from two different laboratories. The boxplot inspired outlier detection technique proposed by De Simone et al. (2011) (see Section 2.6.1) was used and the 95% confidence interval were computed assuming a Student's t-distribution of the scores. Regarding the results for the 3-view configuration, we used the results computed by the MPEG test coordinator on a total of 36 naïve viewers coming from three different laboratories (N12347). In this case, outlier detection was performed by the MPEG test coordinator according to the procedure adopted by VQEG for its Multimedia Project.

10.1.2 Results

Table 10.2 reports the linearity, monotonicity, accuracy, and consistency indexes of the objective video quality models, as defined in Section 10.1.1, for each objective metric separately. The objective metrics are ranked for each objective video quality model and the ranking number is specified below each performance index value.

It can be noticed that the PSNR of the intermediate view (PCC=0.5858, SROCC=0.6234) has a significantly lower correlation with perceived quality than the PSNR of the synthesized view (PCC=0.6668, SROCC=0.6797). PSNR-HVS and PSNR-HVS-M, which are based on PSNR, have a similar behavior. The difference between the intermediate and synthesized views is not significant for the other objective metrics.

For PSNR, PSNR-HVS, PSNR-HVS-M, WSNR, and VSNR, the objective video quality models that take into account the quality of the decoded view have a significantly higher correlation with perceived quality than the other objective video quality models. On the other hand, SSIM, MS-SSIM, VIF, and VQM have similar performance regardless the objective video quality model. A few hypotheses can be raised to explain these observations

i) In terms of perceived quality, the higher quality of the decoded view, which does not contain view synthesis artifacts, tends to mask the lower quality of the synthesized view and/or
ii) Most of the considered objective metrics do not predict well perceived quality of synthesized views.

The first hypothesis is in agreement with the results from Stelmach et al. (2000). The second hypothesis is in agreement with the results from (Bosc et al., 2011a; Bosc et al., 2012b). However, in the studies from Bosc et al., no compression artifacts were considered. It is also known that PSNR has good performance for compression artifacts but rather low performance for other

Table 10.2: Linearity, monotonicity, accuracy, and consistency indexes of the objective metrics under consideration.

Metric	PCC					SROCC				
	Decoded	Intermediate	Synthesized	Decoded and intermediate	Decoded and synthesized	Decoded	Intermediate	Synthesized	Decoded and intermediate	Decoded and synthesized
PSNR	**0.9200** 2	0.5858 9	0.6668 7	0.9028 4	0.8834 8	**0.9114** 7	0.6234 9	0.6797 6	0.8892 9	0.8762 8
SSIM	**0.9130** 7	0.8506 3	0.8460 4	0.8957 6	0.8987 3	**0.9080** 9	0.8655 4	0.8530 4	0.9077 5	0.9022 5
MS-SSIM	**0.9131** 6	0.8507 2	0.8495 3	0.8907 8	0.8977 5	**0.9177** 1	0.8675 3	0.8584 3	0.9094 3	0.9120 2
VSNR	0.9131 5	0.7188 5	0.7532 5	**0.9153** 2	0.8986 4	0.9118 6	0.7500 5	0.7642 5	**0.9201** 1	0.9040 4
VIFp	**0.9094** 8	0.8510 1	0.8544 2	0.8930 7	0.9013 1	**0.9152** 3	0.8740 2	0.8744 2	0.9108 2	0.9128 1
WSNR	**0.9216** 1	0.6530 6	0.6729 6	0.9188 1	0.9002 2	**0.9166** 2	0.6687 6	0.6735 7	0.9082 4	0.8978 6
PSNR-HVS	**0.9194** 3	0.5913 8	0.6491 9	0.9012 5	0.8809 9	**0.9129** 5	0.6279 8	0.6585 9	0.8931 8	0.8747 9
PSNR-HVS-M	**0.9181** 4	0.5962 7	0.6507 8	0.9033 3	0.8836 7	**0.9139** 4	0.6324 7	0.6591 8	0.8964 7	0.8796 7
VQM	**0.8944** 9	0.8466 4	0.8599 1	0.8684 9	0.8874 6	**0.9105** 8	0.8765 1	0.8803 1	0.8966 6	0.9061 3
	RMSE					OR				
	Decoded	Intermediate	Synthesized	Decoded and intermediate	Decoded and synthesized	Decoded	Intermediate	Synthesized	Decoded and intermediate	Decoded and synthesized
PSNR	**0.9334** 2	1.7519 9	1.6795 7	1.0358 5	1.1108 7	**0.0099** 4	0.1013 9	0.1048 7	0.0278 4	0.0340 8
SSIM	**0.9713** 7	1.1942 2	1.2272 3	1.0440 7	1.0381 4	**0.0181** 6	0.0378 1	0.0405 3	0.0252 3	0.0299 4
MS-SSIM	**0.9877** 8	1.2328 4	1.2298 4	1.0981 8	1.0599 6	**0.0285** 8	0.0452 3	0.0511 4	0.0399 9	0.0359 9
VSNR	0.9664 6	1.3329 5	1.4171 5	**0.9450** 2	1.0199 2	**0.0120** 5	0.0587 5	0.0844 5	0.0146 1	0.0159 1
VIFp	**0.9608** 5	1.1656 1	1.1823 2	1.0235 3	0.9918 1	**0.0184** 7	0.0378 2	0.0373 1	0.0289 7	0.0281 3
WSNR	**0.9306** 1	1.6325 6	1.6616 6	0.9449 1	1.0312 3	**0.0087** 3	0.0838 6	0.0993 6	0.0243 2	0.0271 2
PSNR-HVS	**0.9349** 3	1.7377 8	1.7413 9	1.0431 6	1.1265 9	**0.0060** 1	0.1002 7	0.1132 8	0.0278 5	0.0328 6
PSNR-HVS-M	**0.9429** 4	1.7258 7	1.7361 8	1.0307 4	1.1141 8	**0.0074** 2	0.1002 8	0.1132 9	0.0278 6	0.0328 7
VQM	**1.0237** 9	1.2092 3	1.1687 1	1.1212 9	1.0473 5	**0.0285** 9	0.0500 4	0.0395 2	0.0362 8	0.0308 5

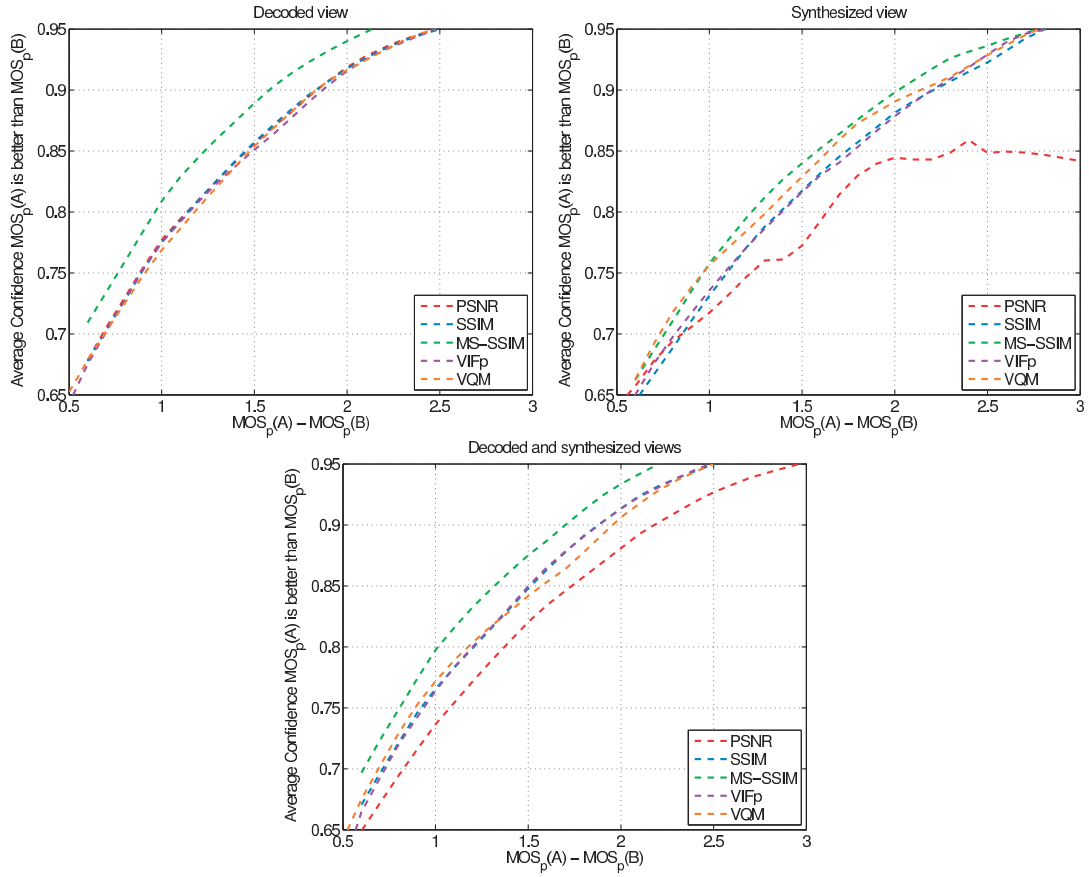


Figure 10.3: Confidence versus ΔMOS_p.

types of degradation or when different types of degradations are combined. All these factors play an important role and should be further investigated to better understand how the viewer perceives quality of a stereo pair formed from a decoded view and a synthesized view and how it can be predicted using objective metrics. A similar study should be conducted using stereo pairs formed from two synthesized views to further investigate these hypotheses.

In general, the objective video quality model based on the quality of the decoded view has the highest correlation with perceived quality. In this case, all objective metrics have a high correlation (PCC≥0.8944, SROCC≥0.9080) with perceived quality. If the objective quality assessment is based on the measured quality of the synthesized view, it is suggested to use VQM, VIF, MS-SSIM, or SSIM since these objective metrics have a significantly higher correlation with perceived quality (PCC≥0.8460, SROCC≥0.8530). Taking into account both views increases correlation with perceived quality as opposed to using the synthesized (intermediate) view only.

Figure 10.3 depicts the probability, p, versus predicted MOS difference, ΔMOS_p, curves for PSNR, SSIM, MS-SSIM, VIFp, and VQM. As it can be observed, when the objective quality assessment is based on the measured quality of the decoded view, MS-SSIM shows a clear

gain in terms of resolving power over the other metrics. For a fixed ΔMOS_p, the confidence than $\Delta MOS_p(A)$ is better than $\Delta MOS_p(B)$ is in average 3.11% higher for MS-SSIM. If both views are taken into account, MS-SSIM shows an average significance level increase of 5.44% over PSNR. When considering the video quality model based on the quality of the synthesized view, MS-SSIM shows, in general, better accuracy than the other metrics. Theoretically, the confidence than $\Delta MOS_p(A)$ is better than $\Delta MOS_p(B)$ should increase as the ΔMOS_p increases (monotonically increasing function). However, this is not the case for PSNR with this video quality model. The explanation for this behavior is the following: for some contents (*Undo Dancer*, *Kendo*, and *Balloons*), a few sequences have a low value for the PSNR of the synthesized view while the corresponding stereo pair has a high MOS. These data points are significantly distant from the trend in the scatter plot of the synthesized view. It is known that one proponent used a different view synthesis algorithm. Our hypothesis is that those results are from this specific proponent. This indicates that PSNR is not a trustfully indicator of quality in this case.

Figures 10.4 to 10.6 depict the metric classification errors for PSNR, SSIM, MS-SSIM, VIFp, and VQM. Even though the results are reported in the native scale of the metric instead of the common scale, it is still possible to compare the classification errors of the different metrics by looking at the relative ΔOM ratio (ΔOM divided by the maximum value of ΔOM) rather than the absolute ΔOM. When considering the video quality model based on the quality of the decoded view, a fast decaying *False Ranking* rate is observed for MS-SSIM, whereas it is significantly slower for VQM. Similarly, the peak of the *Correct Decision* rate is reached at a lower ΔOM ratio for MS-SSIM than for VQM, meaning that MS-SSIM has a higher resolving power. The highest *Correct Decision* rate is obtained with MS-SSIM (0.7839) whereas SSIM has the lowest peak (0.7632). When the objective quality assessment is based on the measured quality of the synthesized view, PSNR shows a significantly lower *Correct Decision* rate and significantly higher *False Ranking* rate than the other metrics. Even for a 1 dB difference in PSNR values, the *False Ranking* rate is around 7.8%. With this video quality model, MS-SSIM also shows better performance over the other metrics in terms of decaying *False Ranking* rate. However, the peak of *Correct Decision* rate is higher for VQM (0.7506) than for MS-SSIM (0.7262). Nevertheless, VQM has a significantly slower decaying *False Ranking* rate. If both views are taken into account, the observations are similar to that of the video quality model based on the quality of the decoded view.

In many applications, the resolving power is not considered and even a small PSNR increase, such as 0.1 dB, is considered as an improvement in perceived quality. Therefore, Table 10.3 reports the *Correct Decision* and *False Ranking* values for a ΔOM of zero. SSIM, MS-SSIM, VIFp, and VQM show similar performance, with non-significant variations. However, PSNR shows a significantly lower reliability when considering the video quality model based on the quality of the synthesized view.

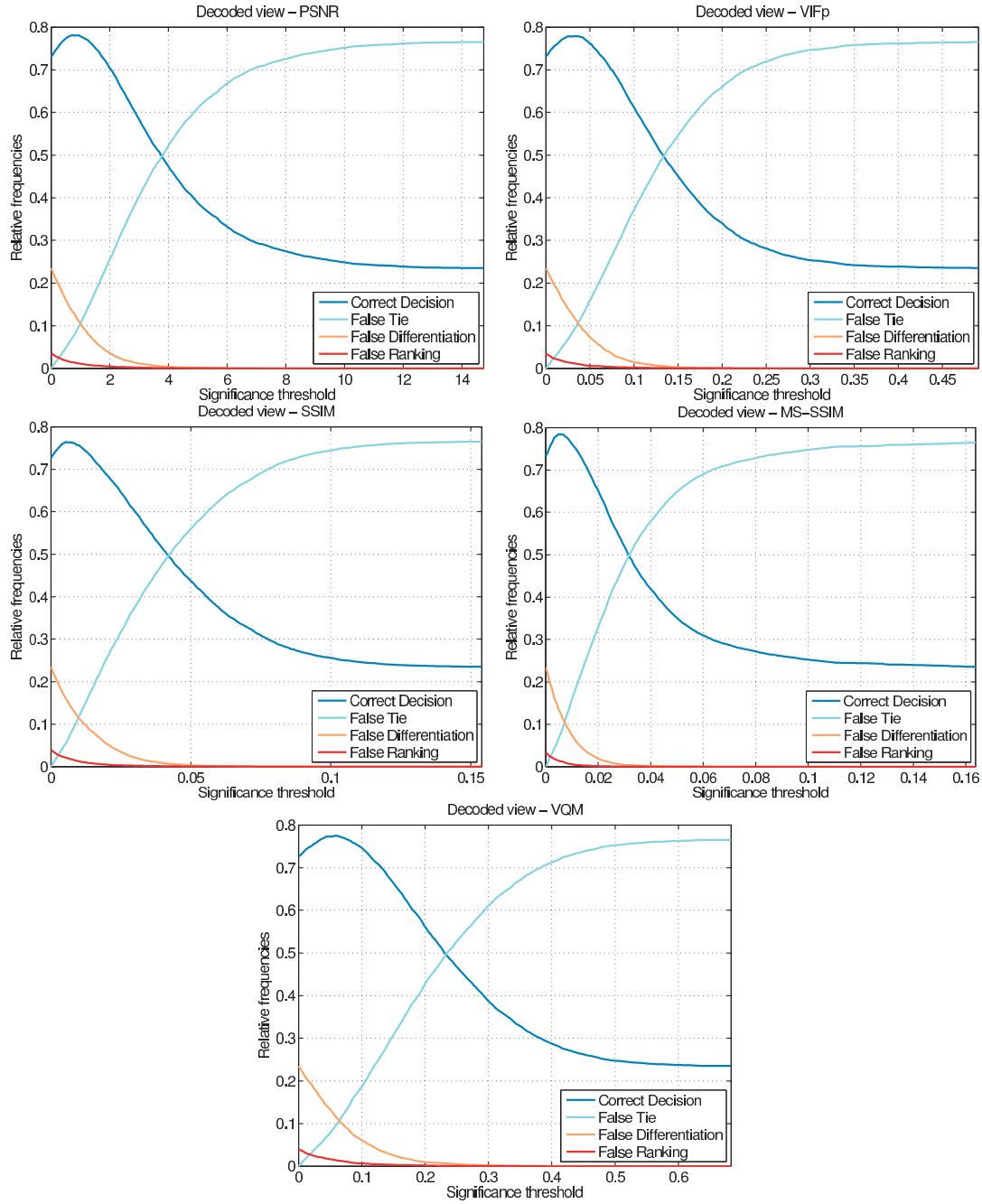


Figure 10.4: Frequencies of classification error: decoded view.

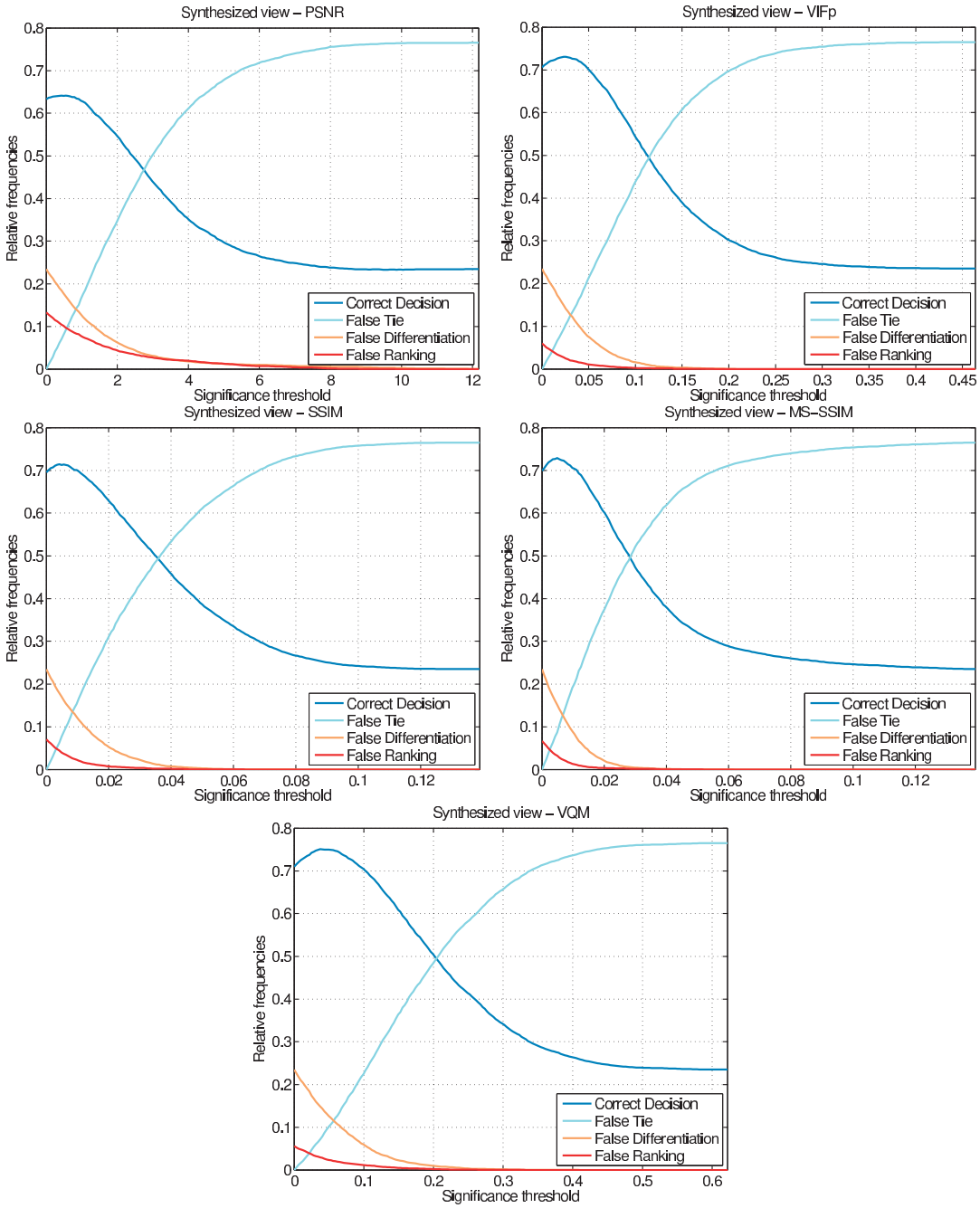


Figure 10.5: Frequencies of classification error: synthesized view.

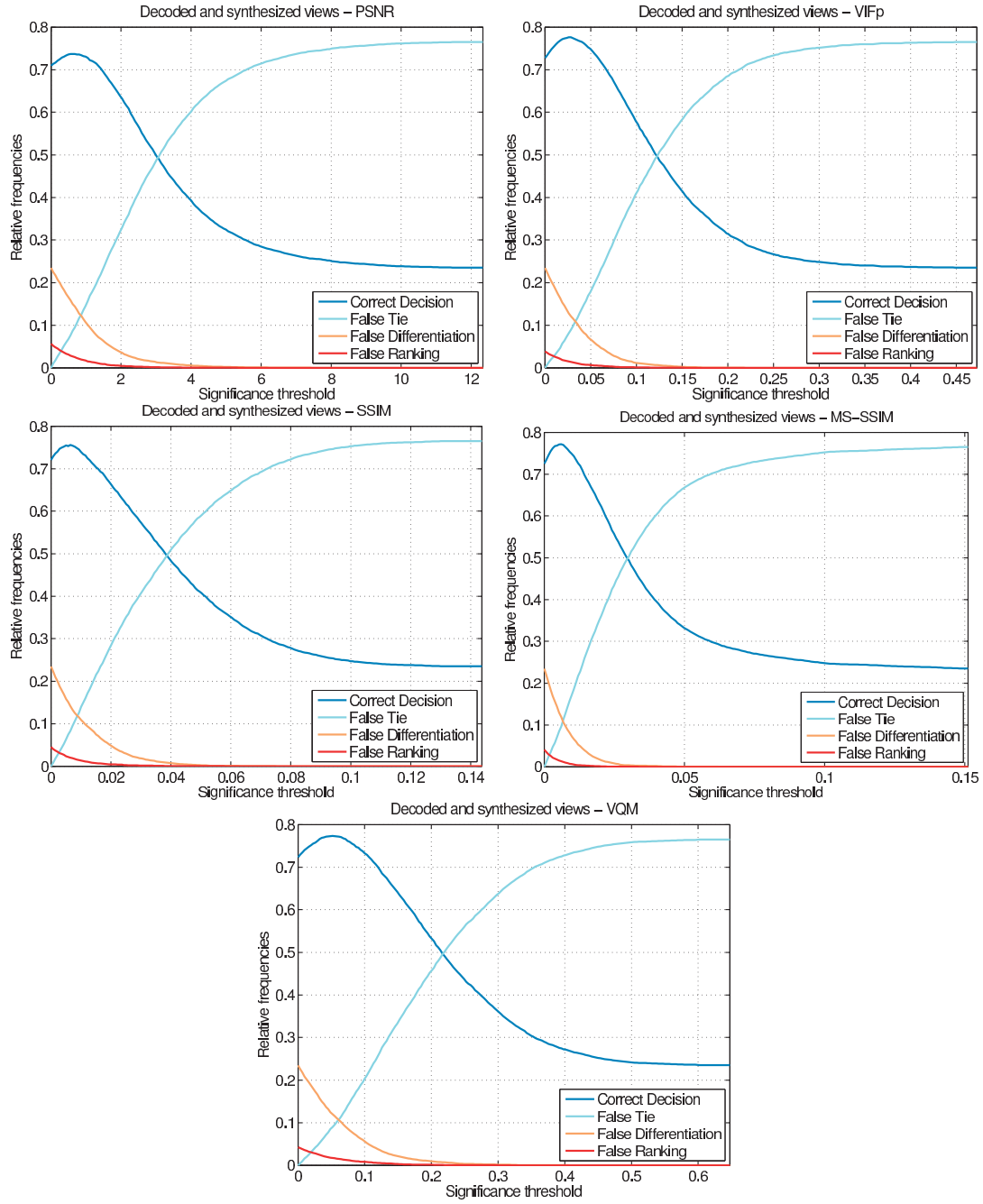


Figure 10.6: Frequencies of classification error: decoded and synthesized views.

Table 10.3: Classification errors for $\Delta OM = 0$.

Metric	Decoded		Synthesized		Decoded & synthesized	
	Correct Decision	False Ranking	Correct Decision	False Ranking	Correct Decision	False Ranking
PSNR	0.7300	0.0357	0.6331	0.1326	0.7094	0.0564
SSIM	0.7267	0.0391	0.6954	0.0703	0.7204	0.0453
MS-SSIM	0.7322	0.0335	0.6990	0.0667	0.7251	0.0406
VIFp	0.7311	0.0346	0.7061	0.0596	0.7270	0.0387
VQM	0.7260	0.0397	0.7096	0.0561	0.7231	0.0427

10.2 Benchmarking of Objective Metrics on Symmetric Stereo Pairs formed from two Synthesized Views

Understanding and measuring the effect of view synthesis on perceived quality, in conjunction with compression, is particularly important for multiview autostereoscopic displays, which usually synthesize N views from a limited number of input views, and stereoscopic displays that modify the baseline to adjust the depth perception based on viewing distance and viewing preferences. Hewage et al. (2009) have investigated objective quality assessment of 3D content represented in video plus depth (2D+Z) format using PSNR, SSIM, and VQM. The objective quality metrics were computed on the 2D video and on the rendered left and right 3D views. It was found that VQM had the highest correlation with perceived quality. The metrics showed lower correlation with perceived quality when using the average quality of the left and right 3D views than when using the quality of the 2D video. This effect was particularly strong for PSNR, where the correlation coefficient dropped from 0.81 to 0.74.

In the previous section, we investigated the correlation between different state-of-the-art 2D quality metrics, including perceptual based metrics, and the perceived quality of a stereo pair formed from a decoded view and a synthesized view. Results showed that the measured quality of the decoded view had the highest correlation in terms of the PCC with perceived quality. When the objective quality assessment was based on the measured quality of the synthesized view, results showed that VIF, VQM, MS-SSIM, and SSIM significantly outperformed other objective metrics. Two hypotheses were raised to explain these observations

i) In terms of perceived quality, the higher quality of the decoded view, which does not contain view synthesis artifacts, tends to mask the lower quality of the synthesized view
ii) Most of the considered objective metrics do not predict well perceived quality of synthesized views.

In this section, we report the results of a different problem, namely when a stereo pair is formed from two synthesized views, which might help us accepting or rejecting the hypotheses formulated in the previous section. Following a similar methodology as in our previous study, we benchmark the same objective metrics using a new set of stereoscopic video sequences and associated subjective quality scores.

10.2.1 Methodology

The same objective quality metrics as in the previous section were used. Three different objective video quality models were considered

i) Quality of the left view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side,
ii) Quality of the right view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side, and
iii) Average quality of both views, computed as the mean value of i) and ii).

The video sequences and ground truth subjective scores corresponding to the 3-view configuration of the dataset described in Section 10.1.1 were used. The objective metrics were benchmarked following the procedure described in Chapter 9. First, a linear function was used to map the objective scores to the subjective ratings. Then, the performance indexes were computed between the predicted MOS values and ground truth MOSs.

10.2.2 Results

Tables 10.4 and 10.5 report the linearity, monotonicity, and accuracy indexes of the objective video quality models for the fixed and random stereo pairs, respectively. The objective metrics are ranked for each objective video quality model and the ranking number is specified below each performance index value. The fixed stereo pair is centered on the central decoded view and both views are equidistant from the central decoded view. Thus, both views should have the same amount of disocclusion and the same strength of view synthesis artifacts. The random stereo pair is located in-between two decoded views; one view of the stereo pair is always located closer to one of the decoded views than the other view of the stereo pair. Thus, we denote them as closer and farther views rather than left and right views. For example, for content *Kendo* (see Table 10.1), view 2.75 is the closer view whereas view 2.25 is the farther view. The closer view has a lower amount of disocclusion than the farther view. Thus, the closer view should contain less view synthesis artifacts than the farther view. However, for the random stereo pair, there is no significant difference between the results for the closer and farther views ($\max|\Delta PCC| = 0.0146$, $\max|\Delta SROCC| = 0.0114$). In general, the objective video quality model based on the average quality of both views has the highest correlation with perceived quality, but the difference between the models is not significant ($\max|\Delta PCC| = 0.0231$, $\max|\Delta SROCC| = 0.0171$).

For stereo pairs formed from a decoded view and a synthesized view (see previous section), the SNR-based metrics (PSNR, PSNR-HVS, PSNR-HVS-M, WSNR, and VSNR) had significantly lower correlation with perceived quality than the perceptual metrics (VIF, VQM, SSIM, and MS-SSIM) when using the synthesized view. In this study, a similar behavior is observed on the three objective video quality models for stereo pairs formed from two synthesized views. The results reported in this study show that PSNR, PSNR-HVS, PSNR-HVS-M, and

Table 10.4: Fixed stereo pair: linearity, monotonicity, and accuracy indexes of the objective metrics under consideration.

Metric	PCC			SROCC			RMSE		
	Left view	Right view	Average	Left view	Right view	Average	Left view	Right view	Average
PSNR	0.78919	0.80849	**0.80869**	0.79579	0.80959	**0.80969**	1.35819	1.30539	**1.30159**
PSNR-HVS	0.79958	0.81908	**0.81908**	0.80388	0.81678	**0.81798**	1.33048	1.27468	**1.27258**
PSNR-HVS-M	0.80167	0.82087	**0.82107**	0.80437	0.81757	**0.81877**	1.32747	1.27117	**1.26897**
WSNR	0.83736	**0.85876**	0.85866	0.83866	0.85266	**0.85366**	1.20876	**1.13106**	1.13186
VSNR	0.90505	**0.92811**	0.92671	0.91685	**0.93393**	0.93245	0.93134	**0.82741**	0.83992
SSIM	0.91893	0.92054	**0.92153**	0.92954	0.93115	**0.93244**	0.88573	0.87694	**0.87214**
MS-SSIM	**0.90744**	0.90465	0.90735	0.93741	0.93592	**0.93881**	**0.94295**	0.95745	0.94495
VIF	0.92141	0.92412	**0.92452**	0.93662	0.93621	**0.93822**	0.85631	**0.83762**	0.83771
VQM	0.91962	**0.92103**	0.92084	0.93353	0.93184	**0.93373**	0.86132	**0.85283**	0.85423

Table 10.5: Random stereo pair: linearity, monotonicity, and accuracy indexes of the objective metrics under consideration.

Metric	PCC			SROCC			RMSE		
	Closer view	Farther view	Average	Closer view	Farther view	Average	Closer view	Farther view	Average
PSNR	0.70779	0.70829	**0.71229**	0.73909	0.74009	**0.74159**	1.59039	1.60419	**1.58809**
PSNR-HVS	0.72168	0.72168	**0.72658**	0.74428	0.74527	**0.74808**	1.55998	1.57548	**1.55648**
PSNR-HVS-M	0.72567	0.72627	**0.73097**	0.74567	0.74528	**0.74977**	1.55427	1.56637	**1.54847**
WSNR	0.75696	0.75876	**0.76336**	0.77356	0.76526	**0.77846**	1.47216	1.47776	**1.46096**
VSNR	0.83685	0.85145	**0.85175**	0.84955	0.84195	**0.85695**	1.16375	1.16745	**1.14365**
SSIM	0.93073	**0.94042**	0.93842	0.93384	**0.94522**	0.94273	0.84523	**0.79492**	0.80562
MS-SSIM	0.90924	0.90504	**0.90994**	0.93383	0.93264	**0.93694**	0.97114	0.99454	**0.97024**
VIF	0.93731	0.94251	**0.94341**	0.94422	0.95001	**0.95111**	0.80981	0.77271	**0.76931**
VQM	0.93142	0.92943	**0.93243**	**0.94661**	0.93923	0.94532	0.83642	0.84483	**0.82793**

Table 10.6: Difference between objective video quality models.

Metric	max|ΔPCC|	max|ΔSROCC|
PSNR	0.2532	0.2317
PSNR-HVS	0.2703	0.2544
PSNR-HVS-M	0.2674	0.2548
WSNR	0.2487	0.2431
VSRN	0.1599	0.1476
SSIM	0.0670	0.0550
MS-SSIM	0.0636	0.0593
VIF	0.0550	0.0408
VQM	0.0345	0.0302

WSNR have a significantly lower correlation with perceived quality than VIF, VQM, SSIM, and MS-SSIM. The difference is particularly strong for the random stereo pair between SNR-based metrics (PCC $\leq$ 0.7633 and SROCC $\leq$ 0.7784) and perceptual metrics (PCC $\geq$ 0.9050 and SROCC $\geq$ 0.9326). In this case, PSNR (PCC $\leq$ 0.7122, SROCC $\leq$ 0.7415) has a significantly lower correlation with perceived quality compared to VIF (PCC $\geq$ 0.9373, SROCC $\geq$ 0.9442). For the fixed stereo pair, all perceptual metrics (PCC $\geq$ 0.9046 and SROCC $\geq$ 0.9295) outperform PSNR (PCC $\leq$ 0.8086 and SROCC $\leq$ 0.8096).

Table 10.6 reports the maximum absolute difference, calculated between the different objective video quality models, of PCC and SROCC values for stereo pairs formed from a decoded view and a synthesized view (see previous section). Only the quality of the decoded view, the quality of the synthesized view, and the average quality of the decoded view and the synthesized view are considered. The difference between the objective video quality models is about four times higher for PSNR, PSNR-HVS, PSNR-HVS-M, and WSNR (max$|\Delta PCC| \geq 0.2487$, max$|\Delta SROCC| \geq 0.2317$) than for the perceptual metrics (max$|\Delta PCC| \leq$ 0.0670, max$|\Delta SROCC| \leq 0.0593$). There is a significant difference in performance between the different objective video quality models for these SNR-based metrics. However, the perceptual metrics have similar performance regardless the objective video quality model.

The results obtained for stereo pairs formed from two synthesized views lead to similar conclusion than the results obtained for stereo pairs formed from a decoded view and a synthesized view. These results indicate that some objective metrics do not predict well perceived quality of synthesized views and we must accept our second hypothesis. This conclusion is in line with the results from (Bosc et al., 2011a; Bosc et al., 2012b).

Let's now consider only the objective metrics that have a high correlation with perceived quality of synthesized views, namely VIF, VQM, SSIM, and MS-SSIM. If there was a masking effect between the decoded view and the synthesized view in the study reported in the previous section, we should have observed a significant difference between the objective video quality model based on the quality of the decoded view and the objective video quality model based on the quality of the synthesized view. However, these metrics have similar performance

regardless the objective video quality model. These results indicate that there is no significant masking effect between a decoded view and a synthesized view and we must reject our first hypothesis.

10.3 Benchmarking of Objective Metrics on Free-Viewpoint Video Sequences

Free-viewpoint systems are meant to provide the viewer with the ability to interactively change his/her viewpoint to enjoy a 3D scene. Among these, FTV is one of the key technologies brought by the development of 3D video applications. It opens the door to new applications in entertainment, post-production, teleconferencing, security applications, etc. These applications are based on a limited number of cameras for recording the 3D scene. Many 3D scene representations have been proposed (Smolic et al., 2009b), among which is the MVD format. The MVD format consists of a set of texture views and associated depth maps acquired at different viewpoints. From color and depth information, new virtual viewpoints can be rendered through DIBR techniques (Fehn, 2004b).

The perceived image quality of free-viewpoint content can be affected at many stages of the processing chain. In particular, the impact of both compression and DIBR algorithms on the quality of virtual viewpoints has been shown (Do et al., 2009; Merkle et al., 2009). Considering compression, VCEG and MPEG have joined their efforts to develop new 3D video coding standards for advanced 3D video applications. These algorithms consider the quality of the synthesized views to optimize compression. As any technology, the performance evaluation of free-viewpoint systems, in terms of quality of user experience, is essential. Therefore, objective quality assessment tools are needed as well. However, very few metrics have been proposed for FTV applications (see Section 8.4) and common full reference 2D metrics, e.g., PSNR and SSIM, are still mostly used (Do et al., 2009; Merkle et al., 2009).

As outlined in (Do et al., 2009; Y. Liu et al., 2009), the 3D warping process involved in the DIBR techniques induces distortions mainly known as *"cracks"* or *"holes"*, which are due to the sampling rate and the discovering of areas not visible from the reference viewpoint, but visible in the new viewpoint, and *"ghosting"*, which is due to the edge resolution in the depth maps. These distortions are different from those commonly encountered in video compression. Moreover, video compression related artifacts are often scattered over the whole image, whereas DIBR related artifacts are mostly located around the disoccluded areas. Most of the commonly used objective quality metrics were initially designed to address video compression related artifacts and are not reliable predictors of perceived quality of monoscopic and stereoscopic video sequences formed from synthesized views, as demonstrated in the previous section and shown by Bosc et al. (2011b). As free-viewpoint systems rely on view synthesis to render new virtual viewpoints, it is legitimate to question the reliability of these metrics to assess the quality of FVV sequences.

In Section 5.1, we reported the results of a subjective quality evaluation designed to assess the quality of FVV sequences corresponding to a smooth camera motion during a time freeze, which were generated through DIBR from 3D content represented in the MVD format. Only depth maps compression was concerned (and not color view compression, as in a classical scenario) since it has been shown that depth compression has a critical impact on the quality of synthesized views. In this section, we analyze and report the performance of several commonly used objective quality metrics using the FVV sequences and corresponding ground truth subjective scores obtained in Section 5.1.

10.3.1 Methodology

In this study, the performance of the following metrics (see Section 8.1) in predicting image quality of FVV sequences was assessed: 1) PSNR, 2) SSIM, 3) MS-SSIM, 4) VIF, 5) VIFp: VIF pixel domain version, 6) UQI, and 7) IFC.

All above objective metrics were computed on the luma component of each frame of the FVV sequence and the resulting values were averaged across the frames to produce a global index for the entire FVV sequence. All objective metrics were computed using MeTriX MuX Visual Quality Assessment Package.

The objective metrics were benchmarked following the procedure described in Chapter 9. First, a cubic function was used to map the objective scores to the subjective ratings. Then, the performance indexes were computed between the predicted MOS values and ground truth MOSs. Finally, to determine whether the difference between two performance index values corresponding to two different objective metrics is statistically significant, statistical tests were performed.

A PCA was also applied between the DMOSs and the objective scores to further investigate the correlation of the objective metrics with perceived quality. As the different metrics have different scales and PCA is sensitive to the relative scaling of the original variables, normalized variables with zero-mean and unit-variance were used.

10.3.2 Results

This section presents the results of statistical analyses that aim at determining the existence of a correlation between the obtained subjective scores and the corresponding objective scores. As stated in Section 10.3.1, a PCA was applied on the DMOS and the objective quality scores of the stimuli. In the following, various aspects regarding the correlation and agreement between subjective and objective scores are discussed.

Table 10.7: Linearity, monotonicity, accuracy, and consistency indexes for the different metrics.

	All contents				Average			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
PSNR	0.2671	0.2945	0.9072	0.5091	0.3284	0.4505	0.5663	0.3452
SSIM	0.0000	0.0000	0.9414	0.5641	0.2202	0.3670	0.6035	0.3741
MS-SSIM	0.0105	0.0611	0.9413	0.5604	0.1870	0.3942	0.6098	0.3960
VIF	0.0584	0.0948	0.9398	0.5714	0.2642	0.3415	0.5836	0.3853
VIFP	0.0798	0.1223	0.9384	0.5678	0.2624	0.3305	0.5847	0.3854
UQI	0.0000	0.0000	0.9414	0.5641	0.2395	0.3441	0.6007	0.3853
IFC	0.1289	0.0657	0.9335	0.5531	0.2808	0.3307	0.5799	0.3741

Correlation between Objective Metrics and Perceived Quality

Table 10.7 reports the linearity, monotonicity, accuracy, and consistency indexes for the cubic mapping. The mapping was applied in two different ways

i) on all contents at once and
ii) on each content separately.

In the latter case, the performance indexes were computed separately on each content and then averaged across contents. When the mapping was applied on all contents at once, the correlation was lower than 0.15 for all metrics, except for PSNR, which showed a correlation around 0.3. The RMSE was around 0.9 for all metrics. The OR was higher than 55% on all metrics. These results show that there is almost no correlation between objective metrics and perceived quality. Note that the correlation for SSIM and UQI is null, which is due to the fact that the cubic mapping function was constrained to be monotonic on the interval of observed quality values whereas the non-fitted scores for these two metrics mostly showed a negative correlation with perceived quality (see Figure 10.7b, which shows the correlation between the obtained subjective scores and the corresponding non-fitted objective scores). When the mapping was applied on each content separately, the obtained performance did marginally improved, as the PCC and SROCC scores are still in the range 0.18-0.33 and 0.33-0.45, respectively. The RMSE and OR decreased below 0.61 and 40%, respectively. However, these results still lead to the conclusion that there is almost no correlation between objective metrics and perceived quality.

When the mapping was applied on all contents at once, PSNR seems to outperform other metrics, even though the correlation was still very low. To determine if the difference between PSNR and the other metrics is significant, statistical tests were performed following the procedures described in Section 9.3. Table 10.8 reports the results of the statistical tests for the cubic mapping. Each entry in the table corresponds to the results of the statistical tests performed on the following performance indexes (from left to right): PCC, SROCC, RMSE, and OR. The statistical tests were performed to determine whether the difference between two performance index values corresponding to two different metrics was statistically

Table 10.8: Statistical analysis of the different metrics.

	PSNR	SSIM	MS-SSIM	VIF	VIFP	UQI	IFC
PSNR		≠≠==	≠≠==	≠≠==	≠≠==	≠≠==	=≠==
SSIM	≠≠==		=≠==	=≠==	=≠==	=≠==	=≠==
MS-SSIM	≠≠==	=≠==		====	====	=≠==	====
VIF	≠≠==	=≠==	====		====	=≠==	====
VIFP	≠≠==	=≠==	====	====		=≠==	====
UQI	≠≠==	=≠==	=≠==	=≠==	=≠==		=≠==
IFC	=≠==	=≠==	====	====	====	=≠==	

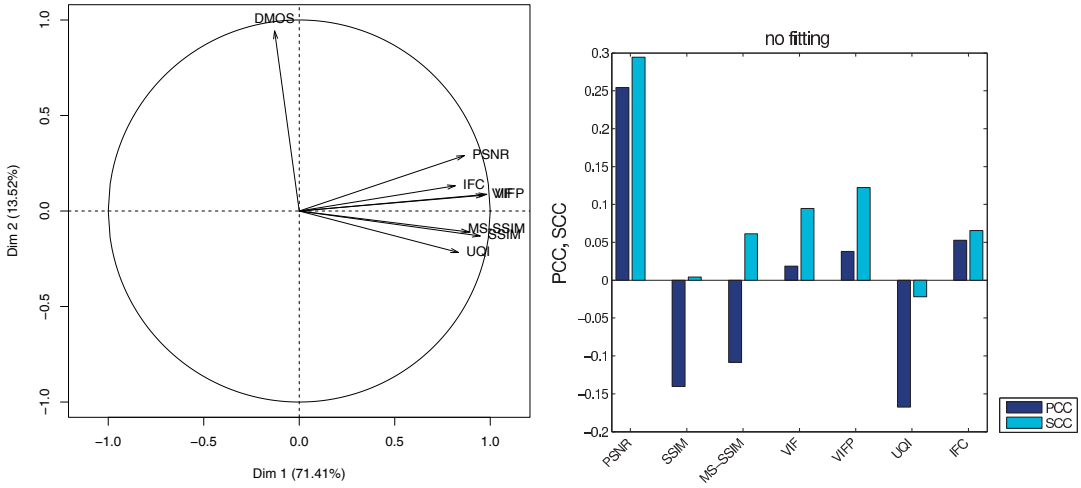
Each entry in the table corresponds to the results of the statistical tests performed on the following performance indexes (from left to right): PCC, SROCC, RMSE, and OR. "=" means that there was no significant difference between the two metrics, whereas "≠" means that the difference was significant. Reading: Line 2, column 4: SSIM and VIF are statistically different according to SROCC, whereas they are similar according to the other performance indexes.

significant: "=" means that there was no significant difference between the two metrics, whereas "≠" means that the difference was significant. The results showed that PSNR was statistically different from the other metrics according to the PCC and SROCC values, except for IFC.

Figure 10.7 shows the correlation between the obtained subjective scores and the corresponding objective scores. Figure 10.7a depicts the circle of correlations derived from the PCA. Figure 10.7b depicts the PCC and SROCC between the DMOS and the objective scores. Only two components had an eigenvalue larger than 1 in the PCA. These components extracted 84% of the inertia.

The circle of correlations allows the observation of correlations between variables and principal components. Each measured variable is represented as a vector. The vector length represents the combined strength of the relationships between measured variable and principal components. The vector direction indicates whether these relationships are positive or negative. Since the data is not perfectly represented by the only two principal components, the variables are positioned inside the circle of correlations. The closer the variable is to the circle, the more important it is to the principal components. The lower the angle between two measured variable's vector representations, the higher their correlation. In Figure 10.7a, it can be observed that the objective metrics are grouped, which shows that they are correlated with each others. However, the angle between most of the objective metrics and DMOS is large (close to $\frac{\pi}{2}$), which indicates that subjective scores are not correlated to objective metrics. This is confirmed by the analyses of PCC and SROCC scores in Figure 10.7b: these correlation scores are very low since they do not reach 0.3.

An other interesting observation concerns the contributions of the variables to the principal components in Figure 10.7a. The variables for which the contribution value is larger than the average contribution for the first component are VIFP, VIF, SSIM, MS-SSIM, and PSNR.



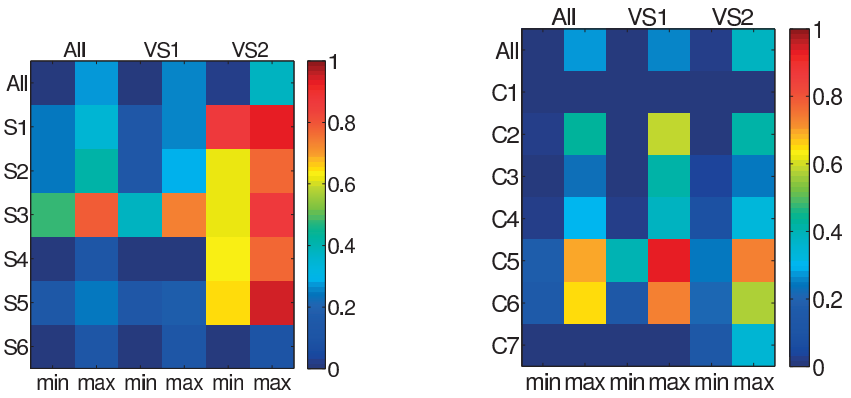
(a) PCA (b) Correlation

Figure 10.7: Circle of correlations and PCC and SROCC between DMOS and objective scores.

The only variable for which the contribution value is larger than the average contribution for the second component is DMOS. Correlation is different from agreement as argued in (Haber and Barnhart, 2006). Considering the directions of the vectors in Figure 10.7a, points in the upper part have large DMOS and those below have low DMOS in the corresponding individual factor map. Points in the left have low objective scores and those in the right have large scores. So points showing the agreement between DMOS and objective scores should be located in upper right part and in the lower left part of the individuals factor map. In the following, qualitative supplementary data (view synthesis modes, contents, and depth map compression algorithms) will be considered and depicted in the individuals factor map. This aspect of agreement will be studied through the results of the PCA applied on the DMOS and the objective quality scores of the stimuli in the individuals factor map, in the following parts.

Scope of Validity of the Objective Metrics

Huynh-Thu and Ghanbari (2008) have shown that even PSNR can be a valid quality measure if the video content and the codec type are not changed. It is well known that objective metrics can better handle some types of degradations and often fail when different types of degradations are combined. In this study, different views synthesis modes, contents, depth map compression algorithms, and bit rates were considered. As it was shown in Section 5.1.3, the view synthesis mode had an impact on perceived quality and modified the behavior of a compression algorithms. Therefore, we benchmarked the different metrics on sub-groups of stimuli, where only one view synthesis mode and one codec were considered. The same analysis was performed with only one view synthesis mode and one compression algorithm. Figure 10.8 shows the minimum and maximum PCC values (across all metrics) for the different sub-groups. It can be observed that the correlation can be quite high when only *VS2* is



(a) Contents (b) Codecs

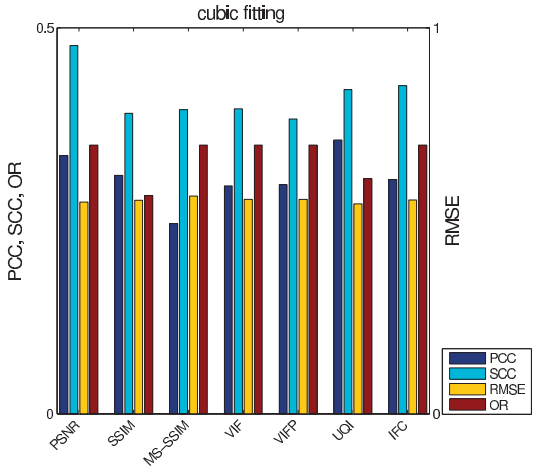
Figure 10.8: Minimum and maximum PCC values across all metrics for the different subgroups. Reading: Contents, line 4, columns 1 and 2: $\min(PCC) \approx 0.5$, $\max(PCC) \approx 0.8$ for content *S3* when all synthesis modes are considered.

considered and the analysis is performed for each content separately (except for content *S6*, where the correlation remains very low).

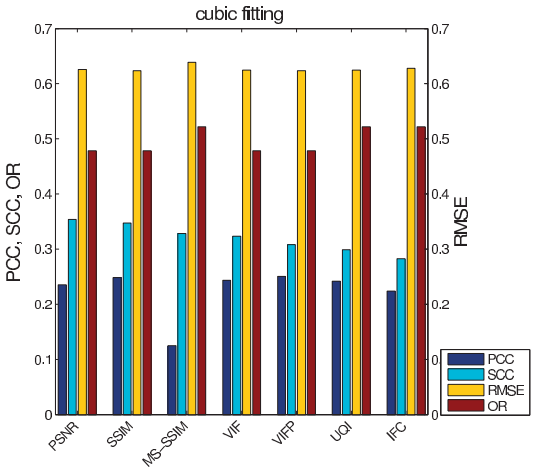
Figure 10.9 depicts the linearity, monotonicity, accuracy, and consistency indexes when considering only content *S1*. The results show that there is almost no correlation between objective metrics and perceived quality when the views are not blended (*VS1*, see Figure 10.9b) as the PCC value is lower than 0.25 on all metrics, whereas the correlation is very strong when the views are blended (*VS2*, see Figure 10.9c) as the PCC value is higher than 0.8 on all metrics. These results show that the objective metrics can achieve a good correlation with perceived quality if content characteristics are considered, but cannot handle the artifacts produced by some view synthesis algorithms.

View Synthesis Modes

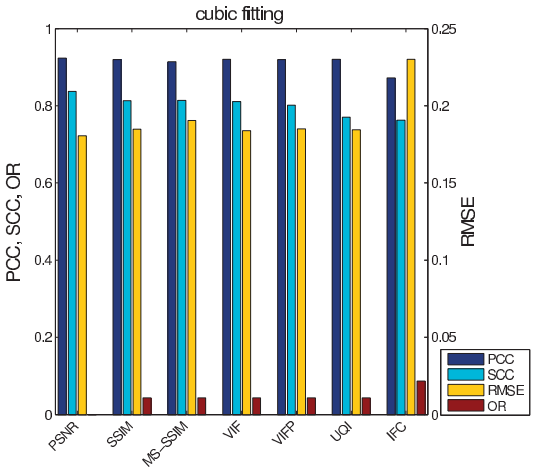
Figure 10.10 shows the individual factor map derived from the PCA with emphasis on the view synthesis modes. In Section 5.1.3, Figures 5.3 and 5.4 showed that *VS1* generally obtained larger DMOS values than *VS2*. Thus, the agreement between DMOS and objective scores regarding the perceived quality of *VS1* and *VS2* related views should be represented as two separated clouds diametrically opposed in the upper right part and in the lower left part of the individual factor map, respectively. However, although the confidence ellipses are clearly non-overlapping and diametrically opposed (upper left part and lower right part), it can be observed that the two clouds are neither located in the expected parts of the plot. This indicates that the objective scores do not correctly express human perception difference between *VS1* and *VS2*.



(a) *VS1* and *VS2*



(b) *VS1* only



(c) *VS2* only

Figure 10.9: Linearity, monotonicity, accuracy, and consistency indexes when considering only content *S1*.

Content Characteristics

Figure 10.11 shows the individual factor map derived from the PCA with emphasis on the contents. The confidence ellipses of contents *S1*, *S4*, and *S5* are clearly located in the upper right part. This indicates that the objective scores obtained with these contents agree with the corresponding subjective scores. In addition, the centroid of content *S2* is close to the center of gravity of the whole set of stimuli. However, two contents seem to involve disagreement between DMOS and objective scores: *S3* (lower right part) and *S6* (upper left part). Content *S3*, in particular, shows interesting results whose explanation can be found in the analysis presented in Section 5.1.3: we observed that contents with highly textured information, negative skew in the distribution of disparity values, and important depth discontinuities might be rated lower by observers. In this case, objective scores disagreed with human perception of visual

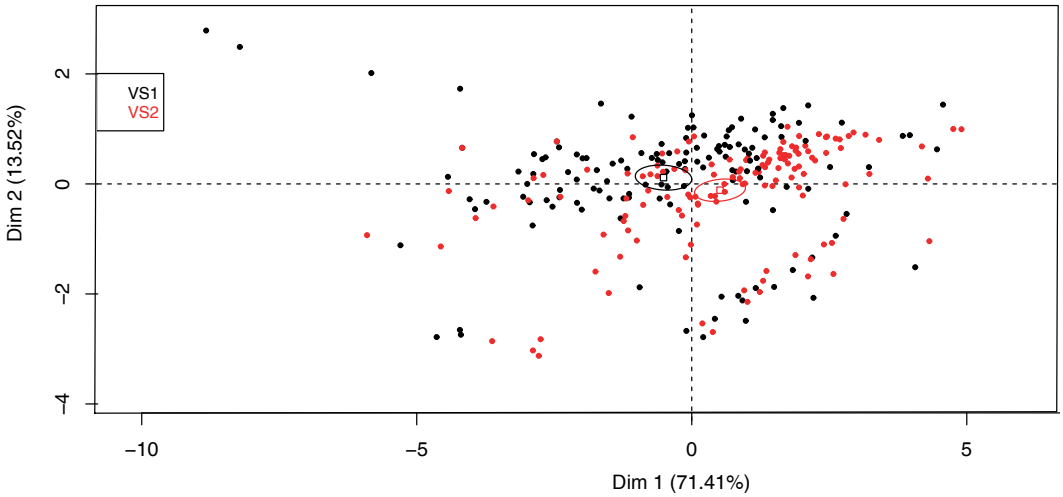


Figure 10.10: PCA plot with graphical emphasis on the synthesis modes.

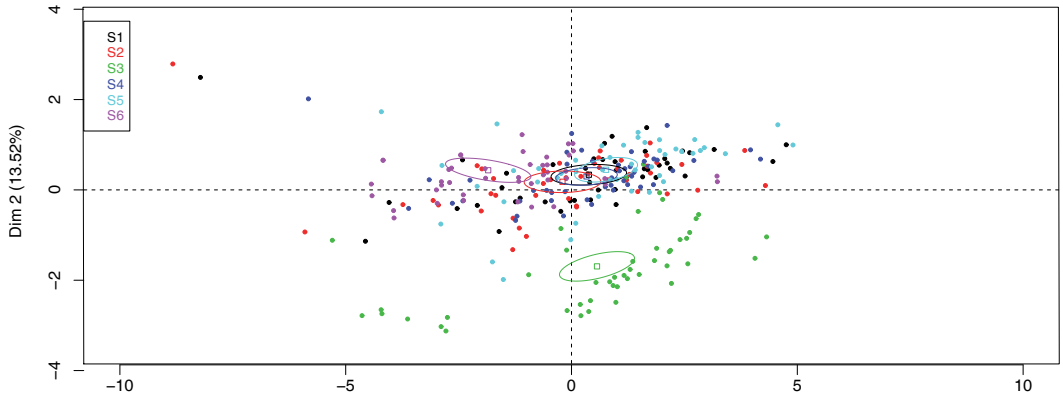


Figure 10.11: PCA plot with graphical emphasis on the contents.

quality regarding *S3*, which contains two highly textured major transversal planes for the wall and floor. As it can be observed, the range of objective scores corresponding to content *S3* is similar to the range of objective scores corresponding to the other contents. However, a significant shift can be observed in the subjective scores. These observations show a clear disagreement between DMOS and objective scores for this content. However, as observed in Section 10.3.2, correlation between objective and subjective scores increased when only specific contents or compression algorithms were considered. In particular, when only the stimuli corresponding to content *S3* were considered, the correlation increased form 0.3 to 0.8 (see Figure 10.8). These observations illustrate the difference between correlation and agreement, as argued in (Haber and Barnhart, 2006).

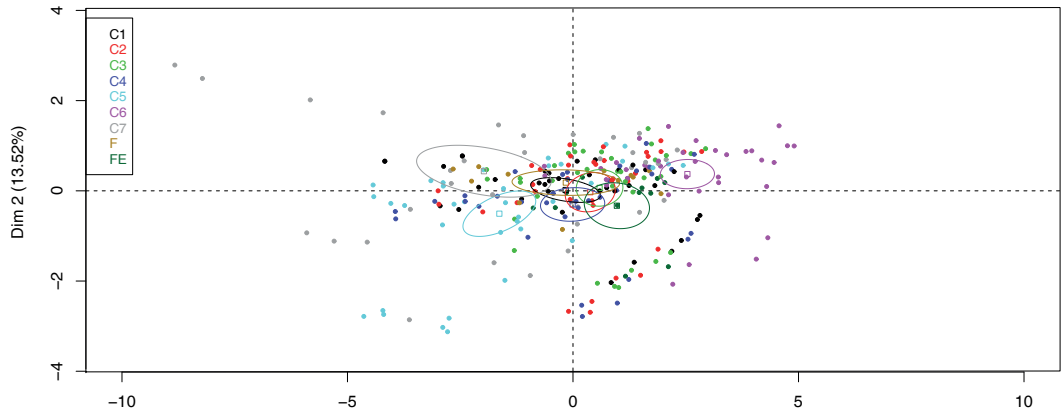


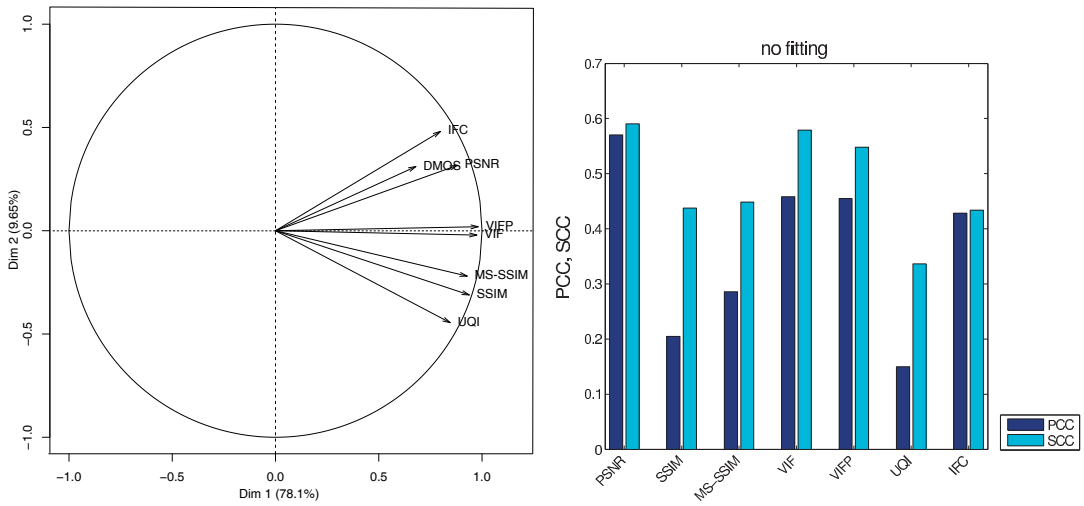
Figure 10.12: PCA plot with graphical emphasis on the compression algorithms.

Depth Map Compression Algorithms

Figure 10.12 shows the individual factor map derived from the PCA with emphasis on the depth map coding methods. The centroids of most of the depth map compression algorithms are located in the upper right part and in the lower left part of the individual factor map except for *C7* (and *F* and *FE*, which are not depth map compression algorithms but additional conditions). *C7* is located in the upper left part of the map. Considering the direction of the relationships between the variables in the circle of correlations, we can argue that generally the subjective and objective scores do not agree on the perceived quality of *C7* related stimuli. As observed in Section 5.1.3, *C7* induces a depth-fading-like distortion to ensure bit rate savings. This flattens the depth maps and involves a global scene shifting rather than motion parallax. Objective quality metrics that are mostly pixel-based methods can hardly predict perceived quality in this case. This explains the location of the centroid related to *C7* in the upper left part of the individual map. Regarding the bit rates, the confidence ellipses for *R0*, *R1*, and *R2* are non-overlapping and located in the upper right part, the center and the lower left part, respectively, as expected (the figure is not presented here).

A Specific Case: *C6*

Figure 10.13 shows the circle of correlations derived from PCA (a) and the PCC and SROCC (b) with subjective and objective scores of *C6* related stimuli only. In Figure 10.13a, the two principal components resumed 87.7% of the total inertia. The variables for which the contribution value is larger than the average contribution for the first component are VIFP, VIF, SSIM, and MS-SSIM. These objective metrics are known to be perception-oriented. The variables for which the contribution value is larger than the average contribution for the second component are IFC, UQI, DMOS, PSNR, and SSIM. In addition, Figure 10.13a shows that the angle between the vectors representing DMOS, IFC, and PSNR are very low, which indicates a large correlation between these variables. This information is confirmed by the results in Figure 10.12 since the ellipse of *C6* is located in the upper right part of the map,



(a) PCA (b) Correlation

Figure 10.13: Circle of correlations and PCC and SROCC scores between DMOS and objective scores when considering subjective and objective scores for compression algorithm *C6*.

which means that DMOS and objective scores agree regarding the image quality of these stimuli. These results are also in line with the obtained correlation scores in Figure 10.13b regarding PSNR: according to PCC and SROCC, PSNR is the most correlated objective metric. Our observation of *C6* related depth maps shows that this coding method distorts only slightly small pixel blocks around the edges. So the quality of the resulting synthesized views is close to that of the reference stimuli, which explains the higher objective scores.

These observations show that objective metrics are strongly content dependent, as previously shown in (Huynh-Thu and Ghanbari, 2008). Therefore, content characteristics should be considered by objective metrics or the benchmarking of objective metrics should be made on a per content basis for fair comparison.

10.4 Benchmarking of Objective Metrics for HDR Image Quality Assessment

For LDR image content, extensive studies have shown that not all metrics can be considered as reliable predictors of perceived quality (Sheikh et al., 2006), whereas only a few recent studies have benchmarked objective metrics for HDR image quality assessment. The study of Valenzise et al. (2014) compared the performance of PSNR and SSIM, computed in the logarithmic and PU (Aydın et al., 2008) domains, and HDR-VDP-2. The authors have concluded that non-uniformity must be corrected for a proper metric application, as most have been designed for perceptual uniform scales. Another subjective study was reported by Mantel et al. (2014). A comparison with objective metrics in physical domain and using a gamma correction to approximate perceptually uniform luminance is also presented, concluding that

the mean relative squared error (MRSE) metric provides best performance in predicting quality. The correlation between thirteen well known full-reference metrics and perceived quality of compressed HDR content is investigated in (Hanhart et al., 2014f). The metrics were applied on the linear domain and results show that only HDR-VDP-2 and FSIM predicted visual quality reasonably well. Finally, Narwaria et al. (2015b) have reported that their HDR-VQM metric performs similar or slightly better than HDR-VDP-2 for HDR image quality assessment.

The main limitation of these studies lie in the small number of images used in their experiments, which was limited to five or six contents. Also, a proper adaptation of the contents to the HDR display and correction of the metrics for non-uniformity were not always considered. Therefore, in this section, we report and analyze the results of an extensive benchmarking of objective quality metrics for HDR image quality assessment. In Section 4.4, we analyzed the performance of JPEG XT on a dataset of 20 HDR image contents compressed at 4 bit rates with profiles A, B, and C. In this section, we analyze and report the performance of 35 objective metrics benchmarked using the 240 HDR images and corresponding ground truth subjective scores obtained in Section 4.4. The objective metrics were computed in the linear, logarithmic, PU (Aydın et al., 2008), and Dolby PQ (Miller et al., 2013) domains. Additionally, the metrics were computed both on the luminance channel alone and as the average quality score of the Y, C_b, and C_r channels. For each metric, objective scores were fitted to subjective scores using logistic fitting. Performance indexes were computed to assess the linearity, monotonicity, accuracy, and consistency of the metrics estimation of subjective scores. Finally, statistical analysis was performed on the performance indexes computed from the 240 data points to discriminate small differences between two metrics.

10.4.1 Methodology

In this study, the performance of the following metrics (see Sections 8.1 and 8.3) in predicting HDR image quality was assessed

A) Full-reference metrics
   i) HDR metrics
      1) HDR-VDP and 2) HDR-VQM.
   ii) Difference measures and statistical-oriented metrics
      3) MSE, 4) PSNR, and 5) SNR.
   iii) Structural similarity measures
      6) UQI, 7) SSIM, 8) MS-SSIM, 9) M-SVD, and 10) QILV.
   iv) Visual information measures
      11) IFC, 12) VIF, 13) VIFP: VIF pixel domain version, and 14) FSIM.
   v) Information weighted metrics
      15) IW-MSE, 16) IW-PSNR, and 17) IW-SSIM.
   vi) HVS inspired metrics
      18) JND_st, 19) WSNR, and 20) DN.

vii) Objective color difference measures
21) CIE1976, 22) CIE94, 23) CMC, and 24) CIEDE2000.

B) No-reference metrics
25) JND, 26) VAR: variance, 27) LAP: laplacian, 28) GRAD: gradient, 29) FTM: frequency threshold metric, 30) HPM: HP Metric, 31) Marziliano: marziliano blurring metric, 32) KurtZhang: kurtosis based metric, 33) KurtWav: kurtosis of wavelet coefficients, 34) AutoCorr: auto correlation, and 35) RTBM: Riemannian tensor based metric.

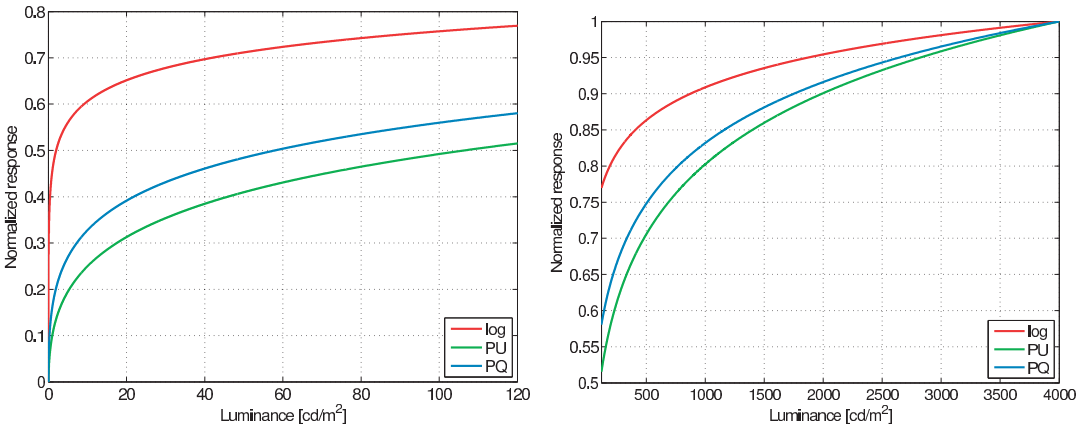
Metrics Computation and Transform Domains

LDR metrics are designed for gamma encoded images, typically having luminance values in the range 0.1-100 cd/m^2, whereas HDR images have linear values and are meant to capture a much wider range of luminance. Therefore, in this study, metrics were computed not only in the linear space but also in transformed spaces that provide a more perceptual uniformity. This space transformation was not applied to HDR-VDP-2 and HDR-VQM, which are calibrated metrics and require absolute luminance values as input. The color difference metrics, i.e., CIE1976, CIE94, CMC, and CIEDE2000, were also not computed in transformed spaces. These color difference measures require a conversion from the RGB representation to the CIELAB color space, considering a D65 100 cd/m^2 reflective white point as reference white point.

Before any metric was computed, images were clipped to the range $[0.001, 4000]$ cd/m^2 (theoretical range of luminance values that the HDR monitor used in the subjective tests can render) to mimic the physical clipping performed by the HDR display. To compute the metrics in the linear domain, these luminance values were normalized to the interval $[0,1]$. This normalization was not applied to HDR metrics and to color difference metrics.

The remaining metrics were computed in three transform domains: the log domain, the PU domain (Aydın et al., 2008), and the PQ domain (Miller et al., 2013). The PU transform is derived using the threshold-integration method (Wilson, 1980). The transform is constrained such that luminance values in the range 0.1-80 cd/m^2, as produced by a typical CRT display, are mapped to the range 0-255 to mimic the sRGB non-linearity. The PQ transform is derived from the Barten contrast sensitivity function (Barten, 1999). The PQ curve has a square-root and log behavior at the darkest and highest light levels, respectively, while it exhibits a slope similar to the gamma non-linearities between those extreme luminance regions. Figure 10.14 depicts the normalized response of the log, PU, and PQ responses in the range $[0, 4000]$ cd/m^2.

These transformations were applied before any normalization and only after their application the resulting color components were normalized to the interval $[0,1]$. After the normalizations, the values considered to be in the RGB color space were transformed to the YC_bC_r color space (ITU-R BT.709, 2015). The exception is the DN metric, which uses directly these RGB components. The metrics were computed on each of these components separately and two final metrics were considered: the quality score computed on the luminance channel alone and the average quality score of the Y, C_b, and C_r channels.



(a) $[0, 120]$ cd/m^2 (b) $[120, 4000]$ cd/m^2

Figure 10.14: Comparison of responses for the transformation functions in two different luminance ranges.

Statistical Analysis

The objective metrics were benchmarked following the procedures described in Chapter 9. First, a 4-parameter logistic function was used to map the objective scores to the subjective ratings. Then, the performance indexes were computed between the predicted MOS values and ground truth MOSs. Finally, to determine whether the difference between two performance index values corresponding to two different objective metrics is statistically significant, statistical tests were performed.

10.4.2 Results

Tables 10.9 to 10.12 report the linearity, monotonicity, accuracy, and consistency indexes for the metrics computed in the different domains. The metrics are sorted from best (top) to least (bottom) performing, based on the different performance indexes (higher PCC/SROCC and lower RMSE/OR values indicate better performance). As HDR-VDP-2 and HDR-VQM require absolute luminance values as input, these metrics were computed neither on the chrominance channels nor in the transform domains. Similarly, the different color difference metrics were computed only in the linear domain, after converting the absolute RGB values to the CIELAB color space. The DN metric was computed on the RGB components, considering all three channels together. Finally, the remaining 28 metrics were computed both on the luminance channel alone (_Y suffix) and as the average quality score of the luminance, blue-difference, and red-difference channels (_M suffix). The statistical analysis results are reported in the same tables. This analysis was performed on the performance indexes computed from 240 data points to discriminate small differences between two metrics. Metrics whose performance indexes are connected by a line are considered statistically not significantly different. For example, in the linear domain, according to PCC, there is no statistical evidence

to show performance differences between IFC and FSIM computed on the luminance channel, but they are statistically different from HDR-VDP-2 (see Table 10.9).

Best Performing Metrics

As expected, HDR-VDP-2 and HDR-VQM, which are the only true HDR quality metrics considered in this study, computed on absolute luminance values, are the best performing metrics when compared to all other metrics and domains. Both metrics have a correlation above 0.95 and a particularly low RMSE (around 0.35) and low OR, whereas all other metrics have an OR above 0.48. HDR-VDP-2 (OR = 0.35) has a slightly lower OR than HDR-VQM (OR = 0.4083), but there is no statistical evidence to show a significant difference. However, HDR-VQM is over three times faster than HDR-VDP-2 (Narwaria et al., 2015b), which makes it a suitable alternative to HDR-VDP-2.

The results for HDR-VDP-2 are in line with the finding of (Hanhart et al., 2014f), slightly better than that of (Valenzise et al., 2014), but in contradiction with Mantel et al. (2014), who reported a much lower correlation. However, Mantel et al. used unusual combinations of parameters for the base and extension layers, especially for content *BloomingGorse*. Narwaria et al. (2015b) found that HDR-VQM was performing significantly better than HDR-VDP-2 for both video and still image content. However, our results show that both metrics have similar performance, while it was reported in (Hanhart et al., 2015c) that HDR-VQM performs lower than HDR-VDP-2 for HDR video compression. The divergence between these findings might be due to the contents and types of artifacts considered in the different studies.

In contrast to the HDR metrics, the NR metrics show the worst performance with PCC and SROCC values below 0.5 and RMSE and OR values above 1 and 0.8, respectively, independently of the domain in which the metric was computed. These results show that NR metrics are not sufficient to reach satisfactory prediction accuracy considering a perceptual domain and that specific NR metrics should be designed for HDR image quality assessment.

Difference Measures and Statistical-Oriented Metrics

Results show that MSE-based metrics, i.e., MSE, SNR, and PSNR, are not very reliable predictors of perceived quality when computed in the linear domain, with correlation between 0.65 and 0.75. Higher PCC values were reported in (Hanhart et al., 2014f) for MSE and SNR (PCC=0.88), but the study was performed considering only five contents. These metrics are know to be very content dependent (Huynh-Thu and Ghanbari, 2008), which might explain the drop in performance when considering 20 images. The correlation of MSE-based metrics computed on the luminance channel alone can be improved by about 0.1 by considering a more perceptual domain than the linear domain, which does not take into account the contrast sensitivity response of the HVS. In the log and PU domains, the correlation is about 0.83 and 0.84, respectively, which is in line with the results from (Valenzise et al., 2014). Nevertheless, the

Table 10.9: Linearity, monotonicity, accuracy, and consistency indexes for each objective metric computed in the linear space.

PLCC		SROCC		RMSE		OR	
HDRVDP2	0.9604	HDRVDP2	0.9564	HDRVDP2	0.3498	HDRVDP2	0.3500
HDRVQM	0.9602	HDRVQM	0.9564	HDRVQM	0.3506	HDRVQM	0.4083
IFC_Y	0.9140	IFC_Y	0.9205	IFC_Y	0.5109	IFC_Y	0.5458
FSIM_Y	0.8938	FSIM_Y	0.9160	FSIM_Y	0.5643	UQI_Y	0.5667
UQI_Y	0.8873	IFC_M	0.8952	UQI_Y	0.5792	IWPSNR_Y	0.6167
IFC_M	0.8855	DN	0.8926	IFC_M	0.5845	IWSSIM_Y	0.6167
DN	0.8814	WSNR_Y	0.8791	DN	0.5936	IWPSNR_M	0.6417
WSNR_Y	0.8786	MSSSIM_Y	0.8776	WSNR_Y	0.5995	IWSSIM_M	0.6417
FSIM_M	0.8571	FSIM_M	0.8768	FSIM_M	0.6476	IFC_M	0.6458
MSSSIM_Y	0.8545	UQI_Y	0.8737	VIF_Y	0.6521	WSNR_Y	0.6500
VIF_Y	0.8545	VIF_Y	0.8617	MSSSIM_Y	0.6527	VIF_Y	0.6583
IWPSNR_Y	0.8352	IWMSE_Y	0.8374	IWPSNR_Y	0.6907	WSNR_M	0.6667
IWSSIM_Y	0.8352	IWPSNR_Y	0.8374	IWSSIM_Y	0.6907	VIFP_Y	0.6667
VIFP_Y	0.8275	IWSSIM_Y	0.8374	VIFP_Y	0.7050	DN	0.6667
WSNR_M	0.8178	VIFP_Y	0.8344	WSNR_M	0.7225	FSIM_M	0.6750
UQI_M	0.8082	IWMSE_M	0.8298	UQI_M	0.7396	FSIM_Y	0.6833
IWPSNR_M	0.8052	WSNR_M	0.8273	IWPSNR_M	0.7446	UQI_M	0.6917
IWSSIM_M	0.8052	IWPSNR_M	0.8145	IWSSIM_M	0.7446	CMC	0.6917
CMC	0.8045	IWSSIM_M	0.8145	CMC	0.7458	VIF_M	0.7000
CIE94	0.7987	UQI_M	0.8112	CIE94	0.7558	IWMSE_M	0.7042
CIEDE00	0.7951	CIE94	0.8031	CIEDE00	0.7615	MSSSIM_M	0.7042
IWMSE_Y	0.7951	CMC	0.8019	IWMSE_Y	0.7622	CIEDE00	0.7083
IWMSE_M	0.7907	VIF_M	0.7946	IWMSE_M	0.7689	CIE94	0.7083
VIF_M	0.7813	CIEDE00	0.7908	VIF_M	0.7836	SNR_M	0.7208
MSVD_Y	0.7629	MSSSIM_M	0.7877	MSVD_Y	0.8120	SNR_Y	0.7333
MSSSIM_M	0.7610	MSVD_Y	0.7758	MSSSIM_M	0.8146	PSNR_M	0.7375
SNR_Y	0.7535	VIFP_M	0.7657	SNR_Y	0.8253	CIE1976	0.7458
VIFP_M	0.7520	MSVD_M	0.7600	VIFP_M	0.8275	VIFP_M	0.7500
MSVD_M	0.7466	SNR_Y	0.7574	MSVD_M	0.8355	IWMSE_Y	0.7542
SSIM_Y	0.7374	SSIM_Y	0.7559	SSIM_Y	0.8482	MSSSIM_Y	0.7583
CIE1976	0.7254	MSE_Y	0.7262	CIE1976	0.8642	MSVD_M	0.7667
PSNR_Y	0.7152	PSNR_Y	0.7262	PSNR_Y	0.8774	SSIM_M	0.7708
MSE_Y	0.7050	CIE1976	0.7256	MSE_Y	0.8924	PSNR_Y	0.7792
SNR_M	0.6872	MSE_M	0.6978	SNR_M	0.9120	QILV_M	0.7792
MSE_M	0.6777	SNR_M	0.6969	MSE_M	0.9232	MSVD_Y	0.7833
PSNR_M	0.6525	PSNR_M	0.6628	PSNR_M	0.9513	MSE_M	0.8000
QILV_Y	0.6183	QILV_Y	0.6615	QILV_Y	0.9867	QILV_Y	0.8042
QILV_M	0.6164	QILV_M	0.6535	QILV_M	0.9886	SSIM_Y	0.8125
SSIM_M	0.5904	SSIM_M	0.5948	SSIM_M	1.0132	MSE_Y	0.8167
Marziliano_Y	0.4551	Marziliano_Y	0.4160	Marziliano_Y	1.1178	Marziliano_M	0.8292
Marziliano_M	0.3669	HPM_M	0.3341	Marziliano_M	1.1678	Marziliano_Y	0.8333
HPM_Y	0.3625	HPM_Y	0.3211	HPM_Y	1.1700	JND_St_M	0.8375
HPM_M	0.3503	Marziliano_M	0.3093	HPM_M	1.1758	RTBM_Y	0.8375
JND_St_Y	0.2913	JND_St_Y	0.2393	JND_St_Y	1.2009	RTBM_M	0.8375
JND_St_M	0.2509	JND_St_M	0.1599	JND_St_M	1.2152	JND_St_Y	0.8500
GRAD_M	0.1060	LAP_M	0.1197	GRAD_M	1.2483	VAR_Y	0.8500
GRAD_Y	0.1012	LAP_Y	0.1030	GRAD_Y	1.2489	VAR_M	0.8500
KurtZhang_M	0.0829	GRAD_M	0.0784	KurtZhang_M	1.2513	LAP_M	0.8500
KurtWav_M	0.0709	GRAD_Y	0.0742	KurtWav_M	1.2522	GRAD_M	0.8500
RTBM_Y	0.0709	KurtWav_M	0.0607	RTBM_Y	1.2522	FTM_Y	0.8500
AutoCorr_Y	0.0675	KurtZhang_Y	0.0518	AutoCorr_Y	1.2525	HPM_Y	0.8500
AutoCorr_M	0.0664	KurtZhang_M	0.0366	AutoCorr_M	1.2526	KurtZhang_Y	0.8500
RTBM_M	0.0623	AutoCorr_Y	0.0301	RTBM_M	1.2529	KurtWav_Y	0.8500
LAP_M	0.0537	AutoCorr_M	0.0297	LAP_M	1.2535	KurtWav_M	0.8500
LAP_Y	0.0458	KurtWav_Y	0.0265	LAP_Y	1.2540	JND_Y	0.8542
FTM_Y	0.0324	JND_M	0.0138	FTM_Y	1.2547	JND_M	0.8542
KurtWav_Y	0.0290	JND_Y	0.0110	KurtWav_Y	1.2548	LAP_Y	0.8542
JND_Y	0.0200	RTBM_Y	−0.0095	JND_Y	1.2551	FTM_M	0.8542
JND_M	0.0198	VAR_Y	−0.0221	JND_M	1.2551	KurtZhang_M	0.8542
KurtZhang_Y	0.0194	VAR_M	−0.0354	KurtZhang_Y	1.2551	AutoCorr_Y	0.8542
FTM_M	0.0082	FTM_Y	−0.0396	FTM_M	1.2553	AutoCorr_M	0.8542
VAR_M	0.0068	RTBM_M	−0.0421	VAR_M	1.2553	GRAD_Y	0.8625
VAR_Y	0.0067	FTM_M	−0.0748	VAR_Y	1.2553	HPM_M	0.8625

Table 10.10: Linearity, monotonicity, accuracy, and consistency indexes for each objective metric computed in the logarithm space.

PLCC		SROCC		RMSE		OR	
VIFP_Y	0.9230	VIFP_Y	0.9200	VIFP_Y	0.4832	VIFP_Y	0.4833
VIF_Y	0.9185	VIF_Y	0.9174	VIF_Y	0.4974	IFC_Y	0.5500
IFC_Y	0.9051	IFC_Y	0.9112	IFC_Y	0.5355	VIF_Y	0.5583
MSSSIM_Y	0.8971	MSSSIM_Y	0.9091	MSSSIM_Y	0.5560	UQI_Y	0.5917
IFC_M	0.8928	IFC_M	0.9037	IFC_M	0.5672	SSIM_Y	0.6125
SSIM_Y	0.8900	SSIM_Y	0.8952	SSIM_Y	0.5727	PSNR_Y	0.6208
UQI_Y	0.8780	FSIM_Y	0.8817	UQI_Y	0.6009	IFC_M	0.6250
FSIM_Y	0.8553	UQI_Y	0.8603	FSIM_Y	0.6516	MSSSIM_Y	0.6292
WSNR_Y	0.8404	UQI_M	0.8441	WSNR_Y	0.6803	MSE_Y	0.6375
UQI_M	0.8373	WSNR_Y	0.8416	UQI_M	0.6870	MSVD_Y	0.6500
PSNR_Y	0.8348	MSE_Y	0.8399	PSNR_Y	0.6911	IWPSNR_Y	0.6542
MSVD_Y	0.8316	PSNR_Y	0.8399	MSVD_Y	0.6975	SNR_Y	0.6542
MSE_Y	0.8272	MSVD_Y	0.8370	MSE_Y	0.7055	IWSSIM_Y	0.6542
SNR_Y	0.8269	SNR_Y	0.8333	SNR_Y	0.7060	UQI_M	0.6583
IWPSNR_Y	0.8160	IWPSNR_Y	0.8165	IWPSNR_Y	0.7256	WSNR_Y	0.6625
IWSSIM_Y	0.8160	IWSSIM_Y	0.8165	IWSSIM_Y	0.7256	DN	0.6625
VIF_M	0.8079	IWMSE_Y	0.8165	VIF_M	0.7401	FSIM_Y	0.6708
VIFP_M	0.7986	VIF_M	0.8152	VIFP_M	0.7556	MSE_M	0.6958
IWMSE_Y	0.7912	VIFP_M	0.8082	IWMSE_Y	0.7680	VIF_M	0.7000
DN	0.7877	DN	0.7993	DN	0.7737	IWPSNR_M	0.7083
MSSSIM_M	0.7482	FSIM_M	0.7603	MSSSIM_M	0.8330	IWSSIM_M	0.7083
FSIM_M	0.7363	MSSSIM_M	0.7584	FSIM_M	0.8498	VIFP_M	0.7125
WSNR_M	0.7252	WSNR_M	0.7324	WSNR_M	0.8644	MSSSIM_M	0.7125
QILV_Y	0.6918	QILV_Y	0.6913	QILV_Y	0.9088	WSNR_M	0.7208
SSIM_M	0.6855	SSIM_M	0.6847	SSIM_M	0.9139	PSNR_M	0.7250
MSE_M	0.6785	MSVD_M	0.6772	MSE_M	0.9222	SSIM_M	0.7250
MSVD_M	0.6779	IWPSNR_M	0.6580	MSVD_M	0.9229	IWMSE_Y	0.7333
IWPSNR_M	0.6646	IWSSIM_M	0.6580	IWPSNR_M	0.9380	SNR_M	0.7417
IWSSIM_M	0.6646	PSNR_M	0.6409	IWSSIM_M	0.9380	IWMSE_M	0.7458
SNR_M	0.6415	SNR_M	0.6394	SNR_M	0.9630	MSVD_M	0.7583
PSNR_M	0.6412	MSE_M	0.6360	PSNR_M	0.9633	QILV_Y	0.7583
IWMSE_M	0.6162	IWMSE_M	0.5931	IWMSE_M	0.9890	FSIM_M	0.8125
HPM_Y	0.4900	QILV_M	0.5265	HPM_Y	1.0944	KurtWav_M	0.8333
Marziliano_Y	0.4855	HPM_Y	0.4874	Marziliano_Y	1.0975	HPM_Y	0.8417
Marziliano_M	0.4059	Marziliano_Y	0.4303	Marziliano_M	1.1473	AutoCorr_Y	0.8458
GRAD_Y	0.3736	HPM_M	0.3518	GRAD_Y	1.1645	JND_St_Y	0.8542
GRAD_M	0.2844	Marziliano_M	0.3056	GRAD_M	1.2035	JND_St_M	0.8542
JND_St_M	0.2591	GRAD_Y	0.2570	JND_St_M	1.2125	JND_Y	0.8542
LAP_Y	0.2153	GRAD_M	0.1915	LAP_Y	1.2270	JND_M	0.8542
VAR_M	0.1654	LAP_M	0.1736	VAR_M	1.2381	VAR_Y	0.8542
QILV_M	0.1427	LAP_Y	0.1642	KurtWav_M	1.2425	LAP_M	0.8542
KurtWav_M	0.1425	VAR_M	0.1548	VAR_Y	1.2426	FTM_Y	0.8542
VAR_Y	0.1423	FTM_M	0.1314	QILV_M	1.2427	KurtZhang_M	0.8542
KurtZhang_Y	0.1053	KurtZhang_Y	0.1268	KurtZhang_Y	1.2484	KurtWav_Y	0.8542
FTM_M	0.0736	VAR_Y	0.1226	FTM_M	1.2520	AutoCorr_M	0.8542
HPM_M	0.0599	KurtWav_M	0.0961	HPM_M	1.2531	RTBM_Y	0.8542
AutoCorr_M	0.0560	AutoCorr_Y	0.0754	AutoCorr_M	1.2534	RTBM_M	0.8542
KurtZhang_M	0.0429	JND_St_M	0.0752	KurtZhang_M	1.2542	QILV_M	0.8583
JND_M	0.0402	KurtZhang_M	0.0726	JND_M	1.2543	FTM_M	0.8583
JND_Y	0.0401	KurtWav_Y	0.0543	JND_Y	1.2543	HPM_M	0.8583
AutoCorr_Y	0.0379	JND_M	0.0516	AutoCorr_Y	1.2545	Marziliano_Y	0.8583
RTBM_M	0.0315	JND_Y	0.0498	RTBM_M	1.2547	KurtZhang_Y	0.8583
KurtWav_Y	0.0312	RTBM_M	0.0429	KurtWav_Y	1.2547	LAP_Y	0.8625
LAP_M	0.0306	FTM_Y	0.0230	LAP_M	1.2548	GRAD_Y	0.8625
FTM_Y	0.0227	AutoCorr_M	−0.0019	FTM_Y	1.2551	VAR_M	0.8667
RTBM_Y	0.0173	JND_St_Y	−0.0482	RTBM_Y	1.2553	Marziliano_M	0.8750
JND_St_Y	0.0038	RTBM_Y	−0.0494	JND_St_Y	1.2553	GRAD_M	0.8792

Table 10.11: Linearity, monotonicity, accuracy, and consistency indexes for each objective metric computed in the PU space.

PLCC		SROCC		RMSE		OR	
MSSSIM_Y	0.9447	MSSSIM_Y	0.9501	MSSSIM_Y	0.4132	VIF_Y	0.4833
FSIM_Y	0.9376	FSIM_Y	0.9470	FSIM_Y	0.4377	IWPSNR_Y	0.5167
VIF_Y	0.9291	VIF_Y	0.9276	VIF_Y	0.4649	IWSSIM_Y	0.5167
VIFP_Y	0.9288	VIFP_Y	0.9228	VIFP_Y	0.4656	VIFP_Y	0.5208
IWPSNR_Y	0.9130	IFC_Y	0.9170	IWPSNR_Y	0.5121	IFC_Y	0.5375
IWSSIM_Y	0.9130	IWMSE_Y	0.9109	IWSSIM_Y	0.5121	MSSSIM_Y	0.5417
IFC_Y	0.9110	IWPSNR_Y	0.9109	IFC_Y	0.5196	WSNR_Y	0.5583
DN	0.9078	IWSSIM_Y	0.9109	DN	0.5275	FSIM_Y	0.5625
SSIM_Y	0.9060	DN	0.9090	SSIM_Y	0.5316	SSIM_Y	0.5792
WSNR_Y	0.8959	SSIM_Y	0.9072	WSNR_Y	0.5577	UQI_Y	0.5833
IFC_M	0.8928	IFC_M	0.9043	IFC_M	0.5670	DN	0.5875
IWMSE_Y	0.8841	WSNR_Y	0.8950	IWMSE_Y	0.5878	PSNR_Y	0.5917
UQI_Y	0.8777	MSVD_Y	0.8638	UQI_Y	0.6016	SNR_Y	0.5958
MSVD_Y	0.8612	UQI_Y	0.8610	MSVD_Y	0.6392	IWMSE_Y	0.6250
PSNR_Y	0.8526	MSE_Y	0.8564	PSNR_Y	0.6562	IFC_M	0.6375
SNR_Y	0.8472	PSNR_Y	0.8564	SNR_Y	0.6669	MSVD_Y	0.6375
MSE_Y	0.8352	SNR_Y	0.8556	MSE_Y	0.6915	VIF_M	0.6625
FSIM_M	0.8310	FSIM_M	0.8484	FSIM_M	0.6991	VIFP_M	0.6667
UQI_M	0.8278	MSSSIM_M	0.8442	UQI_M	0.7049	UQI_M	0.6833
VIF_M	0.8275	VIF_M	0.8373	VIF_M	0.7053	IWPSNR_M	0.6917
MSSSIM_M	0.8273	UQI_M	0.8335	MSSSIM_M	0.7059	WSNR_M	0.6917
VIFP_M	0.8242	VIFP_M	0.8327	VIFP_M	0.7109	IWSSIM_M	0.6917
WSNR_M	0.8163	WSNR_M	0.8233	WSNR_M	0.7254	MSE_Y	0.7000
IWPSNR_M	0.7848	QILV_Y	0.8047	IWPSNR_M	0.7781	QILV_M	0.7042
IWSSIM_M	0.7848	IWPSNR_M	0.7937	IWSSIM_M	0.7781	SSIM_M	0.7083
MSVD_M	0.7837	IWSSIM_M	0.7937	MSVD_M	0.7804	MSSSIM_M	0.7083
QILV_Y	0.7779	MSVD_M	0.7862	QILV_Y	0.7922	IWMSE_M	0.7458
IWMSE_M	0.7414	IWMSE_M	0.7700	IWMSE_M	0.8426	SNR_M	0.7458
MSE_M	0.7280	MSE_M	0.7444	MSE_M	0.8651	FSIM_M	0.7458
SSIM_M	0.7194	SSIM_M	0.7324	SSIM_M	0.8719	PSNR_M	0.7542
SNR_M	0.7085	SNR_M	0.7147	SNR_M	0.8859	MSVD_M	0.7542
PSNR_M	0.7033	PSNR_M	0.7088	PSNR_M	0.8925	QILV_Y	0.7542
QILV_M	0.6789	QILV_M	0.6739	QILV_M	0.9218	MSE_M	0.7708
Marziliano_Y	0.5114	HPM_Y	0.4442	Marziliano_Y	1.0788	HPM_Y	0.8375
HPM_Y	0.4548	Marziliano_Y	0.4179	HPM_Y	1.1181	Marziliano_M	0.8375
Marziliano_M	0.4217	HPM_M	0.3679	Marziliano_M	1.1383	JND_St_Y	0.8417
HPM_M	0.4004	Marziliano_M	0.3378	HPM_M	1.1503	LAP_Y	0.8417
JND_St_Y	0.2975	GRAD_Y	0.2040	JND_St_Y	1.1985	AutoCorr_M	0.8458
LAP_Y	0.1824	GRAD_M	0.1869	LAP_Y	1.2343	RTBM_M	0.8458
VAR_Y	0.1736	VAR_M	0.1387	VAR_Y	1.2363	VAR_Y	0.8500
GRAD_M	0.1618	VAR_Y	0.1258	GRAD_M	1.2389	GRAD_M	0.8500
GRAD_Y	0.1599	RTBM_Y	0.1223	GRAD_Y	1.2397	FTM_Y	0.8500
VAR_M	0.1031	KurtZhang_Y	0.1044	VAR_M	1.2487	FTM_M	0.8500
LAP_M	0.0948	LAP_M	0.0858	LAP_M	1.2497	AutoCorr_Y	0.8500
RTBM_Y	0.0946	RTBM_M	0.0744	RTBM_Y	1.2498	JND_St_M	0.8542
AutoCorr_Y	0.0860	LAP_Y	0.0713	AutoCorr_Y	1.2507	JND_Y	0.8542
KurtZhang_Y	0.0803	KurtWav_M	0.0634	KurtZhang_Y	1.2513	JND_M	0.8542
AutoCorr_M	0.0609	KurtWav_Y	0.0596	AutoCorr_M	1.2530	GRAD_Y	0.8542
RTBM_M	0.0577	FTM_Y	0.0578	RTBM_M	1.2533	HPM_M	0.8542
FTM_Y	0.0560	AutoCorr_Y	0.0518	FTM_Y	1.2534	KurtZhang_M	0.8542
JND_St_M	0.0545	JND_M	0.0515	JND_St_M	1.2538	KurtWav_Y	0.8542
KurtWav_M	0.0422	JND_Y	0.0499	JND_M	1.2545	KurtWav_M	0.8542
JND_M	0.0361	KurtZhang_M	0.0357	JND_Y	1.2545	RTBM_Y	0.8542
JND_Y	0.0360	AutoCorr_M	0.0356	KurtWav_Y	1.2552	LAP_M	0.8625
KurtWav_Y	0.0143	JND_St_M	0.0321	KurtZhang_M	1.2553	KurtZhang_Y	0.8625
KurtZhang_M	0.0093	JND_St_Y	0.0313	FTM_M	1.2553	VAR_M	0.8667
FTM_M	0.0090	FTM_M	−0.0193	KurtWav_M	1.2553	Marziliano_Y	0.8708

Table 10.12: Linearity, monotonicity, accuracy, and consistency indexes for each objective metric computed in the PQ space.

PLCC		SROCC		RMSE		OR	
MSSSIM_Y	0.9380	MSSSIM_Y	0.9435	MSSSIM_Y	0.4366	VIF_Y	0.4917
VIFP_Y	0.9301	FSIM_Y	0.9361	VIFP_Y	0.4613	VIFP_Y	0.4958
VIF_Y	0.9292	VIF_Y	0.9272	VIF_Y	0.4646	IFC_Y	0.5333
FSIM_Y	0.9240	VIFP_Y	0.9242	FSIM_Y	0.4812	IWPSNR_Y	0.5458
SSIM_Y	0.9107	IFC_Y	0.9151	SSIM_Y	0.5188	IWSSIM_Y	0.5458
IFC_Y	0.9093	SSIM_Y	0.9117	IFC_Y	0.5243	MSSSIM_Y	0.5542
IWPSNR_Y	0.9025	IFC_M	0.9039	IWPSNR_Y	0.5407	WSNR_Y	0.5750
IWSSIM_Y	0.9025	IWPSNR_Y	0.9024	IWSSIM_Y	0.5407	SNR_Y	0.5792
IFC_M	0.8930	IWSSIM_Y	0.9024	IFC_M	0.5667	SSIM_Y	0.5792
WSNR_Y	0.8893	IWMSE_Y	0.9022	WSNR_Y	0.5743	UQI_Y	0.5875
DN	0.8887	DN	0.8917	DN	0.5768	PSNR_Y	0.5917
UQI_Y	0.8767	WSNR_Y	0.8890	UQI_Y	0.6039	FSIM_Y	0.6042
IWMSE_Y	0.8730	MSE_Y	0.8656	IWMSE_Y	0.6132	IWMSE_Y	0.6250
MSVD_Y	0.8604	PSNR_Y	0.8656	PSNR_Y	0.6400	DN	0.6375
PSNR_Y	0.8603	MSVD_Y	0.8646	MSVD_Y	0.6409	MSVD_Y	0.6417
SNR_Y	0.8511	UQI_Y	0.8603	SNR_Y	0.6592	IFC_M	0.6500
MSE_Y	0.8451	SNR_Y	0.8589	MSE_Y	0.6721	VIF_M	0.6542
UQI_M	0.8285	MSSSIM_M	0.8359	UQI_M	0.7037	VIFP_M	0.6583
VIF_M	0.8282	VIF_M	0.8358	VIF_M	0.7039	WSNR_M	0.6833
VIFP_M	0.8224	UQI_M	0.8344	VIFP_M	0.7143	UQI_M	0.6875
MSSSIM_M	0.8181	FSIM_M	0.8315	MSSSIM_M	0.7224	QILV_Y	0.6917
FSIM_M	0.8129	VIFP_M	0.8302	FSIM_M	0.7318	QILV_M	0.6917
WSNR_M	0.8047	WSNR_M	0.8127	WSNR_M	0.7455	MSE_M	0.7083
QILV_Y	0.7744	QILV_Y	0.7964	QILV_Y	0.7946	IWPSNR_M	0.7083
MSVD_M	0.7671	IWPSNR_M	0.7734	MSVD_M	0.8058	SSIM_M	0.7083
IWPSNR_M	0.7653	IWSSIM_M	0.7734	IWPSNR_M	0.8081	IWSSIM_M	0.7083
IWSSIM_M	0.7653	MSVD_M	0.7690	IWSSIM_M	0.8081	MSE_Y	0.7167
SSIM_M	0.7243	IWMSE_M	0.7368	SSIM_M	0.8655	MSSSIM_M	0.7292
MSE_M	0.7219	MSE_M	0.7362	MSE_M	0.8688	PSNR_M	0.7417
IWMSE_M	0.7125	SSIM_M	0.7359	IWMSE_M	0.8810	SNR_M	0.7417
SNR_M	0.7041	SNR_M	0.7117	SNR_M	0.8914	IWMSE_M	0.7458
PSNR_M	0.7007	PSNR_M	0.7088	PSNR_M	0.8956	MSVD_M	0.7500
QILV_M	0.6601	QILV_M	0.6411	QILV_M	0.9430	FSIM_M	0.7708
Marziliano_Y	0.5065	HPM_Y	0.4685	Marziliano_Y	1.0824	LAP_Y	0.8375
HPM_Y	0.4717	Marziliano_Y	0.4199	HPM_Y	1.1069	HPM_Y	0.8375
Marziliano_M	0.4213	HPM_M	0.3486	Marziliano_M	1.1385	Marziliano_M	0.8417
HPM_M	0.4108	Marziliano_M	0.3267	HPM_M	1.1445	AutoCorr_M	0.8458
LAP_Y	0.1929	GRAD_Y	0.2241	LAP_Y	1.2318	RTBM_M	0.8458
VAR_M	0.1797	GRAD_M	0.1917	VAR_M	1.2349	JND_St_Y	0.8500
GRAD_Y	0.1733	VAR_M	0.1483	GRAD_Y	1.2368	FTM_Y	0.8500
JND_St_Y	0.1603	VAR_Y	0.1273	GRAD_M	1.2403	HPM_M	0.8500
GRAD_M	0.1556	RTBM_Y	0.1177	JND_St_Y	1.2411	JND_St_M	0.8542
VAR_Y	0.1048	KurtZhang_Y	0.1062	VAR_Y	1.2485	JND_Y	0.8542
LAP_M	0.0978	LAP_M	0.1004	LAP_M	1.2493	JND_M	0.8542
KurtZhang_Y	0.0867	LAP_Y	0.0846	KurtZhang_Y	1.2506	GRAD_Y	0.8542
AutoCorr_Y	0.0724	KurtWav_M	0.0780	AutoCorr_Y	1.2521	GRAD_M	0.8542
RTBM_Y	0.0651	FTM_Y	0.0642	RTBM_Y	1.2527	FTM_M	0.8542
KurtWav_M	0.0587	RTBM_M	0.0639	KurtWav_M	1.2532	KurtZhang_M	0.8542
JND_Y	0.0373	AutoCorr_Y	0.0578	JND_Y	1.2545	KurtWav_Y	0.8542
FTM_Y	0.0372	KurtWav_Y	0.0546	FTM_Y	1.2545	KurtWav_M	0.8542
JND_M	0.0362	JND_M	0.0511	JND_M	1.2545	AutoCorr_Y	0.8542
AutoCorr_M	0.0333	JND_Y	0.0506	AutoCorr_M	1.2547	RTBM_Y	0.8542
RTBM_M	0.0324	KurtZhang_M	0.0358	RTBM_M	1.2547	VAR_M	0.8583
KurtWav_Y	0.0166	JND_St_Y	0.0310	KurtWav_Y	1.2552	Marziliano_Y	0.8583
JND_St_M	0.0128	AutoCorr_M	0.0264	JND_St_M	1.2552	KurtZhang_Y	0.8625
KurtZhang_M	0.0119	JND_St_M	−0.0238	KurtZhang_M	1.2553	LAP_M	0.8667
FTM_M	−0.0086	FTM_M	−0.0521	FTM_M	1.2553	VAR_Y	0.8708

performance of the MSE-based metrics computed as the average quality score of the Y, C_b, and C_r channels did not improve when considering perceptual domains. These observations indicate that the log, PU, and PQ domains can better represent the luminance sensitivity of the HVS than the linear domain, but they might not be optimal for the chrominance sensitivity.

Objective Color Difference Measures

In the linear domain, the color difference metrics, with the exception of the original CIE1976 color difference metric, are the best performing pixel-based metrics. They outperform the MSE-based metrics, but there is no statistical evidence to show a significant improvement over SNR computed on the luminance alone. Nevertheless, their correlation with perceived visual quality is only about 80%, with an OR above 69%, which cannot be considered as reliable prediction. Since the release of the CIE1976 color difference metric, two extensions have been developed in 1994 and 2000 to better address perceptual non-uniformities of the HVS. But, according to the benchmarking results, further improvements might be necessary for HDR images to handle non-uniformities in low and high luminance ranges, outside of the typical range of LDR displays. The color difference metrics are computed in the CIELAB color space, which considers relative luminance values with respect to a reference white point, typically a reflective D65 white point about 100-120 cd/m^2. This reference white point is similar to the targeted peak luminance that is typically considered when calibrating LDR reference monitors. Therefore, for HDR images, one would be tempted to set the luminance of the reference white point considered in the color conversion equal to the peak luminance of the HDR monitor. However, this leads to lower performance of the color difference metrics and the reflective white point should also be used for HDR content instead.

Structural Similarity and Visual Information Measures

The performance of SSIM and its multiscale extension, MS-SSIM, is improved by considering logarithm instead of linear values, and are even further improved by considering the PU or PQ transform. In particular, on the luminance channel, the correlation of SSIM is increased by about 0.15 from linear to logarithm, while MS-SSIM improved by only about 0.03. From log to PU/PQ, improvements are relatively low for SSIM, whereas MS-SSIM exhibits a gain of about 0.04. Results show that MS-SSIM (luminance only) performs the best in PU and PQ spaces according to the PCC, SROCC, and RMSE indexes. The correlation obtained for SSIM in the log and PU domains is similar to the results of Valenzise et al. (2014). On the other hand, UQI, which corresponds to the special case of SSIM when the constants C_1 and C_2 are set to 0, does not perform better in the log, PU, or PQ space than in the linear domain. Similar correlation results for SSIM and MS-SSIM are reported in (Hanhart et al., 2014f) as in this paper (for the linear domain). However, it is reported that the relative change between the worst and best qualities for SSIM and MS-SSIM was less than 0.003% and 0.0003%, respectively. In this study, the average relative change computed over all domains is 16.5% and 11.5% for SSIM and MS-SSIM, respectively. One major difference between the two works is the use of

absolute luminance values in (Hanhart et al., 2014f), whereas luminance values were linearly mapped from the theoretical display range to the range [0,1] in this paper. For LDR content, SSIM uses different values for C_1 and C_2 depending on whether the images are in the range [0,1] or [0,255]. For HDR content, our findings suggest that the value of these constants should be adjusted according the luminance range and depending on whether scaling of the values is performed or not.

Metrics that quantify the loss of image information, i.e., VIF, its pixel-based version, VIFP, and its predecessor, IFC, also show good performance. In particular, IFC (luminance only) is the second best performing metric in the linear domain. While the performance of IFC is not influenced by the domain in which the metric is computed, the performance of VIF(P) is significantly improved when considering a more perceptual domain than the linear space. In the log domain, results show that VIF computed on the luminance alone is the best performing metric. Note that the correlation reported for VIF(P) in this paper is significantly better than the one reported in (Hanhart et al., 2014f). Similarly to (MS-)SSIM, the difference might be due to the scaling procedure. Among the other HVS-based metrics, FSIM also shows good performance, especially in the PU and PQ space (RMSE below 0.5). In the linear domain, results are similar to our previous work.

Statistical Analyses

To determine how the best metrics of each domain compare to each others, a direct benchmarking of the two HDR metrics, which are the best performing metrics in the linear space, and the best performing metric of the log, PU, and PQ spaces was performed. The PSNR metric computed on the luminance channel in the log space was added to this comparison, as this metric is widely used in HDR compression studies. Table 10.13 reports the results of the statistical analysis of the six metrics. To identify metrics computed in the log, PU, and PQ spaces, the *LOG_*, *PU2*, and *PQ2* prefixes are used, respectively. According to PCC and SROCC, there is no statistical evidence to show performance differences between HDR-VDP-2, HDR-VQM, and MS-SSIM computed on the luminance channel in the PU space. However, HDR-VDP-2 and HDR-VQM have a significantly lower RMSE than all other metrics. Figure 10.15 depicts the scatter plots of subjective versus objective results for these metrics. As it can be observed, the data points are well concentrated near the fitting curve for HDR-VDP-2, as well as for HDR-VQM, while they are more scattered for the other metrics, especially in the case of LOG_PSNR_Y, which shows higher content dependency. These findings indicate that HDR-VDP-2 and HDR-VQM have a very high consistency when compared to the other metrics. Nevertheless, HDR-VDP-2 is complex and requires heavy computational resources, which limits its use in many applications. HDR-VQM and MS-SSIM computed in the PU space are lower complexity alternatives to HDR-VDP-2.

The statistical analysis was also used to understand whether there is a statistically significant difference between the performance of each metric when computed on the luminance component alone and when computed on all components. Only results from the analysis

Table 10.13: Statistical analysis comparing the HDR metrics and best performing metric of each domain.

PLCC		SROCC		RMSE		OR	
HDRVDP2	0.9604	HDRVDP2	0.9564	HDRVDP2	0.3498	HDRVDP2	0.3500
HDRVQM	0.9602	HDRVQM	0.9564	HDRVQM	0.3506	HDRVQM	0.4083
PU2MSSSIM_Y	0.9447	PU2MSSSIM_Y	0.9501	PU2MSSSIM_Y	0.4132	LOG_VIFP_Y	0.4833
PQ2MSSSIM_Y	0.9380	PQ2MSSSIM_Y	0.9435	PQ2MSSSIM_Y	0.4366	PU2MSSSIM_Y	0.5417
LOG_VIFP_Y	0.9230	LOG_VIFP_Y	0.9200	LOG_VIFP_Y	0.4832	PQ2MSSSIM_Y	0.5542
LOG_PSNR_Y	0.8348	LOG_PSNR_Y	0.8399	LOG_PSNR_Y	0.6911	LOG_PSNR_Y	0.6208

Table 10.14: Comparison of the 28 metrics computed on the Y and YC_bC_r channels. Comparison of the metrics computed as the average quality score of the Y channel alone and as the average quality score of the YC_bC_r channels.

	lin				log				PU				PQ			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
Y is better	6	7	8	3	16	14	16	8	16	14	15	14	14	14	15	14
similar	22	21	20	25	11	14	12	20	12	14	13	14	14	14	13	14
YC_bC_r is better	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

performed on the 28 metrics that were computed both on the Y channel alone and as the average quality score of the Y, C_b, and C_r channels were considered. Table 10.14 reports the number of metrics for which one approach was significantly better than the other one, as well as when no significant difference between the two approaches was observed. The analysis was performed individually for each performance index and domain. In the linear domain, there is no statistical evidence to show performance differences between the two approaches for about 80% of the metrics. However, in the log, PU, and PQ space, roughly half of the metrics perform significantly better when computed on the luminance channel alone. According to PCC, the JND metric, FR version, computed in the log domain, is the only case for which better performance is achieved when considering all channels. As HDR is often considered in combination with WCG, it is expected that the fidelity of color reproduction will play a more important role in the context of HDR when compared to LDR. We believe that improvements can be achieved by considering different domains for computing the metrics on the chrominance channels and by using better pooling strategies.

Similarly, the statistical analysis was also used to understand whether there is a statistically significant difference between the performance of a particular metric computed in one domain and another domain. Only results from the analysis performed on the 57 metrics that were computed in all domains were considered. Table 10.15 reports the number of times a metric computed in the domain i performs significantly better than when computed in the domain j, where i and j are the row and column of the table. Results show that most metrics perform the best in the PU and PQ spaces when compared to the lin and log spaces, which is in line with our previous observations. Note that results based on PCC, SROCC, and RMSE are in agreement, while the OR metric shows fewer cases where statistically significant difference are observed. Additionally, there are also metrics for which computations performed in the

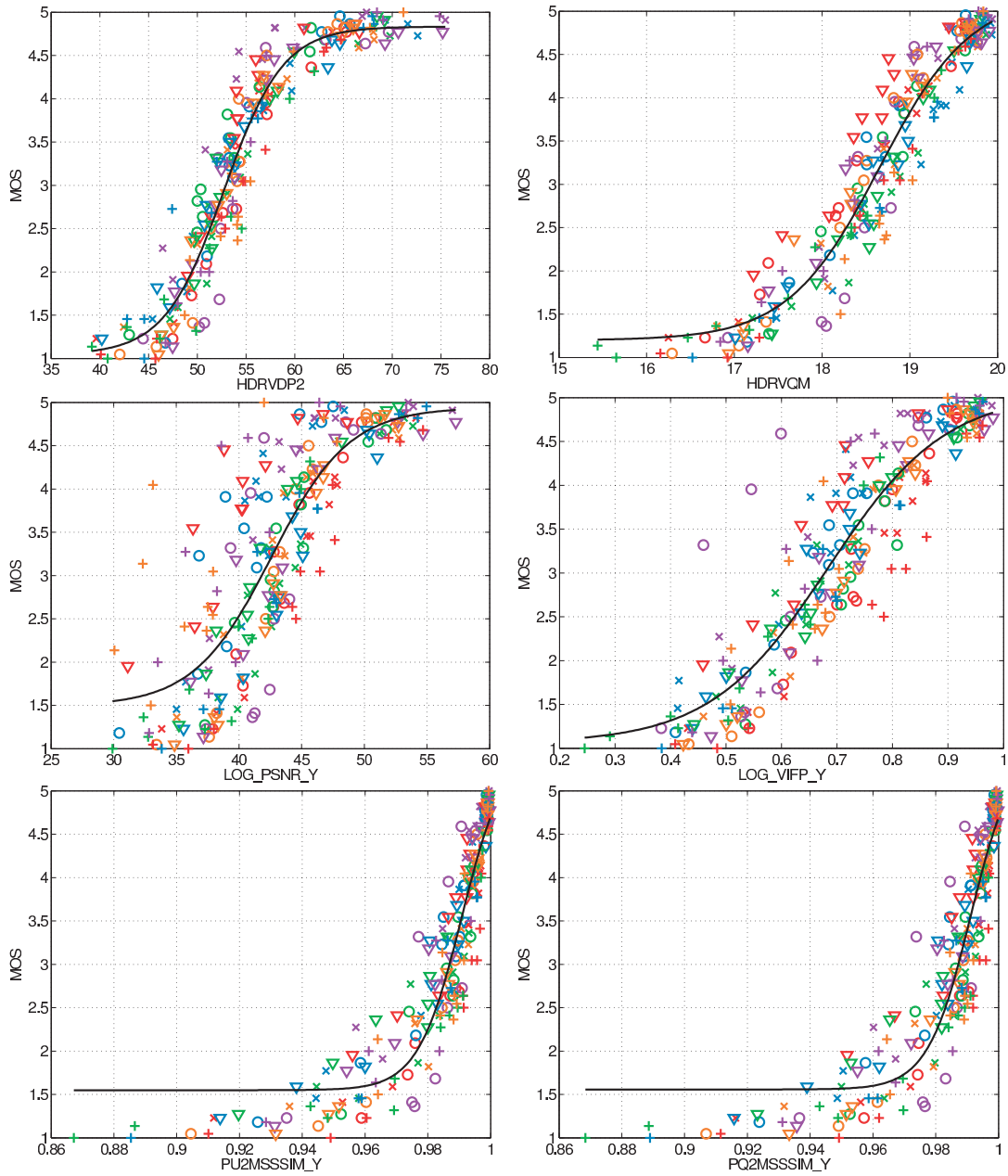


Figure 10.15: Subjective versus objective evaluations results for the HDR metrics and best performing metric of each domain. Each symbol, i.e., combination of marker and color, corresponds to a specific content.

linear and logarithm domains perform better than in the PU and PQ space. Overall, there is no optimal domain that performs the best for all metrics. Instead, different metrics should use different domains to maximize the correlation with perceived quality.

Table 10.15: Comparison of the 57 metrics computed on all domains. Results represent the number of times a metric computed in the domain i performs significantly better than when computed in the domain j, where i and j are the row and column of the table.

	PCC				SROCC				RMSE				OR			
	lin	log	PU	PQ	lin	log	PU	PQ	lin	log	PU	PQ	lin	log	PU	PQ
lin	0	9	5	6	0	9	6	8	0	8	5	5	0	2	1	1
log	10	0	2	2	10	0	4	4	8	0	0	0	8	0	3	3
PU	14	17	0	11	13	15	0	11	11	16	0	10	9	7	0	7
PQ	13	17	8	0	11	15	8	0	11	16	6	0	9	7	3	0

10.5 Effectiveness of Objective Metrics for HDR Video Quality Assessment

For LDR video content, extensive studies have shown that not all metrics can be considered as reliable predictors of perceived quality (Seshadrinathan et al., 2010), whereas only a few recent studies have benchmarked objective metrics for HDR video quality assessment. Azimi et al. (2014) have investigated the performance of HDR-VDP-2, PSNR, SSIM, and VIF, where the last three metrics were computed both in the PU domain (Aydın et al., 2008) and using multi-exposure (Munkberg et al., 2006). Five HDR video contents were impaired with additive white Gaussian noise, salt & pepper noise, low pass filter, mean intensity shift, and compression artifacts using HEVC Main 10 profile at four QP values, leading to a total of 40 HDR impaired video sequences. Results showed that HDR-VDP-2 had the highest correlation when considering only compression artifacts, but the lowest correlation when considering the other artifacts. When considering the non-compression artifacts or all artifacts, the VIF metric computed in the PU domain had the highest correlation. Rerabek et al. (2015a) investigated the quality of five HDR video sequences compressed with HEVC at four bit rates using the PC method. The subjective results were then converted to MOS-like scores using the Thurstone Case V model to benchmark seven objective metrics, which were computed in different variations, e.g., color components and transformed domains. Results showed that HDR-VDP-2 had by far the highest correlation with perceived quality. PSNR computed in the PQ domain (Miller et al., 2013) on the luma component, as well as MSE, MS-SSIM, and VIF (pixel domain version) computed in the PU domain on the luma component, also had correlation above 0.7. Narwaria et al. (2015b) found that HDR-VQM is the best metric, far beyond HDR-VDP-2. However, in their later study (Narwaria et al., 2015c), HDR-VQM was found to be only slightly better than HDR-VDP-2. The divergence between these findings might be due to the contents and types of artifacts considered in the different studies. Indeed, mostly computer generated contents were considered in (Narwaria et al., 2015b), whereas video sequences captured using HDR video cameras were considered in the other studies.

The previous studies on HDR video quality assessment were focused on the correlation between objective and subjective scores. However, none of these studies investigated the effectiveness of objective metrics to discriminate between quality levels when comparing two video sequences, which is essential for codec comparison and optimization. To address this

problem, we computed the classification errors of several metrics applied on compressed HDR video sequences. We used as ground truth the results from the 176 paired comparisons obtained in Section 4.5, where we compared the performance of potential technologies for HDR video compression against HEVC on a dataset of five HDR video contents compressed at four bit rates. This section reports the details and results of this performance evaluation.

10.5.1 Methodology

In this study, the performance of the following objective metrics (see Sections 8.1 and 8.3) was assessed

A) Metrics computed in linear domain
   1) PSNR_DEx: PSNR of mean of absolute value of deltaE2000 metric, derived with x as reference luminance value,
   2) PSNR_Lx: PSNR of mean square error of L component of the CIELab color space used for the deltaE2000 metric, derived with x as reference luminance value,
   3) HDR-VDP-2, and
   4) HDR-VQM.

B) Metrics computed in PQ domain (Miller et al., 2013)
   5) tPSNR-x: PSNR computed on x component,
   6) PQ2SSIM,
   7) PQ2MS-SSIM, and
   8) PQ2VIFP: VIF pixel domain version.

C) Metrics computed using multi-exposure (Munkberg et al., 2006)
   9) mPSNR.

SSIM, MS-SSIM, and VIFP were computed using MeTriX MuX Visual Quality Assessment Package. For these three metrics, the luminance information was extracted from the RGB values, clipped to the range $[0.005, 4000]$ cd/m^2, transformed using the PQ EOTF, and normalized to the interval $[0, 255]$ before computing the metric. The MATLAB implementations of HDR-VDP-2 and HDR-VQM were used. The remaining metrics were computed using HDRTool version 0.9 (M35471). For contents *ShowGirl2* and *WarmNight*, the top and bottom black borders were discarded when computing the metrics.

To benchmark the objective metrics, the classification errors were computed following the procedure described in Section 9.5. To compute the classification errors, the one-tailed binomial test described in Section 4.5.2 was performed at 5% significance level to determine which video sequence in a pair has the best visual quality or if no difference is perceived. The classification errors were computed considering all contents together.

10.5.2 Results

Figure 10.16 reports the classification errors for each metric separately. Even though the results are reported in the native scale of the metric instead of a common scale, it is still possible to compare the classification errors of the different metrics by looking at the relative ΔOM ratio (ΔOM divided by the maximum value of ΔOM) rather than the absolute ΔOM.

Subjective results reported in Section 4.5.3 showed that there were many cases where the Proponent version was providing similar quality when compared to the Anchor. More precisely, in 55% of the cases, no statistically significant difference was observed between Proponent and Anchor, while the difference was statistically significant in 45% of the cases. These values determine the plateau for the *Correct Decision* and *False Tie* frequencies, i.e., if the threshold on ΔOM is set to infinite, all pairs of video sequences are considered as equal for the objective metric, which will lead to a *Correct Decision* frequency of 55%, as 55% of the pairs were evaluated as not statistically different in the subjective evaluations. Similarly, the plateau for the *False Tie* frequency is 45%.

On Figure 10.16, dashed lines indicate ΔOM that maximizes the *Correct Decision* frequency. As it can be observed, the maximum of *Correct Decision* is between 0.55 and 0.71. In particular, for HDR-VQM and mPSNR, the highest *Correct Decision* frequency corresponds to the plateau, i.e., the metric cannot distinguish quality. The results for HDR-VQM are quite surprising, as this is the only metric designed to assess quality of HDR video sequences and it was reported to have a relatively low OR (Narwaria et al., 2015b). The reason might be due to the data used by Narwaria *et al.*to train and validate the metric in their experiments on video quality. In particular, they used seven computer-generated contents and only three real scenes, while it is known that computer generated content has very different noise characteristics. Additionally, they used their own backward-compatible HDR compression scheme to generate distortions, which might be very different from that of the algorithms considered in the CfE evaluations.

The PSNR metric provides similar results on the different components considered in this study. In all cases, the highest *Correct Decision* frequency is about 60%, which means that it cannot reliably detect visible differences. Additionally, the *False Ranking* frequency decay is very slow, i.e., the probability of making the wrong decision remains, even for large relative ΔOM ratio. It is known that PSNR is not good at handling different types of artifacts (Huynh-Thu and Ghanbari, 2008), which explains the relatively low performance when comparing compression algorithms based on different schemes.

Regarding SSIM, MS-SSIM, and VIFP computed in the PQ domain, results show that these metrics achieve similar results to PSNR in terms of *Correct Decision*. They have a faster decay for the *False Ranking* frequency, but a slower for the *False Differentiation* frequency. Surprisingly, MS-SSIM shows slightly lower performance than SSIM in terms of *Correct Decision*, while the multiscale approach usually improves performance for SDR content.

PSNR-DE1000 shows the highest *Correct Decision* frequency with a peak at about 0.71, but

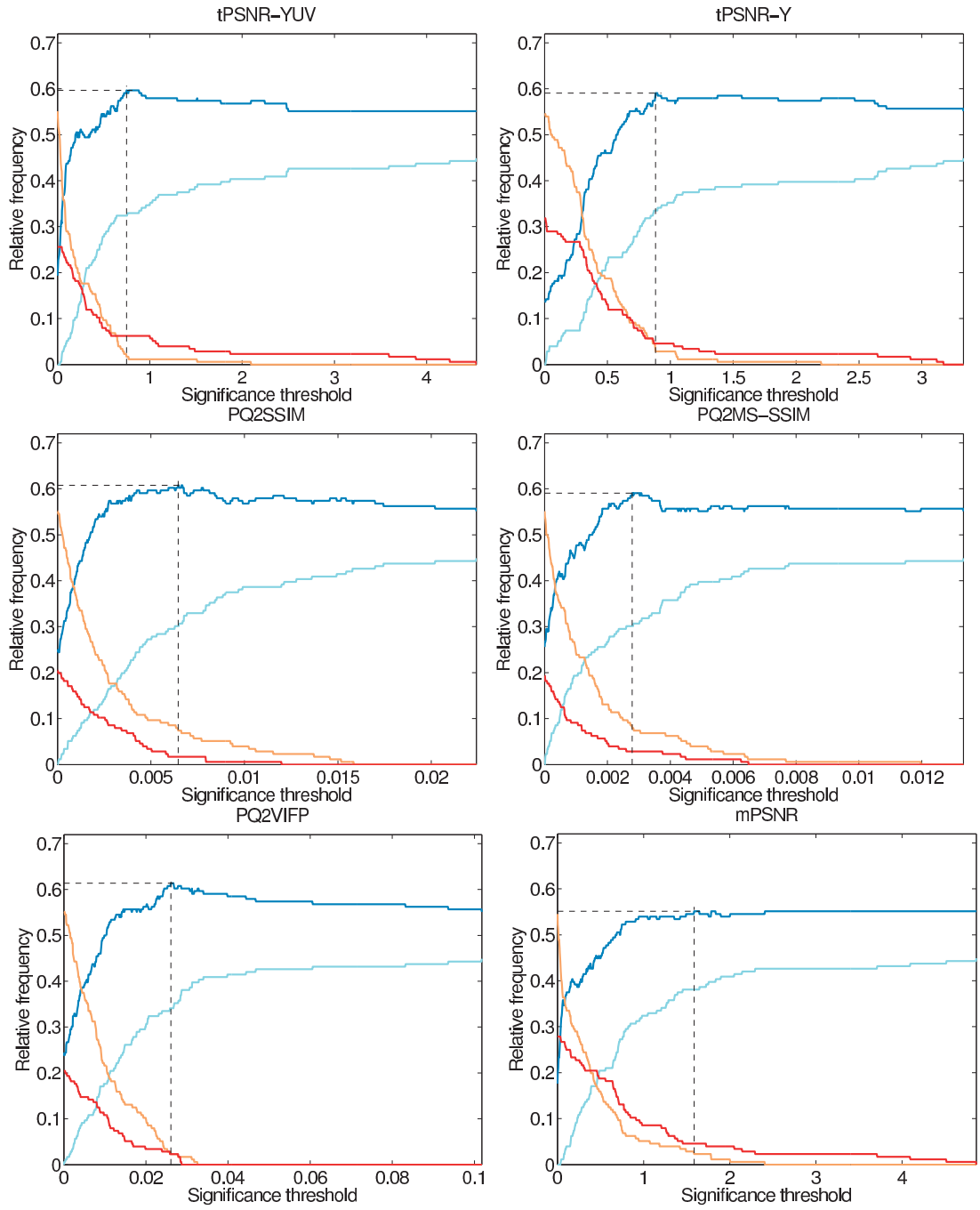


Figure 10.16: Frequencies of classification error.

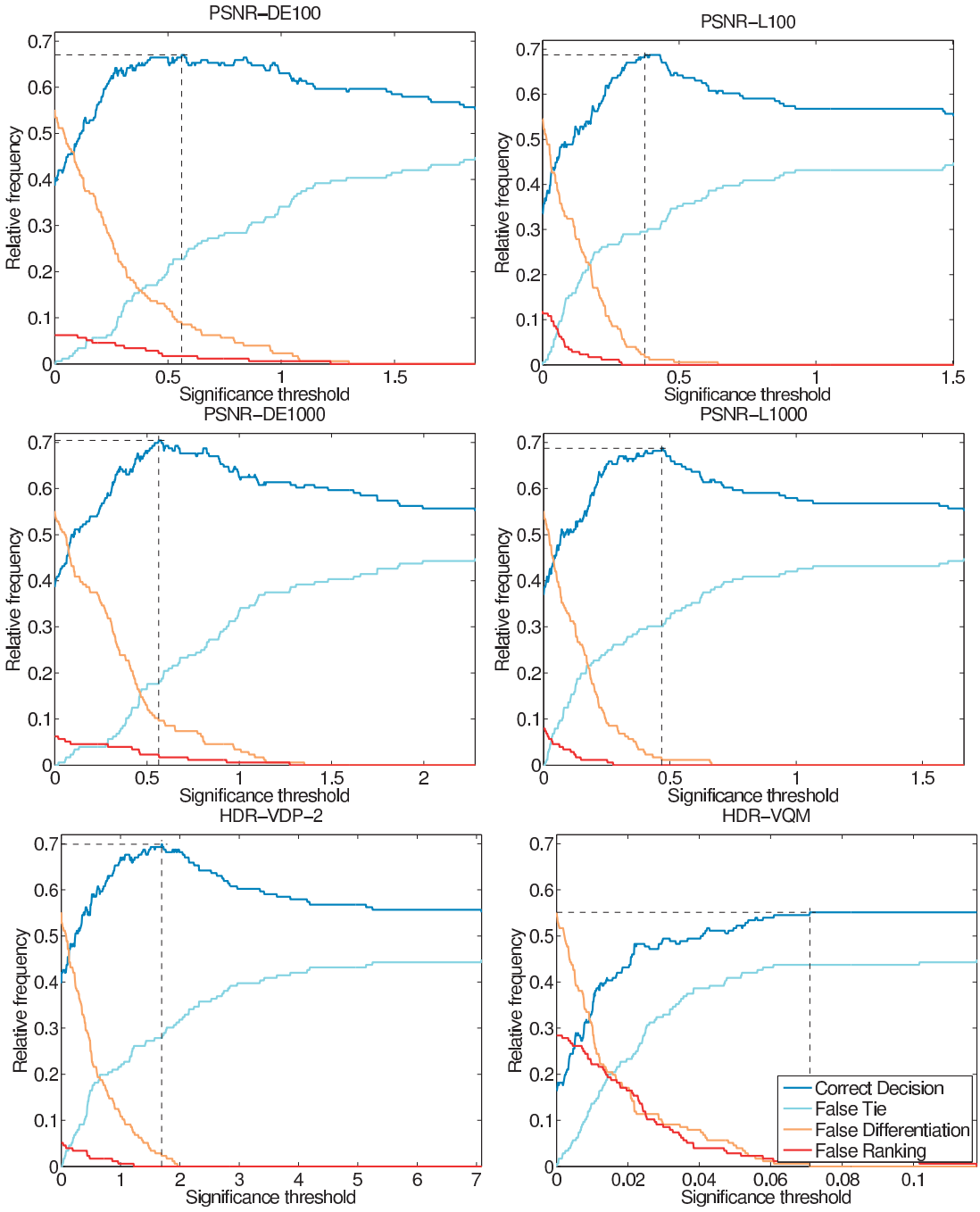


Figure 10.16: Frequencies of classification error *(Continued)*.

the *False Ranking* and *False Differentiation* frequencies are not null at the peak. PSNR-L100, PSNR-L1000, and HDR-VDP-2 seem to be better alternatives, as they have a faster decay for the *False Ranking* frequency and reach similar *Correct Decision* frequency for a *False Ranking* frequency of 0. In particular, PSNR-L100 and PSNR-L1000 show slightly less *False Differentiation* compared to HDR-VDP-2. Considering that HDR-VDP-2 has a very high complexity and requires a lot of processing time when compared to the other metrics, PSNR-Lx seems to be a good alternative.

10.6 Conclusion

This chapter reported the results of performance evaluation of several objective metrics in different scenarios. In particular, we investigated the correlation between different state-of-the-art objective 2D metrics and the perceived quality of a stereo pair formed from a decoded view and a synthesized view. Results showed that, in general, the measured quality of the decoded view has the highest correlation in terms of the Pearson correlation coefficient with perceived quality. Similar performance can be achieved when considering the average quality of both views. However, if the objective quality assessment is based on the measured quality of the synthesized view, it is suggested to use VIF, VQM, MS-SSIM, or SSIM since they significantly outperform other objective metrics, including PSNR. These four objective metrics have similar performance when using the decoded view, the synthesized view, and both views.

We also investigated the correlation between different state-of-the-art objective 2D metrics and the perceived quality of a stereo pair formed from two synthesized views. Results showed that PSNR, PSNR-HVS, PSNR-HVS-M, and WSNR have a significantly lower correlation with perceived quality than VIF, VQM, SSIM, and MS-SSIM. From these observations and those of the study on stereo pair formed from a decoded view and a synthesized view, we concluded that some objective metrics do not predict well perceived quality of synthesized views and that there is no significant masking effect between a decoded view and a synthesized view.

Additionally, we analyzed the performance of several commonly used objective quality metrics on FVV sequences. The considered FVV sequences were generated from decompressed data and simulating a smooth camera motion during a time freeze. The results showed that objective metrics achieved low correlation with subjective scores when various conditions were considered. However, the correlation with perceived quality improved when content characteristics were considered. In addition, the artifacts produced by some view synthesis algorithms might not be correctly handled by the objective quality metrics. These results motivate the need to design better objective metrics that can accurately assess the specific artifacts generated by the view synthesis process.

In this chapter, we also benchmarked 35 objective metrics on a database of 240 compressed HDR images. Additionally to the linear space, metrics were computed in the logarithm, PU, and PQ domains to mimic non-linearities of the HVS. Results showed that the performance of most full-reference metrics could be improved by considering perceptual transforms when

compared to linear values. On the other hand, our findings suggested that a lot of work remains to be done for no-reference quality assessment of HDR content. Our benchmark demonstrated that HDR-VDP-2 and HDR-VQM are ultimately the most reliable predictors of perceived quality. Nevertheless, HDR-VDP-2 is complex and requires heavy computational resources, which limits its use in many applications. HDR-VQM is over three times faster, which makes it a suitable alternative to HDR-VDP-2. Alternatively, MS-SSIM computed in the PU space is another lower complexity substitute, as there is no statistical evidence to show performance differences between these metrics in terms of PCC and SROCC. Even though the numbers of contents and compressed images considered in the experiments are quite large, different performance might be observed for other contents and types of artifacts.

Finally, we investigated the effectiveness of nine objective quality metrics to discriminate between quality levels when comparing two HDR video sequences. We computed the classification errors using as ground truth the results from 176 paired comparisons from a dataset of 5 HDR video contents compressed at 4 bit rates with 10 different compression algorithms. Results showed that PSNR-DE1000, HDR-VDP-2, and PSNR-Lx can reliably detect visible differences between two HDR video sequences, whereas HDR-VQM could not distinguish quality differences.

11 Predicting 3D Quality based on Content Analysis

With the rapid growth of 3D video technologies, the design of objective quality assessment methods, i.e., metrics, that can reliably predict the quality of 3D content as perceived by the end user, is of crucial importance. Subjective tests are time consuming, expensive, and not always feasible. Therefore, objective measurements are needed, especially to assess advances in the design of coding technologies. Despite the efforts of the scientific community in recent years, 3D video quality assessment is still an open challenge. There are no metrics that are widely recognized as reliable predictors of human 3D quality perception. PSNR is commonly accepted and used by video coding experts to evaluate the performance of coding algorithms, even though its correlation with human perception of visual quality is known to be limited (see Sections 10.1 and 10.2).

PSNR values below 25 dB and over 40 dB are often considered as bad and excellent quality, respectively. However, the exact relationship between PSNR values and perceived quality has not been established yet. This relationship should consider non-linearities and saturation effect of the HVS. As it was shown that PSNR is strongly content dependent (Huynh-Thu and Ghanbari, 2008), this relationship should also be determined for each content separately.

Korhonen and You (2010) have found a strong correlation between the parameters of an exponential function, which was used to map PSNR values to MOSs, and the spatial and temporal activity of a set of six 2D video sequences. Based on this finding, they have used a linear regression to estimate the parameters of the mapping function based on the spatial and temporal activity of the six contents.

Liao et al. (2013) have shown how the QoE of a set of 2D video sequences was correlated with objective quality metrics, video content characteristics, and device features. From these results, a linear mapping between MS-SSIM and QoE was proposed. The authors assumed that the parameters of the linear mapping can be accurately estimated from the amount of spatial details, motion level, display resolution, and device type, but this assumption was not investigated.

In the study reported in this chapter, we investigate the prediction of perceived quality of stereoscopic video sequences based on PSNR and content analysis. We propose a model based on a logistic function to map the PSNR values to perceived quality, which should better represent the saturation effect of the HVS when compared to linear or exponential mapping. The parameters of the mapping function were predicted using 2D and 3D content features, which were extracted from the original sequences. Each parameter of the logistic function was predicted from two content features. To select the most relevant features for each parameter, the dataset was split into training and testing sets and the model was trained on the training set. To evaluate how well the proposed model predicts perceived quality, the trained model was applied to the testing set.

A subset of the MPEG 3DV dataset presented in Section 10.1.1 was used to train and evaluate the proposed model. Only the results for the 3-view configuration, fixed stereo pair, of the two best AVC proposals and two best HEVC proposals were used as ground truth. The PSNR was computed as the average PSNR of the left and right views of the displayed stereo pair.

11.1 Proposed Model

This section describes the feature extraction and feature selection processes used to predict the parameters of the mapping function of the proposed model.

11.1.1 Feature Extraction

Both 2D and 3D features were extracted from the original video sequences. For the 2D features, the well-known SI and TI (see Section 2.2) are often used to characterize the amount of spatial detail of a picture and temporal changes of a video sequence, respectively. These two features were used by Korhonen and You (2010) to map PSNR values to perceived quality in the case of 2D video sequences. In this study, the temporal perceptual information and a modified version of the spatial perceptual information, referred to as $\tilde{SI}$, were used. $\tilde{SI}$ was computed using a Sobel kernel multiplied by $\frac{1}{8}$. The 2D features were computed on the luminance component of each content.

Mittal et al. (2011) have proposed that 3D images have certain statistical properties that can be captured using simple statistical measures of the disparity distribution. They used statistical features from disparity and disparity gradient maps to predict the QoE of 3D images and video sequences. Thus, the following 3D features were computed on the disparity map D of each content, according to (Mittal et al., 2011)

1) mean disparity $\mu = \mathrm{E}[D]$,
2) median disparity $med = median(D)$,
3) disparity standard deviation $\sigma = \sqrt{\mathrm{E}[(D-\mu)^2]}$,
4) kurtosis of disparity $\kappa = \frac{\mathrm{E}[(D-\mu)^4]}{(\mathrm{E}[(D-\mu)^2])^2}$,

5) skewness of disparity $skew = \frac{\mathrm{E}[(D-\mu)^3]}{(\mathrm{E}[(D-\mu)^2])^{(3/2)}}$,
6) mean differential disparity $\mu_d = \mathrm{E}[\delta D]$,
7) differential disparity standard deviation $\sigma_d = \sqrt{\mathrm{E}[(\delta D - \mu_d)^2]}$,
8) kurtosis of differential disparity $\kappa_d = \frac{\mathrm{E}[(\delta D-\mu_d)^4]}{(\mathrm{E}[(\delta D-\mu_d)^2])^2}$, and
9) skewness of differential disparity $skew_d = \frac{\mathrm{E}[(\delta D-\mu_d)^3]}{(\mathrm{E}[(\delta D-\mu_d)^2])^{(3/2)}}$

where the differential disparity (δD) was computed using a Laplacian operator on the disparity map. The 3D features were computed on a frame-by-frame basis and then averaged across frames.

Therefore, a total of eleven features, two 2D features and nine 3D features, were extracted for each content.

11.1.2 Mapping Function

To consider non-linearities and saturation effect of the HVS, a 4-parameter logistic function was used to predict perceived quality from PSNR values

$$MOS_p(PSNR) = a + \frac{b-a}{1+\exp\left[-c\left(PSNR-d\right)\right]} \tag{11.1}$$

where the parameters c and d are related to the slope and translation of the logistic function, respectively, and can be controlled independently. The parameters a and b were determined as follows. The subjective scores range R is typically divided into five parts of equal lengths, which are associated with distinct quality levels. By varying the bit rate, the quality of the video sequence varies from the lowest quality level to the highest quality level. Therefore, we assumed that the horizontal asymptotes of the logistic function are associated with the lowest and highest quality levels for the lowest and highest bit rates, respectively

$$\lim_{PSNR\to 0} MOS_p(PSNR) = a = R_{10\%} \qquad \lim_{PSNR\to +\infty} MOS_p(PSNR) = b = R_{90\%} \tag{11.2}$$

To determine the optimal values c_o and d_o for each content of the dataset, a fitting using Equation (11.1), partially constrained by Equation (11.2), was performed between the PSNR values and ground truth MOSs, for each content separately.

11.1.3 Feature Selection

The total number of extracted features (see Section 11.1.1) is higher than the number of contents in the dataset. Therefore, the number of features used to predict the parameters c and d of the mapping function in Equation (11.1) needs to be restricted. To avoid the risk of over-fitting, only two features out of eleven were used to predict each parameter of the logistic

function

$$c = \alpha f_1 + \beta f_2 + \gamma \tag{11.3}$$

$$d = \delta f_3 + \epsilon f_4 + \zeta \tag{11.4}$$

where f_1, f_2, f_3, and f_4 are content features, and α, β, γ, δ, ϵ, and ζ are coefficients.

To determine which extracted features should be used to predict the parameters of the mapping function, the proposed model was trained on a subset of contents of the dataset. For each combination of two features, a least square regression was performed to determine the coefficients of 11.3. The pair of features which obtained the best correlation with the optimal parameters c_o of the contents in the training set was chosen to predict the parameter c of the contents in the testing set. Similarly, for each combination of two features, a least square regression was performed to determine the coefficients of 11.4. The pair of features which obtained the best correlation with the optimal parameters d_o of the contents in the training set was chosen to predict the parameter d of the contents in the testing set.

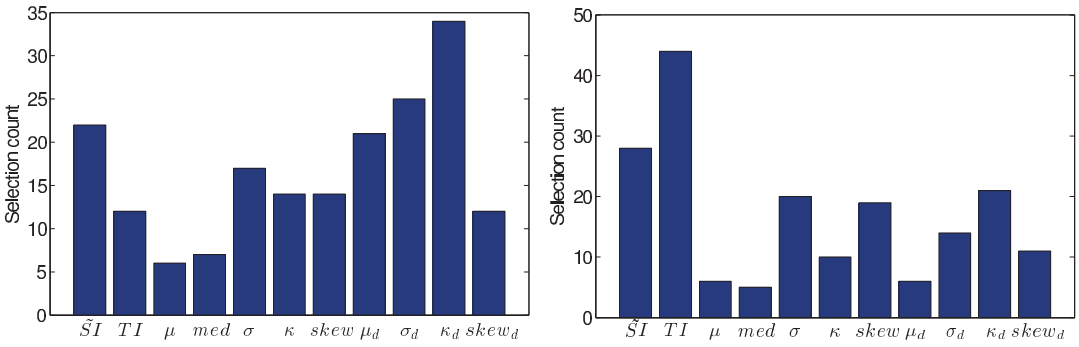
11.2 Performance Evaluation

To evaluate the performance of the proposed model, a dataset of 3D video sequences with associated ground truth subjective scores, containing a total of $n = 8$ contents, was used. The contents were divided into a training set and a testing set. The size of the training set was varied between five and seven contents to evaluate the influence of the training set size. For a fixed training set of size k, all possible $\binom{n}{k}$ combinations to split the contents into training and testing sets were realized to evaluate the robustness of the proposed model across contents. For each train-test trial, the model was trained on the training set according to Section 11.1.3 and the performance of the trained model was evaluated on the testing set.

The proposed model was benchmarked following the procedure described in Chapter 9. In particular, the performance indexes were computed between the predicted MOS values and ground truth MOSs. Note that no mapping between the scores predicted by the proposed model and the ground truth subjective scores was applied, as this is already considered in the proposed model.

11.2.1 Selected Features

Figures 11.1a and 11.1b show the histograms of features selected across $\binom{8}{7} + \binom{8}{6} + \binom{8}{5} = 92$ train-test trials to predict the parameters c and d, respectively. To predict the parameter c, no feature, except κ_d, was selected in more than a third of the train-test trials. Features extracted from the differential disparity map were more often selected than features extracted from the disparity map. This result is intuitive since the differential disparity map is related to occluded areas. Whereas the temporal activity was used to model the slope of the exponential function



(a) Parameter c (b) Parameter d

Figure 11.1: Histogram of features selected across all train-test trials.

in (Korhonen and You, 2010), the TI feature was selected only 12 times out of 92 train-test trials. Regarding the prediction of the parameter d, the TI and $\tilde{SI}$ features were selected in almost half and a third of the train-test trials, respectively. However, the translation of the exponential function in (Korhonen and You, 2010) was modeled using the spatial activity. This difference might come from the fact that the training contents only covered a limited range of spatial activity and no general trend could be drawn.

11.2.2 Anchors

To compare the performance of the proposed model to useful reference points, a fitting using Equation (11.1), unconstrained, was performed between the PSNR values and ground truth MOSs to determine all four parameters (a, b, c, and d). In this case, no prediction was performed and all eight contents were used. The fitting was applied in two different ways

A) on all contents at once and
B) on each content separately.

In the latter case, the performance indexes were computed separately on each content and then averaged across contents. Anchor A does not consider content characteristics since all contents are mixed. Therefore, the proposed model must show better performance than anchor A to be considered as valid. However, anchor B does consider all contents characteristics as the fitting is applied on each content separately. Thus, this anchor should provide upper bounds on PCC and SROCC as well as lower bounds on RMSE, mean absolute error (MAE), and OR for comparison with the proposed model. Table 11.1 reports the performance indexes of the two anchors.

Table 11.1: Performance indexes of the anchors.

Anchor	PCC	SROCC	RMSE	OR
A	0.3926	0.3973	1.4592	0.7344
B	0.9462	0.9015	0.3723	0.2109

Table 11.2: Performance indexes of the proposed model.

	Mean value				Standard deviation			
	PCC	SROCC	RMSE	OR	PCC	SROCC	RMSE	OR
$k = 7$	0.9341	0.9015	1.2181	0.6250	0.0595	0.0888	1.0914	0.3204
$k = 6$	0.8743	0.8711	2.0893	0.7143	0.2552	0.2486	1.7016	0.2395
$k = 5$	0.7815	0.7863	2.2065	0.7437	0.4559	0.4500	1.7753	0.2351

11.2.3 Results

Table 11.2 reports the mean value and standard deviation of the performance indexes across $\binom{n}{k}$ train-test trials of the proposed model for different training set sizes. For each train-test trial, the best features selected on the training set (with a frequency shown in Figures 11.1a and 11.1b) were used to predict the parameters of the mapping function for the testing set. Whereas the PCC and SROCC were quite high over the different training set sizes, the RMSE and MAE increased significantly for $k < 7$. Since the mapping function was applied on each content separately in the proposed model, the PCC and SROCC values were quite high when compared to anchor A. Nevertheless, if the mapping function had a wrong slope or translation, namely if there was an error in the prediction of c or d, the RMSE, MAE, and OR values increased significantly compared to anchor B. For $k = 7$, the standard deviation of the PCC and SROCC was quite low, which indicates that the proposed model was quite robust across contents when the training set contained various contents. However, in some cases for $k < 7$, the predicted quality scores had a negative correlation with the ground truth MOSs, which explains the high standard deviation for PCC and SROCC. This indicates that the training set should contain different contents covering a wide range of spatiotemporal characteristics. In general, predicted quality always achieved a high correlation with perceived quality when compared to anchor A, which does not consider content characteristics in the fitting process. This result indicates that content analysis can improve the accuracy of the mapping of PSNR values to perceived quality.

11.3 Conclusion

In this chapter, we proposed a model to predict perceived quality of stereoscopic video sequences based on content analysis. A logistic function was used to map the PSNR values to perceived quality. The parameters of the mapping function were predicted using 2D and 3D content features. The model was trained and evaluated on a dataset of stereoscopic video sequences with associated ground truth MOS. Results showed that the proposed model achieved

high correlation with perceived quality and was quite robust across contents when the training set contained various contents. This finding indicates that perceived quality can be predicted from PSNR values based on content analysis and that subjective tests might not be always required.

Part III Improving Quality of Experience

12 Improving 3D Quality of Experience

Quality assessment in the conventional video processing chain takes into account many characteristic 2D artifacts (Yuen and H. Wu, 2005). When extended to 3D video, the HVS further processes additional monocular and binocular stimuli. Thus, the resulting video quality at the end of the 3D video processing chain depends also on the level of stereoscopic artifacts or binocular impairments affecting the depth perception. In fact, stereo artifacts can cause unnatural changes in structure, motion, and color vision of the scene and distort the binocular depth cues, which result in visual discomfort and eyestrain.

At each particular stage of a 3D video processing chain (see Figure 12.1), different stereoscopic artifacts can be identified and described by Boev et al. (2009a). According to the HVS layer interpretation of the visual pathway, all artifacts can be divided into four groups: structural, color, motion, and binocular. Whereas structural artifacts affect contours and textures and distort spatial vision, the motion and color artifacts negatively influence motion and color perception. The most important binocular artifacts that have an impact on depth perception are discussed below. Nevertheless, it is very important to consider the interaction between individual groups and artifacts.

During 3D capture, several distortions occur (keystone distortion, depth field curvature, cardboard effect, etc.), which are mostly optical and due to camera setup and parameters. Assuming a toed-in convergent camera setup where each camera has a different perspective of the scene, the quality of binocular depth perception is affected by the vertical and unnatural horizontal parallax caused by *keystone distortion* and *depth plane curvature*, respectively. Keystone distortion, as a result of the position of the two cameras in slightly different planes, causes a vertical difference between homologous points, called vertical parallax. Vertical parallax is bigger in the corners of the image, proportional to camera separation, and inversely proportional to convergence distance and focal length (Woods et al., 1993). The same principle leads, in the horizontal plane, to depth field curvature. This causes an unnatural horizontal parallax, which results in wrong representation of relative object distances on the screen. Keystone distortion and depth plane curvature can be suppressed by using a parallel camera configuration in stereoscopic video acquisition.

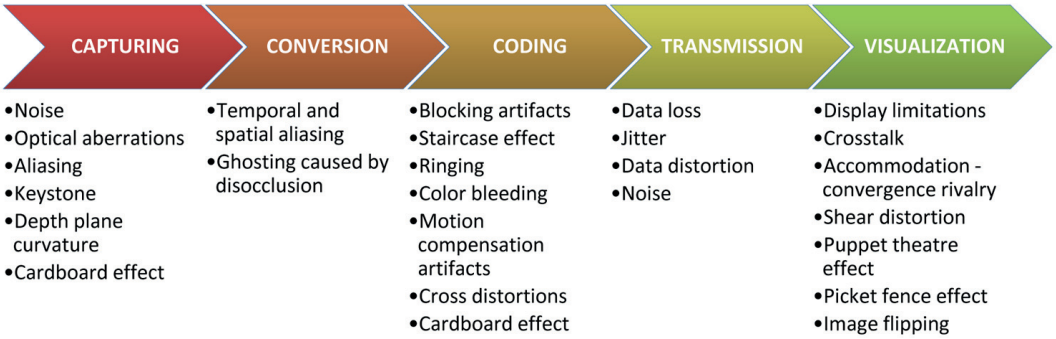


Figure 12.1: 3D video processing chain artifacts.

The *cardboard effect* is related to the availability of the proper disparity information, without which the viewer sees wrong objects–screen distance and, therefore, the perceived size and depth of objects do not correspond one to another. This results in objects appearing flat, as if the scene was divided into discrete depth planes (Chapiro et al., 2014). While keystone distortion and depth plane curvature are introduced only at the capture stage, the cardboard effect occurs also at the conversion and coding phase due to the sparse depth quantization.

Depending on the capture and rendering formats used and their mutual adaptation, various 3D artifacts occur in the conversion stage. The most common artifact here is *ghosting*, which is caused by disocclusion. It occurs when video plus depth representation and rendering are used, mainly due to the interpolation of occluded areas needed for view synthesis.

For individual coding of stereo image pairs, binocular suppression (Julesz, 1971) as a property of the HVS can be exploited. This ability to compensate the loss of information in one of the stereo views is particularly suitable for asymmetric coding (Fehn et al., 2007). In asymmetric coding, the left and right views are encoded with different quality. However, if the qualities of two encoded stereo channels differ too much, the resulting spatial (resolution) and temporal (frame rate) mismatch between them, commonly referred to as a *cross-distortion*, produce wrong depth perception.

Crosstalk, also known as image ghosting, is caused by visualization of 3D content and is one of the most annoying distortions in stereoscopic displays. It comes from imperfect left and right image separation when the view for the left eye is partially visible by the right eye, and vice versa. Crosstalk usually results in ghosting, shadowing, and double contours perception. Another issue is the unnatural decoupling of vergence and accommodation: when watching a stereoscopic display, the eyes converge to the location of the virtual object while the accommodation is always set for the screen surface. This effect is referred to as *vergence-accommodation rivalry*. *Picket fence*, *image flipping*, *pseudoscopy*, and *shear distortion*, as other artifacts arising during visualization, are related to the display technology used and when an observer changes their position. Shear distortion occurs with a change of position of the viewer resulting in wrong head parallax and distorted perspective vision. Whereas shear distortion is typical for stereoscopic displays allowing only one correct viewing position,

picket fence, image flipping, and pseudoscopy are experienced with autostereoscopic displays exclusively. More specifically, picket fence is introduced by the spatial multiplexing of parallax barrier-based autostereoscopic displays only. It is noticeable as vertical banding when the observer moves laterally in front of the screen. Image flipping is basically the consequence of parallax discretization. It is observed as a leap transition between viewing zones. Psuedoscopy or reverse stereo only occurs with 2-view autostereoscopic displays when the viewer moves away from the sweet spot and reaches a point where the left (right) eye only perceives the right (left) image.

All these artifacts impact picture and depth quality, as well as visual comfort. In particular, crosstalk is one of the stereo artifacts with the largest influence on image quality and visual comfort (Meesters et al., 2004; Seuntiëns et al., 2005). Vergence-accommodation rivalry is believed to increase visual discomfort (Hoffman et al., 2008). All the problems related to the sweet spot position in autostereoscopic displays also considerably reduce the overall 3D QoE.

This chapter investigates different systems to reduce stereo artifacts generated at the visualization stage to improve QoE on 3D displays. Section 12.1 proposes two different approaches that exploit visual attention to mitigate crosstalk and vergence-accommodation rivalry on stereoscopic displays: an offline system, which uses a computational model of visual attention to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions. The gaze points are used in conjunction with the disparity map to extract the disparity of the object-of-interest. Then, horizontal image translation is performed to bring the fixated object on the screen plane. Section 12.2 proposes a complete active crosstalk reduction system for mobile autostereoscopic displays. To determine the crosstalk level at each position, a full display characterization is performed. Based on the user position and crosstalk profile, the system first helps the user to find the sweet spot using visual feedback. If the user moves away from the sweet spot, then the active crosstalk compensation is performed and reverse stereo phenomenon is corrected. Section 12.3 proposes a similar system for multiview autostereoscopic displays. First, the display is characterized in terms of luminance distribution. Then, the luminance profiles are modeled using a limited set of parameters. A Kinect sensor is used to determine the viewer position in front of the display. Finally, the proposed system performs an intelligent on the fly allocation of the output views to minimize the perceived crosstalk. All these systems were evaluated in term of picture and depth quality, as well as visual comfort, to demonstrate the improvements on 3D QoE.

12.1 Improving 3D QoE on Stereoscopic Displays

To improve the QoE provided by stereoscopic displays, researchers have proposed to exploit visual attention (Huynh-Thu et al., 2011b). Since two decades, researchers have investigated different solutions based on visual attention to improve the rendering of stereoscopic displays. Several systems have been developed using gaze tracking to determine the virtual object fixated by the user. The proposed solutions have been implemented either by hardware

means, to control the stereoscopic display (Shiwa et al., 1996; Talmi and J. Liu, 1999), or by software means, to adapt the rendering of the 3D content (Peli et al., 2001; R. Yang and Z. Zhang, 2004).

To reduce the vergence-accommodation rivalry, the 3D content should be reconverged such that the fixated object lies on the screen plane. This can be achieved by performing view synthesis to generate a new stereo pair, which requires depth information and typically introduces visible artifacts due to imperfect depth data, or simply by applying a horizontal image translation, i.e., shifting horizontally the left and right views of the original stereo pair (Mendiburu, 2009). By bringing the fixated object on the screen plane, perceived crosstalk is also reduced as the virtual object is translated to the zero disparity plane (ZDP). D. Xu et al. (2012) have shown that horizontal parallax adjustment significantly increases the overall 3D QoE.

Chamaret et al. (2010) have proposed a system to adapt the 3D rendering based on the region-of-interest. To predict the most salient region, a computational model of visual attention was used to compute the saliency map considering spatial, temporal, and depth features. Then, the most relevant disparity of the region-of-interest was extracted. Finally, a shift was applied to the left and right views of the stereo pair, based on filtered disparity values, to translate the region-of-interest to the screen plane. They have reported that the proposed system was more pleasant to watch than the original video sequence, but no results of a proper subjective evaluation are reported to support their claim. Moreover, this system relies on the accuracy of the computational model of visual attention. To address these problems, we propose and evaluate two different approaches that exploit visual attention to improve 3D QoE on stereoscopic displays: an offline system, which uses a computational model of visual attention to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions. The user preference between standard 3D mode and the two proposed systems is evaluated in terms of image quality, depth quality, and visual discomfort through a subjective evaluation. This section describes in details the proposed systems and reports the details and results of the subjective evaluation.

12.1.1 System Description and Implementation

This subsection describes the saliency map computation and most salient disparity extraction used in the saliency map based system, as well as the gaze points filtering and disparity extraction used in the eye tracking based system. Details are provided regarding the shift filtering, horizontal image translation used to reconverge the 3D scene, and implementation of the application.

Saliency Map Based System

One of the approaches considered in this paper to improve 3D QoE relies on a visual attention model to determine the most salient region-of-interest and its corresponding disparity. Since we are considering stereoscopic video sequences, the following features should be exploited to compute the saliency map: spatial, temporal, and depth features.

Saliency Map Computation To compute the spatial and temporal saliency maps, the Graph-Based Visual Saliency (Harel et al., 2006) algorithm was used as it is reported to be one of the best algorithms (Judd et al., 2012) and it considers temporal features. The algorithm was applied twice on the left view of the stereo pair: once to compute the spatial saliency map using Derrington-Krauskopf-Lennie (DKL) color space, intensity, and orientation features; once to compute the temporal saliency map using motion features only.

To compute depth saliency, a few models have been proposed in the recent years, but there is no publicly available implementation. Readers can refer to (J. Wang et al., 2013) for a recent review of some of these models. Most visual attention models considering depth features perform a simple weighting of the 2D saliency map with the depth map, based on the assumption that pixels located closer to the observers and in front of the screen are more salient. However, J. Wang et al. (2013) have shown that combining 2D saliency and depth contrast provides better results. Therefore, in this paper, depth contrast was used to compute the depth saliency map. To compute depth contrast, a difference of gaussians (DoG) filter was applied to the left depth map (J. Wang et al., 2013).

First, the perceived depth map, which represents the distance between the observer and the virtual object, was computed from the left disparity map considering viewing conditions (J. Wang et al., 2013). The relationship between the perceived depth D in meters and the disparity d in pixel is given by

$$D = \frac{V}{1 + \frac{d \cdot W}{I \cdot R_x}} \tag{12.1}$$

where I is the interocular distance, V is the viewing distance, and W and R_x are the width and horizontal resolution of the screen, respectively. In our experiments, the interocular distance was set to 65 mm and the screen property parameters were set according to the setup of the subjective evaluation (see Section 12.1.2).

Then, the depth contrast was computed by filtering the perceived depth map with the DoG filter defined as

$$f(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) - \frac{1}{2\pi K^2\sigma^2} \exp\left(-\frac{x^2+y^2}{2K^2\sigma^2}\right) \tag{12.2}$$

where (x, y) is the location in the filter, σ and K are used to control the scales of DoG and the

ratio between the "center" area and "surround" area, respectively. According to (J. Wang et al., 2013), only one scale of DoG was applied, with $\sigma = 32$ pixels and $K = 1.6$. Finally, the depth saliency map was computed as the absolute value of the depth contrast, normalized by its maximum value.

Similarly to the straightforward approach used by J. Wang et al. (2013), the final saliency map SM was computed as a weighted sum of the spatial saliency map SM_s, temporal saliency map SM_t, and depth saliency map SM_d

$$SM(i,j) = \omega_1 SM_s + \omega_2 SM_t + \omega_3 SM_d \tag{12.3}$$

where $\omega_1 = \omega_2 = \omega_3 = \frac{1}{3}$.

Most Salient Disparity Extraction To extract the most salient region-of-interest and to determine its most relevant disparity, a similar approach to that presented by Chamaret et al. (2010) was used. First, the N most salient pixels were determined using the saliency map computed as described above. In our experiments, N was empirically set to 2% of the total number of pixels. Then, a connected-component analysis, considering an 8-connected neighborhood, was performed to determine the M regions formed from the N most salient pixels. Finally, the region containing the highest number of most salient pixels was considered as the most salient region-of-interest.

Once the most salient region-of-interest was determined, the most relevant disparity of this area was extracted. First, a histogram was constructed from the disparity values inside the most salient region-of-interest to estimate the spreading of disparity values. A step of 5 disparity units was used to construct the histogram. Then, the bin with the highest frequency count was extracted from the histogram. Finally, the median disparity of this bin was considered as the most salient disparity.

Eye Tracking Based System

The second approach considered in this paper to improve 3D QoE relies on an eye tracking system to determine the gaze positions on the screen, and therefore, the object-of-interest watched by the viewer and its corresponding disparity.

The left and right gaze points were computed independently using the gaze origin and gaze direction measured by the eye tracking system and the position of the eye tracker with respect to the screen. Additionally, the gaze direction quality estimated by the eye tracking system was used to discard low quality measurements. In our experiments, if the gaze direction quality was below 0.1, the gaze point was discarded. This process was performed independently for the left and right eyes.

The eye tracking system used in our experiments (see Section 12.1.2) provides both unfiltered

and filtered versions of the gaze origin and gaze direction measures. In our experiments, the unfiltered measurements were used as no details are provided regarding the filtering technique applied by the eye tracking system. However, to discard erroneous measurements, a median filter was applied separately on the x- and y-components of the left and right gaze points. To provide robust but reactive gaze positions, a tradeoff has to be made on the size of the kernel. The delay introduced by the filter should be lower than the minimum time required to fuse stereoscopic stimuli, which is around 400 to 500 ms (Hoffman et al., 2008). In our experiments, a kernel of size $L = 13$ was used, which corresponds to a delay of approximately $\frac{1}{r} \cdot \frac{L+1}{2} \approx 120$ ms, where r is the sample rate of the eye tracking system (60 fps in our experiments).

If the gaze points of the left and right eyes could be determined with 100% accuracy, the disparity of the virtual object would be simply given by the difference of the x-coordinates of the left and right gaze points. However, due to the limited precision of the eye tracking system, such a trivial approach would not provide good results in practice. Therefore, the left and right disparity maps of the stereoscopic video sequence were used in conjunction with the filtered gaze points. For the same reason, the shift cannot simply be determined by using the disparity values at the gaze points. For example, if the gaze position is near the boundary between foreground and background, the measured gaze point might be located on the background, whereas the actual gaze position was located on the foreground. Based on the assumption that foreground objects are more salient than background objects, the disparity of the object-of-interest was determined as the maximum disparity in a neighborhood of $N \times N$ pixels around the filtered gaze point. In our experiments, N was empirically set to 15 pixels. This process was performed independently for the left and right eyes and the most salient disparity was computed as the maximum value of the disparity values extracted from the left and right disparity maps.

Shifting

To bring the fixated object on the screen plane, horizontal image translation was performed by shifting horizontally the left and right views of the original stereo pair (Mendiburu, 2009). The shift parameter is given by the disparity value of the fixated object, which was determined by the saliency map or eye tracking system. However, if the shift difference between two successive frames is too large or oscillates at a high frequency, the process may become visible and induce visual discomfort. Therefore, filtering of the shift values was performed.

Shift Filtering Chamaret et al. (2010) have determined that a maximum shift difference of 1.5 pixel can be applied without any visual notification. However, a non-integer shift requires interpolation of the pixel values, which is time consuming and not suitable for a real time application. Therefore, in our experiments, only integer shifts were applied and the maximum shift difference between two successive frames was set to 1 pixel.

Algorithm 1 Determination of current and future shift values

```
if |s_t(n) − s(F)| > δ then
  if |s_t(n) − s(n−1)| ≤ δ then
    s(n,...,F) ← s(n−1)
  else
    k ← 0
    shift ← s(n−1)
    if s_t(n) > s(n−1) then
      while shift < s_t(n) do
        shift ← shift + 1
        s(n+k) ← shift
        k ← k+1
      end while
    else
      while shift > s_t(n) do
        shift ← shift − 1
        s(n+k) ← shift
        k ← k+1
      end while
    end if
    s(n+k,...,F) ← s_t(n)
  end if
end if
```

To improve robustness, the disparity values determined by the saliency map and eye tracking system were filtered using a median filter

$$d_f(n) = \text{median}\,(d(n-K),\ldots,d(n)) \tag{12.4}$$

where d and d_f are the raw and filtered disparity values, respectively, and n is the frame index. In our experiments, K was empirically set to 4.

The target shift value s_t is given by the filtered disparity value

$$s_t(n) = d_f(n) \tag{12.5}$$

Then, the current and future shift values are determined according to Algorithm 1, where s is the shift, F is the number of frames of the video sequence, and n is the frame index. To avoid flickering due to the shifting process, a threshold δ was considered before updating the current and future shift values. In our experiments, δ was empirically set to 1.

Horizontal Image Translation Both views were shifted half way

$$s_l = \left\lceil \frac{s}{2} \right\rceil \qquad\qquad s_r = \left\lfloor \frac{s}{2} \right\rfloor \tag{12.6}$$

where s_l and s_r are the shift parameters for the left and right views, respectively. The horizontal image translation was performed by adding black borders on the left or right side of the picture and cropping the other side by the same amount to preserve the size of the picture (Mendiburu, 2009). To reduce potential stereoscopic window violation (Poulakos et al., 2015) that might occur due to horizontal image translation, the floating window (Mendiburu, 2009) technique was applied as follows: the pixel positions corresponding to the black border in the left view were also set to black in the right view, and vice versa. Therefore, both pictures had black borders on both sides.

Implementation

For our experiments, a dedicated video player was implemented in C++ using the OpenCV library. The application implements the saliency map and eye tracking systems described above. The video player displays a stereoscopic video sequence and performs horizontal image translation on the fly. To ensure real time processing, the application uses multithreading, which was implemented using the Boost library. In particular, one thread per input data, i.e., left view, right view, left disparity map, and right disparity map, was launched to load the data. One thread was dedicated to the eye tracking system, to collect the measurements that were sent by UDP from a remote computer and to compute the gaze points. Finally, one thread was dedicated to the remaining processes, such as horizontal image translation, interlacing (the stereoscopic monitor used in the experiments was line-interleaved), synchronization, filtering, logging, etc.

For each stereoscopic video sequence, the left and right views were converted to RGB 4:4:4, downsampled vertically by two (due to the line-interleaved monitor), and stored as a stack of bitmap image files. The conversion to RGB 4:4:4 allows shifting by odd values and is necessary for the interleaving process. The downsampling process avoids aliasing, which could occur when the interleaving is done by taking every other line without any pre-filtering. The depth maps were stored in half resolution as a stack of monochrome bitmap image files. This solution was found to provide the best results in terms of loading time and quality when compared to other alternatives, such as using compressed video sequences for each stream. With this solution, a frame rate of 30 fps was achieved with the developed video player.

12.1.2 Subjective Evaluation

This subsection presents the details of the subjective evaluation conducted to assess user preference between standard 3D mode and the two proposed systems.

Dataset

Eight stereoscopic video sequences with associated depth maps were used in the experiments (see Table 12.1). Sequences *Boxers*, *Hall*, *Lab*, *News report*, and *Phone call* were obtained

Table 12.1: Stereoscopic video sequences used in the experiments.

Sequence	Frames	Views	$d_{\min}$	$d_{\max}$
Boxers	1-250	0-1	-14	29
Hall	1-250	0-1	-15	20
Lab	151-400	0-1	-100	44
News report	1-250	0-1	-45	71
Phone call	151-400	0-1	-35	39
Musicians	1-250	0-1	0	176
Poker	1-250	0-1	0	176
Poznan Hall2	1-200	7-6	16	118

from the NAMA3DS1 database (Urvoy et al., 2012). These sequences are available as raw video files, progressively scanned, with YCbCr 4:2:2 color sampling, and 8 bit per sample. Additionally, the left disparity map is available as 16 bit floating values, half resolution. The right disparity map was generated by warping the left disparity map to the right view, filling holes using background propagation, and applying 3×3 median filtering. Both disparity maps were converted to 8 bit and scaled to the range [0, 255] using the minimum and maximum disparity values observed in the whole video (see Table 12.1). Sequences *Musicians* and *Poker* were obtained from the European FP7 Research Project MUSCADE (M23703). Sequence *Poznan Hall2* was obtained from the Poznań multiview video database (M17050). These video sequences and associated depth maps are available as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bit per sample. The camera parameters are provided to convert the disparity values scaled in the range [0, 255] to real disparity values in pixels.

Since the cameras used for recording sequences *Musicians*, *Poker*, and *Poznan Hall2* were set in a parallel direction, they are assumed to converge at an infinite point. This setup leads to stereoscopic window violation (Mendiburu, 2009) and does not sufficiently exploits the depth range, as the 3D content appears only in front of the screen plane. Therefore, when displaying the original version of these three sequences, horizontal image translation was applied with a shift defined as (J. Wang et al., 2013)

$$s = \frac{d_{\min} - d_{\max}}{2} \tag{12.7}$$

where $d_{\min}$ and $d_{\max}$ are the minimum and maximum disparity values computed from the camera parameters (see Table 12.1), respectively. Note that no shift was applied to the sequences from the NAMA3DS1 database, as the cameras were already converged during recording.

Test Environment

The experiments were conducted at the MMSPG quality test laboratory, which fulfills the recommendations for the subjective evaluation of visual data issued by ITU. The test room is

equipped with a controlled lighting system with a 6500 K color temperature and an ambient luminance at 15% of the maximum screen luminance, whereas the color of all the background walls and curtains present in the test area were in mid grey. The test room was separated in two by a curtain to isolate the subject and equipment from the test operator, which was present during the test session to supervise the recording of the eye tracking data. The laboratory setup was intended to ensure the reproducibility of the subjective tests results by avoiding unintended influence of external factors.

To display the test stimuli, a full HD 46" Hyundai S465D polarized stereoscopic monitor was used. The monitor was calibrated using an X-Rite i1Display Pro color calibration device according to the following profile: sRGB gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level.

A Smart Eye Pro 5.8 remote eye tracking system was used to determine the gaze position on the screen of the left and right eyes independently. The system was equipped with three Sony HR-50 cameras at a frame rate of 60 fps and two infrared flashes. All measurements were recorded on a separate computer and sent by UDP on a local network.

The experiment involved one subject per test session. The subjects were seated in line with the center of the monitor, at a distance of 3.2 times the picture height (see Table 2.1), corresponding to roughly 1.8 meters from the stereoscopic monitor. The eye tracking system was placed at 1.28 meters from the stereoscopic monitor such that the face was well captured by the cameras.

Test Method

The PC method with a ternary scale (see Section 2.4.3) was chosen as judging the quality of different imaging systems individually may be quite difficult. Pairs of video sequences, "A" and "B", which resulted from different imaging systems, were presented in succession order on the same display. Subjects were asked to judge which video sequence in a pair ("A" or "B") is preferred in terms of picture quality, depth quality, and visual comfort. The option "same" was included to avoid random preference selections. For each of the 8 test sequences, all the possible combinations of the 3 test conditions (original, saliency map, and eye tracking) were considered, leading to a total of $8 \times \binom{3}{2} = 24$ comparisons.

Training Session

Before the experiment, oral instructions were provided to the subjects to explain their tasks. Additionally, a training session was organized to allow subjects to familiarize with the assessment procedure. The training materials shown in the training session were selected by expert viewers to include examples of all evaluated aspects. More particularly, the *Umbrella* and *Basket* sequences (Urvoy et al., 2012) were used with different horizontal image translations to illustrate picture quality. To illustrate depth quality, the *Soccer* sequence (Urvoy et al.,

2012) was shown in 2D and 3D viewing conditions. Finally, from the EPFL 3D Video Database (Goldmann et al., 2010), the *Notebook* sequence, with 10 cm and 20 cm baselines, was used to illustrate visual comfort. The training materials were presented to subjects exactly as for the test materials.

Test Session

Before the test session, the aperture and focus settings of the eye tracker cameras were adjusted for optimal conditions and a full camera calibration was performed to maximize the accuracy of the measurements. For each subject, a personal profile was created by recording several head poses and gaze calibration using four calibration points close to the screen corners. Subjects were instructed to hold their head still while watching the video sequences to ensure good tracking of their gaze.

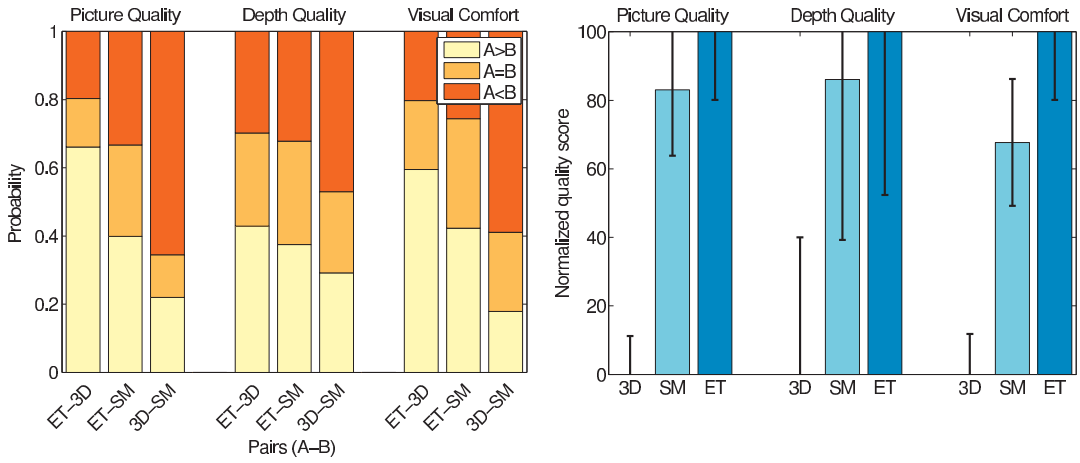
A trial was initiated by the presentation of a message showing the letter "A", at zero disparity and with a mid-grey background, for 2 s. Then, the sequences to be compared were presented. The sequences were temporally separated by the presentation of a message showing the letter "B" for 2 s. The trial ended with a message showing the words "Vote now" without any restriction in time. Votes were collected by the test operator such that the subject was always seated optimally with respect to the eye tracking system.

Two dummy pairs, whose scores were not included in the results, were included at the beginning of the session to stabilize the subjects' ratings. To reduce contextual effects, the stimuli orders of display, both within and between trials, were randomized applying different permutation for each subject, whereas the same content was never shown consecutively. In total, the test session lasted for about 16 minutes.

A total of 21 naïve subjects (5 females and 16 males) took part in the experiment. They were between 18 and 31 years old with an average of 21.8 years of age. All subjects were screened for correct visual acuity, color vision, and stereo vision using Snellen chart, Ishihara chart, and Randot test, respectively.

Data Processing

To analyze user preference for the different imaging systems, statistical tools were applied to the individual ratings. No outlier detection was performed since there is no international recommendation or a commonly used outlier detection technique for PC results. Before estimating MOS values for PC results, the winning frequency and the tie frequency are computed from the obtained subjective ratings for each pair of stimuli. This can be done individually for each test video content or jointly over all contents. To compute the preference matrix, only wins were taken into account, whereas ties were discarded. Then, the Bradley-Terry-Luce model (see Section 2.6.3) was used to estimate MOSs via maximum likelihood estimation. Ties were considered as half way between the two preference options and equally distributed.



(a) Preference probabilities.

(b) Normalized quality scores.

Figure 12.2: Subjective results for 3D mode, saliency map based system (SM), and eye tracking based system (ET).

The CIs for the maximum likelihood estimates of the scores were obtained using the Hessian matrix of the log-likelihood function. Results were normalized to the range $[0, 100]$ for a better representation.

12.1.3 Results

Figure 12.2a shows the preference and tie probabilities obtained over all test sequences for picture quality, depth quality, and visual comfort. As it can be observed, the visual attention based systems significantly improve picture quality when compared to the original video sequence, as they are preferred in about 65% of the test stimuli, whereas the original video sequence is preferred in only about 20% of the test stimuli. The eye tracking and saliency map based systems are quite competitive, with a preference probability of 40% and 33%, respectively, which shows a slight advantage for real time gaze determination.

Similar results can be observed for visual comfort. However, the preference probabilities for the visual attention based systems over the original video sequence have decreased by about 6%, whereas the ties have increased, especially for the original versus saliency map comparison. Regarding the eye tracking versus saliency map comparison, the probability of choosing the saliency map based system has decreased to 25%, whereas the probability of choosing the eye tracking based system is about 42%, which means that the difference between the two systems is slightly bigger in terms of visual comfort than in terms of picture quality.

Regarding depth quality, the preference between the different conditions is not obvious, as the preference probabilities are between 29% and 47%, which is close to the random 33%. However, the results show a slight preference for the two visual attention based systems, with

a preference probability for the eye tracking and saliency map based systems over the original video of 43% and 47%, respectively.

Results for picture quality and visual comfort show similar behavior, which was expected as the individual ratings for these two aspects were usually highly correlated. The link between picture quality and visual comfort is quite intuitive, since objects located far in front or behind the screen plane may induce vergence-accommodation rivalry, which creates visual discomfort, and are perceived with double edges due to imperfect separation of the left and right views, which reduces picture quality. However, for some of the subjects, the depth quality ratings usually agreed with the picture quality and visual comfort ratings, whereas they were opposed for the other subjects. This means that horizontal image translation had a positive impact on depth quality for some of the subjects, whereas it had a negative impact for others. As the 3D picture was shifted to reconverge the scene, not recalculated, this process translated the whole 3D scene along the z-axis, which could induce a scale-down effect (Mendiburu, 2009) and negatively impact depth perception. To overcome this drawback, a new stereo pair should be synthesized, with different camera parameters, but this process cannot be performed in real time. Ideally, depth blur should be considered as well to strengthen monocular depth cues.

Figure 12.2b shows the MOSs and CIs obtained over all test sequences for picture quality, depth quality, and visual comfort. As it can be observed, the visual attention based systems significantly enhance picture quality and visual comfort. However, the difference between the saliency map and eye tracking based systems is not significant, as expected by largely overlapping CIs. For some of the test sequences, the most salient region-of-interest predicted using the saliency map highly correlated with the object-of-interest watched by the subject. However, assuming that the eye tracking system can accurately measure real time gaze position, this system is expected to perform better than the saliency map system, which relies on an algorithm to predict visual saliency. Considering that the saliency map system does not require a specific hardware setup and can be performed off line, which would allow more advanced techniques than horizontal image translation to reconverge the 3D picture, this system shows promising results. In a future study, we plan to benchmark the performance of different 3D saliency models using the ground truth eye tracking data recorded in our experiments.

Regarding depth quality, the visual attention based systems provide slightly better results when compared to the original video sequence, but the difference is not as significant as for picture quality and visual comfort. However, the difference between the eye tracking based system and original sequence can be expected as statistically significant, as the CIs do not overlap. To further investigate the influence on depth quality, additional statistical analysis should be performed, considering ties to determine the CIs.

12.2 Improving 3D QoE on Mobile Autostereoscopic Displays

In the past few years, the entertainment industry expanded into mobile technologies. Constant improvement of the processing power and rendering technology enabled the consumption of 3D media on mobile devices via autostereoscopic displays using parallax barrier. Parallax barrier, as a less intrusive technology (glassesless), exploits a special LCD layer as a spatial filter directing two different views towards the eyes of the viewer. There are certain limitations related to parallax barrier technology based systems. The viewing areas are limited by specific viewing zones, and as soon as the viewer moves his head outside the sweet spot, a significant visual impairment is perceived. QoE of such devices is therefore limited as well as the viewing freedom of the user. However, there were no particular attempts to improve the QoE for such devices and no commercially available hardware exists with embedded algorithms improving the visual quality.

To improve the QoE provided by autostereoscopic displays, researchers have proposed to perform active crosstalk reduction based on user position (Boev et al., 2008; Boev et al., 2009b; Park et al., 2011). The user is tracked using face and eye detection and its position related to sweet spot is evaluated. Outside the sweet spots, crosstalk is canceled by simply switching back to 2D mode (Boev et al., 2009b) or reduced by performing an intelligent assignment of the pixel values based on the visibility profiles of different views (Park et al., 2011). In case of reverse stereo, the left and right images are swapped (Park et al., 2011).

Combining face detection and eye tracking with dynamic 3D rendering to improve QoE attracts the attention of researchers in autostereoscopic community. In (Y.-S. Chen et al., 2001; J.-C. Yang et al., 2008), authors used the eye tracking systems coupled with autostereoscopic two-view displays. They simulated motion parallax and synthesized the left and right views in real time according to the user position. Other displays (Chae et al., 2011; Yi et al., 2008), such as the famous Free2C display from Fraunhofer HHI, controls the parallax barrier or the lenticular sheet such that the sweet spot follows the user head. Several works also proposed to combine face/eye tracking with multiview autostereoscopic display. Dodgson (2006) analyzed an ideal 3-view display, where only two views are actually displayed, to better deal with the transition of one eye between two adjacent zones. Boev et al. (2008) presented a single-viewer system based on user tracking. The system performs on-the-fly visual optimization to achieve continuous head parallax, i.e., to avoid the repetition effect between the lobes, mitigate crosstalk, and improve brightness. There are also some autostereoscopic displays, e.g., some Toshiba laptops, that use head tracking to get better 3D experience. In all above-mentioned cases, a powerful hardware including a pair of cameras used for better tracking is needed.

Similar approach has been rarely addressed for mobile devices due to their limitation in processing power and/or real-time face/eye tracking capabilities. Boev et al. (2009b) relied on the Open Multimedia Applications Platform (OMAP) embedded in a smartphone and using one camera for eye tracking. Correction of the pseudoscopy was performed, whereas the crosstalk was not reduced and when the viewer experienced it, the screen was switched

to 2D mode. Moreover, the good viewing zones were measured empirically and the viewer was supposed to be at one fixed optimal distance from the screen. In another words, fixed interpupillary distance (IPD) was assumed. Another special hardware prototype of mobile multiview autostereoscopic display based on a method providing clear images appropriate to the positions of the observer's eyes was proposed by Park et al. (2011). However, to the best of our knowledge, no subjective assessment demonstrating the effectiveness of an active crosstalk reduction system on a mobile device has been yet reported. To overcome this lack, we propose and evaluate a complete active crosstalk reduction system running on an HTC EVO 3D smartphone. The user preference between standard 2D and 3D modes and the proposed system is evaluated in terms of image quality and depth quality through a subjective evaluation. This section describes in details the proposed system and reports the details and results of the subjective evaluation.

12.2.1 Display Characterization

The main goal of the display characterization is to define the sweet spots in terms of viewing angle and OVD. Knowing the technical details of the autostereoscopic display, such as barrier slit or pixel-barrier separation, allows exploiting simple optical laws in order to characterize the display. In common practice, display technical details are not provided and more complex approach must be used.

The autostereoscopic display characterization is usually performed as a measure of crosstalk at different positions in space providing 3D crosstalk map. Significant effort has been devoted to autostereoscopic display characterization over past few years. The International Committee for Display Metrology has recently proposed a standardized way to measure crosstalk at a given point in space (International Committee for Display Metrology, 2012). This approach is, however, time-consuming and expensive as dedicated hardware, such as luminance meters and Fourier Optics (Abileah, 2011; Boher et al., 2009; Boher et al., 2012), is often required. Another simple yet effective approach, which we adopted for our purposes, was proposed by Sykora et al. (2011) and Hong (2012). The main idea is to project a predefined pattern rendered by the autostereoscopic display and acquire a crosstalk map at a given distance using a DSLR camera. The measurement setup, assuming a semi-transparent projection surface is shown in Figure 12.3.

The mobile device with the autostereoscopic display was placed on a rail on one side of the projection surface, whereas the DSLR camera was placed on the other. Both devices, i.e., mobile phone and DSLR, were controlled remotely by a computer which assures the setting of shooting parameters as well as the user distance (distance between display and projection surface) and the changes of test pattern. The predefined patterns, i.e., black-black (KK), black-white (KW), and white-black (WK) (see Figure 12.4a), were displayed and recorded at different user distances and the 3D crosstalk map was then reconstructed. In our experiments, the distance of the mobile phone varied from 220 mm to 420 mm with a step of 5 mm. A dark

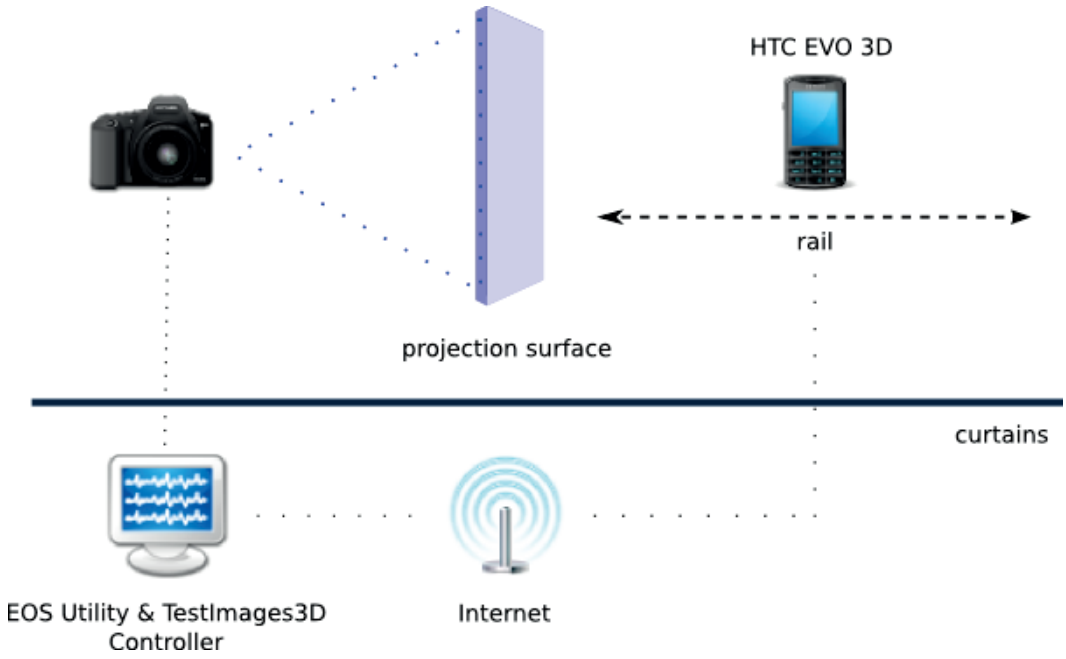


Figure 12.3: Display characterization setup.

room environment, totally isolated from any external light source, was needed to assure the high contrast of the projected pattern. A sheet of tracing paper placed between two glass plates providing efficient flatness, transparency, uniformity, and light scattering was used as a projection surface.

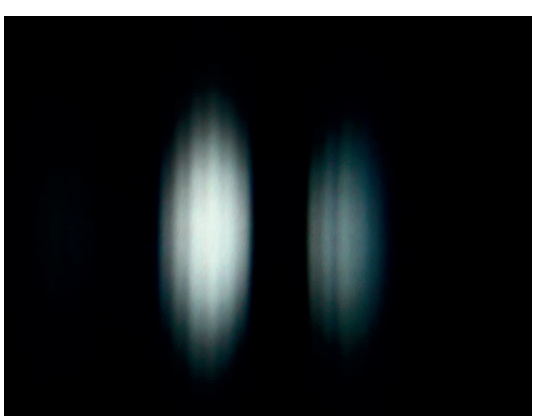
Crosstalk Computation

The first step towards crosstalk computation is the preprocessing of the DSLR JPEG images. For each test patterns and distance, five images were captured and averaged in order to suppress the noise. Averaged images were further cropped to the region of interest (the area of the projection surface). The luminance information was then extracted by using sRGB to XYZ conversion matrix as follows

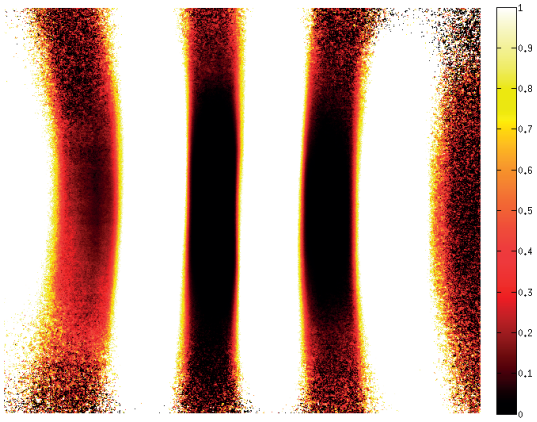
$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{bmatrix} \begin{bmatrix} R_{linear} \\ G_{linear} \\ B_{linear} \end{bmatrix} \tag{12.8}$$

The linear sRGB tristimulus values were extracted by using the following gamma correction formula

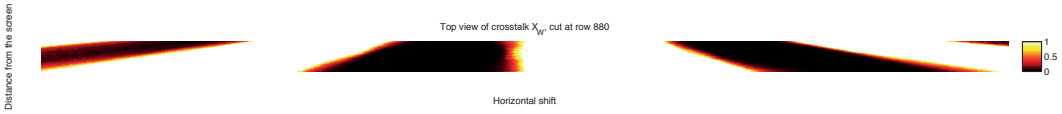
$$C_{linear} = \begin{cases} \frac{C_{srgb}}{12.92} & \text{if } C_{srgb} \leq 0.04045 \\ \left(\frac{C_{srgb}+\alpha}{1+\alpha}\right)^{2.4} & \text{if } C_{srgb} > 0.04045, \end{cases} \tag{12.9}$$



(a) Example of the projected image of the WK test pattern.



(b) Crosstalk profile in the middle of the display at distance $310mm$.



(c) Top view of the crosstalk map for the left eye. Captures of the projection surface were made in the range $[220,420]mm$ in steps of $5mm$.

Figure 12.4: Crosstalk profiles.

where C_{linear} is one of the three linear sRGB tristimulus values (R_{linear}, G_{linear}, B_{linear}), C_{srgb} is one of the RGB values of input image in the range $[0,1]$, and $\alpha = 0.055$. The final crosstalk value is computed by using the luminance Y value only. The crosstalk X_L for the left eye is

$$X_L = \frac{\max(KW - KK, 0)}{\max(WK - KK, 0)} \tag{12.10}$$

where KK is the luminance of the black-black pattern, KW the luminance of the black-white pattern, and WK the luminance of the white-black pattern. The maximum function is used to avoid negative luminance values. Right eye crosstalk is simply the inverse of X_L.

Figure 12.4b illustrates the profile of computed crosstalk at distance 310 mm. Zero means no crosstalk, whereas 1 means maximum crosstalk. Values higher than 100% have been set to 1. The regions with high crosstalk are clearly separated from regions with no crosstalk.

Crosstalk appears as bent stripes whose widths increase with user distance as can be see in Figure 12.4c. Optimal eyes locations, defined as sweet spots, are those areas where the crosstalk is minimum. Sweet spots can be found by a thorough analysis of crosstalk profiles at a given user distance for the horizontal center of the display (see Figure 12.5).

Crosstalk values higher than 1 are ignored, and intensities of $WK - KK$ and $KW - KK$ representing the luminance information are shown with a dashed line. So, when one eye sees

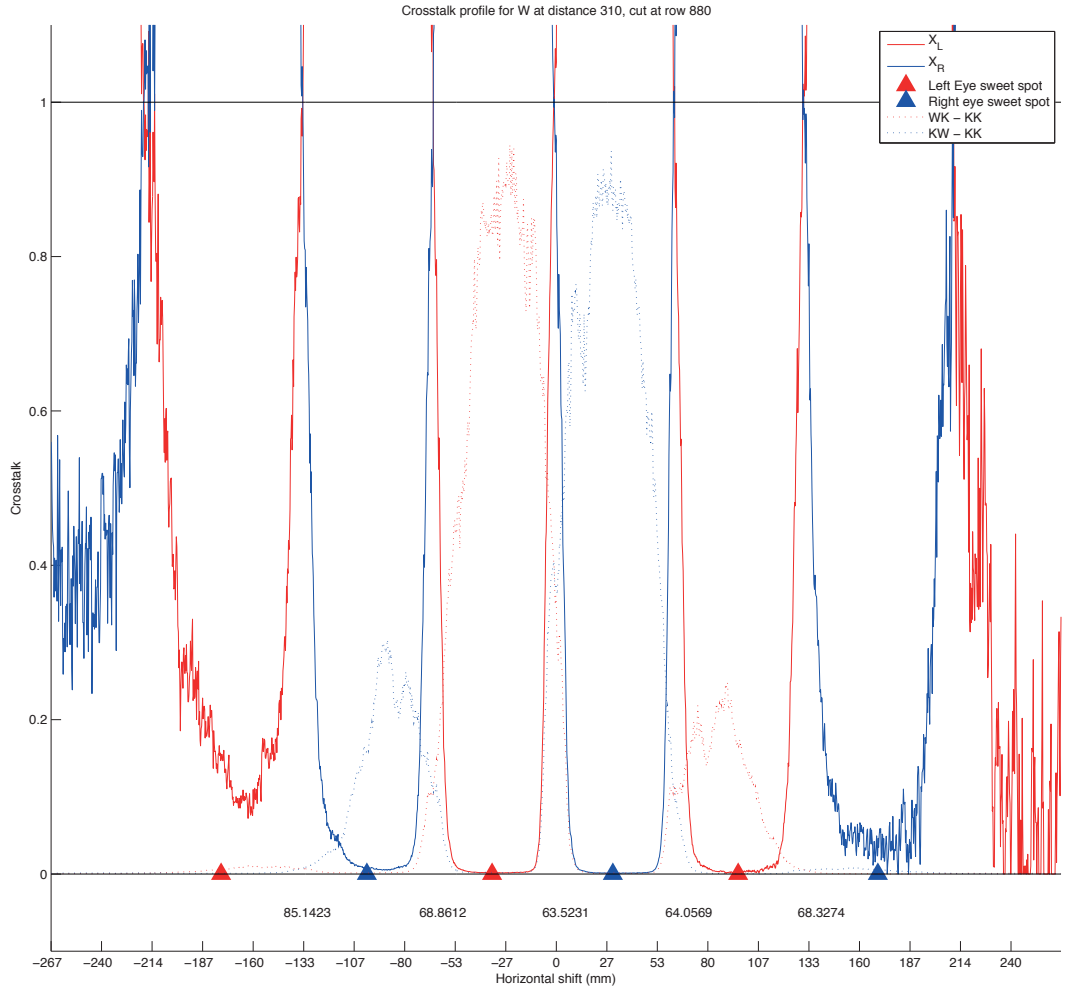
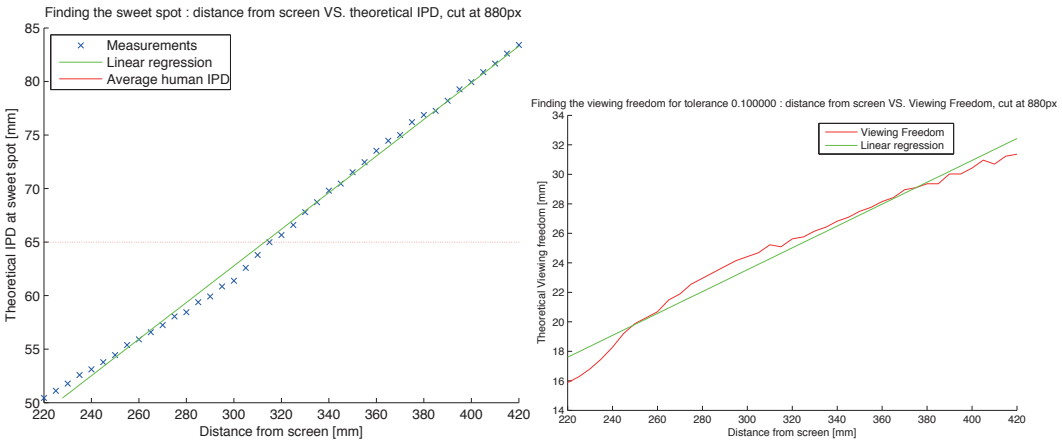


Figure 12.5: Crosstalk profile in the middle of the screen, at distance 310 mm.

as much from the left view as from the right view, the crosstalk value is 1. The sweet spot position is determined with respect to a tolerated error ϵ as a middle point between crosstalk values $X_L = X_R = 1 \pm \epsilon$. The left and right eye sweet spots are marked by red and blue triangles, respectively. Having the crosstalk profile for a whole range of distances, the relation between IPD and OVD can be estimated (Salmimaa and Järvenpää, 2008). To achieve the best viewing condition, the distance between the sweet spots for left and right eye should correspond to the IPD. Figure 12.6a illustrates the mutual dependence between IPD and OVD, which can be approximated by a linear regression

$$\text{OVD (mm)} = 5.848203 \times \text{IPD (mm)} - 67.143329 \tag{12.11}$$

For a standard male IPD (65 mm) the OVD corresponds to 313 mm. Although the crosstalk stripes are bent (see Figure 12.4b), the variation corresponding to possible user movements along the vertical axis is relatively small. Therefore, an assumption of spatial invariance of crosstalk along vertical axis can be made.



(a) Sweet spot for a given IPD.

(b) Viewing freedom versus user distance.

Figure 12.6: Sweet spot and viewing freedom.

Crosstalk tolerance	Linear regression
5%	VF (mm) = 0.068600 × distance (mm) + 1.314246
10%	VF (mm) = 0.074119 × distance (mm) + 1.295925
15%	VF (mm) = 0.078044 × distance (mm) + 1.120717
20%	VF (mm) = 0.080536 × distance (mm) + 1.195488
25%	VF (mm) = 0.081833 × distance (mm) + 1.447605

Table 12.2: Regression models for different crosstalk tolerances.

Considering certain crosstalk threshold, assuring good viewing quality, the viewing freedom in horizontal axis for a given user distance can be determined. For the autostereoscopic displays, the viewing freedom is limited by diamond-shaped regions. However, as the measurements in our experiment were done only within a certain range from 220 mm to 420 mm, the relationship between user distance and viewing freedom must be linear. Moreover, when the crosstalk tolerance decreases, so does the viewing freedom. An example of graph with 10% crosstalk tolerance is illustrated in Figure 12.6b. A linear regression is used for fitting purposes (see Table 12.2). Note that the viewing freedom is assumed to be constant along the vertical axis.

12.2.2 System Description

In this subsection, we propose an implementation in form of a mobile application for parallax barrier based autostereoscopic devices, improving 3D user experience. The core principle of this application is based on an active crosstalk reduction dependent on user position. The actual user position is tracked using face and eye detection and an efficiently designed pixel-wise image operation is performed based on the luminance profiles of each view. To determine the concrete luminance profiles, characterization of the mobile display was performed. Final application subsequently exploits both, the information of eye position and the display char-

acterization, and applies the rendering methods enhancing the QoE. Moreover, the following requirements were taken into account for the final application development

i) real time face detection and eye tracking,
ii) real time crosstalk reduction and pseudoscopy correction,
iii) intuitive, user-friendly feedback for the user,
iv) user-oriented, allowing the parameters change according to user characteristics, and
v) low battery/CPU consumption.

Because 3D imaging is quite a recent topic for mobile displays, there are only two 3D-capable Android phones available, the LG Optimus 3D and the HTC Evo 3D. Both use a 3D display with parallax-barrier that can be switched off to get 2D mode, but differ in terms of CPU, memory, software version and 3D API. For our development, the HTC Evo 3D with the latest available version of Android (Android Ice Scream Sandwich) system was used assuming the landscape displaying mode. The main parameters related to display and front camera are

i) screen size of the smart phone: 960 × 540 pixels,
ii) front camera preview size: 960 × 544 pixels, and
iii) front camera picture size: 1280 × 720 pixels.

In this section, technical aspects of above mentioned individual parts, as well as of the entire mobile application, are described.

Face Detection and Eye Tracking

There are several different face detection algorithms running on Android devices: the Android built-in face detector algorithm, the OpenCV face detection algorithm, and the CamShift algorithm. In order to get as real-time face detection feedback as possible, the Android built-in algorithm *FaceDetectionListener* was used together with OpenCV and HAAR cascade classifiers for eye detection.

The general diagram of the eye tracking algorithm is summarized in Figure 12.7. First, the cascade of HAAR classifiers for eye detection is stored in an XML file and loaded when application boots. Each frame is processed independently and once a face is detected, the current frame is converted from YUV (Android camera format) to RGB using OpenCV conversion functions. Converted frame is then cropped to the region of interest (the upper half of the face) and the result is sharpened by subtracting a blurry (Gaussian blur is assumed) image from the cropped region. Cropping significantly decreases the computational time. Eye detection is performed on the cropped image using the OpenCV library only if a face is previously found.

The frame rate of the front-facing camera varies a lot according to ambient light. When the environment around the user is too dark, the frame rate of the camera decreases and face detection becomes slow. For good lighting conditions, the real-time face detection achieves a frame rate of 15 fps, which is the maximal frame rate of the front-facing camera, whereas

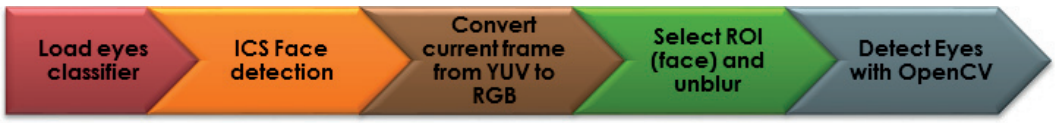


Figure 12.7: Face and eye detection: General diagram.

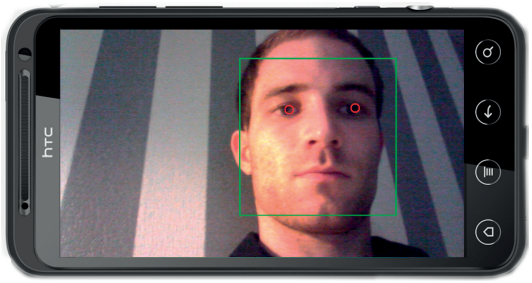


Figure 12.8: Face and eye detection: Preview.

eye detection is performed at maximum of 10 fps. Figure 12.8 shows an example of camera preview with detected face and eyes.

Camera Calibration

To find the distance of the user with respect to the autostereoscopic display, and to determine the sweet spot location, calibration of the front camera was performed. An ideal pinhole camera model, providing the approximation of the relationship between a scene and its projection on a sensor, is assumed. The real-world scene is projected on the image plane, which lies at a focal distance f measured along the optical axis. The focal length f is divided in two components, f_x and f_y, because usual CMOS sensors have rectangular dimensions. The point of intersection between the optical axis and the image plane is called the principal point and, in an ideal case, it corresponds to the center of the image plane. However, this is not the case for most camera models, so two additional parameters, c_x and c_y, related to directional shifts must be introduced. These four parameters were estimated using well known algorithms and a calibration pattern (Bradski and Kaehler, 2008). Distortion parameters of the camera were ignored, as they don't significantly influence the user position in space.

Once the parameters f_x and f_y are found, one can easily compute the distance of the user with respect to the display. Assuming, that the positions of left (y_1) and right eye (y_2) on the horizontal axis of the image plane are known, then

$$y_1 = f_y \frac{Y_1}{Z} + c_y \qquad y_2 = f_y \frac{Y_2}{Z} + c_y \tag{12.12}$$

and their difference, representing the known eye distance D in pixels in the image plane is

$$D = y_1 - y_2 = f_y \frac{Y_1 - Y_2}{Z} \tag{12.13}$$

Figure 12.9: Different colors for feedback circles. After some time in the sweet spots, the camera preview is made invisible and 3D content is shown.

where the difference $Y_1 - Y_2$ represents the IPD of the user in mm. The only unknown variable in Equation (12.13) is Z, which is the user distance with respect to the image plane.

The front camera is located in the upper left corner of the phone, when used in landscape orientation. Therefore, an offset depending on the user position correcting the face and eye position with respect to the horizontal center of the display must be set. Actual offset value for given user distance can be computed according to following linear regression

$$\text{offset (pixels)} = \max(0, -0.410263 \times \text{user distance (mm)} + 318.707571) \tag{12.14}$$

which was found empirically for seventeen different user distances. Moreover, the ratio between the image size (720×1280) and the screen dimension (540×960) is taken into account, thus the offset value corresponds to an accurate shift on the screen. For the unrealistic user distances bigger than 776.8 mm, the offset is set to zero. In the final system, the horizontal offset was computed only for the OVD, whereas the vertical offset was ignored.

An intuitive and non-intrusive feedback helping the user to find a sweet spot was implemented in the final system and its mechanisms are illustrated in Figure 12.9. First, the circles denoting a position where the viewer's left and right eyes should be placed are shown together with camera preview. The position and radius of the circles are based on the iterative combination of four equations

i) Equation (12.11) is used to compute OVD according to the user IPD (internal parameter of the system),
ii) Equation (12.14) is used to shift the circles with an adequate offset corresponding to the OVD,
iii) the OVD is then mapped to a theoretical eye distance in pixels by using the camera calibration (see Equation (12.13)), and
iv) the diameter of the circles depends on the viewing freedom, which is computed using Table 12.2 according to viewer preference for crosstalk tolerance.

The color of the circles is changed from red to green when the eyes of the viewer enter in the sweet spot feedback circles. When good viewing conditions are detected for several consequent frames, feedback circles disappear and 3D content is displayed.

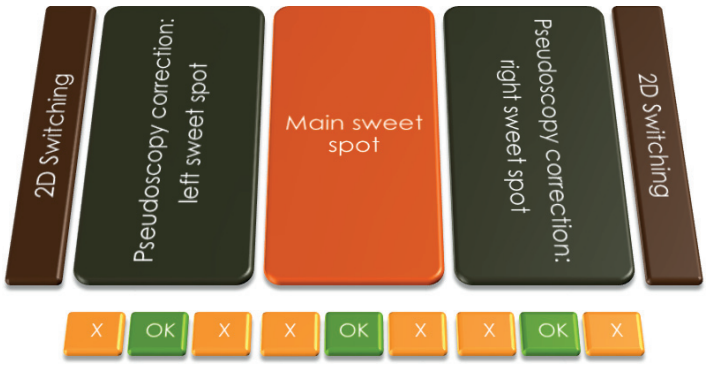


Figure 12.10: Overview of artifacts correction.

Artifacts Correction

Relative user position with respect to the screen determines the particular corrections of the visual information. Corrections of crosstalk visual artifact are carried out in three different ways: pseudoscopy correction, crosstalk compensation and 2D/3D switching, as illustrated in Figure 12.10.

Whereas main sweet spot position implies no processing of the visual information, in one of the side sweet spot position the pseudoscopy must be corrected. For the position between sweet spots, the crosstalk compensation is performed depending on the crosstalk tolerance threshold. At the edge positions, when the viewing angle is too large, the display mode is switched back to 2D.

Pseudoscopy is corrected by swapping left and right views when the user is in a "pseudoscopic" sweet spot, i.e., when the left eye sees the image intended for the right eye and vice versa. More specifically, the left and right views are swapped when the crosstalk for left or right eye is higher than 50%. Pseudoscopy correction considers four central sweet spots illustrated as blue and red triangles in Figure 12.5.

The simple yet effective method for crosstalk reduction has been implemented according to Daly et al. (2011) and is illustrated in Figure 12.11. First, the crosstalk values X_L and X_R are retrieved from the crosstalk map for the left and right eye. Subsequently, the unintended visual information $X_L R$ and $X_R L$ is subtracted from the left L and right view R, respectively. Each RGB channel is finally normalized independently into the range $[0,255]$. All operations are carried out on matrices, instead of bitmaps, in 16 bits representation to allow negative values and normalization. The average time to process a pair of views whose size is 540×480 pixels is 200 ms, which is slower than real-time (15 fps $\rightarrow$ 67 ms), however still sufficient to improve the 3D rendering.

Switching to 2D mode has been implemented when crosstalk information is not available for a user position. To get a real-time response and to keep the same level of the display luminosity when switching from 2D to 3D rendering mode, the 3D device capability is always enabled and

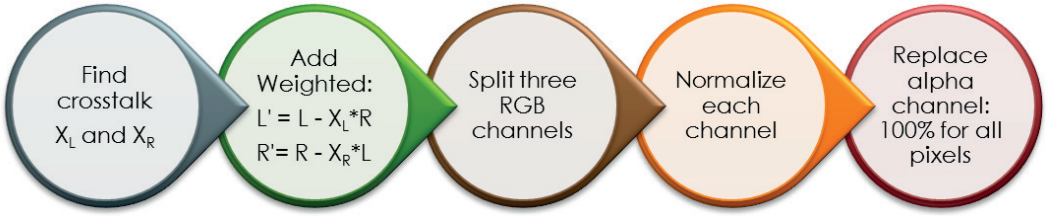


Figure 12.11: Active crosstalk compensation process.

(a) *Bridge* (b) *Cactus* (c) *Dog*

(d) *Flowers* (e) *Landscape* (f) *Mountains*

Figure 12.12: Left view of the images used for subjective evaluation, image (a) was used for training.

only one image (left or right) is displayed for both views. In 2D mode, left image was chosen to be displayed for both views by default.

12.2.3 Subjective Evaluation

In this subsection, the proposed crosstalk reduction system is evaluated through a subjective assessment. More specifically, the user preference between two standard modes (2D and 3D) of the mobile phone and the proposed system is analyzed in terms of two different aspects: image and depth quality.

Dataset

The dataset consists of six JPS images in side-by-side stereo format with different disparities and colors. Resolution of the content was set to 540×960 pixels (resolution of display) for efficiency reasons. One content was used for training (see Figure 12.12a) and the rest for the test purposes (see Figures 12.12b to 12.12f).

Test Methodology

The PC method with a ternary scale (see Section 2.4.3) was chosen as judging the quality of different 2D and 3D rendering systems individually may be quite difficult. Pairs of images, "A" and "B", which resulted from different imaging systems, were presented in succession order on the mobile phone to the viewer using the developed Android application. Subjects were asked to judge which video sequence in a pair ("A" or "B") is preferred in terms of image and depth quality. The option "same" was included to avoid random preference selections. For each of the 5 test images (*Cactus*, *Dog*, *Flowers*, *Landscape*, and *Mountains*), all the possible combinations of the 3 test conditions (2D mode, 3D mode, and proposed system) were considered, leading to a total of $5 \times \binom{3}{2} = 15$ comparisons. The two asked questions for each test condition were closely specified as

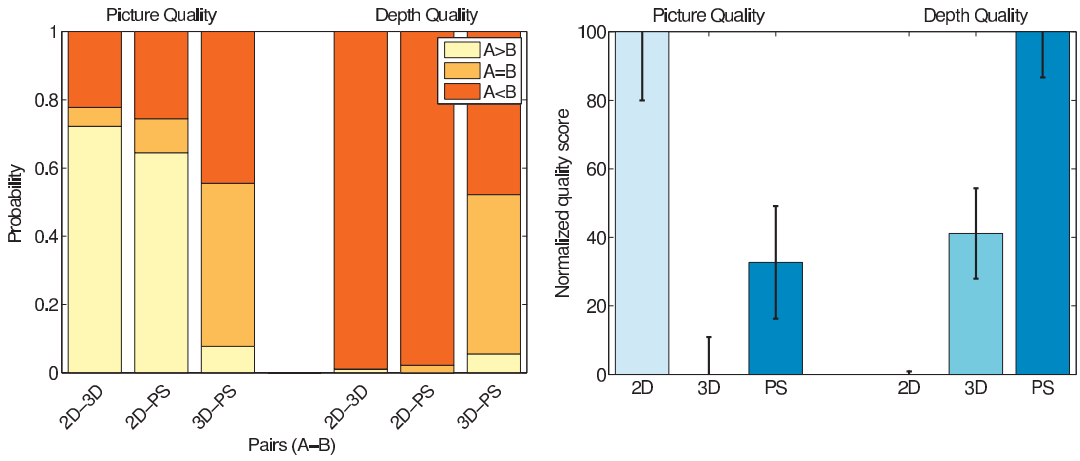
- Image quality: tilt the phone a little bit! Which scenario leads to the best image quality?
- Depth quality: which scenario gives better depth or 3D perception?

A total of 18 naïve subjects (2 females and 16 males) took part in the experiment. All subjects were screened for correct visual acuity, color vision, and stereo vision using Snellen chart, Ishihara chart, and Randot test, respectively.

Before the test, oral and written instructions were given to the participants to explain their tasks and different aspects of the evaluation, such as short description of 3D technology and common artifacts. Then, the subject's IPD was measured and set as a parameter of the application manually. Additionally, a training session was organized to allow participants to familiarize with the assessment procedure. The experiment was performed in normal daylight conditions. The crosstalk tolerance threshold for the proposed 3D rendering scheme was set to 5%.

Data Processing

To analyze user preference for the different imaging systems, statistical tools were applied to the individual ratings. No outlier detection was performed since there is no international recommendation or a commonly used outlier detection technique for PC results. Before estimating MOS values for PC results, the winning frequency and the tie frequency are computed from the obtained subjective ratings for each pair of stimuli. This can be done individually for each test video content or jointly over all contents. To compute the preference matrix, only wins were taken into account, whereas ties were discarded. Then, the Bradley-Terry-Luce model (see Section 2.6.3) was used to estimate MOSs via maximum likelihood estimation. Ties were considered as half way between the two preference options and equally distributed. The CIs for the maximum likelihood estimates of the scores were obtained using the Hessian matrix of the log-likelihood function. Results were normalized to the range $[0, 100]$ for a better representation.



(a) Preference probabilities.

(b) Normalized quality scores.

Figure 12.13: Subjective results for 2D mode, 3D mode, and proposed system (PS).

12.2.4 Results

Figure 12.13a shows the preference and tie probabilities obtained over all test images for picture and depth quality. In terms of picture quality, the proposed 3D rendering outperforms the 3D mode with a ratio of preference probability 5 : 1. The general preference of 2D mode over proposed 3D rendering and 3D mode is again demonstrated with a preference probability of 65% and 75%, respectively. The results for the perceived depth quality are, as expected, in favor of both 3D modes. The proposed 3D rendering is much better when compared to the normal 3D mode. This can be explain by the fact that the crosstalk and pseudoscopy reduce perceived depth in general, as showed by Tsirlin et al. (2012).

Figure 12.13b shows the MOSs and CIs obtained over all test images for picture and depth quality. The comparison of depth quality for the different rendering modes shows that the proposed 3D rendering clearly outperforms the others. Moreover, it offers significant improvement in terms of picture quality in comparison to normal 3D rendering mode.

12.3 Improving 3D QoE on Multiview Autostereoscopic Displays

To improve the QoE provided by multiview autostereoscopic displays, researchers have proposed to exploit viewer tracking. Dodgson (2006) has analyzed an ideal 3-view display, where only two views are actually displayed, to better deal with the transition of one eye between two adjacent zones. Boev et al. (2008) have developed a single-viewer system based on user-tracking. The system performs on-the-fly visual optimization to achieve continuous head parallax, i.e., to avoid the repetition effect between the lobes, mitigate crosstalk, and improve brightness. Kooima et al. (2010) have proposed three techniques to improve the user experience: perspective tracking, channel tracking, and channel reassignment. Similar works were

also reported by other researchers (S.-K. Kim et al., 2012; S.-K. Kim et al., 2013; J. Liu et al., 2015a). Nam et al. (2011) have proposed another approach to actively reduce crosstalk based on the user position. This technique reduces the crosstalk level form 19.1% to only 2.6% for a multiview display using sub-pixel rendering. Advanced multi-user autostereoscopic displays have been developed within the European Union-funded projects MUTED and HELIUM 3D (Surman et al., 2010). These displays utilize multi-user head-tracking to provide a proper 3D image to each viewer, based on the eyes position.

Most of the previous works only describe a proposed system without evaluating its performance, whereas other works were performed on very expensive technologies, e.g., laser projection and low loss transparent display screen, which are far from mass production. Except for the MUTED and HELIUM 3D projects, other works were performed on multiview autostereoscopic displays having a rather limited number of views (typically eight to nine), whereas most advanced multiview autostereoscopic displays, e.g., the Dimenco displays, typically have around 30 views. With fewer views, the separation between the different luminance profiles is more pronounced and crosstalk compensation is relatively easy, whereas this problem is much more difficult as the number of views increases since the overlap between the luminance profiles is more sever. Additionally, none of these works provides a full description and subjective evaluation of a complete active crosstalk reduction system for current multiview autostereoscopic display technology. To address these problems, we propose and evaluate a system to improve the QoE provided by current and future multiview autostereoscopic display technologies. In particular, our solution aims to reduce the amount of crosstalk perceived by the viewer. The idea is to determine the viewers' position, hence the views they can see, and to adjust the different displayed views in real time such that the QoE is maximized for each viewer. We implemented our solution considering a single viewer scenario for a 52" full HD 28-view Dimenco BDL5231V autostereoscopic display with slanted lenticular sheet. The user preference between standard 2D and 3D modes and the proposed system is evaluated in terms of image quality, depth quality, and visual discomfort through an informal subjective evaluation conducted with five expert viewers. This section describes in details the proposed system and reports the details and results of the subjective evaluation.

12.3.1 Display Characterization

The characterization of multiview autostereoscopic displays is usually performed by measuring the luminance emitted by each view at different positions in front of the monitor, which is commonly known as luminance profiles. Significant efforts have been devoted to multiview autostereoscopic display characterization over the recent years. The International Committee for Display Metrology has recently proposed a standardized way to measure crosstalk at a given point in space (International Committee for Display Metrology, 2012). However, this approach is time-consuming and expensive, as explained in Section 12.2.1. Consequently, we adopted the same approach as in Section 12.2.1. The main idea is to display a specific test pattern and acquire an estimation of the luminance profiles at a given distance using a DSLR

Table 12.3: Dimenco BDL5231V display characteristics provided by the manufacturer.

Diagonal	52” / 132.0 cm
Pixel Pitch	0.60 mm
Resolution	1920 × 1080 dots
Color	1.07 billion colors
Brightness	700 cd/m^2 (typ.)
Contrast ratio	4000:1 (typ.)
Response time	8 ms
View angle	Up and Down 178°
	Left and Right 178° (typ.)

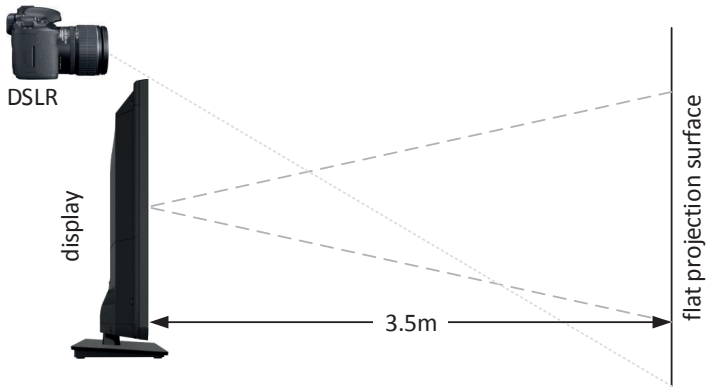


Figure 12.14: Schematic of the display characterization setup.

camera. In this study, a 52” full HD 28-view Dimenco BDL5231V autostereoscopic display with slanted lenticular sheet was used.

Setup

The luminance was measured on a vertical flat projection surface, which was parallel to the display and placed at a fixed distance of 3.5 m from the display. This distance is chosen to be the OVD of the display. The measurements were performed in a dark room environment. Since the camera cannot be placed at the center of the display without interfering with the measurements on the projection surface, the camera was placed on top of the monitor and controlled remotely. Figure 12.14 illustrates the setup. We ensured that the camera was parallel to the 3D display and to the projection screen to minimize any distortion. All camera parameters were kept constant during the experiments. The test patterns displayed on the monitor were generated by setting one particular view to white and all other views to black. This process was repeated for each view to measure the luminance profile of the corresponding view. Figure 12.15 depicts the resulting luminance at 3.5 m from the display. As it can be observed, the luminance distribution consists of five slanted cones, due to the use of a slanted lenticular sheet. The luminance distribution is similar for all views, up to a horizontal shift.

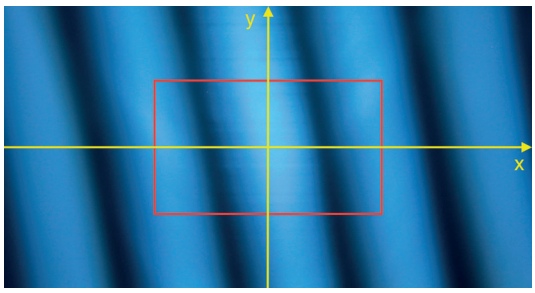


Figure 12.15: Resulting luminance at 3.5 m from the display, captured by the camera placed on top of the monitor, when one view is set to white and all other views are set to black. The red box represents the display area.

Luminance Extraction

For each view, four images were captured and averaged to reduce noise. The averaged images were further cropped to the region of interest (the area of the projection surface). For computation ease, only 30% of the initial size of the picture was kept and a median filter of size 10×10 pixels was applied to further reduce artifacts due to noise. The luminance information was then extracted by converting the gamma-encoded sRGB to linear XYZ values and by keeping only the Y channel. Note that the luminance values are defined up to a scale factor, as no reference luminance value was measured.

Luminance Profile Fitting

Figure 12.16 depicts the variation of all luminance profiles along the horizontal axis, i.e., the x-axis, at the center of the display ($y = 0$). This corresponds to a cut along the x-axis on Figure 12.15, repeated on the luminance distribution generated by each view. As it can be observed, the global intensity is maximum at the center of the display and decreases as the distance from the center of the screen increases. Within the boundaries limited by the display frame (indicated by two red lines on Figure 12.16), the global intensity seems to have a Gaussian envelope. For each view, the envelope seems modulated by a squared cosine (since the luminance values are always positive), with five maxima corresponding to the five cones. The luminance profiles of the different views are similar up to a translation, which corresponds to a phase factor in the cosine modulation. A cut along the vertical axis, i.e., the y-axis, also reveals a Gaussian shape (see Figure 12.17). In this work, we limited the study of the luminance profiles to an area corresponding to the display area.

Based on the above analysis, the luminance profile, $L(x, y)$, was modeled as a 2D Gaussian envelope modulated by a squared cosine function

$$L(x,y) = A\cos^2\left(\omega x + \tau y + \varphi\right) e^{-\left[a(x-x_c)^2 + 2b(x-x_c)(y-y_c) + c(y-y_c)^2\right]} + o \tag{12.15}$$

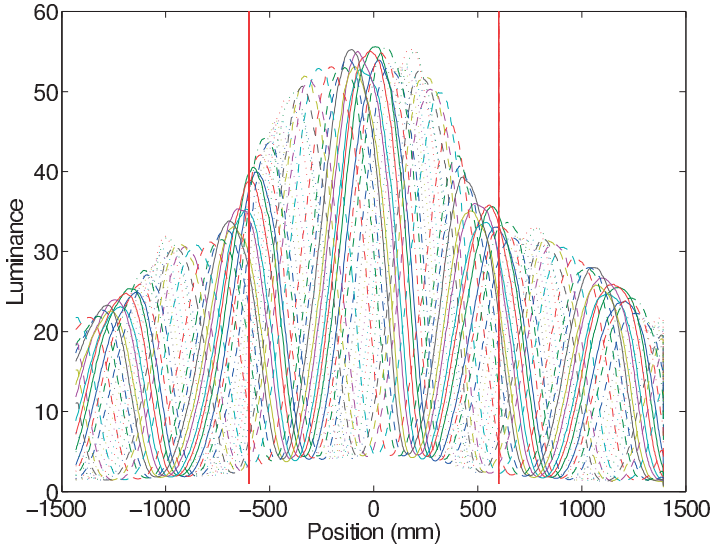


Figure 12.16: Variation of all luminance profiles along the horizontal axis at the center of the display ($y = 0$).

with

$$a = \frac{\cos^2\phi}{2\sigma_x^2} + \frac{\sin^2\phi}{2\sigma_y^2} \qquad b = -\frac{\cos 2\phi}{4\sigma_x^2} + \frac{\sin 2\phi}{4\sigma_y^2} \qquad c = \frac{\sin^2\phi}{2\sigma_x^2} + \frac{\cos^2\phi}{2\sigma_y^2} \tag{12.16}$$

where A and o are the amplitude and offset of the 2D Gaussian, respectively, ω represents the frequency of the cosine modulation, τ is phase factor to represent the slanted nature of the luminance distribution, φ is the phase factor representing the translation between the different views, (x_c, y_c) is the center of the 2D Gaussian, σ_x and σ_y represent the horizontal and vertical standard deviations of the 2D Gaussian, respectively, and ϕ is a tilt factor of the 2D Gaussian added to improve the fitting.

Figure 12.18 depicts the result of the surface fitting of the luminance distribution of view 5 using Equation (12.15), i.e., with 10 parameters.

Parameters Reduction

Each luminance profile of the 28 views was fitted independently using Equation (12.15), yielding to a total of $28 \times 10 = 280$ parameters. All parameters exhibited small variations, except for φ, which evolved linearly with the view number (up to a period π). These results are in line with the observations reported above. Based on these observations, the parameter set was further reduced by computing the average value of the different parameters, except for φ. For the parameter φ, a linear regression was performed

$$\varphi = \alpha v + \beta \tag{12.17}$$

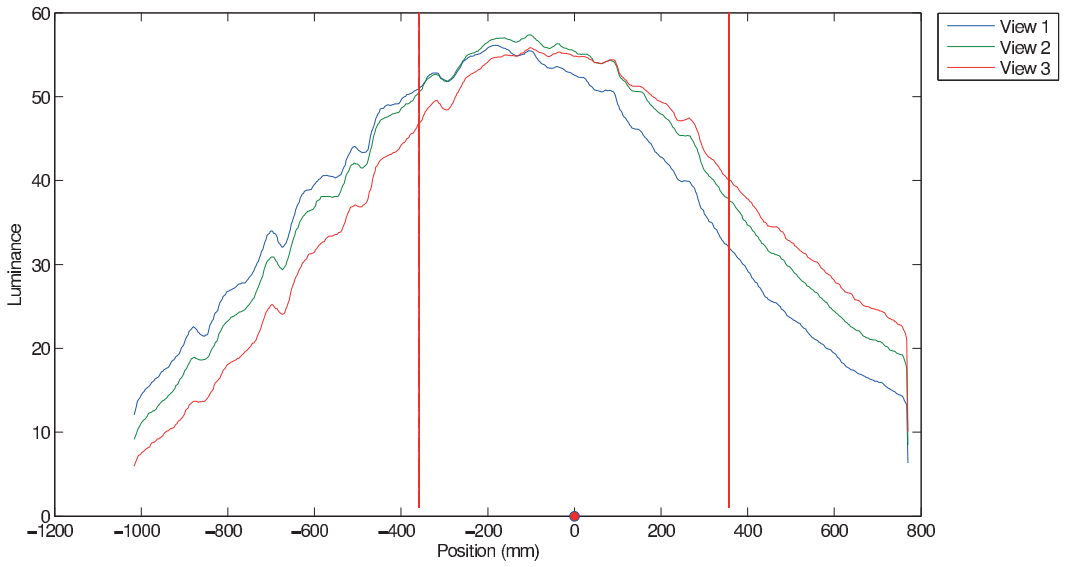


Figure 12.17: Variation of the luminance of views 1, 2, and 3 along the vertical axis at the center of the display ($x = 0$).

Table 12.4: Comparison of the RMSE and coefficient of determination (R^2) for the individual fitting and reduced set.

	RMSE	R^2
Individual fitting	1.9666	0.9848
Reduced set	2.3113	0.9789

where v is the view number and α and β are the parameters of the linear regression.

Table 12.4 reports the RMSE and coefficient of determination (R^2) averaged over the 28 views. As it can be observed, reducing the set of parameters from 280 (individual fitting) to 11 (reduced set) parameters had little impact on the error between the measured and fitted values.

12.3.2 System Description

This subsection describes the on-the-fly intelligent view assignment and multiview shuffling used in the active crosstalk reduction system. Details are provided regarding the user tracking system and implementation of the application.

User Tracking

The Microsoft Kinect and Face tracking SDKs were used to track the face and face features. In particular, the features corresponding to the left and right corners of each eye were used. The center of the eye was computed as the mid-point between the left and right corners, as

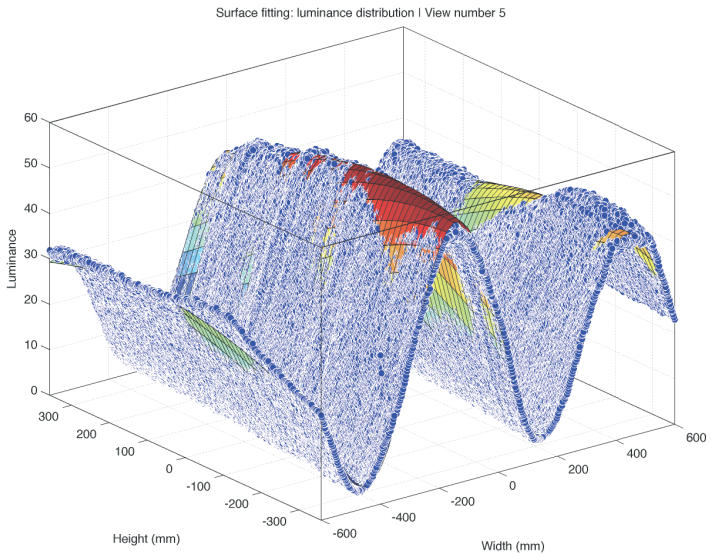


Figure 12.18: The result of the surface fitting of the luminance distribution of view 5. The dots represent the measurements.

this feature is not directly provided by the Face tracking SDK. The face tracking application developed is highly reliable and robust, but sensitive to lightning conditions. The face tracking was performed in real time, with a frame rate varying between 25 and 30 fps, depending on lighting conditions.

Intelligent View Assignment

Typically, an N-view autostereoscopic system takes $M \ll N$ views as input, due to limitations imposed when using physical cameras. From the limited input views, the missing $N - M$ views are synthesized, for examples by using DIBR. In the most common approach, each view corresponds to a slightly different viewpoint. The reasons behind this approach are multiple: providing a motion parallax effect when the observer moves his/her head in front of the display, coping with different viewing distances, coping with different IPD, providing 3D effect for different viewers located at different positions, etc. However, this approach might not be optimal in some cases, for example when only one subject is watching the display and standing still, and introduces crosstalk, as the profiles of the different views overlap quite significantly (see Figure 12.16).

To reduce perceived crosstalk, our idea consists in performing an intelligent assignment of the different views based on the luminance profiles and the observer's position. Let us assume that a single user is positioned such that his/her left and right eyes see only views 3 and 7, respectively. In this case, the optimal solution would be to assign to views 3 and 7 the content intended for the left and right eyes, respectively. Unfortunately, in a practical scenario, the separation is not that clear and each view is perceived by both eyes, at a different level. However, each view is typically perceived more by one eye than by the other. Therefore, the

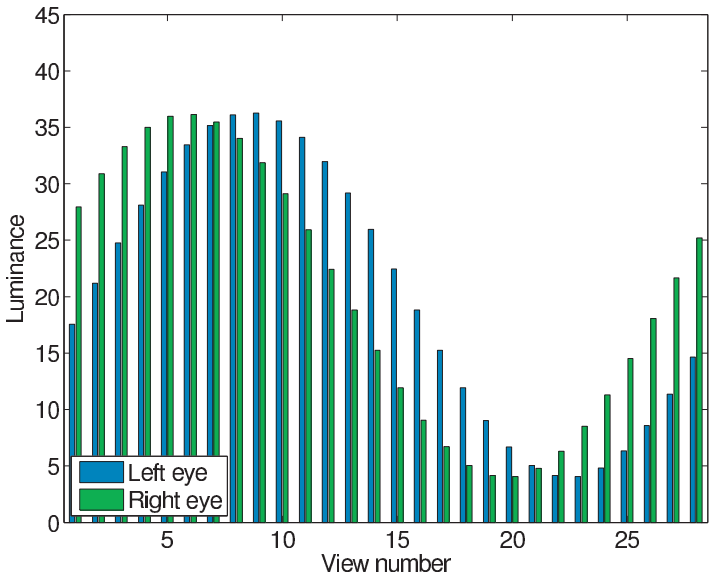


Figure 12.19: Luminance perceived by each eye at a given position.

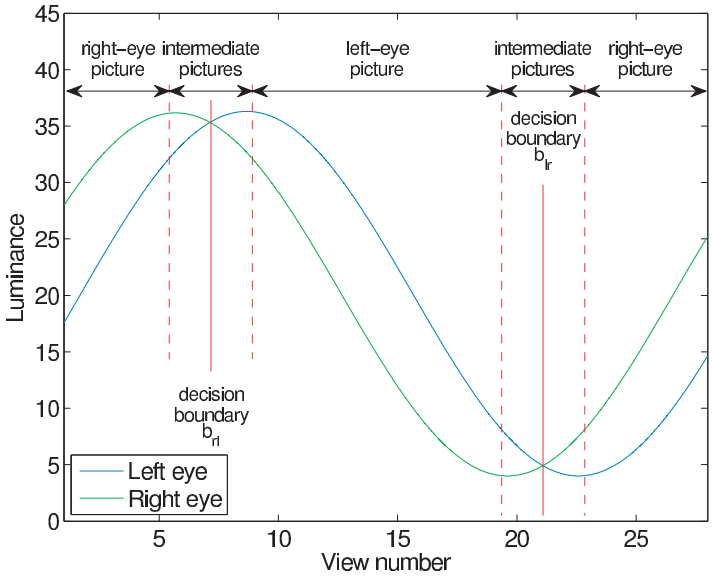


Figure 12.20: View assignment for a given position.

content that should be assigned to each view can be determined by the eye that sees the most this specific view.

From the eyes position determined by the user tracking and luminance profiles, it is possible to determine for each view the luminance perceived by each eye. Figure 12.19 illustrates the luminance perceived by each eye at a given position, as a function of the view number. These values are obtained by evaluating Equation (12.15) at the eyes positions for each view independently. From this information, the eye for which the luminance is maximum would determine the content assigned to each view. In this case, a direct comparison of the lumi-

nance perceived by each eye would be performed for each view. However, Figure 12.19 can be seen as the sampled version at fixed integer positions, corresponding to the view numbers, of a continuous function, as if the view numbering was continuous instead of discrete. Since the luminance profiles were fitted with a limited set of parameters where only φ was depending on the view number, Equation (12.15) can be evaluated at non-integer view numbers. Figure 12.20 illustrates the luminance perceived by each eye, as a continuous function of the view number. Views for which the luminance curve corresponding to the right eye lies above the luminance curve corresponding to the left eye should display the right eye picture, and vice-versa. The decision boundaries can easily be determined by computing the two points at which the curves intersect.

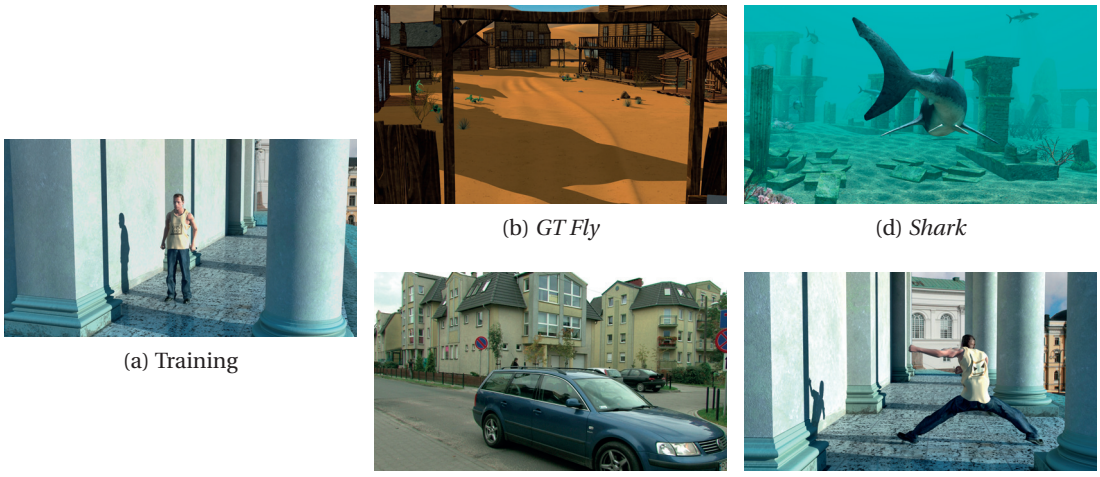
Multiview Shuffling

Assigning only two different images, i.e., the left and right eye pictures, following the methodology described here above did not look very pleasant on the display for two reasons. First, the luminance profiles have a significant overlap: the view that maximizes the luminance perceived by the right eye leaks quite significantly into the left eye, and vice-versa (see Figure 12.19). Second, the edges at the objects' boundaries corresponding to sharp depth transitions did not look very pleasant because of the sub-pixel interlacing. This effect does not appear in standard 3D mode, because the multiple views contain somewhat similar information, which tends to smooth out the depth transitions and blur objects' boundaries. To overcome these issues, three intermediate pictures, corresponding to equidistant viewpoints, located in between the left and right eye pictures, were used for the views near the decision boundary (see Figure 12.20). For example, near the right eye picture to left eye picture decision boundary (b_{rl}), the center-right (p_{cr}), center (p_c), and center-left (p_{cl}) intermediate pictures are assigned as

$$v_n = \begin{cases} p_{cr} & \text{if } n \in [b_{rl} - \delta_e, b_{rl} - \delta_c[\\ p_c & \text{if } n \in [b_{rl} - \delta_c, b_{rl} + \delta_c] \\ p_{cl} & \text{if } n \in]b_{rl} + \delta_c, b_{rl} + \delta_e] \end{cases} \tag{12.18}$$

where v_n is the n-th view. Therefore, 5 pictures were assigned to the 28 views, but with a different spacing for each picture, whereas 28 different pictures are used in the 3D mode with regular spacing, as each view uses a different picture. The number of intermediate pictures and parameters ($\delta_c = 1$ and $\delta_e = 4$) were determined empirically to achieve the best rendering. This solution smooth the image and enhances the visual comfort when compared to using only two pictures.

The Dimenco BDL5231V monitor uses an LCD panel composed of 1920×1080 pixels. However, the shuffling of the 28 views is done at a sub-pixel level. Dimenco provides a software for shuffling 28 full HD video sequences corresponding to the 28 views into a single full HD video to be displayed on the monitor. This tool was reversed engineered, by using simple input patterns, to determine the sub-pixel arrangement, i.e., to determine which sub-pixel



(a) Training

(b) *GT Fly*

(d) *Shark*

(c) *Poznan Street*

(e) *Undo Dancer*

Figure 12.21: MVD contents used in the experiments.

corresponds to which view. This information allows us to perform the multiview shuffling in our application, which is also much faster than the original software provided by Dimenco.

Final System and Implementation

To reduce the impact of the user tracking imprecision and increase visual comfort, small head movements (less than 2 cm in any direction) were discarded. Additionally, to avoid flickering when a new picture is assigned, fading was performed between two successive renderings. The fading was performed by computing three intermediate images using weighted addition of the old and new pictures to display, whose weights increased gradually in favor of the new picture. For our experiments, the system was implemented in C++ using the OpenCV library and achieved a rendering at about 30 fps.

12.3.3 Subjective Evaluation

To evaluate the performance of the proposed system over the 2D and 3D modes of the display, an informal subjective evaluation was performed with five expert viewers.

Dataset

Five MVD contents were used in the experiments: one for training and four for testing (see Figure 12.21). These contents are used by the JCT-3V of VCEG and MPEG (JCT3V-E1100). *Poznan Street* is a real scene with estimated depth maps, whereas the three remaining contents are computer-generated scenes with ground truth depth maps. One key frame, which maximizes the amount of depth, was selected for each content.

Test Methodology

The PC method with a ternary scale (see Section 2.4.3) was chosen as judging the quality of different 2D and 3D rendering systems individually may be quite difficult. Pairs of images, "A" and "B", which resulted from different imaging systems, were presented in succession order on the display. Subjects were asked to judge which video sequence in a pair ("A" or "B") is preferred in terms of image and depth quality. The option "same" was included to avoid random preference selections. For each of the 4 test contents, all the possible combinations of the 3 conditions (2D mode, 3D mode, and proposed system) were considered, leading to a total of $4 \times \binom{3}{2} = 12$ comparisons.

Subjects were allowed to move freely (within a range defined by the monitor frame) along a line parallel to the display, at the OVD of 3.5 m, which corresponded to the measurement distance (see Section 12.3.1).

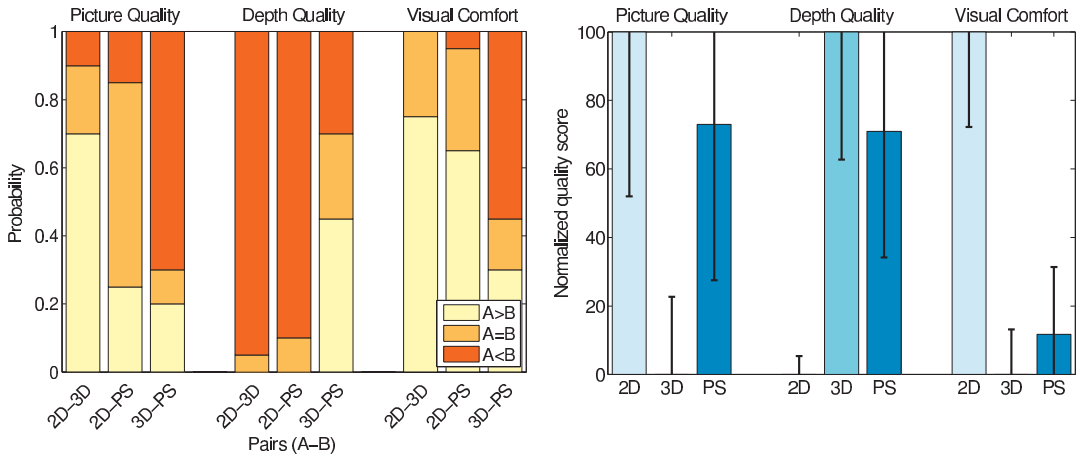
Data Processing

To analyze user preference for the different imaging systems, statistical tools were applied to the individual ratings. No outlier detection was performed since this was an informal subjective evaluation with expert viewers. Before estimating MOS values for PC results, the winning frequency and the tie frequency are computed from the obtained subjective ratings for each pair of stimuli. This can be done individually for each test video content or jointly over all contents. Then, the Bradley-Terry-Luce model (see Section 2.6.3) was used to estimate MOSs via maximum likelihood estimation. Ties were considered as half way between the two preference options and equally distributed. The CIs for the maximum likelihood estimates of the scores were obtained using the Hessian matrix of the log-likelihood function. Results were normalized to the range $[0, 100]$ for a better representation.

12.3.4 Results

Figure 12.22a shows the preference and tie probabilities obtained over all test images for picture quality, depth quality, and visual comfort. As it can be observed, the proposed system significantly improves picture quality when compared to the 3D mode, as it has a preference probability of 70%, whereas the 3D mode has a preference probability of only 20%. With the proposed system, less crosstalk was perceptible and there was no unpleasant transition between the different viewing cones. The 2D mode and proposed system were perceived as similar in 60% of the test stimuli, which shows that the proposed system provided a picture quality comparable to that of the 2D mode.

Regarding depth quality, the 3D mode showed a clear advantage over the 2D mode. Results show a slight preference for the 3D mode over the proposed system, with a preference probability of 45%. Nevertheless, the depth quality of the proposed system is still much better than that of the 2D mode, despite the absence of motion parallax depth cues when compared to



(a) Preference probabilities.

(b) Normalized quality scores.

Figure 12.22: Subjective results for 2D mode, 3D mode, and proposed system (PS).

the 3D mode. In terms of visual comfort, 2D mode is preferred most of the time. The proposed system also improves visual comfort when compared to the 3D mode, as it is preferred in 55% of the test stimuli, whereas the 3D mode is preferred in only 30% of the test stimuli. From the comments of the viewers, this can be explained by the fact that they had some difficulties to predict the behavior of the system as they moved when compared to the 3D mode, where they could find a predictable and fixed sweet-spot.

Figure 12.22b shows the MOSs and CIs obtained over all test images for picture quality, depth quality, and visual comfort. As it can be observed, the proposed system significantly enhances picture quality when compared to the 3D mode and provides similar depth perception, as the CIs overlap significantly. However, the improvement in terms of visual comfort is not significant.

12.4 Conclusion

This chapter investigated different systems to reduce stereo artifacts generated at the visualization stage to improve QoE on 3D displays. First, we proposed and evaluated two different approaches that exploit visual attention to improve 3D QoE on stereoscopic displays: an offline system, which uses a computational model of visual attention to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions. From the saliency map, which was computed using a 3D visual attention model, the region-of-interest and its disparity were extracted. From the eye tracking measurements, filtered gaze points were used in conjunction with the disparity map to extract the disparity of the object-of-interest. Horizontal image translation was performed to bring the fixated object on the screen plane. The shift was determined based on the extracted disparity values and filtered in time to have smooth transitions that do no create visual discomfort. The user

preference between standard 3D mode and the two proposed systems was evaluated in terms of image quality, depth quality, and visual discomfort through a subjective evaluation. Results showed that exploiting visual attention significantly improves image quality and visual comfort, with a slight advantage for real time gaze determination. Depth quality is also improved, but the difference is not significant.

Second, we proposed and evaluated an active crosstalk reduction system for mobile autostereoscopic displays. The proposed system was implemented on a HTC EVO 3D smartphone. To determine the crosstalk level at each position, a full display characterization was performed. Furthermore, the localization of sweet spot and computation of the viewing freedom was performed. A special Android application was implemented to track the user face and eyes, and to correct artifacts in real-time according to his/her position. The proposed system was designed in the way that it first helps the user to find the sweet spot and then compensates for crosstalk artifacts and/or pseudoscopy. The user preference between standard 2D and 3D modes and the proposed system was evaluated in terms of image quality and depth quality through a subjective evaluation. Results showed that in terms of depth perception, the proposed system clearly outperforms the 3D and 2D modes. In terms of image quality, 2D mode was found to be best, but the proposed system outperforms 3D mode. The evaluation of the Android application showed that there are also limitations in terms of processing speed and power usage.

Third, we proposed and evaluated an active crosstalk reduction system for multiview autostereoscopic displays. The proposed system was implemented considering a 52” full HD 28-view Dimenco BDL5231V autostereoscopic display with slanted lenticular sheet. The display was characterized in terms of luminance distribution and the luminance profiles were modeled using a limited set of parameters. A Kinect sensor was used to determine the viewer position in front of the display. The proposed system performs an intelligent on the fly allocation of the output views to minimize the perceived crosstalk. The user preference between standard 2D and 3D modes and the proposed system was evaluated in terms of image quality, depth quality, and visual discomfort through an informal subjective evaluation with five expert viewers. Results showed that picture quality is significantly improved when compared to the standard 3D mode, for a similar depth perception and visual comfort.

13 Conclusion

A major part of this thesis was dedicated to measuring QoE in immersive video technologies via subjective visual quality experiments, preference studies, and eye tracking experiments. Because user studies are time consuming, expensive, and not always feasible, objective models that can predict QoE are needed as well. Therefore, in the second part of this thesis, we evaluated the performance of objective quality models for predicting QoE in immersive video technologies. Finally, image and video processing techniques can be used to improve QoE by reducing visible artifacts that impact the processing chain from capture to display. In particular, the last part of this thesis aimed at improving QoE on 3D displays by reducing stereo artifacts generated at the visualization stage. The following sections enumerate the technical contributions of this thesis, as well as the contributions to reproducible research, and give an outlook for future research.

13.1 Technical Contributions

In this thesis, we have built a rigorous framework for subjective evaluation of new types of image and video content. We have proposed different procedures and analysis tools for measuring QoE in immersive technologies. We have put essential concepts of multimedia QoE under this framework. These concepts not only are of fundamental nature, but also have shown their impact in very practical applications. In particular, the JPEG, MPEG, and VCEG standardization bodies have adopted these concepts to select technologies that were proposed for standardization and to validate the resulting standards in terms of compression efficiency.

This thesis has tackled the problems of measuring, predicting, and improving QoE in immersive technologies. To address these problems, we have applied our rigorous framework through several in-depth investigations. The following subsections describe in details the technical contributions of the thesis in each of these categories.

13.1.1 Measuring Quality of Experience

We have proposed a novel method to estimate CIs for the Thurstone Case V model for paired comparison experiments with ternary scale. The proposed model relies on the assumption that ties convey information about significant differences between two stimuli being compared. This model can be used to better interpret PC data and takes advantage of the ternary scale, which is more natural since it doesn't force subjects to randomly choose one option when they don't perceive any difference between the two stimuli.

We have proposed two novel procedures to compare MOS values of two subjective experiments. The first procedure computes the estimation errors, i.e., the number of times the other experiment underestimates or overestimates the results of the reference experiment. The second procedure computes the classification errors, i.e., the number of times the other experiment leads to a different conclusion on a pair of images or video sequences than the reference experiment. These procedures were inspired from the procedures used to benchmark objective quality metrics and can be used to better analyze potential differences between two subjective experiments than a simple correlation analysis.

We have proposed two extensions of the Bjøntegaard model to calculate the coding efficiency between different codecs. First, we have proposed a model to calculate the coding efficiency for two-layer coding systems. The proposed model extends the Bjøntegaard model from R-D curve fitting to R^2-D surface fitting. It uses a cubic surface as fitting function and a more complex characterization of the domain formed by the data points to compute a more realistic estimate of the compression efficiency. The proposed model aims at investigating the impact on quality of the interaction of the base and enhancement layers bit rates, but it can also be used for other applications, e.g., to optimize the bit rate allocation between texture and depth in 3D video coding. Second, we have proposed a model to calculate the average coding efficiency based on subjective quality scores. To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the human visual system, a logistic function is used to fit the R-D values. The average MOS and bit rate differences are computed between the fitted R-D curves. To consider the statistical property of subjective scores, the 95% CIs associated with the MOSs are considered to estimate corresponding confidence intervals on the calculated average MOS and bit rate differences. The proposed model is expected to report more realistic estimation of coding efficiency than the Bjøntegaard model, as it relies on subjective quality scores instead of PSNR measurements.

We have evaluated the performance of HEVC over AVC for 4K UHD sequences. The test results clearly exhibit a substantial improvement in compression performance for HEVC. In most cases, a significant difference was observed between HEVC and AVC for a similar bit rate. For the natural contents considered in this study, a bit rate reduction ranging from 53% to 59% can be achieved based on subjective results, whereas the predicted reduction based on PSNR values was only between 27% and 38%. This difference is mostly due to the fact that PSNR doesn't take into account the saturation effect of the HVS. PSNR also doesn't capture

the full nature of the artifacts: AVC compressed sequences exhibit blockiness whereas HEVC compression tends to smooth out the content, which is less annoying. For the synthetic content considered in this study, a 73% bit rate reduction can be achieved based on subjective results, whereas the predicted reduction based on PSNR values was 68%.

We have evaluated the performance of HEVC intra coding for still image compression and benchmarked it again JPEG and JPEG 2000 for high resolution images. Subjective results showed that all compression standards performed equally at the highest bit rates, whereas HEVC performed usually better, or at least equal, when compared to all other standards at bit rates lower than 1.00 bpp. Based on PSNR measurement, the average bit rate reduction for HEVC when compared to JPEG, JPEG 2000 4:2:0, and JPEG 2000 4:4:4 is 54%, 19%, and 44%, respectively. Based on the subjective evaluation results, the estimated bit rate saving for HEVC relative to JPEG, JPEG 2000 4:2:0, and JPEG 2000 4:4:4 is about 44%, 31%, and 17%, respectively. The ranking difference between the two chroma sampling formats for JPEG 2000 is due to visual weighting enabled in JPEG 2000 4:4:4, which is not captured by PSNR based metric. These results show the importance of subjective tests to determine a more realistic estimation of the achievable bit rate reduction.

We have evaluated the performance of VP9 against AVC and HEVC considering a real-time Internet-based streaming scenario. The test results showed that HEVC offers improvements in compression performance when compared to VP9 and AVC, if one considers a wide range of bit rates form low to high, corresponding to video with low to transparent quality. VP9 achieved better visual quality than AVC, except for two contents, where CIs overlap significantly. However, in some cases (in particular, at high bit rates), HEVC and VP9 had similar ratings and there is no sufficient statistical evidence indicating differences in performance between these codecs at these bit rates. Objective based measurements showed that HEVC achieves average bit rate savings of 57.3% versus AVC and 33.6% versus VP9, whereas VP9 achieves a bit rate reduction of 40.4% over AVC. Based on perceived quality, the average bit rate reduction of HEVC reaches 59.5% versus AVC and 42.4% versus VP9, whereas VP9 achieves a bit rate reduction of 33.3% over AVC.

We have evaluated the performance of JPEG XT profiles A, B, and C for HDR image compression. In most cases, there was not sufficient statistical evidence to indicate differences in performance between profiles. Some variations were observed at the lowest bit rates, but all three profiles reach transparent quality at the highest bit rates. Overall, we observed that Profile A exhibits a lot of block coding artifacts in flat areas, similar to JPEG, but usually preserves colors, except at very low bit rates. Profile B suffered from color bleeding on areas of uniform colors, but exhibited less block coding artifacts when compared to Profile A. In addition, Profile C performed better on flat uniform areas, but exhibited a checkerboard style color pattern on non-flat areas and introduced color noise near edges at low and medium bit rates, depending on content.

We have evaluated the performance of potential HDR coding technologies against HEVC Main

HDR profile. Subjective results showed that some coding technologies can do better than the HEVC Anchor in a statistically significant way. Regarding the selection of contents, bright scenes are better to perceive color artifacts, especially in whitish parts, and loss of details and high frequencies, especially in textured areas. On the other hand, sequences with a wide dynamic range and strong luminance temporal changes, although good for demonstrating HDR, may not be necessarily best to assess HDR compression performance. Dark scenes are important too, as HDR is not only about high brightness, but it might be hard to see the improvements in these sequences, especially if the previous test sequence was bright, due to the adaptation time of the human eye.

We have evaluated the performance of 3D-AVC over MVC+D for MVD compression in three laboratories across Europe. Analyses showed that laboratories employing different displays and different subjects could still produce highly correlated results, if the test plan is well designed and the tests are conducted following the same guidelines. The subjective results of the three participating laboratories showed high correlation, even though the laboratories used different subjects and different 3D displays having different sizes. Finally, the subjective results aggregated from the three laboratories showed that 3D-AVC achieves bit rate savings ranging from 6% up to 21%, with an average of 14%, when compared to MVC+D.

We have proposed an experimental protocol to evaluate the impact of depth compression on perceived image quality in a FTV scenario. A specific use case was considered to allow a reliable comprehension of the impact of depth coding: a smooth camera motion during a time freeze. The analyses of the resulting subjective scores revealed that the proposed experimental protocol allows the evaluation of different compression and view synthesis algorithms. The use of statistical tools (ANOVA and PCA) to analyze the subjective scores showed particular behaviors such as the influence of different view synthesis modes on the performance of a specific compression algorithm. These results show the originality and the effectiveness of the proposed assessment protocol as well as the importance of subjective quality assessment. This methodology can be considered to evaluate the performances of various depth compression algorithms and can be extended to the assessment of MVD compression schemes and view synthesis algorithms.

We have proposed two possible approaches to crowd-based quality assessment of MVD content on 2D displays: by using a virtual view and by using a FVV, which corresponds to a smooth camera motion during a time freeze. To demonstrate the feasibility of the proposed approaches, a reference ground truth was obtained via a subjective evaluation of stereo pairs on a stereoscopic monitor in a laboratory environment and the two proposed 2D representations were generated and evaluated in a crowdsourcing environment. The crowdsourcing results showed high correlation with the ground truth results. No statistically significant differences between the two approaches were found. In our experiments, 2D impairments were mostly visible in the test material, even though depth maps were also compressed, and the strength of the spatial impairments was similar across time. Therefore, it is reasonable to have high correlation with ground truth results in both approaches. However, if the test material

mostly contains depth impairments, the FVV approach is expected to be more suitable.

We have proposed one possible approach to crowd-based quality assessment of HDR content on LDR displays: by using LDR versions of original HDR content obtained with TMOs. To demonstrate the feasibility of the proposed approach, a reference ground truth was obtained via subjective quality evaluation of HDR images on a HDR monitor in a laboratory environment and LDR versions were generated for each HDR image using different TMOs and evaluated in a crowdsourcing experiment. A second crowdsourcing evaluation was conducted using semantic differentiators to better understand the characteristics of the different TMOs. The crowdsourcing evaluations showed that some TMOs are more suitable for evaluation of HDR image compression.

We have investigated immersive video presentation experience via explicit subjective rating analysis for 2D and 3D multimedia contents of various quality levels. The subjective ratings analysis revealed that the effects of the rendering mode, actual quality level, and content on perceived depth and on sensation of reality were significant. It also revealed that there is a strong correlation between perceived depth and sensation of reality, as well as between sensation of reality and perceived overall quality. Finally, for a given quality level perceived depth and sensation of reality were both higher for 3D when compared to 2D stimuli. Similarly, high quality sequences generally obtained higher ratings for perceived depth quantity, sensation of reality, and perceived overall quality when compared to their corresponding low quality versions. However, the difference in terms of perceived depth and sensation of reality between 3D low quality stimuli and 2D high quality stimuli was not significant.

We have measured the added value of higher dynamic range to user preference for peak luminance levels of 100, 400, 1000, and 4000 cd/m^2. The analysis of the subjective results demonstrates that the increase in maximum luminance level at which higher dynamic range video is displayed is preferred by average viewers, with a steady increase in preference as the maximum luminance increases. The results showed a significant increase in the perceptual experience when viewing HDR content at 4000 cd/m^2 peak luminance compared to the current standards in TV and cinema. When considering pairs with 4000 cd/m^2 only, the quality score values exhibit a convex shape, whereas they exhibit a concave shape when considering pairs with 100 cd/m^2 only, with a maximum at 1000 cd/m^2. The analysis of the results for different test methods demonstrates that a full paired comparison provides more detailed information about viewing preferences. Hence, this methodology is preferred when there are a reasonable number of pairs. In cases when the number of stimuli is too large for full paired comparison, a limited subset of pairs can be considered instead.

We have measured the impact of UHD on visual attention by conducting eye tracking experiments with both HD and UHD resolution images covering a wide variety of scenes. The analysis of the FDMs of HD and UHD images demonstrated that (i) UHD resolution images can grab the focus of attention more than HD images, (ii) humans tend to look at a few attentive regions in the images with more intent when viewing UHD, and (iii) viewing strategy is

different for HD and UHD.

We have measured the impact of HDR on visual attention by conducting eye tracking experiments with both HDR image generated from multiple exposure pictures and a single exposure LDR image of the same scene. The eye tracking test demonstrated that FDMs of HDR images for some scenes are significantly different from the FDMs of the corresponding LDR versions. Three clusters of HDR images were then identified: (i) with FDMs having different visual attention pattern compared to FDMs of LDR versions, (ii) with FDMs showing different distribution of fixation intensities when compared to FDMs of LDR versions, and (iii) with FDMs that are similar to FDMs of LDR images. The similarity metric demonstrated that these clusters are dissimilar in statistically significant way. However, the similarity scores for clusters (i) and (ii) are not as small compared to cluster (iii) as it was expected, which means the metric did not capture the difference between FDMs adequately. Therefore, the impact of HDR on human visual attention is scene-dependent and it is hard to measure it using existing statistical evaluation metrics.

13.1.2 Predicting Quality of Experience

We have evaluated the performance of state-of-the-art 2D metrics for quality assessment of stereo pairs formed from decoded and synthesized views. Results showed that, in general, the measured quality of the decoded view has the highest correlation in terms of the Pearson correlation coefficient with perceived quality. Similar performance can be achieved when considering the average quality of both views. However, if the objective quality assessment is based on the measured quality of the synthesized view, it is suggested to use VIF, VQM, MS-SSIM, or SSIM since they significantly outperform other objective metrics, including PSNR. These four objective metrics have similar performance when using the decoded view, the synthesized view, and both views.

We have evaluated the performance of state-of-the-art 2D metrics for quality assessment of stereo pairs formed from two synthesized views Results showed that PSNR, PSNR-HVS, PSNR-HVS-M, and WSNR have a significantly lower correlation with perceived quality than VIF, VQM, SSIM, and MS-SSIM. From these observations and those of the study on stereo pair formed from a decoded view and a synthesized view, we concluded that some objective metrics do not predict well perceived quality of synthesized views and that there is no significant masking effect between a decoded view and a synthesized view.

We have evaluated the performance of state-of-the-art 2D metrics for quality assessment of FVV sequences corresponding to a smooth camera motion during a time freeze. The results showed that objective metrics achieved low correlation with subjective scores when various conditions were considered. However, the correlation with perceived quality improved when content characteristics were considered. In addition, the artifacts produced by some view synthesis algorithms might not be correctly handled by the objective quality metrics. These results motivate the need to design better objective metrics that can accurately assess the

specific artifacts generated by the view synthesis process.

We have evaluated the performance of state-of-the-art HDR and LDR metrics for HDR image quality assessment. Results showed that the performance of most FR HDR metrics could be improved by considering perceptual transforms when compared to linear values. On the other hand, our findings suggested that a lot of work remains to be done for NR quality assessment of HDR content. Our benchmark demonstrated that HDR-VDP-2 and HDR-VQM are ultimately the most reliable predictors of perceived quality. Nevertheless, HDR-VDP-2 is complex and requires heavy computational resources, which limits its use in many applications. HDR-VQM is over three times faster, which makes it a suitable alternative to HDR-VDP-2. Alternatively, MS-SSIM computed in the PU space is another lower complexity substitute, as there is no statistical evidence to show performance differences between these metrics in terms of PCC and SROCC.

We have evaluated the effectiveness of state-of-the-art HDR and LDR metrics to discriminate between quality levels when comparing two HDR video sequences. Results showed that PSNR-DE1000, HDR-VDP-2, and PSNR-Lx can reliably detect visible differences between two HDR video sequences, whereas HDR-VQM and mPSNR could not distinguish quality differences.

We have proposed a model to predict perceived quality of stereoscopic video sequences based on content analysis. A logistic function is used to map the PSNR values to perceived quality. The parameters of the mapping function are predicted using 2D and 3D content features. The model was trained and evaluated on a dataset of stereoscopic video sequences with associated ground truth MOS. Results showed that the proposed model achieved high correlation with perceived quality and was quite robust across contents when the training set contained various contents. This finding indicates that perceived quality can be predicted from PSNR values based on content analysis and that subjective tests might not be always required.

13.1.3 Improving Quality of Experience

We have proposed and evaluated two different approaches that exploit visual attention to improve 3D QoE on stereoscopic displays: an offline system, which uses a computational model of visual attention to predict gaze position, and an online system, which uses a remote eye tracking system to measure real time gaze positions. From the saliency map, which was computed using a 3D visual attention model, the region-of-interest and its disparity were extracted. From the eye tracking measurements, filtered gaze points were used in conjunction with the disparity map to extract the disparity of the object-of-interest. Horizontal image translation was performed to bring the fixated object on the screen plane. The shift was determined based on the extracted disparity values and filtered in time to have smooth transitions that do no create visual discomfort. Subjective evaluation results showed that exploiting visual attention significantly improves image quality and visual comfort, with a slight advantage for real time gaze determination. Depth quality is also improved, but the difference is not significant.

We have proposed and evaluated an active crosstalk reduction system for mobile autostereoscopic displays. The proposed system was implemented on a HTC EVO 3D smartphone. To determine the crosstalk level at each position, a full display characterization was performed. Furthermore, the localization of sweet spot and computation of the viewing freedom was performed. A special Android application was implemented to track the user face and eyes, and to correct artifacts in real-time according to his/her position. The proposed system was designed in the way that it first helps the user to find the sweet spot and then compensates for crosstalk artifacts and/or pseudoscopy. Subjective evaluation results showed that in terms of depth perception, the proposed system clearly outperforms the 3D and 2D modes. In terms of image quality, 2D mode was found to be best, but the proposed system outperforms 3D mode. The evaluation of the Android application showed that there are also limitations in terms of processing speed and power usage.

We have proposed and evaluated an active crosstalk reduction system for multiview autostereoscopic displays. The proposed system was implemented considering a 52" full HD 28-view Dimenco BDL5231V autostereoscopic display with slanted lenticular sheet. The display was characterized in terms of luminance distribution and the luminance profiles were modeled using a limited set of parameters. A Kinect sensor was used to determine the viewer position in front of the display. The proposed system performs an intelligent on the fly allocation of the output views to minimize the perceived crosstalk. Subjective evaluation results showed that picture quality is significantly improved when compared to the standard 3D mode, for a similar depth perception and visual comfort.

13.2 Contributions to Reproducible Research

The availability of public databases is essential for the scientific community. For example, databases of distorted images and video sequences with corresponding ground truth subjective quality scores are essentials to benchmark objective metrics, develop new objective metrics, as well as for cross-lab evaluations to investigate the influence of different parameters, e.g., test method, viewing conditions, monitor, or subjects, on perceived quality. Using the data from the subjective quality evaluations conducted in this thesis, we have released two public databases of distorted images/video sequences with corresponding ground truth subjective quality scores

1) JPEGXTHDR: Subjective quality assessment database of HDR images compressed with JPEG XT (Korshunov et al., 2015). The database is composed of 20 HDR image contents encoded with JPEG XT profiles A, B, and C at 4 bit rates, leading to a total of 240 compressed HDR images. For each image, 24 raw subjective scores, as well as the corresponding MOS and CI, are provided. This database was used in (Artusi et al., 2015).
2) FVVDB: Free-Viewpoint synthesized videos quality database (Bosc et al., 2013). The database is composed of 6 MVD contents, with depth maps compressed by 7 algorithms at 3 bit rates and processed by 2 more algorithms, and rendered using 2 different view

synthesis configurations, leading to a total of 276 processed FVVs. For each video, 27 raw subjective scores, as well as the corresponding DMOS and CI, are provided. This database was used in (Hanhart et al., 2014e).

Additionally, databases of images and video sequences with corresponding ground truth eye movement data are essentials to benchmark computational models of visual attention, develop new computational models of visual attention, understand viewing strategies and visual attention patterns, as well as for cross-lab evaluations to investigate the influence of different parameters, e.g., task, viewing conditions, monitor, or subjects, on visual attention. Using the data from the eye tracking experiments conducted in this thesis, we have released three public databases of images/video sequences with corresponding ground truth eye movement data

1) Ultra-Eye: UHD and HD images eye tracking dataset (Nemoto et al., 2014b). The database is composed of 45 image contents in both 4K UHD and HD resolutions, leading to a total of 90 images. For each image, the raw fixation points recorded from 20 subjects, as well as the computed FDM, are provided. This database was used in (Nemoto et al., 2014a).
2) HDR-Eye: dataset of HDR images with eye tracking data (Nemoto et al., 2015). The database is composed of 46 image contents in both HDR and LDR formats, leading to a total of 92 images. For each image, the raw fixation points recorded from 20 subjects, as well as the computed FDM, are provided. This database was used in (Nemoto et al., 2015).
3) EyeC3D: 3D video eye tracking dataset (Hanhart and Ebrahimi, 2014c). The database is composed of 8 stereoscopic video sequences. For each video sequence, the raw fixation points recorded from 21 subjects, as well as the computed FDM, are provided. This database was used in (Ferreira et al., 2015a; Ferreira et al., 2015b).

Finally, we have released one public database of video contents with explicit subjective ratings and viewers' brain and peripheral physiological signals

1) MIMESIS: Modeling Immersive Media Experiences by Sensing Impact on Subjects (Kroupi et al., 2015). The database is composed of 7 video contents presented in 2D and 3D modes, with low and high quality levels, leading to a total of 28 video stimuli. Because of copyright reasons, the video sequences are not provided, but the raw subjective scores, as well as the corresponding MOS and CI, are provided for perceived quality, depth perception, content preference, and sensation of reality, as well as the EEG and peripheral signals (ECG and respiration) recorded from 16 subjects. This database was used in (Kroupi et al., 2014a; Kroupi et al., 2014b; Kroupi et al., 2014c; Kroupi et al., 2015).

When a model is proposed in a scientific publication, it is best if the authors provide a public reference implementation of their model to ensure that the implemented model will perform as designed. We have released public implementations of the two extensions of the Bjøntegaard model that we have proposed

1) BD2D: Rate-Distortion Evaluation For Two-Layer Coding Systems (Hanhart and Ebrahimi, 2015).
2) SCENIC: Subjective Comparison of ENcoders based on fItted Curves (Hanhart and Ebrahimi, 2014a). This model was used in (Azimi et al., 2015; Hanhart et al., 2014c; Himawan et al., 2015; Rerabek and Ebrahimi, 2014; Rerabek et al., 2015b; Song et al., 2015; Tabatabai et al., 2014; Tan et al., 2016).

13.3 Outlook for Future Research

In Chapter 3, we have proposed two extensions of the Bjøntegaard model to calculate the coding efficiency between different codecs. The Bjøntegaard model can be further improved by incorporating some of the concepts used in the proposed models, such as the computation of the integration bounds to consider the saturation effect of the HVS or the reliability index. Alternatively, other extensions of the Bjøntegaard model can be designed for other objective metrics than PSNR, such as SSIM or HDR-VDP-2 for HDR, which are more correlated with human perception of visual quality. Finally, the model proposed for two-layer coding systems can be generalized for N-layers coding systems.

In Chapter 4, we have evaluated the performance of different image and video compression algorithms. These evaluations can always be extended to investigate the impact of the different coding tools and parameters available in the algorithms, to compare different (optimized) implementations, to consider more contents, more compression algorithms, and more rate points, or to investigate the impact of the viewing and test conditions. In particular, for JPEG XT, the interaction between the base and extension layer bit rates on the HDR image quality could be investigated to determine optimal allocation strategies, which are most likely profile dependent. Similar investigations could be conducted for MVC+D and 3D-AVC to investigate the optimal bit rate allocation between the texture views and depth maps, as well as between the different views. Obviously, new compression algorithms are always under development, e.g., VP10 and H.266, and will need to be assessed as well.

In Chapter 5, we have investigated alternative evaluation protocols. The protocol proposed for evaluation of FVV in a FTV scenario can be applied for evaluating other quality factors brought by 3D. This protocol can be also extended to stereoscopic viewing conditions through the assessment of stereoscopic FVVs. The alternatives protocols proposed for crowdsourcing evaluation of 3D video quality should be further tested with more contents and different distortion types to validate their application in different scenarios than the one investigated. Furthermore, a real crowdsourcing experiment should be conducted instead of the crowd-based evaluation conducted. The alternative protocols proposed for crowdsourcing evaluation of HDR quality assessment should also be further tested with more contents and different distortion types, as well as for HDR video sequences, to validate its application in different scenarios than the one investigated.

In Chapter 6, we have investigated the impact of 3D and HDR and viewers' QoE. These evalu-

ations can be extended in many aspects. For example, we have investigated the differences between video sequences presented in either 2D or 3D, but the amount of depth in the 3D stimuli was fixed. A possible extension would be to have different amounts of depth for each content. Alternatively, a similar study can be conducted on autostereoscopic displays. We have investigated the difference in viewers preference between different peak luminance levels, but the other parameters were fixed. This study can be extended to consider different black levels, different numbers of active zones in the LED backlight, different native contrast ratios of the front LCD panel, or different ambient lighting conditions.

In Chapter 7, we have investigated the impact of UHD and HDR on visual attention. Our study on UHD can be extended to video, as well as to consider more resolutions, including 8K UHD. Regarding our work on HDR visual attention, future work is needed to find an automated way to classify scenes for better understanding of the influence of HDR on visual attention. Different metrics of visual attention need to be investigated to identify the metric that captures the differences in visual attention patterns caused by HDR. The study can also be extended to inverse TMOs, as there will be a need to display legacy HDR content on future HDR monitors. Finally, the impact of HDR imaging on computational models of visual saliency could also be considered.

In Chapter 8, we reviewed some of the most common or state-of-the-art objective quality metrics. Of course, a lot of research can be done on developing new objective quality metrics that better predict perceived visual quality. Also, there is a lack of RR and NR metrics in some applications, e.g., HDR.

In Chapter 10, we have evaluated the performance of several objective quality metrics in different scenarios. These kind of benchmarkings can always be extended to consider larger databases, with more contents and more degradations (both types and strengths), or different databases obtained in different conditions, e.g., viewing conditions or test methods. As the number of objective metrics developed over the last 40 years is quite large, it is impossible to consider all existing metrics. Thus, the benchmarkings can always be extended to compare more objective metrics. In particular, regarding the quality assessment of stereo pairs, we have only benchmarked the performance of 2D metrics, but 3D metrics should be evaluated as well. A similar remark applies for the quality assessment of FVVs. Finally, further analysis tools can always be applied to better understand the limitations of objective metrics and to define their scope of validity or resolving power.

In Chapter 11, we have investigated the prediction of 3D quality based on content analysis. To extend this work, different metrics could be considered instead of PSNR. Additional content features could be used to better predict the parameters of the logistic function. Also, a larger dataset with more contents should be used to further evaluate the performance of the proposed model. Finally, this model could be applied to other scenarios, such as HDR or FTV.

In Chapter 12, we have proposed and evaluated different systems to improve 3D quality of experience. The systems proposed for stereoscopic displays can be further tuned to pro-

vide even better QoE, for example by using more advanced computational models of visual attention, as well as better filtering of the gaze points. Additionally, progressive blur could be added to the stereo pair based on the depth map to better mimic the depth of field of the human eye. Our system for mobile autostereoscopic displays can also be extended in many directions to achieve better QoE. An automatic IPD computation, eye tracking improvement, and transparency of the feedback are possible improvements. Implementation of more complex crosstalk compensation algorithms while keeping low computational complexity is another challenge. Regarding the active crosstalk reduction system we have proposed for multiview autostereoscopic displays, the system should be assessed by naïve subjects. Further improvements include better assignment of the views, especially near the decision boundary, better fading, and better filtering of the user position. The measurements and luminance model can be extended for different viewing distances to allow the user to move back and forth. Finally, the system could be extended to handle several viewers, located at different positions.

In this thesis, we did not consider HFR. Thus, most of the investigations performed on other immersive technologies could also be conducted for HFR. Additionally, WCG was little considered in this thesis and further evaluations can be made to determine its impact on viewers' QoE and visual attention. One of the ultimate goals is also to understand the added value of each of these immersive technologies and their interactions.

Finally, there are emerging imaging technologies, such as plenoptic, light-field, and 360 video, and new applications, such as virtual and augmented reality, that are gaining huge interest in the scientific, entertainment, and arts communities. These new technologies will most likely revolutionize the way we will interact with multimedia content in the future. These new technologies create many challenges and opportunities in different fields of research, such as compression, subjective and objective quality assessment, and visual attention modeling.

A Maximum Likelihood for Two Options

If only two options, 'A' and 'B', are compared, the log-likelihood function (see Equation (2.25)) is

$$\begin{aligned}\mathscr{L}\left(\Delta\mu|C,\mu\right) = {} & C_{AB}^{-}\log\left\{\Phi\left[\left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right)\right]\right\}\\ & + C_{BA}^{-}\log\left\{\Phi\left[\left(\mu_B-\Delta\mu_B^{-}\right)-\left(\mu_A+\Delta\mu_A^{+}\right)\right]\right\}\\ & + C_{AB}^{+}\log\left\{\Phi\left[\left(\mu_A+\Delta\mu_A^{+}\right)-\left(\mu_B-\Delta\mu_B^{-}\right)\right]\right\}\\ & + C_{BA}^{+}\log\left\{\Phi\left[\left(\mu_B+\Delta\mu_B^{+}\right)-\left(\mu_A-\Delta\mu_A^{-}\right)\right]\right\}\end{aligned}$$

To find the maximum likelihood solution quality scale values, one must solve

$$\arg\max_{\Delta\mu}\mathscr{L}\left(\Delta\mu|C,\mu\right) \quad \text{subject to} \quad \Delta\mu \geq 0$$

Let us recall that

$$\Phi(-x) = 1-\Phi(x) \qquad \frac{\mathrm{d}\Phi(x)}{\mathrm{d}x} = \phi(x) \qquad \phi(-x) = \phi(x)$$

where ϕ is the standard normal probability density functions.

The maximum likelihood problem can be solved by setting the partial derivatives of the objective to zero,

$$\begin{aligned}0 = \frac{\partial\mathscr{L}}{\partial\mu_A^{-}} &= -C_{AB}^{-}\frac{\phi\left[\left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right)\right]}{\Phi\left[\left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right)\right]} + C_{BA}^{+}\frac{\phi\left[\left(\mu_B+\Delta\mu_B^{+}\right)-\left(\mu_A-\Delta\mu_A^{-}\right)\right]}{\Phi\left[\left(\mu_B+\Delta\mu_B^{+}\right)-\left(\mu_A-\Delta\mu_A^{-}\right)\right]}\\ &= -C_{AB}^{-}\frac{\phi\left[\left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right)\right]}{\Phi\left[\left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right)\right]} + C_{BA}^{+}\frac{\phi\left[\left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right)\right]}{1-\Phi\left[\left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right)\right]}\end{aligned}$$

$$\Rightarrow \quad \left(\mu_A-\Delta\mu_A^{-}\right)-\left(\mu_B+\Delta\mu_B^{+}\right) = \Phi^{-1}\left(\frac{C_{AB}^{-}}{C_{AB}^{-}+C_{BA}^{+}}\right)$$

$$
\begin{aligned}
0 = \frac{\partial \mathscr{L}}{\partial \mu_A^+} &= -C_{BA}^- \frac{\phi\left[\left(\mu_B - \Delta\mu_B^-\right) - \left(\mu_A + \Delta\mu_A^+\right)\right]}{\Phi\left[\left(\mu_B - \Delta\mu_B^-\right) - \left(\mu_A + \Delta\mu_A^+\right)\right]} + C_{AB}^+ \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{\Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]} \\
&= -C_{BA}^- \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{1 - \Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]} + C_{AB}^+ \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{\Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}
\end{aligned}
$$

$$
\Rightarrow \quad \left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right) = \Phi^{-1}\left(\frac{C_{AB}^+}{C_{AB}^+ + C_{BA}^-}\right)
$$

$$
\begin{aligned}
0 = \frac{\partial \mathscr{L}}{\partial \mu_B^-} &= -C_{BA}^- \frac{\phi\left[\left(\mu_B - \Delta\mu_B^-\right) - \left(\mu_A + \Delta\mu_A^+\right)\right]}{\Phi\left[\left(\mu_B - \Delta\mu_B^-\right) - \left(\mu_A + \Delta\mu_A^+\right)\right]} + C_{AB}^+ \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{\Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]} \\
&= -C_{BA}^- \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{1 - \Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]} + C_{AB}^+ \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{\Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}
\end{aligned}
$$

$$
\Rightarrow \quad \left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right) = \Phi^{-1}\left(\frac{C_{AB}^+}{C_{AB}^+ + C_{BA}^-}\right)
$$

$$
\begin{aligned}
0 = \frac{\partial \mathscr{L}}{\partial \mu_B^+} &= -C_{AB}^- \frac{\phi\left[\left(\mu_A - \Delta\mu_A^-\right) - \left(\mu_B + \Delta\mu_B^+\right)\right]}{\Phi\left[\left(\mu_A - \Delta\mu_A^-\right) - \left(\mu_B + \Delta\mu_B^+\right)\right]} + C_{BA}^+ \frac{\phi\left[\left(\mu_B + \Delta\mu_B^+\right) - \left(\mu_A - \Delta\mu_A^-\right)\right]}{\Phi\left[\left(\mu_B + \Delta\mu_B^+\right) - \left(\mu_A - \Delta\mu_A^-\right)\right]} \\
&= -C_{AB}^- \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{\Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]} + C_{BA}^+ \frac{\phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}{1 - \Phi\left[\left(\mu_A + \Delta\mu_A^+\right) - \left(\mu_B - \Delta\mu_B^-\right)\right]}
\end{aligned}
$$

$$
\Rightarrow \quad \left(\mu_A - \Delta\mu_A^-\right) - \left(\mu_B + \Delta\mu_B^+\right) = \Phi^{-1}\left(\frac{C_{AB}^-}{C_{AB}^- + C_{BA}^+}\right)
$$

which verify that the modified definitions of Thurstone's Law for the lower and upper counts (see Equation (2.24)) yield the maximum likelihood solution if there are only two options.

Bibliography

[3D-HTM] K. Suehring and K. Sharman, 3D-HTM: 3D-HEVC extension test model, `http://hevc.hhi.fraunhofer.de` (visited on 31/3/2016).

[Abileah, 2011] A. Abileah, "3-D displays - Technologies and testing methods," *Journal of the Society for Information Display*, vol. 19, no. 11, pp. 749–763

[Adams, 1980] A. Adams, *The camera*, The Ansel Adams Photography Series, Little, Brown and Company

[Adams, 1981] A. Adams, *The negative*, The Ansel Adams Photography Series, Little, Brown and Company

[Adams, 1983] A. Adams, *The print*, The Ansel Adams Photography Series, Little, Brown and Company

[Ahmed et al., 1974] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. 100, no. 1, pp. 90–93

[Aja-Fernandéz et al., 2006] S. Aja-Fernandéz, R. Estepar, C. Alberola-Lopez, and C.-F. Westin, "Image Quality Assessment based on Local Variance," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*

[Akyüz and Reinhard, 2006] A. O. Akyüz and E. Reinhard, "Color appearance in high-dynamic-range imaging," *SPIE Journal of Electronic Imaging*, vol. 15, no. 3

[Akyüz et al., 2007] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do HDR Displays Support LDR Content?: A Psychophysical Evaluation," *ACM Transactions on Graphics*, vol. 26, no. 3

[Alpert et al., 1997] T. Alpert, V. Baroncini, D. Choi, L. Contin, R. Koenen, F. Pereira, and H. Peterson, "Subjective evaluation of MPEG-4 video codec proposals: Methodological approach and test procedures," *Signal Processing: Image Communication*, vol. 9, no. 4, pp. 305–325

[Andrén et al., 2012] B. Andrén, K. Wang, and K. Brunnström, "Characterizations of 3D TV: Active vs Passive," *SID Symposium Digest of Technical Papers*, vol. 43, no. 1, pp. 137–140

[Annighöfer et al., 2010] B. Annighöfer, T. Tajbakhsh, and R.-R. Grigat, "Prediction of results from subjective evaluation of real-time-capable tone-mapping operators applied to limited high-dynamic-range images," *Journal of Electronic Imaging*, vol. 19, no. 1

[ANSI-T1.801.03, 2003] ANSI-T1.801.03, Digital Transport of one-Way Video Signals - Parameters Forobjective Performance ASSE, American National Standards Institute

[Armstrong et al., 2009] M. Armstrong, D. Flynn, M. Hammond, S. Jolly, and R. Salmon, "High Frame-Rate Television," *SMPTE Motion Imaging Journal*, vol. 118, no. 7, pp. 54–59

[Artusi et al., 2015] A. Artusi, R. K. Mantiuk, T. Richter, P. Hanhart, P. Korshunov, M. Agostinelli, A. Ten, and T. Ebrahimi, "Overview and Evaluation of the JPEG XT HDR Image Compression Standard," *Real Time Image Processing Journal*

[Ashikhmin, 2002] M. Ashikhmin, "A Tone Mapping Algorithm for High Contrast Images," *Eurographics Workshop on Rendering*

[Aydın et al., 2008] T. O. Aydın, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," *Proceedings of SPIE 6806*, Human Vision and Electronic Imaging XIII

[Aydin et al., 2008] T. O. Aydin, K. Mantiuk Rafałand Myszkowski, and H.-P. Seidel, "Dynamic Range Independent Image Quality Assessment," *ACM Transactions on Graphics*, vol. 27, no. 3

[Aydin et al., 2014] T. O. Aydin, N. Stefanoski, S. Croci, M. Gross, and A. Smolic, "Temporally Coherent Local Tone Mapping of HDR Video," *ACM Transactions on Graphics*, vol. 33, no. 6, 196:1–196:13

[Azimi et al., 2014] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the Performance of Existing Full-Reference Quality Metrics on High Dynamic Range (HDR) Video Content," *IEEE International Conference on Multimedia Signal Processing (MMSP)*

[Azimi et al., 2015] M. Azimi, R. Boitard, B. Oztas, S. Ploumis, H. Tohidypour, M. Pourazad, and P. Nasiopoulos, "Compression efficiency of HDR/LDR content," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Banterle et al., 2009] F. Banterle, K. Debattista, A. Artusi, S. Pattanaik, K. Myszkowski, P. Ledda, M. Bloj, and A. Chalmers, "High dynamic range imaging and LDR expansion for generating HDR content," *Eurographics State-of-the-Art Report (STAR)*

[Banitalebi-Dehkordi et al., 2012] A. Banitalebi-Dehkordi, M. Pourazad, and P. Nasiopoulos, "A human visual system-based 3D video quality metric," *International Conference on 3D Imaging (IC3D)*

[Barkowsky et al., 2009] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal Trajectory Aware Video Quality Measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279

[Baroncini and Quackenbush, 2012] V. Baroncini and S. Quackenbush, "MPEG Video/Audio Quality Evaluation," *The MPEG Representation of Digital Media*, ed. by L. Chiariglione, Springer New York

[Barkowsky et al., 2013] M. Barkowsky et al., "Towards standardized 3DTV QoE assessment: cross-lab study on display technology and viewing environment parameters," *Proceedings of SPIE 8648*, Stereoscopic Displays and Applications XXIV

[Barnard, 1945] G. A. Barnard, "A new test for 2× 2 tables," *Nature*, vol. 156, p. 177

[Barten, 1999] P. G. Barten, *Contrast sensitivity of the human eye and its effects on image quality*, Bellingham, Washington, USA: SPIE Optical Engineering Press

[Batten, 2000] C. F. Batten, "Autofocusing and Astigmatism Correction in the Scanning Electron Microscope," Master's thesis, U.K.: University of Cambridge

[Battisti et al., 2015a] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Processing: Image Communication*, vol. 30, pp. 78 –88

[Battisti et al., 2015b] F. Battisti, M. Carli, A. Stramacci, A. Boev, and A. Gotchev, "A perceptual quality metric for high-definition stereoscopic 3D video," *Proceedings of SPIE 9399*, Image Processing: Algorithms and Systems XIII

[BD2D] P. Hanhart and T. Ebrahimi, Rate-Distortion Evaluation For Two-Layer Coding Systems, `http://mmspg.epfl.ch/2dbd` (visited on 31/3/2016).

[Bech and Zacharov, 2006] S. Bech and N. Zacharov, *Perceptual Audio Evaluation – Theory, Method and Application*, John Wiley & Sons, Ltd

[Benzie et al., 2007] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, and C. von Kopylow, "A Survey of 3DTV Displays: Techniques and Technologies," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1647–1658

[Benoit et al., 2008] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality Assessment of Stereoscopic Images," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 659024

[Bensalma and Larabi, 2013] R. Bensalma and M.-C. Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimensional Systems and Signal Processing*, vol. 24, no. 2, pp. 281–316

[Bhaskaran and Konstantinides, 1997] V. Bhaskaran and K. Konstantinides, *Image and Video Compression Standards: Algorithms and Architectures*, The Springer International Series in Engineering and Computer Science, Springer US

[Bjøntegaard, 2001] G. Bjøntegaard, *Calculation of average PSNR differences between RD-curves*, tech. rep. VCEG-M33, Austin, Texas, USA: ITU-T SG16/Q6

[Bjøntegaard, 2008] G. Bjøntegaard, *Improvements of the BD-PSNR model*, tech. rep. VCEG-AI11, Berlin, Germany: ITU-T SG16/Q6

[Boev et al., 2006] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. Akar, "Towards compound stereo-video quality metric: a specific encoder-based framework," *IEEE Southwest Symposium on Image Analysis and Interpretation*

[Boev et al., 2008] A. Boev, M. Georgiev, A. Gotchev, and K. Egiazarian, "Optimized single-viewer mode of multiview autostereoscopic display," *European Signal Processing Conference (EUSIPCO)*

[Boev et al., 2009a] A. Boev, D. Hollosi, A. Gotchev, and K. Egiazarian, "Classification and simulation of stereoscopic artifacts in mobile 3DTV content," *Proceedings of SPIE 7237*, Stereoscopic Displays and Applications XX

[Boev et al., 2009b] A. Boev, M. Georgiev, A. Gotchev, N. Daskalov, and K. Egiazarian, "Optimized visualization of stereo images on an OMAP platform with integrated parallax barrier auto-stereoscopic display," *European Signal Processing Conference (EUSIPCO)*

[Boher et al., 2009] P. Boher, T. Leroux, T. Bignon, and V. Collomb-Patton, "A new way to characterize autostereoscopic 3D displays using Fourier optics instrument," *Proceedings of SPIE 7237*, Stereoscopic Displays and Applications XX

[Boher et al., 2012] P. Boher, T. Leroux, T. Bignon, and V. Collomb-Patton, "Optical characterization of different types of 3D displays," *Proceedings of SPIE 8280*, Advances in Display Technologies II

[Bosc et al., 2011a] E. Bosc, M. Köppel, R. Pépion, M. Pressigout, L. Morin, P. Ndjiki-Nya, and P. Le Callet, "Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols?" *IEEE International Conference on Image Processing (ICIP)*

[Bosc et al., 2011b] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a New Quality Metric for 3-D Synthesized View Assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332–1343

[Bosc et al., 2012a] E. Bosc, P. Le Callet, L. Morin, and M. Pressigout, "An edge-based structural distortion indicator for the quality assessment of 3D synthesized views," *Picture Coding Symposium (PCS)*

[Bosc et al., 2012b] E. Bosc, R. Pepion, P. Le Callet, M. Pressigout, and L. Morin, "Reliability of 2D quality assessment methods for synthesized views evaluation in stereoscopic viewing conditions," *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*

[Bosc, 2012] E. Bosc, "Compression of Multi-View-plus-Depth (MVD) data: from perceived quality analysis to MVD coding tools designing," Doctoral dissertation, INSA de Rennes

[Bosc et al., 2013] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for Free-viewpoint video sequences synthesized from decompressed depth data," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Boyce et al., 2016] J. Boyce, Y. Ye, J. Chen, and A. Ramasubramonian, "Overview of SHVC: Scalable Extensions of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20–34

[Bradski and Kaehler, 2008] G. Bradski and A. Kaehler, *Learning OpenCV*, O'Reilly Media Inc.

[Bradley and Terry, 1952] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3-4, pp. 324–345

[Brill et al., 2004] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 101 –107

[Brunnström et al., 2013] K. Brunnström, I. V. Ananth, C. Hedberg, K. Wang, B. Andrén, and M. Barkowsky, "Comparison between Different Rating Scales for 3D TV," *SID Symposium Digest of Technical Papers*, vol. 44, no. 1, pp. 509–512

[Čadík et al., 2008] M. Čadík, M. Wimmer, L. Neumann, and A. Artusi, "Evaluation of HDR Tone Mapping Methods Using Essential Perceptual Attributes," *Computers & Graphics*, vol. 32, no. 3, pp. 330–349

[Campisi et al., 2007] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," *European Signal Processing Conference (EUSIPCO)*

[D. Chandler and S. Hemami, 2007] D. Chandler and S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298

[Chamaret et al., 2010] C. Chamaret, S. Godeffroy, P. Lopez, and O. Le Meur, "Adaptive 3D rendering based on region-of-interest," *Proceedings of SPIE 7524*, Stereoscopic Displays and Applications XXI

[Chae et al., 2011] B. Chae, Y.-R. You, S.-C. Kwon, and S.-H. Lee, "Three-dimensional display system using a variable parallax barrier and eye tracking," *Optical Engineering*, vol. 50, no. 8, pp. 087401–087401

[Chapiro et al., 2014] A. Chapiro, O. Diamanti, S. Poulakos, C. O'Sullivan, A. Smolic, and M. Gross, "Perceptual Evaluation of Cardboarding in 3D Content Visualization," *Proceedings of the ACM Symposium on Applied Perception*, SAP '14, ACM

[Chappuis et al., 2014] A. Chappuis, M. Rerabek, P. Hanhart, and T. Ebrahimi, "Subjective evaluation of an active crosstalk reduction system for mobile autostereoscopic displays," *Proceedings of SPIE 9011*, Stereoscopic Displays and Applications XXV

[Y.-S. Chen et al., 2001] Y.-S. Chen, C.-H. Su, J.-H. Chen, C.-S. Chen, P. Hung, and C.-S. Fuh, "Video-based eye tracking for autostereoscopic displays," *Optical Engineering*, vol. 40, no. 12, pp. 2726–2734

[M. Chen et al., 2006] M. Chen, G. Qiu, Z. Chen, and C. Wang, "JPEG Compatible Coding of High Dynamic Range Imagery using Tone Mapping Operators," *Picture Coding Symposium (PCS)*

[Y. Chen et al., 2009] Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1

[Z. Chen et al., 2010] Z. Chen, W. Lin, and K. N. Ngan, "Perceptual video coding: Challenges and approaches," *IEEE International Conference on Multimedia and Expo (ICME)*

[W. Chen et al., 2012] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "Quality of experience model for 3DTV," *Proceedings of SPIE 8288*, Stereoscopic Displays and Applications XXIII

[M.-J. Chen et al., 2013a] M.-J. Chen, C.-C. Su, K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143 –1155

[M.-J. Chen et al., 2013b] M.-J. Chen, L. K. Cormack, and A. C. Bovik, "No-Reference Quality Assessment of Natural Stereopairs," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3379–3391

[Y. Chen et al., 2014] Y. Chen, M. M. Hannuksela, T. Suzuki, and S. Hattori, "Overview of the MVC+D 3D video coding standard," *Journal of Visual Communication and Image Representation*, vol. 25, no. 4, pp. 679–688

[Y. Chen and Vetro, 2014] Y. Chen and A. Vetro, "Next-Generation 3D Formats with Depth Map Support," *IEEE MultiMedia*, vol. 21, no. 2, pp. 90–94

[Chikkerur et al., 2011] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182

[Chiu et al., 1993] K. Chiu, M. Herf, P. Shirley, S. Swamy, C. Wang, K. Zimmerman, et al., "Spatially nonuniform scaling functions for high contrast images," *Graphics Interface*

[Christopoulos et al., 2000] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE Transactions on Consumer Electronics*, vol. 46, no. 4, pp. 1103–1127

[Ciaramello and Reibman, 2011] F. M. Ciaramello and A. R. Reibman, "Supplemental subjective testing to evaluate the performance of image and video quality estimators," *Proceedings of SPIE 7865*, Human Vision and Electronic Imaging XVI

[CIE1986] International Commission on Illumination, Colorimetry, tech. rep. 15

[CIE1995] International Commission on Illumination, Industrial Colour-Difference Evaluation, tech. rep. 116

[Clarke et al., 1984] F. J. J. Clarke, R. McDonald, and B. Rigg, "Modification to the JPC79 Colour-difference Formula," *Journal of the Society of Dyers and Colourists*, vol. 100, no. 4, pp. 128–132

[Conze et al., 2012] P.-H. Conze, P. Robert, and L. Morin, "Objective view synthesis quality assessment," *Proceedings of SPIE 8288*, Stereoscopic Displays and Applications XXIII

[Daly et al., 2011] S. J. Daly, R. T. Held, and D. M. Hoffman, "Perceptual Issues in Stereoscopic Signal Processing," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 347–361

[Daly et al., 2013] S. J. Daly, T. Kunkel, X. Sun, S. Farrell, and P. Crum, "Viewer Preferences for Shadow, Diffuse, Specular, and Emissive Luminance Limits of High Dynamic Range Displays," *SID Symposium Digest of Technical Papers*, vol. 44, no. 1, pp. 563–566

[Daly, 1992] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," *Proceedings of SPIE 1666*, Human Vision, Visual Processing, and Digital Display III

[Damera-Venkata et al., 2000] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650

[Daribo et al., 2008] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Motion Vector Sharing and Bitrate Allocation for 3D Video-plus-depth Coding," *EURASIP Journal on Applied Signal Processing*, vol. 2009, 3:1–3:13

[De Simone et al., 2009a] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[De Simone et al., 2009b] F. De Simone, L. Goldmann, V. Baroncini, and T. Ebrahimi, "Subjective evaluation of JPEG XR image compression," *Proceedings of SPIE 7443*, Applications of Digital Image Processing XXXII

[De Simone et al., 2011] F. De Simone, L. Goldmann, J.-S. Lee, and T. Ebrahimi, "Towards high efficiency video coding: Subjective evaluation of potential coding technologies," *Journal of Visual Communication and Image Representation*, vol. 22, no. 8, pp. 734–748

[De Silva et al., 2013] V. De Silva, H. Arachchi, E. Ekmekcioglu, and A. Kondoz, "Toward an Impairment Metric for Stereoscopic Video: A Full-Reference Video Quality Metric to Assess Compressed Stereoscopic Video," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3392–3404

[De Vries, 1943] H. De Vries, "The quantum character of light and its bearing upon threshold of vision, the differential sensitivity and visual acuity of the eye," *Physica*, vol. 10, no. 7, pp. 553–564

[Debevec and Malik, 2008] P. E. Debevec and J. Malik, "Recovering High Dynamic Range Radiance Maps from Photographs," *ACM SIGGRAPH 2008 Classes*, SIGGRAPH '08, ACM

[Derf's dataset] X. Foundation, Xiph.org Video Test Media [derf's collection], `https://media.xiph.org/video/derf/` (visited on 31/3/2016).

[Devlin, 2002] K. Devlin, *A review of tone reproduction techniques*, tech. rep. CSTR-02-005, Computer Science, University of Bristol

[Do et al., 2009] L. Do, S. Zinger, Y. Morvan, and P. de With, "Quality improving techniques in DIBR for free-viewpoint video," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*

[Dodgson, 2006] N. A. Dodgson, "On the number of viewing zones required for head-tracked autostereoscopic display," *Proceedings of SPIE 6055*, Stereoscopic Displays and Virtual Reality Systems XIII

[Dong et al., 2014] Y. Dong, E. Nasiopoulos, M. T. Pourazad, and P. Nasiopoulos, "High Dynamic Range Video Eye Tracking Dataset," *International Conference on Electronics, Signal processing and Communications (ESPCO)*

[Drago et al., 2003] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive Logarithmic Mapping For Displaying High Contrast Scenes," *Computer Graphics Forum*, vol. 22, no. 3, pp. 419–426

[Dufaux et al., 2009] F. Dufaux, G. J. Sullivan, and T. Ebrahimi, "The JPEG XR image coding standard," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 195–199, 204–204

[Durand and Dorsey, 2002] F. Durand and J. Dorsey, "Fast Bilateral Filtering for the Display of High-dynamic-range Images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 257–266

[Egiazarian et al., 2006] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," *International Workshop on Video Processing and Quality Metrics (VPQM)*

[Eilertsen et al., 2013] G. Eilertsen, R. Wanat, R. K. Mantiuk, and J. Unger, "Evaluation of Tone Mapping Operators for HDR-Video," *Computer Graphics Forum*, 7, Wiley Online Library

[Einhäuser and König, 2003] W. Einhäuser and P. König, "Does luminance-contrast contribute to a saliency map for overt visual attention?" *European Journal of Neuroscience*, vol. 17, no. 5, pp. 1089–1097

[Emoto et al., 2006] M. Emoto, K. Masaoka, M. Sugawara, and Y. Nojiri, "The viewing angle dependency in the presence of wide field image viewing and its relationship to the evaluation indices," *Displays*, vol. 27, no. 2, pp. 80 –89

[Emoto et al., 2014] M. Emoto, Y. Kusakabe, and M. Sugawara, "High-Frame-Rate Motion Picture Quality and Its Independence of Viewing Distance," *Journal of Display Technology*, vol. 10, no. 8, pp. 635–641

[EMPA HDR Image Database] P. Zolliker, Z. Baranczuk, D. Kupper, I. Sprow, and T. Stamm, EMPA HDR Image Database, `http://www.empamedia.ethz.ch/hdrdatabase/` (visited on 31/3/2016).

[Engelke and Zepernick, 2007] U. Engelke and H.-J. Zepernick, "Perceptual-based Quality Metrics for Image and Video Services: A Survey," *Conference on Next Generation Internet Networks (EuroNGI)*

[Engelke et al., 2009] U. Engelke, A. Maeder, and H. Zepernick, "Visual attention modelling for subjective image quality databases," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*

[Engelke et al., 2011] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual Attention in Quality Assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50–59

[Engelke et al., 2013] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H.-J. Zepernick, and A. J. Maeder, "Comparative Study of Fixation Density Maps," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1121–1133

[EPFL 3D Video Database] L. Goldmann, F. De Simone, and T. Ebrahimi, 3D Video Quality Assessment, `http://mmspg.epfl.ch/3dvqa` (visited on 31/3/2016).

[EPFL/PoliMI Video Database] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, EPFL/PoliMI Video Quality Assessment Database, `http://vqa.como.polimi.it` (visited on 31/3/2016).

[Erasmus and Smith, 1982] S. J. Erasmus and K. C. A. Smith, "An automatic focusing and astigmatism correction system for the SEM and CTEM," *Journal of Microscopy*, vol. 127, no. 2, pp. 185–199

[Europeana] Europeana Foundation, Europeana: think culture, `http://www.europeana.eu` (visited on 31/3/2016).

[EyeC3D] P. Hanhart and T. Ebrahimi, EyeC3D: 3D video eye tracking dataset, `http://mmspg.epfl.ch/eyec3d` (visited on 31/3/2016).

[Fairchild] M. D. Fairchild, The HDR Photographic Survey, `http://www.rit-mcsl.org/fairchild/HDR.html` (visited on 31/3/2016).

[Fang et al., 2014] Y. Fang, J. Wang, J. Li, R. Pepion, and P. Le Callet, "An eye tracking database for stereoscopic video," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Fattal et al., 2002] R. Fattal, D. Lischinski, and M. Werman, "Gradient Domain High Dynamic Range Compression," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 249–256

[Fehn, 2004a] C. Fehn, "3D-TV using depth-image-based rendering (DIBR)," *Picture Coding Symposium (PCS)*

[Fehn, 2004b] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proceedings of SPIE 5291*, Stereoscopic Displays and Virtual Reality Systems

[Fehn et al., 2007] C. Fehn, P. Kauff, S. Cho, H. Kwon, N. Hur, and J. Kim, "Asymmetric Coding of Stereoscopic Video for Transmission Over T-DMB," *3DTV Conference*

[Fenimore et al., 2004] C. Fenimore, V. Baroncini, T. Oelbaum, and T. K. Tan, "Subjective testing methodology in MPEG video verification," *Proceedings of SPIE 5558*, Applications of Digital Image Processing XXVII

[Ferwerda, 2001] J. Ferwerda, "Elements of early vision for computer graphics," *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 22–33

[Ferzli et al., 2005] R. Ferzli, L. J. Karam, and J. Caviedes, "A robust image sharpness metric based on kurtosis measurement of wavelet coefficients," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Ferzli and Karam, 2007] R. Ferzli and L. J. Karam, "A no-reference objective sharpness metric using Riemannian tensor," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Ferreira et al., 2015a] L. Ferreira, L. A. da Silva Cruz, and P. Assuncao, "A generic framework for optimal 2D/3D key-frame extraction driven by aggregated saliency maps," *Signal Processing: Image Communication*, vol. 39, Part A, pp. 98 –110

[Ferreira et al., 2015b] L. Ferreira, L. da Silva Cruz, and P. Assuncao, "A method to compute saliency regions in 3D video based on fusion of feature maps," *IEEE International Conference on Multimedia and Expo (ICME)*

[Fiedler et al., 2010] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *Network, IEEE*, vol. 24, no. 2, pp. 36–41

[FVVDB] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, Free-Viewpoint synthesized videos quality database, `http://ivc.univ-nantes.fr/en/databases/Free-Viewpoint_synthesized_videos/` (visited on 31/3/2016).

[Garbas and Thoma, 2011] J.-U. Garbas and H. Thoma, "Temporally coherent luminance-to-luma mapping for high dynamic range video coding with H.264/AVC," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[Gautier et al., 2012] J. Gautier, O. Le Meur, and C. Guillemot, "Efficient depth map compression based on lossless edge coding and diffusion," *Picture Coding Symposium (PCS)*

[Ghanbari, 1989] M. Ghanbari, "Two-layer coding of video signals for VBR networks," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 5, pp. 771–781

[Glickman, 1999] M. E. Glickman, "Parameter Estimation in Large Dynamic Paired Comparison Experiments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 3, pp. 377–394

[Goldmann et al., 2010] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *Proceedings of SPIE 7526*, Three-Dimensional Image Processing (3DIP) and Applications

[Grois et al., 2013] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, "Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders," *Picture Coding Symposium (PCS)*

[Grois et al., 2014] D. Grois, D. Marpe, T. Nguyen, and O. Hadar, "Comparative assessment of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders for low-delay video applications," *Proceedings of SPIE 9217*, Applications of Digital Image Processing XXXVII

[Gulliksen, 1956] H. Gulliksen, "A least squares solution for paired comparisons with incomplete data," *Psychometrika*, vol. 21, no. 2, pp. 125–134

[Haber and Barnhart, 2006] M. Haber and H. Barnhart, "Coefficients of agreement for fixed observers," *Statistical Methods in Medical Research*, vol. 15, no. 3, p. 255

[Hanhart et al., 2012a] P. Hanhart, F. De Simone, and T. Ebrahimi, "Quality Assessment of Asymmetric Stereo Pair Formed From Decoded and Synthesized Views," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Hanhart and Ebrahimi, 2012] P. Hanhart and T. Ebrahimi, "Quality Assessment of a Stereo Pair Formed From Decoded and Synthesized Views Using Objective Metrics," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*

[Hanhart et al., 2012b] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," *Proceedings of SPIE 8499*, Applications of Digital Image Processing XXXV

[Hanhart and Ebrahimi, 2013a] P. Hanhart and T. Ebrahimi, "Quality Assessment of a Stereo Pair Formed From Two Synthesized Views Using Objective Metrics," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Hanhart et al., 2013] P. Hanhart, M. Rerabek, P. Korsunov, and T. Ebrahimi, "Subjective evaluation of HEVC intra coding for still image compression," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Hanhart and Ebrahimi, 2013b] P. Hanhart and T. Ebrahimi, "Predicting 3D quality based on content analysis," *IEEE IVMSP Workshop*

[Hanhart and Ebrahimi, 2013c] P. Hanhart and T. Ebrahimi, "On the evaluation of 3D codecs on multiview autostereoscopic display," *IEEE International Workshop on Hot Topics in 3D (Hot3D)*

[Hanhart and Ebrahimi, 2014a] P. Hanhart and T. Ebrahimi, "Calculation of average coding efficiency based on subjective quality scores," *Journal of Visual Communication and Image Representation*, vol. 25, no. 3, pp. 555–564

[Hanhart et al., 2014a] P. Hanhart, P. Korshunov, T. Ebrahimi, Y. Thomas, and H. Hoffmann, "Subjective Quality Evaluation Of High Dynamic Range Video And Display For Future TV," *International Broadcasting Convention (IBC)*

[Hanhart and Ebrahimi, 2014b] P. Hanhart and T. Ebrahimi, "Subjective evaluation of two stereoscopic imaging systems exploiting visual attention to improve 3D quality of experience," *Proceedings of SPIE 9011*, Stereoscopic Displays and Applications XXV

[Hanhart et al., 2014b] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective Evaluation of Higher Dynamic Range Video," *Proceedings of SPIE 9217*, Applications of Digital Image Processing XXXVII

[Hanhart et al., 2014c] P. Hanhart, N. Ramzan, V. Baroncini, and T. Ebrahimi, "Cross-lab subjective evaluation of the MVC+D and 3D-AVC 3D video coding standards," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Hanhart et al., 2014d] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowdsourcing evaluation of high dynamic range compression," *Proceedings of SPIE 9217*, Applications of Digital Image Processing XXXVII

[Hanhart and Ebrahimi, 2014c] P. Hanhart and T. Ebrahimi, "EyeC3D: 3D video eye tracking dataset," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Hanhart et al., 2014e] P. Hanhart, E. Bosc, P. Le Callet, and T. Ebrahimi, "Free-viewpoint video sequences: A new challenge for objective quality metrics," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*

[Hanhart et al., 2014f] P. Hanhart, M. Bernardo, P. Korshunov, M. Pereira, A. Pinheiro, and T. Ebrahimi, "HDR image compression: A new challenge for objective quality metrics," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Hanhart et al., 2014g] P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowd-based quality assessment of multiview video plus depth coding," *IEEE International Conference on Image Processing (ICIP)*

[Hanhart et al., 2015a] P. Hanhart, P. Korshunov, T. Ebrahimi, Y. Thomas, and H. Hoffmann, "Subjective Quality Evaluation of High Dynamic Range Video and Display for Future TV," *SMPTE Motion Imaging Journal*, vol. 124, no. 4, pp. 1–6

[Hanhart et al., 2015b] P. Hanhart, C. di Nolfo, and T. Ebrahimi, "Active crosstalk reduction system for multiview autostereoscopic displays," *IEEE International Conference on Multimedia and Expo (ICME)*

[Hanhart et al., 2015c] P. Hanhart, M. Rerabek, and T. Ebrahimi, "Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies," *Proceedings of SPIE 9599*, Applications of Digital Image Processing XXXVIII

[Hanhart and Ebrahimi, 2015] P. Hanhart and T. Ebrahimi, "Rate-distortion evaluation for two-layer coding systems," *IEEE International Conference on Image Processing (ICIP)*

[Hanhart et al., 2015d] P. Hanhart, M. Bernardo, M. Pereira, A. G. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, p. 39

[Harel et al., 2006] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Proceedings of Neural Information Processing Systems (NIPS)*

[Haskell et al., 1997] B. Haskell, A. Puri, and A. Netravali, *Digital Video: An Introduction to MPEG-2*, Digital multimedia standards series, Springer

[Hayward et al., 2004] V. Hayward, O. R. Astley, M. Cruz-Hernandez, D. Grant, and G. Robles-De-La-Torre, "Haptic interfaces and devices," *Sensor Review*, vol. 24, no. 1, pp. 16–29

[HdM-HDR-2014] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, Hdm-HDR-2014 Project, `https://hdr-2014.hdm-stuttgart.de` (visited on 31/3/2016).

[HDR-Eye] H. Nemoto, P. Korshunov, P. Hanhart, and T. Ebrahimi, HDR-Eye: dataset of high dynamic range images with eye tracking data, `http://mmspg.epfl.ch/hdr-eye` (visited on 31/3/2016).

[HDR Toolbox] F. Banterle, HDR Toolbox for Matlab, `http://www.github.com/banterle/HDR_Toolbox` (visited on 31/3/2016).

[HDR-VDP-2] R. Mantiuk, HDR-VDP-2: High Dynamic Range Visible Difference Predictor, version 2.2.1, `http://hdrvdp.sourceforge.net` (visited on 31/3/2016).

[HDR-VQM] M. Narwaria, HDR-VQM: An objective quality measure for high dynamic range video, version 1, `http://sites.google.com/site/narwariam/hdr-vqm/` (visited on 31/3/2016).

[Hewage et al., 2009] C. Hewage, S. Worrall, S. Dogan, S. Villette, and A. Kondoz, "Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304–318

[Hibbard, 2008] P. B. Hibbard, "Binocular energy responses to natural images," *Vision Research*, vol. 48, no. 12, pp. 1427 –1439

[Himawan et al., 2015] I. Himawan, W. Song, and D. Tjondronegoro, "Impact of automatic region-of-interest coding on perceived quality in mobile video," *Multimedia Tools and Applications*, pp. 1–29

[HM] F. Bossen, D. Flynn, K. Sharman, and K. Sühring, HM: H.265/HEVC reference software, `http://hevc.hhi.fraunhofer.de/` (visited on 31/3/2016).

[Hoffman et al., 2008] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, no. 3

[Hong, 2012] H. Hong, "Simple method of characterizing the spatial luminance distribution at the user position for autostereoscopic 3-D display," *Journal of the Society for Information Display*, vol. 20, no. 2, pp. 118–122

[Hossfeld et al., 2014a] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558

[Hossfeld and Keimel, 2014] T. Hossfeld and C. Keimel, "Crowdsourcing in QoE Evaluation," *Quality of Experience*, ed. by S. Möller and A. Raake, T-Labs Series in Telecommunication Services, Springer International Publishing, pp. 315–327

[Hossfeld et al., 2014b] T. Hossfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, "Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment," *International Workshop on Multimedia Signal Processing (MMSP)*

[Huang et al., 2012] H. Huang, B. Zhang, S.-H. Chan, G. Cheung, and P. Frossard, "Coding and replication co-design for interactive multiview video streaming," *IEEE Conference on Computer Communications (INFOCOM)*

[Hunt, 1952] R. W. G. Hunt, "Light and Dark Adaptation and the Perception of Color," *Journal of the Optical Society of America*, vol. 42, no. 3, pp. 190–199

[Huynh-Thu et al., 2007] Q. Huynh-Thu, M. Brotherton, D. Hands, K. Brunnström, and M. Ghanbari, "Examination of the SAMVIQ methodology for the subjective assessment of multimedia quality," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Huynh-Thu and Ghanbari, 2008] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801

[Huynh-Thu et al., 2011a] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 1–14

[Huynh-Thu et al., 2011b] Q. Huynh-Thu, M. Barkowsky, and P. Le Callet, "The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 421–431

[Huynh-Thu and Ghanbari, 2012] Q. Huynh-Thu and M. Ghanbari, "The accuracy of PSNR in predicting video quality for different video scenes and frame rates," *Telecommunication Systems*, vol. 49, no. 1, pp. 35–48

[International Committee for Display Metrology, 2012] International Committee for Display Metrology, Information Display Measurements Standard, version 1.03a

[IRCCyN/IVC 1080i Database] S. Péchard, R. Pépion, and P. Le Callet, IRCCyN/IVC 1080i Video Quality Database, `http://ivc.univ-nantes.fr/en/databases/1080i_Videos/` (visited on 31/3/2016).

[IRCCyN/IVC Image Database] P. Le Callet and F. Autrusseau, Subjective quality assessment IRCCyN/IVC database, `http://ivc.univ-nantes.fr/en/databases/Subjective_Database/` (visited on 31/3/2016).

[Ito, 2010] T. Ito, "Future television - Super Hi-Vision and beyond," *IEEE Asian Solid State Circuits Conference (A-SSCC)*

[Itti, 2004] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318

[Itti et al., 1998] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259

[ITU-T Tutorial, 2004] ITU-T Tutorial, Objective Perceptual Assessment of Video Quality: Full Reference Television, International Telecommunication Union

[ITU-T J.144, 2004] ITU-T J.144, Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, International Telecommunication Union

[ITU-T J.149, 2004] ITU-T J.149, Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM), International Telecommunication Union

[ITU-T BT.1683, 2006] ITU-T BT.1683, Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference, International Telecommunication Union

[ITU-R BT.1788, 2007] ITU-R BT.1788, Methodology for the subjective assessment of video quality in multimedia applications, International Telecommunication Union

[ITU-T P.910, 2008] ITU-T P.910, Subjective video quality assessment methods for multimedia applications, International Telecommunication Union

[ITU-T E.800, 2008] ITU-T E.800, Definitions of terms related to quality of service, International Telecommunication Union

[ITU-R BT.500-13, 2012] ITU-R BT.500-13, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union

[ITU-T P.1401, 2012] ITU-T P.1401, Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models, International Telecommunication Union

[ITU-R BT.2022, 2012] ITU-R BT.2022, General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays, International Telecommunication Union

[ITU-R BT.2021, 2012] ITU-R BT.2021, Subjective methods for the assessment of stereoscopic 3DTV systems, International Telecommunication Union

[ITU-R BT.709, 2015] ITU-R BT.709, Parameter values for the HDTV standards for production and international programme exchange, International Telecommunication Union

[ITU-R BT.2020, 2015] ITU-R BT.2020, Parameter values for ultra-high definition television systems for production and international programme exchange, International Telecommunication Union

[ITU-T P.911, 1998] ITU-T P.911, Subjective audiovisual quality assessment methods for multimedia applications, International Telecommunication Union

[Iwanami et al., 2009] T. Iwanami, A. Kikuchi, T. Kaneko, K. Hirai, N. Yano, T. Nakaguchi, N. Tsumura, Y. Yoshida, and Y. Miyake, "The relationship between ambient illumination and psychological factors in viewing of display Images," *Color Imaging XIV: Displaying, Processing, Hardcopy, and Applications*

[JCT3V-E1100] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Common Test Conditions of 3DV Core Experiments, Doc. JCT3V-E1100, Vienna, Austria

[JCT3V-F0094] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, AHG9: 3D-AVC Coding Results, Doc. JCT3V-F0094, Geneva, Switzerland

[JCT3V-F1011] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 3D Video Subjective Quality Assessment Test Plan, Doc. JCT3V-F1011, Geneva, Switzerland

[JCT3V-G1003] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 3D-AVC Test Model 9, Doc. JCT3V-G1003, San Jose, USA

[JCT3V-G1005] ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Test Model 7 of 3D-HEVC and MV-HEVC, Doc. JCT3V-G1005, San Jose, USA

[JCTVC-I0461] ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, On HEVC still picture coding performance, Doc. JCTVC-I0461, Geneva, Switzerland

[JCTVC-I0595] ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Performance Comparison of HM 6.0 with Existing Still Image Compression Schemes Using a Test Set of Popular Still Images, Doc. JCTVC-I0595, Geneva, Switzerland

[Jermann et al., 2012] P. Jermann, M.-A. Nuessli, and K. Sharma, "Attentional Episodes and Focus," *Dual Eye-Tracking workshop in ACM Conference on Computer Supported Cooperative Work*

[Jin et al., 2011] L. Jin, A. Boev, A. Gotchev, and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video," *IEEE International Conference on Image Processing (ICIP)*

[JM] A. M. Tourapis, A. Leontaris, K. Sühring, and G. J. Sullivan, JM: H.264/AVC reference software, `http://iphome.hhi.de/suehring/tml/` (visited on 31/3/2016).

[JPEGXTHDR] P. Korshunov, P. Hanhart, T. Ricther, A. Artusi, R. Mantiuk, and T. Ebrahimi, HDR image dataset with results of JPEG XT subjective evaluation HDR image dataset with results of JPEG XT subjective evaluation Subjective quality assessment database of HDR images compressed with JPEG XT, `http://mmspg.epfl.ch/jpegxt-hdr` (visited on 31/3/2016).

[Judd et al., 2012] T. Judd, F. Durand, and A. Torralba, *A Benchmark of Computational Models of Saliency to Predict Human Fixations*, tech. rep. MIT-CSAIL-TR-2012-001, MIT Computer Science and Artificial Intelligence Laboratory

[Julesz, 1971] B. Julesz, *Foundations of cyclopean perception*, University of Chicago Press

[Jung et al., 2015] Y. Jung, H. Kim, and Y. Ro, "Critical binocular asymmetry measure for perceptual quality assessment of synthesized stereo 3D images in view synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1

[Kakinuma et al., 2007] K. Kakinuma, M. Shinoda, T. Arai, H. Shibata, T. Shirakuma, M. Kawase, T. Uba, T. Kumakura, S. Haga, and T. Matsumoto, "Technology of Wide Color Gamut Backlight with RGB Light-Emitting Diode for Liquid Crystal Display Television," *SID Symposium Digest of Technical Papers*, vol. 38, no. 1, pp. 1232–1235

[Kakadu] D. Taubman, Kakadu software, `http://www.kakadusoftware.com` (visited on 31/3/2016).

[Kalva et al., 2007] H. Kalva, L. Christodoulou, and B. Furht, "Evaluation of 3DTV Service using Asymmetric View Coding Based on MPEG-2," *3DTV Conference*

[Kauff et al., 2007] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Processing: Image Communication*, vol. 22, no. 2, Special issue on three-dimensional video and television, pp. 217 –234

[Keimel et al., 2012] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd - A framework for crowd-based quality evaluation," *Picture Coding Symposium (PCS)*

[Khattak et al., 2012] S. Khattak, R. Hamzaoui, S. Ahmad, and P. Frossard, "Low-complexity multiview video coding," *Picture Coding Symposium (PCS)*

[Khattak et al., 2013] S. Khattak, R. Hamzaoui, S. Ahmad, and P. Frossard, "Fast encoding techniques for Multiview Video Coding," *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 569 –580

[Khan Md et al., 2015] S. Khan Md, B. Appina, and S. Channappayya, "Full-Reference Stereo Image Quality Assessment Using Natural Stereo Scene Statistics," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1985–1989

[Kilner et al., 2009] J. Kilner, J. Starck, J. Y. Guillemaut, and A. Hilton, "Objective quality assessment in free-viewpoint video production," *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 3–16

[D. Kim et al., 2011] D. Kim, S. Choi, and K. Sohn, "Depth adjustment for stereoscopic images and subjective preference evaluation," *Journal of Electronic Imaging*, vol. 20, no. 3

[S.-K. Kim et al., 2012] S.-K. Kim, S.-K. Yoon, and H. Yoon, "Crosstalk minimization in autostereoscopic multiveiw 3D display by eye tracking and fusion (overlapping) of viewing zones," *Proceedings of SPIE 8384*, Three-Dimensional Imaging, Visualization, and Display

[S.-K. Kim et al., 2013] S.-K. Kim, S.-K. Yoon, and H. Yoon, "Generation of flat viewing zone in DFVZ autostereoscopic multiview 3D display by weighting factor," *Proceedings of SPIE 8738*, Three-Dimensional Imaging, Visualization, and Display

[Kimoto and Kato, 2014] T. Kimoto and C. Kato, "Novel Evaluation of Digital Halftone Image Qualities by Psychological Analysis," *Advances in Image and Video Processing*, vol. 2, no. 2, pp. 8–25

[Kooima et al., 2010] R. Kooima, A. Prudhomme, J. Schulze, D. Sandin, and T. DeFanti, "A multi-viewer tiled autostereoscopic virtual reality display," *Symposium on Virtual Reality Software and Technology*

[Korhonen and You, 2010] J. Korhonen and J. You, "Improving Objective Video Quality Assessment with Content Analysis," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Korhonen et al., 2012] J. Korhonen, N. Burini, J. You, and E. Nadernejad, "How to evaluate objective video quality metrics reliably," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Korhonen and You, 2012] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Korshunov and Ebrahimi, 2012] P. Korshunov and T. Ebrahimi, "A JPEG backward-compatible HDR image compression," *Proceedings of SPIE 8499*, Applications of Digital Image Processing XXXV

[Korshunov and Ebrahimi, 2013] P. Korshunov and T. Ebrahimi, "Context-dependent JPEG backward-compatible high-dynamic range image compression," *Optical Engineering*, vol. 52, no. 10

[Korshunov et al., 2014] P. Korshunov, H. Nemoto, A. Skodras, and T. Ebrahimi, "Crowdsourcing-based Evaluation of Privacy in HDR Images," *Proceedings of SPIE 9138*, Optics, Photonics, and Digital Technologies for Multimedia Applications III

[Korshunov et al., 2015] P. Korshunov, P. Hanhart, T. Ricther, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of HDR images compressed with JPEG XT," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Krawczyk et al., 2007] G. Krawczyk, R. Mantiuk, D. Zdrojewska, and H.-P. Seidel, "Brightness adjustment for HDR and tone mapped images," *Conference on Computer Graphics and Applications*, IEEE

[Kroupi et al., 2014a] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "EEG correlates during video quality perception," *European Signal Processing Conference (EUSIPCO)*

[Kroupi et al., 2014b] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "Predicting Subjective Sensation of Reality During Multimedia Consumption Based on EEG and Peripheral Physiological Signals," *IEEE International Conference on Multimedia and Expo (ICME)*

[Kroupi et al., 2014c] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "User-independent classification of 2D versus 3D multimedia experiences through EEG and physiological signals," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Kroupi et al., 2015] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "Modeling immersive media experiences by sensing impact on subjects," *Multimedia Tools and Applications*, pp. 1–21

[Kuang et al., 2007a] J. Kuang, H. Yamaguchi, C. Liu, G. M. Johnson, and M. D. Fairchild, "Evaluating HDR rendering algorithms," *ACM Transactions on Applied Perception*, vol. 4, no. 2, 9:1–9:27

[Kuang et al., 2007b] J. Kuang, G. M. Johnson, and M. D. Fairchild, "iCAM06: A Refined Image Appearance Model for HDR Image Rendering," *Journal of Visual Communication and Image Representation*, vol. 18, no. 5, pp. 406–414

[Kubota et al., 2007] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV," *IEEE Signal Processing Magazine*, vol. 24, no. 6, p. 10

[Kulkarni et al., 2012] S. D. Kulkarni, M. A. Minor, M. W. Deaver, E. R. Pardyjak, and J. M. Hollerbach, "Design, sensing, and control of a scaled wind tunnel for atmospheric display," *IEEE/ASME Transactions on Mechatronics*, vol. 17, no. 4, pp. 635–645

[Kulyk et al., 2013] V. Kulyk, S. Tavakoli, M. Folkesson, K. Brunnstrom, K. Wang, and N. Garcia, "3D video quality assessment with multi-scale subjective method," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Lainema et al., 2012] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, "Intra Coding of the HEVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1792–1801

[Lang et al., 2012] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth Matters: Influence of Depth Cues on Visual Saliency," *Computer Vision*, Lecture Notes in Computer Science, Springer Berlin Heidelberg, pp. 101–115

[Laparra et al., 2010] V. Laparra, J. Muñoz-Marí, and J. Malo, "Divisive normalization image quality metric revisited," *Journal of the Optical Society of America A*, vol. 27, no. 4, pp. 852–864

[Le Meur et al., 2010a] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?" *Signal Processing: Image Communication*, vol. 25, no. 8, pp. 597 –609

[Le Meur et al., 2010b] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Processing: Image Communication*, vol. 25, no. 7, Special Issue on Image and Video Quality Assessment, pp. 547 –558

[Le Callet et al., 2013] P. Le Callet, S. Möller, and A. Perkis, *Qualinet White Paper on Definitions of Quality of Experience*, tech. rep., European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)

[Ledda et al., 2005] P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, "Evaluation of tone mapping operators using a High Dynamic Range display," *ACM SIGGRAPH 2005 Papers*, SIGGRAPH'05, ACM

[C. Lee and C.-S. Kim, 2008] C. Lee and C.-S. Kim, "Rate-distortion optimized compression of high dynamic range videos," *European Signal Processing Conference (EUSIPCO)*

[J.-S. Lee et al., 2011] J.-S. Lee, L. Goldmann, and T. Ebrahimi, "A New Analysis Method for Paired Comparison and Its Application to 3D Quality Assessment," *ACM International Conference on Multimedia (ACMMM)*

[X. Li et al., 2010] X. Li, M. Wien, and J.-R. Ohm, "Rate-complexity-distortion evaluation for hybrid video coding," *IEEE International Conference on Multimedia and Expo (ICME)*

[J. Li et al., 2013a] J. Li, M. Barkowsky, and P. Le Callet, "Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs," *Proceedings of SPIE 8648*, Stereoscopic Displays and Applications XXIV

[J. Li et al., 2013b] J. Li, M. Barkowsky, and P. Le Callet, "Subjective assessment methodology for preference of experience in 3DTV," *IEEE IVMSP Workshop*

[J. Li et al., 2013c] J. Li, O. Kaller, F. De Simone, J. Hakala, D. Juszka, and P. Le Callet, "Cross-lab study on preference of experience in 3DTV: Influence from display technology and test environment," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Liao et al., 2013] Y. Liao, A. Younkin, J. Foerster, and P. Corriveau, "Achieving High QoE Across the Compute Continuum: How Compression, Content, and Devices Interact," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[W. Lin and Kuo, 2011] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297 –312

[H. Lin and L. Wu, 2014] H. Lin and L. Wu, "Quality Assessment of Stereoscopic 3D Image Compression by Binocular Integration Behaviors," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1527–1542

[Lischinski et al., 2006] D. Lischinski, Z. Farbman, M. Uyttendaele, and R. Szeliski, "Interactive Local Adjustment of Tonal Values," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 646–653

[Y. Liu et al., 2009] Y. Liu, S. Ma, Q. Huang, D. Zhao, W. Gao, and N. Zhang, "Compression-Induced Rendering Distortion Analysis for Texture/Depth Rate Allocation in 3D Video Compression," *Data Compression Conference (DCC)*

[H. Liu and Heynderickx, 2011] H. Liu and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982

[J. Liu et al., 2013] J. Liu, F. Hassan, and J. Carletta, "Embedding high dynamic range tone mapping in JPEG compression," *Proceedings of SPIE 8655*, Image Processing: Algorithms and Systems XI

[J. Liu et al., 2015a] J. Liu, T. Malzbender, S. Qin, B. Zhang, C.-A. Wu, and J. Davis, "Dynamic mapping for multiview autostereoscopic displays," *Proceedings of SPIE 9391*, Stereoscopic Displays and Applications XXVI

[J. Liu et al., 2015b] J. Liu, N. Stefanoski, T. O. Aydin, A. Grundhofer, and A. Smolic, "Chromatic calibration of an HDR display using 3D octree forests," *IEEE International Conference on Image Processing (ICIP)*

[LIVE Image Database] H. R. Sheikh, Z. Wang, L. K. Cormack, and A. C. Bovik, LIVE Image Quality Assessment Database Release 2, `http://live.ece.utexas.edu/research/quality/subjective.htm` (visited on 31/3/2016).

[LIVE Video Database] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, LIVE Video Quality Database, `http://live.ece.utexas.edu/research/quality/live_video.html` (visited on 31/3/2016).

[Lu et al., 2013] T. Lu, H. Ganapathy, G. Lakshminarayanan, T. Chen, W. Husak, and P. Yin, "Orthogonal Muxing Frame Compatible Full Resolution technology for multi-resolution frame-compatible stereo coding," *IEEE International Conference on Multimedia and Expo (ICME)*

[Luce, 1959] R. D. Luce, "Individual Choice Behaviours: A Theoretical Analysis"

[Luo et al., 2001] M. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Research & Application*, vol. 26, no. 5, pp. 340–350

[M17050] ISO/IEC JTC1/SC29/WG11, Poznań Multiview Video Test Sequences and Camera Parameters, Doc. M17050, Xian, Chine

[M22988] ISO/IEC JTC1/SC29/WG11, [AHG8] Objective and subjective evaluation of HM5.0, Doc. M22988, San Jose, USA

[M23703] ISO/IEC JTC1/SC29/WG11, Proposed Stereo Test Sequences for 3D Video Coding, Doc. M23703, San Jose, USA

[M23863] ISO/IEC JTC1/SC29/WG11, JCT-VC AHG report: HM subjective quality investigation (AHG22), Doc. M23863, San Jose, USA

[M35471] ISO/IEC JTC1/SC29/WG11, HDRTools: Software updates, Doc. M35471, Geneva, Switzerland

[M35480] ISO/IEC JTC1/SC29/WG11, Selected test content and timeline for HDR single layer anchors generation for 111th MPEG meeting, Doc. M35480, Geneva, Switzerland

[M35852] ISO/IEC JTC1/SC29/WG11, Report on the anchors generation for HDR /WCG video coding, Doc. M35852, Geneva, Switzerland

[Maalouf and Larabi, 2011] A. Maalouf and M.-C. Larabi, "CYCLOP: A stereo color image quality assessment metric," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[Mai et al., 2011a] Z. Mai, H. Mansour, R. Mantiuk, P. Nasiopoulos, R. Ward, and W. Heidrich, "Optimizing a tone curve for backward-compatible high dynamic range image and video compression," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1558–1571

[Mai et al., 2011b] Z. Mai, C. Doutre, P. Nasiopoulos, and R. Ward, "Subjective evaluation of tone-mapping methods on 3D images," *International Conference on Digital Signal Processing (DSP)*

[Mantiuk et al., 2004] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "Perception-motivated High Dynamic Range Video Encoding," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 733–741

[Mantiuk et al., 2005] R. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," *Proceedings of SPIE 5666*, Human Vision and Electronic Imaging X

[Mantiuk et al., 2006a] R. Mantiuk, K. Myszkowski, and H. Seidel, "A perceptual framework for contrast processing of high dynamic range images," *ACM Transactions on Applied Perception*, vol. 3, no. 3, pp. 286–308

[Mantiuk et al., 2006b] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel, "Backward Compatible High Dynamic Range MPEG Video Compression," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 713–723

[Mantiuk et al., 2007] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel, "High Dynamic Range Image and Video Compression - Fidelity Matching Human Visual Performance," *IEEE International Conference on Image Processing (ICIP)*

[Mantiuk et al., 2008] R. Mantiuk, S. J. Daly, and L. Kerofsky, "Display Adaptive Tone Mapping," *ACM Transactions on Graphics*, vol. 27, no. 3, 68:1–68:10

[Mantiuk et al., 2011] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions," *ACM Transactions on Graphics*, vol. 30, no. 4

[Mantel et al., 2014] C. Mantel, S. Ferchiu, and S. Forchhammer, "Comparing subjective and objective quality assessment of HDR images compressed with JPEG-XT," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*

[Mantel et al., 2015a] C. Mantel, J. Korhonen, J. M. Pedersen, S. Bech, J. D. Andersen, and S. Forchhammer, "Subjective quality of video sequences rendered on LCD with local backlight dimming at different lighting conditions," *Proceedings of SPIE 9396*, Image Quality and System Performance XII

[Mantel et al., 2015b] C. Mantel, J. Korhonen, S. Forchhammer, J. Pedersen, and S. Bech, "Subjective quality of videos displayed with local backlight dimming at different peak white and ambient light levels," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Mannos and Sakrison, 1974] J. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536

[Mann and Picard, 1995] S. Mann and R. W. Picard, "On Being 'undigital' With Digital Cameras: Extending Dynamic Range By Combining Differently Exposed Pictures," *IS&T Annual Conference*

[Masaoka et al., 2006] K. Masaoka, M. Emoto, M. Sugawara, and Y. Nojiri, "Contrast effect in evaluating the sense of presence for wide displays," *Journal of the Society for Information Display*, vol. 14, no. 9, pp. 785–791

[Masaoka et al., 2010] K. Masaoka, Y. Nishida, M. Sugawara, and E. Nakasu, "Design of Primaries for a Wide-Gamut Television Colorimetry," *IEEE Transactions on Broadcasting*, vol. 56, no. 4, pp. 452–457

[Maugey and Frossard, 2011] T. Maugey and P. Frossard, "Interactive multiview video system with low decoding complexity," *IEEE International Conference on Image Processing (ICIP)*

[Maugey and Frossard, 2013] T. Maugey and P. Frossard, "Interactive Multiview Video System With Low Complexity 2D Look Around at Decoder," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1070–1082

[Maugey et al., 2013] T. Maugey, I. Daribo, G. Cheung, and P. Frossard, "Navigation Domain Representation For Interactive Multiview Imaging," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3459–3472

[Meesters et al., 2004] L. Meesters, W. IJsselsteijn, and P. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 381–391

[Mendiburu, 2009] B. Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*, Focal Press/Elsevier

[Merkle et al., 2006] P. Merkle, K. Müller, A. Smolic, and T. Wiegand, "Efficient Compression of Multi-View Video Exploiting Inter-View Dependencies Based on H.264/MPEG4-AVC," *IEEE International Conference on Multimedia and Expo (ICME)*

[Merkle et al., 2007a] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient Compression of Multi-View Depth Data Based on MVC," *3DTV Conference*

[Merkle et al., 2007b] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-View Video Plus Depth Representation and Coding," *IEEE International Conference on Image Processing (ICIP)'*

[Merkle et al., 2007c] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient Prediction Structures for Multiview Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473

[Merkle et al., 2009] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, and T. Wiegand, "The effects of multiview depth video compression on multiview rendering," *Singal Processing: Image Communication*, vol. 24, no. 1-2, pp. 73–88

[MeTriX MuX] M. Gaubatz, MeTriX MuX Visual Quality Assessment Packag version 1.1, `http://foulard.ece.cornell.edu/gaubatz/metrix_mux/` (visited on 31/3/2016).

[Miller et al., 2013] S. Miller, M. Nezamabadi, and S. J. Daly, "Perceptual Signal Coding for More Efficient Usage of Bit Codes," *SMPTE Motion Imaging Journal*, vol. 122, no. 4, pp. 52–59

[MIMESIS] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, MIMESIS: Modeling Immersive Media Experiences by Sensing Impact on Subjects, `http://mmspg.epfl.ch/mimesis` (visited on 31/3/2016).

[Mittal et al., 2011] A. Mittal, A. K. Moorthy, J. Ghosh, and A. C. Bovik, "Algorithmic assessment of 3D quality of experience for images and videos," *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop*

[Mitsa and Varkur, 1993] T. Mitsa and K. L. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*

[Mitchell, 1997] J. L. Mitchell, *MPEG video compression standard*, Springer Science & Business Media

[Mitsunaga and Nayar, 1999] T. Mitsunaga and S. Nayar, "Radiometric self calibration," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*

[Mocanu et al., 2014] D. Mocanu, G. Exarchakos, and A. Liotta, "Deep learning for objective quality assessment of 3D images," *IEEE International Conference on Image Processing (ICIP)*

[Möller and Raake, 2014] S. Möller and A. Raake, *Quality of Experience*, T-Labs Series in Telecommunication Services, Springer International Publishing

[Morrissey, 1955] J. H. Morrissey, "New Method for the Assignment of Psychometric Scale Values from Incomplete Paired Comparisons," *Journal of the Optical Society of America*, vol. 45, no. 5, pp. 373–378

[Mukherjee et al., 2013] D. Mukherjee, J. Bankoski, A. Grange, J. Han, J. Koleszar, P. Wilkins, Y. Xu, and R. Bultje, "The latest open-source video codec VP9-An overview and preliminary results.," *Picture Coding Symposium (PCS)*

[Mukherjee et al., 2015a] D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins, and Y. Xu, "A Technical Overview of VP9 - The Latest Open-Source Video Codec," *SMPTE Motion Imaging Journal*, vol. 124, no. 1, pp. 44–54

[Mukherjee et al., 2015b] D. Mukherjee, H. Su, J. Bankoski, A. Converse, J. Han, Z. Liu, and Y. Xu, "An overview of new video coding tools under consideration for VP10: the successor to VP9," *Proceedings of SPIE 9599*, Applications of Digital Image Processing XXXVIII

[Muller et al., 2008] K. Muller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View Synthesis for Advanced 3D Video Systems," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 438148

[Muller et al., 2009] K. Muller, A. Smolic, K. Dix, P. Merkle, and T. Wiegand, "Coding and intermediate view synthesis of multiview video plus depth," *IEEE International Conference on Image Processing (ICIP)*

[Muller et al., 2013] K. Muller et al., "3D High-Efficiency Video Coding for Multi-View Video and Depth Data," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3366–3378

[Munkberg et al., 2006] J. Munkberg, P. Clarberg, J. Hasselgren, and T. Akenine-Möller, "High Dynamic Range Texture Compression for Graphics Hardware," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 698–706

[Murthy and Karam, 2010] A. V. Murthy and L. J. Karam, "A MATLAB-based framework for image and video quality evaluation," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[N11113] ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/16, Joint Call for Proposals on Video Compression Technology, Doc. N11113, Kyoto, Japan

[N12036] ISO/IEC JTC1/SC29/WG11, Call for Proposals on 3D Video Coding Technology, Doc. N12036, Geneva, Switzerland

[N12347] ISO/IEC JTC1/SC29/WG11, Report of Subjective Test Results from the Call for Proposals on 3D Video Coding Technology, Doc. N12347, Geneva, Switzerland

[N15083] ISO/IEC JTC1/SC29/WG11, Call for Evidence (CfE) for HDR and WCG Video Coding, Doc. N15083, Geneva, Switzerland

[Nam et al., 2011] D. Nam, J. Park, D. Park, and C. Y. Kim, "Autostereoscopic 3D - How can we move to the next step?" *Euro-American Workshop on Information Optics*

[NAMA3DS1 database] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutierrez, and N. Garcia, Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences, `http://ivc.univ-nantes.fr/en/databases/NAMA3DS1` (visited on 31/3/2016).

[Narwaria et al., 2014] M. Narwaria, M. P. D. Silva, P. L. Callet, and R. Pepion, "Tone mapping based HDR compression: Does it affect visual experience?" *Signal Processing: Image Communication*, vol. 29, no. 2, Special Issue on Advances in High Dynamic Range Video Research, pp. 257 –273

[Narwaria et al., 2015a] M. Narwaria, R. Mantiuk, M. Perreira Da Silva, and P. Le Callet, "HDR-VDP-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1

[Narwaria et al., 2015b] M. Narwaria, M. Perreira Da Silva, and P. Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60

[Narwaria et al., 2015c] M. Narwaria, M. Perreira Da Silva, and P. Le Callet, "Study of High Dynamic Range Video Quality Assessment," *Proceedings of SPIE 9599*, Applications of Digital Image Processing XXXVIII

[Nasiopoulos et al., 2014] E. Nasiopoulos, Y. Dong, and A. Kingstone, "Evaluation of High Dynamic Range Content Viewing Experience Using Eye-Tracking Data," *International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine)*

[Nemoto et al., 2014a] H. Nemoto, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Impact of Ultra High Definition on Visual Attention," *ACM International Conference on Multimedia*

[Nemoto et al., 2014b] H. Nemoto, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Ultra-Eye: UHD and HD images eye tracking dataset," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Nemoto et al., 2015] H. Nemoto, P. Korshunov, P. Hanhart, and T. Ebrahimi, "Visual attention in LDR and HDR images," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Netravali, 2013] A. Netravali, *Digital pictures: representation and compression*, Springer Science & Business Media

[Ninassi et al., 2006] A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, and A. Tirel, "Task impact on the visual attention in subjective image quality assessment," *European Signal Processing Conference (EUSIPCO)*

[Ninassi et al., 2007] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba, "Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric," *IEEE International Conference on Image Processing (ICIP)*

[Oelbaum et al., 2004] T. Oelbaum, V. Baroncini, T. K. Tan, and C. Fenimore, "Subjective quality assessment of the emerging AVC/H. 264 video coding standard," *International Broadcasting Conference (IBC)*

[Oelbaum et al., 2008] T. Oelbaum, H. Schwarz, M. Wien, and T. Wiegand, "Subjective performance evaluation of the SVC extension of H.264/AVC," *IEEE International Conference on Image Processing (ICIP)*

[Oh et al., 2010] K.-J. Oh, S. Yea, A. Vetro, and S. Ho, "Virtual view synthesis method and self-evaluation metrics for free viewpoint television and 3D video," *International Journal of Imaging Systems and Technology*, vol. 20, no. 4, pp. 378–390

[Ohm et al., 2012] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC)," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684

[Oppenheim et al., 1968] A. Oppenheim, R. Schafer, and J. Stockham T.G., "Nonlinear filtering of multiplied and convolved signals," *Proceedings of the IEEE*, vol. 56, no. 8, pp. 1264–1291

[Park et al., 2011] J. Park, D. Nam, G. Sung, Y. Kim, D. Park, and C. Kim, "Active Crosstalk Reduction on Multi-View Displays Using Eye Detection," *SID Symposium Digest of Technical Papers*, vol. 42, no. 1, pp. 920–923

[Pasteau et al., 2011] F. Pasteau, C. Strauss, M. Babel, O. Déforges, and L. Bédat, "Adaptive colour decorrelation for predictive image codecs," *European Signal Processing Conference (EUSIPCO)*

[Pattanaik and Hughes, 2005] S. Pattanaik and C. Hughes, "High-Dynamic-Range Still-Image Encoding in JPEG 2000," *IEEE Computer Graphics and Applications*, vol. 25, no. 6, pp. 57–64

[Péchard et al., 2008] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," *International Workshop on Image Media Quality and its Applications (IMQA)*

[Pedersen and Hardeberg, 2009] M. Pedersen and J. Y. Hardeberg, *Survey of full-reference image quality metrics*, tech. rep., Høgskolen i Gjøviks rapportserie

[Peli et al., 2001] E. Peli, T. R. Hedges, J. Tang, and D. Landmann, "A Binocular Stereoscopic Display System with Coupled Convergence and Accommodation Demands," *SID Symposium Digest of Technical Papers*, vol. 32, no. 1, pp. 1296–1299

[Perkis et al., 2012] A. Perkis et al., "Towards certification of 3D video quality assessment," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Perera et al., 2014] G. Perera, V. De Silva, A. Kondoz, and S. Dogan, "An improved model of binocular energy calculation for full-reference stereoscopic image quality assessment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

[Pictureaunt 3.2] M. Mehl and C. Bloch, Pictureaunt 3.2 software, `http://www.hdrlabs.com/picturenaut/` (visited on 31/3/2016).

[Pinson and Wolf, 2004] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322

[Pinheiro et al., 2014] A. Pinheiro, K. Fliegel, P. Korshunov, L. Krasula, M. Bernardo, M. Pereira, and T. Ebrahimi, "Performance evaluation of the emerging JPEG XT image compression standard," *International Workshop on Multimedia Signal Processing (MMSP)*

[Ponomarenko et al., 2007] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Ponomarenko et al., 2009] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45

[Ponomarenko et al., 2015] N. Ponomarenko et al., "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57 –77

[Poulakos et al., 2015] S. Poulakos, R. Monroy, T. Aydin, O. Wang, A. Smolic, and M. Gross, "A computational model for perception of stereoscopic window violations," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Poynton, 2012] C. Poynton, *Digital video and HD: Algorithms and Interfaces*, Burlington, Vermont, USA: Elsevier/Morgan Kaufmann

[Qi et al., 2015] F. Qi, D. Zhao, and W. Gao, "Reduced Reference Stereoscopic Image Quality Assessment Based on Binocular Perceptual Information," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2338–2344

[Qualinet Databases] K. Fliegel, Qualinet Databases, `http://dbq.multimediatech.cz` (visited on 31/3/2016).

[Rabbani and Jones, 1991] M. Rabbani and P. Jones, *Digital Image Compression Techniques*, Books in the Spie Tutorial Texts Series, SPIE Optical Engineering Press

[Redi et al., 2009] J. Redi, H. Liu, P. Gastaldo, R. Zunino, and I. Heynderickx, "How to apply spatial saliency into objective metrics for JPEG compressed images?" *IEEE International Conference on Image Processing (ICIP)*

[Redi et al., 2013] J. Redi, T. Hossfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel, "Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal," *ACM International Workshop on Crowdsourcing for Multimedia*

[Reinhard et al., 2002] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic Tone Reproduction for Digital Images," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 267–276

[Reinhard et al., 2005] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*, The Morgan Kaufmann Series in Computer Graphics, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[Rempel et al., 2009] A. G. Rempel, W. Heidrich, H. Li, and R. Mantiuk, "Video Viewing Preferences for HDR Displays Under Varying Ambient Illumination," *Symposium on Applied Perception in Graphics and Visualization*

[Rerabek and Ebrahimi, 2014] M. Rerabek and T. Ebrahimi, "Comparison of compression efficiency between HEVC/H.265 and VP9 based on subjective assessments," *Proceedings of SPIE 9217*, Applications of Digital Image Processing XXXVII

[Rerabek et al., 2015a] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and objective evaluation of HDR video compression," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Rerabek et al., 2015b] M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Quality evaluation of HEVC and VP9 video compression in real-time applications," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Ribeiro et al., 2011] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," *IEEE International Conference on Image Processing (ICIP)*

[Richter, 2013] T. Richter, "Backwards Compatible Coding of High Dynamic Range Images with JPEG," *Data Compression Conference (DCC)*

[Richter, 2014] T. Richter, "On the integer coding profile of JPEG XT," *Proceedings of SPIE 9217*, Applications of Digital Image Processing XXXVII

[Rimac-Drlje et al., 2009] S. Rimac-Drlje, M. Vranjes, and D. Zagar, "Influence of temporal pooling method on the objective video quality evaluation," *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*

[Robertson et al., 1999] M. Robertson, S. Borman, and R. Stevenson, "Dynamic range improvement through multiple exposures," *IEEE International Conference on Image Processing (ICIP)*

[Rogers and Graham, 1979] B. Rogers and M. Graham, "Motion parallax as an independent cue for depth perception," *Perception*, vol. 8, no. 2, pp. 125–34

[Rohaly et al., 1999] A. M. Rohaly, N. R. Lu Jiuhuai andFranzen, and M. K. Ravel, "Comparison of temporal pooling methods for estimating the quality of complex video sequences," *Proceedings of SPIE 3644*, Human Vision and Electronic Imaging IV

[Rose, 1948] A. Rose, "The sensitivity performance of the human eye on an absolute scale," *Journal of the Optical Society of America A*, vol. 38, no. 2, pp. 196–208

[Ryu et al., 2012] S. Ryu, D. H. Kim, and K. Sohn, "Stereoscopic image quality metric based on binocular perception model," *IEEE International Conference on Image Processing (ICIP)*

[Ryu and Sohn, 2014] S. Ryu and K. Sohn, "No-Reference Quality Assessment for Stereoscopic Images Based on Binocular Quality Perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 591–602

[Salmimaa and Järvenpää, 2008] M. Salmimaa and T. Järvenpää, "Optical characterization and measurements of autostereoscopic 3D displays," *Proceedings of SPIE 7001*, Photonics in Multimedia II

[Sanchez-Vives and Slater, 2005] M. V. Sanchez-Vives and M. Slater, "From presence to consciousness through virtual reality," *Nature Reviews Neuroscience*, vol. 6, no. 4, pp. 332–339

[Sazzad et al., 2010] Z. Sazzad, S. Yamanaka, and Y. Horita, "Spatio-temporal segmentation based continuous no-reference stereoscopic video quality prediction," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[SCENIC] P. Hanhart and T. Ebrahimi, SCENIC: Subjective Comparison of ENcoders based on fItted Curves SCENIC: Subjective Comparison of ENcoders based on fItted Curves SCENIC: Subjective Comparison of ENcoders based on fItted Curves, `http://mmspg.epfl.ch/scenic` (visited on 31/3/2016).

[Schwarz et al., 2007] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120

[Schelkens et al., 2009] P. Schelkens, A. Skodras, and T. Ebrahimi, *The JPEG 2000 Suite*, vol. 15, John Wiley & Sons

[Seetzen et al., 2003] H. Seetzen, L. A. Whitehead, and G. Ward, "A High Dynamic Range Display Using Low and High Resolution Modulators," *SID Symposium Digest of Technical Papers*, vol. 34, no. 1, pp. 1450–1453

[Seetzen et al., 2004] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs, "High Dynamic Range Display Systems," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 760–768

[Seetzen et al., 2006] H. Seetzen, H. Li, L. Ye, W. Heidrich, L. Whitehead, and G. Ward, "Observations of Luminance, Contrast and Amplitude Resolution of Displays," *SID Symposium Digest of Technical Papers*, vol. 37, no. 1, pp. 1229–1233

[Seshadrinathan and Bovik, 2010] K. Seshadrinathan and A. C. Bovik, "Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350

[Seshadrinathan et al., 2010] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441

[Seuntiëns et al., 2005] P. Seuntiëns, L. Meesters, and W. IJsselsteijn, "Perceptual attributes of crosstalk in 3D images," *Displays*, vol. 26, no. 4-5, pp. 177–183

[Seufert et al., 2013] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, ""To pool or not to pool": A comparison of temporal pooling methods for HTTP adaptive video streaming," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Shaked and Tastl, 2005] D. Shaked and I. Tastl, "Sharpness measure: towards automatic image enhancement," *IEEE International Conference on Image Processing (ICIP)*

[Shao et al., 2013] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual Full-Reference Quality Assessment of Stereoscopic Images by Considering Binocular Visual Characteristics," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1940–1953

[Shao et al., 2014] F. Shao, G.-y. Jiang, M. Yu, F. Li, Z. Peng, and R. Fu, "Binocular energy response based quality assessment of stereoscopic images," *Digital Signal Processing*, vol. 29, pp. 45 –53

[Shao et al., 2015a] F. Shao, K. Li, W. Lin, G. Jiang, M. Yu, and Q. Dai, "Full-Reference Quality Assessment of Stereoscopic Images by Learning Binocular Receptive Field Properties," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 2971–2983

[Shao et al., 2015b] F. Shao, K. Li, W. Lin, G. Jiang, and M. Yu, "Using Binocular Feature Combination for Blind Quality Assessment of Stereoscopic Images," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1548–1551

[Shevell, 2003] S. K. Shevell, *The science of color*, Boston, Massachusetts, USA: Elsevier

[Sheikh et al., 2005] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128

[Sheikh and Bovik, 2006] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444

[Sheikh et al., 2006] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451

[Shiwa et al., 1996] S. Shiwa, K. Omura, and F. Kishino, "Proposal for a 3-D display with accommodative compensation: 3DDAC," *Journal of the Society for Information Display*, vol. 4, no. 4, pp. 255–261

[Shnayderman et al., 2006] A. Shnayderman, A. Gusev, and A. Eskicioglu, "An SVD - based grayscale image quality measure for local and global assessment," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 422–429

[Silva et al., 2015] A. R. Silva, M. E. Vizcarra Melgar, and M. C. Q. Farias, "A no-reference stereoscopic quality metric," *Proceedings of SPIE 9393*, Three-Dimensional Image Processing, Measurement (3DIPM), and Applications

[Skodras et al., 2001] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58

[Slater, 2009] M. Slater, "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3549–3557

[Slater and Wilbur, 1997] M. Slater and S. Wilbur, "A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616

[Smolic et al., 2004] A. Smolic, K. Mueller, P. Merkle, T. Rein, M. Kautzner, P. Eisert, and T. Wiegand, "Free viewpoint video extraction, representation, coding, and rendering," *IEEE International Conference on Image Processing (ICIP)*

[Smolic and Kauff, 2005] A. Smolic and P. Kauff, "Interactive 3-D Video Representation and Coding Technologies," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 98–110

[Smolic et al., 2006] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D Video and Free Viewpoint Video - Technologies, Applications and MPEG Standards," *IEEE International Conference on Multimedia and Expo (ICME)*

[Smolic et al., 2007] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. Akar, G. Triantafyllidis, and A. Koz, "Coding Algorithms for 3DTV - A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1606–1621

[Smolic et al., 2008] A. Smolic, K. Muller, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "Intermediate view interpolation based on multiview video plus depth for advanced 3D video systems," *IEEE International Conference on Image Processing (ICIP)*

[Smolic et al., 2009a] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand, "An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution," *Picture Coding Symposium (PCS)*

[Smolic et al., 2009b] A. Smolic, K. Mueller, P. Merkle, and A. Vetro, "Development of a new MPEG standard for advanced 3D video applications," *International Symposium on Image and Signal Processing and Analysis (ISPA)*

[Smolic, 2011] A. Smolic, "3D video and free viewpoint video - From capture to display," *Pattern Recognition*, vol. 44, no. 9, Computer Analysis of Images and Patterns, pp. 1958 –1968

[Snedecor and Cochran, 1989] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Iowa State University Press

[Someya et al., 2006] J. Someya, Y. Inoue, H. Yoshii, M. Kuwata, S. Kagawa, T. Sasagawa, A. Michimori, H. Kaneko, and H. Sugiura, "Laser TV: Ultra-Wide Gamut for a New Extended Color-Space Standard, xvYCC," *SID Symposium Digest of Technical Papers*, vol. 37, no. 1, pp. 1134–1137

[Song et al., 2015] W. Song, Y. Xiao, D. Tjondronegoro, and A. Liotta, "QoE Modelling for VP9 and H.265 Videos on Mobile Devices," *ACM International Conference on Multimedia*

[Spaulding et al., 2003] K. Spaulding, G. J. Woolfe, and R. L. Joshi, "Using a residual image to extend the color gamut and dynamic range of an sRGB image," *IS&T PICS Conference*

[Stelmach et al., 2000] L. Stelmach, W. Tam, D. Meegan, A. Vincent, and P. Corriveau, "Human Perception of Mismatched Stereoscopic 3D Inputs," *IEEE International Conference on Image Processing (ICIP)*

[Stefan Winkler's website] S. Winkler, Image and Video Quality Resources, `http://stefan.winkler.net/resources.html` (visited on 31/3/2016).

[Stoakley et al., 1995] R. Stoakley, M. J. Conway, and R. Pausch, "Virtual reality on a WIM: interactive worlds in miniature," *SIGCHI Conference on Human Factors in Computing Systems*

[Sugiura et al., 2003] H. Sugiura, H. Kaneko, S. Kagawa, M. Ozawa, K. Niki, and H. Tanizoe, "Prototype of a Wide Gamut Monitor Adopting an LED-Backlighting LCD Panel," *SID Symposium Digest of Technical Papers*, vol. 34, no. 1, pp. 1266–1269

[Sugawara et al., 2003] M. Sugawara, M. Kanazawa, K. Mitani, H. Shimamoto, T. Yamashita, and F. Okano, "Ultrahigh-Definition Video System with 4000 Scanning Lines," *SMPTE Motion Imaging Journal*, vol. 112, no. 10-11, pp. 339–346

[Sullivan and Ohm, 2010] G. J. Sullivan and J.-R. Ohm, "Recent developments in standardization of high efficiency video coding (HEVC)," *Proceedings of SPIE 7798*, Applications of Digital Image Processing XXXIII

[Sullivan et al., 2012] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668

[Sullivan et al., 2013] G. J. Sullivan, J. Boyce, Y. Chen, J.-R. Ohm, C. Segall, and A. Vetro, "Standardized Extensions of High Efficiency Video Coding (HEVC)," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 1001–1016

[Surman et al., 2010] P. Surman, R. S. Brar, I. Sexton, and K. Hopf, "MUTED and HELIUM3D autostereoscopic displays," *IEEE International Conference on Multimedia and Expo (ICME)*

[Sykora et al., 2011] M. Sykora, J. Schultz, and R. Brott, "Optical characterization of autostereoscopic 3D displays," *Proceedings of SPIE 7863*, Stereoscopic Displays and Applications XXII

[Tabatabai et al., 2014] A. Tabatabai, T. Suzuki, P. Hanhart, P. Korshunov, T. Ebrahimi, M. Horowitz, F. Kossentini, and H. Tmar, "Compression Performance Analysis in HEVC," *High Efficiency Video Coding (HEVC)*, ed. by V. Sze, M. Budagavi, and G. J. Sullivan, Integrated Circuits and Systems, Springer International Publishing, pp. 275–302

[Talmi and J. Liu, 1999] K. Talmi and J. Liu, "Eye and gaze tracking for visually controlled interactive stereoscopic displays," *Signal Processing: Image Communication*, vol. 14, no. 10, pp. 799–810

[Tanimoto, 2006] M. Tanimoto, "Overview of free viewpoint television," *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461

[Tan et al., 2016] T. Tan, R. Weerakkody, M. Mrak, N. Ramzan, V. Baroncini, J. Ohm, and G. J. Sullivan, "Video Quality Evaluation Methodology and Verification Testing of HEVC Compression Performance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 76–90

[Thurstone, 1927] L. L. Thurstone, "A law of comparative judgment," *Psychological review*, vol. 34, no. 4, p. 273

[TID2008] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, Tampere Image Database 2008 version 1.0, `http://www.ponomarenko.info/tid2008.htm` (visited on 31/3/2016).

[TID2013] N. Ponomarenko et al., Tampere Image Database 2013 version 1.0, `http://www.ponomarenko.info/tid2013.htm` (visited on 31/3/2016).

[Toni et al., 2013] L. Toni, N. Thomos, and P. Frossard, "Interactive free viewpoint video streaming using prioritized network coding," *IEEE International Workshop on Multimedia Signal Processing (MMSP)*

[Tsirlin et al., 2012] I. Tsirlin, R. Allison, and L. Wilcox, "Crosstalk reduces the amount of depth seen in 3D images of natural scenes," *Proceedings of SPIE 8288*, Stereoscopic Displays and Applications XXIII

[Tsukida and Gupta, 2011] K. Tsukida and M. R. Gupta, *How to analyze paired comparison data*, tech. rep. UWEETR-2011-0004, Seattle, Washington, USA: Department of Electrical Engineering, University of Washington

[Tukey, 1977] J. Tukey, *Exploratory data analysis*, Reading: Addison-Wesley

[Tumblin and Rushmeier, 1993] J. Tumblin and H. Rushmeier, "Tone reproduction for realistic images," *IEEE Computer Graphics and Applications*, vol. 13, no. 6, pp. 42–48

[Ultra-Eye] H. Nemoto, P. Hanhart, P. Korshunov, and T. Ebrahimi, Ultra-Eye: UHD and HD images eye tracking dataset, `http://mmspg.epfl.ch/ultra-eye` (visited on 31/3/2016).

[Urey et al., 2011] H. Urey, K. Chellappan, E. Erden, and P. Surman, "State of the Art in Stereoscopic and Autostereoscopic Displays," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 540–555

[Urvoy et al., 2012] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutierrez, and N. Garcia, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Valenzise et al., 2014] G. Valenzise, F. De Simone, P. Lauga, and F. Dufaux, "Performance evaluation of objective quality metrics for HDR image compression," *Proceedings of SPIE 9217*, Applications of Digital Image Processing XXXVII

[Vetro et al., 2008] A. Vetro, S. Yea, and A. Smolic, "Toward a 3D video format for auto-stereoscopic displays," *Proceedings of SPIE 7073*, Applications of Digital Image Processing XXXI

[Vetro et al., 2011] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642

[Vetro and Tian, 2012] A. Vetro and D. Tian, "Analysis of 3D and multiview extensions of the emerging HEVC standard," *Proceedings of SPIE 8499*, Applications of Digital Image Processing XXXV

[Von der Pütten et al., 2012] A. M. von der Pütten, J. Klatt, S. T. Broeke, R. McCall, N. C. Krämer, R. Wetzel, L. Blum, L. Oppermann, and J. Klatt, "Subjective and behavioral presence measurement and interactivity in the collaborative augmented reality game TimeWarp," *Interacting with Computers*, vol. 24, pp. 317–325

[VQEG, 2003] VQEG, Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II (FR_TV2)

[VQM] M. Pinson and S. Wolf, Video Quality Metric, `http://vqm.its.bldrdoc.gov` (visited on 31/3/2016).

[VQMT] P. Hanhart, Video Quality Measurement Tool, `http://mmspg.epfl.ch/vqmt/` (visited on 31/3/2016).

[VSNR] D. M. Chandler and S. S. Hemami, VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images, `http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html` (visited on 31/3/2016).

[Vu et al., 2003] K. Vu, K. A. Hua, S. Member, and W. Tavanapong, "Image Retrieval Based on Regions of Interest," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, pp. 1045–1049

[Wallace, 1991] G. Wallace, "The JPEG Still Picture Compression Standard," *IEEE Transactions on Consumer Electronics*

[Z. Wang and Bovik, 2002] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84

[Z. Wang et al., 2003] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Asilomar Conference on Signals, Systems and Computers*

[Z. Wang et al., 2004] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612

[Z. Wang and Bovik, 2006] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, Morgan & Claypool, p. 156

[Z. Wang and Q. Li, 2011] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198

[D. Wang et al., 2011] D. Wang, G. Li, W. Jia, and X. Luo, "Saliency-driven Scaling Optimization for Image Retargeting," *Visual Computing*, vol. 27, no. 9, pp. 853–860

[X. Wang et al., 2011] X. Wang, S. Kwong, and Y. Zhang, "Considering binocular spatial sensitivity in stereoscopic image quality assessment," *IEEE Visual Communications and Image Processing (VCIP)*

[K. Wang et al., 2013] K. Wang, K. Brunnström, M. Barkowsky, M. Urvoy, M. Sjöström, P. Le Callet, S. Tourancheau, and B. Andrén, "Stereoscopic 3D video coding quality evaluation with 2D objective metrics," *Proceedings of SPIE 8648*, Stereoscopic Displays and Applications XXIV

[J. Wang et al., 2013] J. Wang, M. Da Silva, P. Le Callet, and V. Ricordel, "Computational Model of Stereoscopic 3D Visual Saliency," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2151–2165

[Ward and Simmons, 2006] G. Ward and M. Simmons, "JPEG-HDR: A Backwards-compatible, High Dynamic Range Extension to JPEG," *ACM SIGGRAPH 2006 Courses*

[Ward, 1994] G. Ward, "A contrast-based scalefactor for luminance display," *Graphics Gems IV*, pp. 415–421

[Ward, 1998] G. Ward, "LogLuv Encoding for Full-Gamut, High-Dynamic Range Images," *Journal of Graphics Tools*, vol. 3, no. 1, pp. 15–31

[Watson, 1993] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," *Digest of Technical Papers*, Society for Information Display

[Weerakkody et al., 2014] R. Weerakkody, M. Mrak, V. Baroncini, J.-R. Ohm, T. K. Tan, and G. J. Sullivan, "Verification testing of HEVC compression performance for UHD video," *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*

[Wiegand et al., 2003a] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576

[Wiegand et al., 2003b] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703

[Wiegand and Schwarz, 2010] T. Wiegand and H. Schwarz, *Source coding: Part I of fundamentals of source and video coding*, Now Publishers Inc

[Wilson, 1980] H. Wilson, "A transducer function for threshold and suprathreshold human vision," *Biological Cybernetics*, vol. 38, no. 3, pp. 171–178

[Winkler, 2005] S. Winkler, *Digital Video Quality: Vision Models and Metrics*, John Wiley & Sons

[Winkler, 2009] S. Winkler, "On the properties of subjective ratings in video quality experiments," *International Workshop on Quality of Multimedia Experience (QoMEx)*

[Winkler, 2012] S. Winkler, "Analysis of Public Image and Video Databases for Quality Assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625

[Winkler and Subramanian, 2013] S. Winkler and R. Subramanian, "Overview of Eye tracking Datasets," *International Workshop on Quality of Multimedia Experience (QoMEX)*

[Woods et al., 1993] A. J. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," *Proceedings of SPIE 1915*, Stereoscopic Displays and Applications IV

[H. R. Wu and Rao, 2005] H. R. Wu and K. R. Rao, *Digital video image quality and perceptual coding*, CRC press

[D. Xu et al., 2012] D. Xu, L. Coria, and P. Nasiopoulos, "Quality of experience for the horizontal pixel parallax adjustment of stereoscopic 3D videos," *IEEE International Conference on Consumer Electronics (ICCE)*

[J. Xu et al., 2016] J. Xu, R. Joshi, and R. Cohen, "Overview of the Emerging HEVC Screen Content Coding Extension," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 50–62

[R. Yang and Z. Zhang, 2004] R. Yang and Z. Zhang, "Eye gaze correction with stereovision for video-teleconferencing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 956–960

[X. Yang et al., 2005] X. Yang, W. Ling, Z. Lu, E. Ong, and S. Yao, "Just noticeable distortion model and its applications in video coding," *Signal Processing: Image Communication*, vol. 20, no. 7, pp. 662–680

[J.-C. Yang et al., 2008] J.-C. Yang, C.-S. Wu, C.-H. Hsiao, R. Y. Tsai, and P. Hung, "Evaluation of an Eye Tracking Technology for 3D Display Applications," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*

[J. Yang et al., 2015] J. Yang, Y. Liu, Z. Gao, R. Chu, and Z. Song, "A perceptual stereoscopic image quality assessment model accounting for binocular combination behavior," *Journal of Visual Communication and Image Representation*, vol. 31, pp. 138 –145

[Yi et al., 2008] S.-Y. Yi, B. Chae, and S.-H. Lee, "Moving Parallax Barrier Design for Eye-Tracking Autostereoscopic Displays," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*

[Yoshida et al., 2005] A. Yoshida, V. Blanz, K. Myszkowski, and H.-P. Seidel, "Perceptual Evaluation of Tone Mapping Operators with Real-World Scenes," *Proceedings of SPIE 5666*, Human Vision and Electronic Imaging X

[You et al., 2010] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual Quality Assessment for Stereoscopic Images Based on 2D Image Quality Metrics and Disparity Analysis," *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*

[Yuen and H. Wu, 2005] M. Yuen and H. Wu, "Coding artifacts and visual distortions," *Digital video image quality and perceptual coding*, pp. 87–122

[Yuen and H. R. Wu, 1998] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, no. 3, pp. 247–278

[N. Zhang et al., 2003] N. Zhang, A. Vladar, M. Postek, and B. Larrabee, "A kurtosis-based statistitcal measure for two-dimensional processes and its application to image sharpness," *Proceedings Section of Physical and Engineering Sciences of American Statistical Society*

[L. Zhang et al., 2011] L. Zhang, D. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386

[Zhou et al., 2014] W. Zhou, G. Jiang, M. Yu, F. Shao, and Z. Peng, "Reduced-reference stereoscopic image quality assessment based on view and disparity zero-watermarks," *Signal Processing: Image Communication*, vol. 29, no. 1, pp. 167 –176

Philippe Hanhart
Rte de Converney 91
CH-1093 La Conversion
Switzerland
Phone: +41 (0)79 356 84 66
Email: philippe.hanhart@gmail.com

Date of birth: February 2, 1987
Nationality: Swiss
Marital status: Single

Summary

Researcher in the fields of image and video processing. Interested in measuring, predicting, and improving quality of experience in immersive video technologies, such as 3D, high dynamic range, and ultra high definition. Aspired to work on cutting edge technologies to make an impact on future multimedia systems and services.
Specialties: image/video processing, subjective/objective quality assessment, image/video compression, UHDTV, 3DTV, HDR imaging. C/C++, OpenCV, MATLAB, Python, HTML, PHP, SQL. Linux, Mac OS X, Windows.

Education

Dec. 2011 - Apr. 2016	**Doctor of Philosophy in Electrical Engineering** Multimedia Signal Processing Group (MMSPG) Swiss Federal Institute of Technology Lausanne (EPFL) Conducting research on quality of experience in immersive video technologies under the supervision of Prof. Touradj Ebrahimi
Sept. 2009 - Oct. 2011	**Master of Science in Electrical and Electronics Engineering** (GPA: 5.86/6) Major in Information Technologies, Minor in Biomedical Engineering Swiss Federal Institute of Technology Lausanne (EPFL)
Sept. 2006 - July 2009	**Bachelor of Science in Electrical and Electronics Engineering** (GPA: 5.71/6) Swiss Federal Institute of Technology Lausanne (EPFL)
Oct. 2005 - June 2006	**Special Mathematics Course** Swiss Federal Institute of Technology Lausanne (EPFL)
Sept. 2002 - July 2005	**Federal Certificate of Proficiency in Electronics and Professional Baccalaureate** College for Technicians and Vocational School of Lausanne (ETML)

Experience

June 2014 - Oct. 2014	**Dolby Laboratories Inc**, Sunnyvale, CA, USA **Video Processing Intern** in the Applied Vision Science Group - Design and implementation of a psychophysical experiment to investigate the impact of screen size on perceived luminance - Implementation of HDR-VDP-2 visual metric in C++
Dec. 2011 - Mar. 2016	**Swiss Federal Institute of Technology Lausanne (EPFL)** **Doctoral assistant** - Assistant of Image & Video Processing and Media Security courses - Supervision of Bachelor and Master students

Dec. 2011 - Mar. 2016	**Swiss Federal Institute of Technology Lausanne (EPFL)** **IT Manager** of the Multimedia Signal Processing Group (MMSPG) - Management of several Windows workstations, Linux servers, and virtual machines - Implementation and management of private file server, public file server, and web server
Feb. 2011 - Aug. 2011	**Dolby Laboratories Inc**, Burbank, CA, USA **Research assistant** in the Image Technology Group - Development of a fast block based stereo matching algorithm - Investigation of a novel view synthesis technique using sparse disparity maps
Oct. 2007 - present	**EDSI-Tech Sàrl**, Startup, Web hosting, web design, and app design, Lausanne **Co-founder, Web developer** (Oct. 2007 - July 2010)**, Consultant** (July 2010 - present) - Web design for SMEs and individuals - Development of a Unix based web hosting control panel
Sept. 2007 - Jan. 2011	**Swiss Federal Institute of Technology Lausanne (EPFL)** **Teaching assistant** - Assistant of Electrotechnics, Electronics, Electrical Systems & Electronics, Measuring Systems, Computer-aided Engineering, and Circuits & Systems courses

Projects

Partially funded by the following projects:

3DMASTER	3D Multiview Auto Stereoscopy Through Enhanced Rendering, Swiss National Foundation for Scientific Research (FN 200021-143696-1)
QoE in 3DTV	Quality of Experience in 3DTV, State Secretariat for Education, Research and Innovation (SERI C10.0132) European Network on Quality of Experience in Multimedia Systems and Services - QUALINET, COST Action IC1003
HDRi	Compression and Evaluation of High Dynamic Range Image and Video, State Secretariat for Education, Research and Innovation (SERI C12.0081) HDRi: The digital capture, storage, transmission and display of real-world lighting, COST Action IC1005
TOFuTV	Transcoders Of the Future TeleVision, Eureka Eurostars project (E!8307)
QoE-Net	innovative Quality Of Experience maNagement in Emerging mulTimedia services, Marie Skłodowska-Curie Initial Training Network (H2020-MSCA-ITN-2014 Contract n. 643072)
ImmersiaTV	Immersive Experiences around TV, an integrated toolset for the production and distribution of immersive and interactive content across devices (H2020-ICT-19-2015-IA)
VideoSense	Virtual Centre of Excellence for Ethically-guided and Privacy-respecting Video-analytics in Security, EC funded Network of Excellence (Grant Agreement Number 261743)

Honors & Awards

Oct. 2014	Top 10% paper award in IEEE International Conference on Image Processing (ICIP)
Sep. 2014	Top 10% paper award in International Workshop on Multimedia Signal Processing (MMSP)
Oct. 2011	Anna Barbara Reinhard Prize for Student Excellence from the Institution of Engineering and Technology (IET)
Sept. 2009 - Aug. 2011	EPFL Excellence Fellowship

Service

Reviewer for IEEE Transactions on Circuits and Systems for Video Technology 2016/2015/2014/2012, IEEE Transactions on Multimedia 2015/2014, IEEE Transactions on Image Processing 2016/2015/2014/2012, IEEE Journal of Selected Topics in Signal Processing 2012, IEEE Signal Processing Letters 2015, ACM Transactions on Applied Perception 2015, SPIE Journal of Electronic Imaging 2015/2014, IS&T Journal of Imaging Science and Technology 2015, Elsevier Signal Processing: Image Communication 2016/2015/2014/2013, Elsevier Journal of Visual Communication and Image Representation 2013, Springer Multimedia Systems 2015, EURASIP Journal on Image and Video Processing 2015/2014/2012, Wiley Computer Graphics Forum 2015, Journal of the Society for Information Display 2015, QoMEX 2015, ICIP 2014, ACM Multimedia 2014, GlobalSIP 2014, DSP 2014, EUSIPCO 2014, ACM NOSSDAV 2014, IVMSP 2013
Webmaster of the electrical engineering student association of EPFL and co-organizer of various activities (concert, ski week-end, ...), Sept. 2009 - July 2011

Computer Skills

Programming:	C/C++, Java, Python, HTML, CSS, Javascript, PHP, SQL
Specific Software:	MATLAB, OpenCV, FFmpeg, AviSynth, LaTeX
Operating Systems:	Linux, Mac OS X, Windows

Languages

French (Native), English (Fluent), German (Elementary)

Interests

Electronics, programming, movies, and sports: skiing, hiking, and swimming

Journal Papers

A. Artusi, R. Mantiuk, T. Ricther, P. Korshunov, P. Hanhart, T. Ebrahimi, and M. Agostinelli, "JPEG XT: A Compression Standard for HDR and WCG Images [Standards in a Nutshell]," IEEE Signal Processing Magazine, vol. 33(2), pp. 118-124, March 2016.

A. Artusi, R. Mantiuk, T. Richter, P. Hanhart, P. Korshunov, M. Agostinelli, A. Ten, and T. Ebrahimi, "Overview and Evaluation of the JPEG XT HDR Image Compression Standard," Journal of Real-Time Image Processing, pp. 1-16, December 2015.

P. Hanhart, M. Bernardo, M. Pereira, A. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for HDR image quality assessment," EURASIP Journal on Image and Video Processing, vol. 2015(39), pp. 1-18, December 2015.

E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "Modeling Immersive Media Experiences by Sensing Impact on Subjects," Multimedia Tools and Applications, pp. 1-21, October 2015.

P. Hanhart, P. Korshunov, T. Ebrahimi, Y. Thomas, and H. Hoffmann, "Subjective Quality Evaluation Of High Dynamic Range Video And Display For Future TV," in SMPTE Motion Imaging Journal, vol. 124(4), pp. 1-6, May 2015.

P. Hanhart and T. Ebrahimi, "Calculation of average coding efficiency based on subjective quality scores," Journal of Visual Communication and Image Representation, vol. 25(3), pp. 555-564, April 2014.

Conference Papers

P. Hanhart and T. Ebrahimi, "Rate-Distortion Evaluation For Two-Layer Coding Systems," 22nd IEEE International Conference on Image Processing (ICIP), Québec City, Canada, September 2015.

P. Hanhart, M. Rerabek, and T. Ebrahimi, "Towards high dynamic range extensions of HEVC: subjective evaluation of potential coding technologies," SPIE Optical Engineering + Applications, Applications of Digital Image Processing XXXVIII, San Diego, California, USA, August 2015.

P. Hanhart, C. di Nolfo, and T. Ebrahimi, "Active crosstalk reduction system for multiview autostereoscopic displays," IEEE International Conference on Multimedia and Expo (ICME), Torino, Italy, June-July 2015 (select for the Best Paper Award).

M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Quality Evaluation of HEVC and VP9 Video Compression in Real-Time Applications," 7th International Workshop on Quality of Multimedia Experience (QoMEX), Costa Navarino, Messinia, Greece, May 2015.

P. Korshunov, P. Hanhart, T. Richter, A. Artusi, R. Mantiuk, and T. Ebrahimi, "Subjective quality assessment database of HDR images compressed with JPEG XT," 7th International Workshop on Quality of Multimedia Experience (QoMEX), Costa Navarino, Messinia, Greece, May 2015.

M. Rerabek, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective and objective evaluation of HDR video compression," 9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Chandler, Arizona, USA, February 2015.

H. Nemoto, P. Korshunov, P. Hanhart, and T. Ebrahimi, "Visual attention in LDR and HDR images," 9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Chandler, Arizona, USA, February 2015.

H. Nemoto, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Impact of Ultra High Definition on Visual Attention," ACM International Conference on Multimedia, Orlando, Florida, USA, November 2014.

P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowd-based quality assessment of multiview video plus depth coding," 21st IEEE International Conference on Image Processing (ICIP), Paris, France, October 2014 (nominated as Top 10% Papers).

P. Hanhart, E. Bosc, P. Le Callet, and T. Ebrahimi, "Free-Viewpoint Video Sequences: a New Challenge for Objective Quality Metrics," 16th International Workshop on Multimedia Signal Processing (MMSP), Jakarta, Indonesia, September 2014.

T. Hossfeld, M. Hirth, P. Korshunov, P. Hanhart, B. Gardlo, C. Keimel, and C. Timmerer, "Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment," 16th International Workshop on Multimedia Signal Processing (MMSP), Jakarta, Indonesia, September 2014 (nominated as Top 10% Papers).

P. Hanhart, M. Bernardo, P. Korshunov, M. Pereira, A. Pinheiro, and T. Ebrahimi, "HDR image compression: a new challenge for objective quality metrics," 6th International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, September 2014.

P. Hanhart, N. Ramzan, V. Baroncini, and T. Ebrahimi, "Cross-lab Subjective Evaluation of the MVC+D and 3D-AVC 3D Video Coding Standards," 6th International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, September 2014.

H. Nemoto, P. Hanhart, P. Korshunov, and T. Ebrahimi, "Ultra-Eye: UHD and HD images eye tracking dataset," 6th International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, September 2014.

P. Hanhart and T. Ebrahimi, "EyeC3D: 3D video eye tracking dataset," 6th International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, September 2014.

P. Hanhart, P. Korshunov, T. Ebrahimi, Y. Thomas, and H. Hoffmann, "Subjective Quality Evaluation Of High Dynamic Range Video And Display For Future TV," International Broadcasting Convention (IBC), Amsterdam, Netherlands, September 2014.

E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "EEG correlates during video quality perception," 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, September 2014.

P. Hanhart, P. Korshunov, and T. Ebrahimi, "Subjective Evaluation of Higher Dynamic Range Video," SPIE Optical Engineering + Applications, Applications of Digital Image Processing XXXVII, San Diego, California, USA, August 2014.

P. Hanhart, P. Korshunov, and T. Ebrahimi, "Crowdsourcing evaluation of high dynamic range compression," SPIE Optical Engineering + Applications, Applications of Digital Image Processing XXXVII, San Diego, California, USA, August 2014.

E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "Predicting Subjective Sensation of Reality During Multimedia Consumption Based on EEG and Peripheral Physiological Signals," IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, July 2014.

P. Hanhart and T. Ebrahimi, "Subjective evaluation of two stereoscopic imaging systems exploiting visual attention to improve 3D quality of experience," SPIE Electronic Imaging, Stereoscopic Displays and Applications XXV, San Francisco, USA, February 2014.

A. Chappuis, M. Rerabek, P. Hanhart, and T. Ebrahimi, "Subjective evaluation of an active crosstalk reduction system for mobile autostereoscopic displays," SPIE Electronic Imaging, Stereoscopic Displays and Applications XXV, San Francisco, USA, February 2014.

E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi, "User-independent classification of 2D versus 3D multimedia experiences through EEG and physiological signals," 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, USA, January 2014.

P. Hanhart, M. Rerabek, I. Ivanov, A. Dufaux, C. Jones, A. Delidais, and T. Ebrahimi, "Automatic defect detection in video archives - Application to Montreux Jazz Festival Digital Archives," SPIE Optical Engineering + Applications, Applications of Digital Image Processing XXXVI, San Diego, USA, August 2013.

P. Hanhart and T. Ebrahimi, "On the evaluation of 3D codecs on multiview autostereoscopic display," 4th IEEE International Workshop on Hot Topics in 3D (Hot3D), San Jose, USA, July 2013.

P. Hanhart, P. Korsunov, M. Rerabek, and T. Ebrahimi, "JPEG backward compatible format for 3D content representation," 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS), Paris, France, July 2013.

E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi, "A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed data," 5th International Workshop on Quality of Multimedia Experience (QoMEX), Klagenfurt am Wörthersee, Austria, July 2013.

P. Hanhart, P. Korshunov, and T. Ebrahimi, "Benchmarking of quality metrics on ultra-high definition video sequences," 18th International Conference on Digital Signal Processing (DSP), Santorini, Greece, July 2013.

P. Hanhart and T. Ebrahimi, "Predicting 3D Quality based on Content Analysis," 11th IEEE IVMSP Workshop: 3D Image/Video Technologies and Applications, Seoul, Korea, June 2013.

P. Hanhart, M. Rerabek, P. Korshunov, and T. Ebrahimi, "Subjective evaluation of HEVC intra coding for still image compression," 7th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, USA, January 2013.

P. Hanhart and T. Ebrahimi, "Quality Assessment of a Stereo Pair Formed From Two Synthesized Views Using Objective Metrics," 7th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, USA, January 2013.

P. Hanhart and T. Ebrahimi, "Quality Assessment of a Stereo Pair Formed From Decoded and Synthesized Views Using Objective Metrics," 3DTV-Conference: The True Vision Capture, Transmission and Display of 3D Video (3DTV-CON), Zurich, Switzerland, October 2012.

P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of the upcoming HEVC video compression standard," SPIE Optical Engineering + Applications, Applications of Digital Image Processing XXXV, San Diego, USA, August 2012.

P. Hanhart, F. De Simone, and T. Ebrahimi, "Quality Assessment of Asymmetric Stereo Pair Formed From Decoded and Synthesized Views," 4th International Workshop on Quality of Multimedia Experience (QoMEX), Yarra Valley, Australia, July 2012.

Contributions to Standardization

P. Hanhart and T. Ebrahimi, "New Test Images for JPEG XS Evaluations," ISO/IEC JTC1/SC29/WG1, Doc. M71032, La Jolla, USA, February 2016.

P. Hanhart and T. Ebrahimi, "HDR CE3: Benchmarking of objective metrics for HDR video quality assessment," JCT-VC, Doc. JCTVC-W0091, San Diego, USA, February 2016.

M. Rerabek, P. Hanhart, and T. Ebrahimi, "HDR CE3: Results of subjective evaluations conducted with the DSIS method," JCT-VC, Doc. JCTVC-W0090, San Diego, USA, February 2016.

P. Hanhart, M. Rerabek, and T. Ebrahimi, "Results of HDR CfE subjective evaluations conducted at EPFL," ISO/IEC JTC1/SC29/WG11, Doc. M36728, Warsaw, Poland, July 2015.

P. Hanhart and T. Ebrahimi, "HDR CfE Subjective Evaluations at EPFL," ISO/IEC JTC1/SC29/WG11, Doc. M36168, Lausanne, Switzerland, May 2015.

P. Hanhart and T. Ebrahimi, "Benchmarking of LDR metrics for HDR video quality assessment," ISO/IEC JTC1/SC29/WG11, Doc. M35469, Geneva, Switzerland, February 2015.

P. Hanhart, T. Ebrahimi, S. Daly, and W. Husak, "Benchmarking of HDR-VDP-2 for HDR video quality assessment," ISO/IEC JTC1/SC29/WG11, Doc. M35307, Strasbourg, France, October 2014.

M. Rerabek, P. Korshunov, P. Hanhart, and T. Ebrahimi, "Correlation of subjective scores and objective metrics for HDR video quality assessment," ISO/IEC JTC1/SC29/WG11, Doc. M35273, Strasbourg, France, October 2014.

V. Baroncini, G. S. Blasi, T. Ebrahimi, P. Hanhart, N. Razam, and I. Zupancic, "Report of the formal subjective assessment of the Submission in response to the Joint Call for Proposal (JCfP) for new technologies in the area of Screen Content Coding (SCC)," JCT-VC, Valencia, Spain, March 2014.

P. Hanhart and T. Ebrahimi, "AHG9: On the evaluation on stereoscopic and multiview autostereoscopic displays," JCT-3V, Doc. JCT3V-D0217, Incheon, Korea, April 2013.

P. Hanhart, M. Rerabek, P. Korshunov, and T. Ebrahimi, "AhG4: Subjective evaluation of HEVC intra coding for still image compression," JCT-VC, Doc. JCTVC-L0380, Geneva, Switzerland, January 2013.

P. Hanhart and T. Ebrahimi, "AHG9: Accuracy of 2D metrics for 3D video quality assessment," JCT-3V, Doc. JCT3V-B0180, Shanghai, China, October 2012.

P. Hanhart and T. Ebrahimi., "3DV: Quality assessment of stereo pairs formed from two synthesized views," JCT-3V, Doc. JCT3V-A0150, Stockholm, Sweden, July 2012.

P. Hanhart, F. De Simone, and T. Ebrahimi, "3DV: Alternative metrics to PSNR," ISO/IEC JTC1/SC29/WG11, Doc. M24807, Geneva, Switzerland, May 2012.

P. Hanhart, F. De Simone, M. Rerabek, and T. Ebrahimi, "3DV: Objective quality measurement for the 2-view case scenario," ISO/IEC JTC1/SC29/WG11, Doc. M23908, San Jose, USA, February 2012.

Book Chapters

A. Tabatabai, T. Suzuki, P. Hanhart, P. Korshunov, T. Ebrahimi, M. Horowitz, F. Kossentini, and H. Tmar, "Compression Performance Analysis in HEVC," High Efficiency Video Coding (HEVC) - Algorithms and Architectures, p. 275-302, Springer, 2014.

P. Hanhart, F. De Simone, M. Rerabek, and T. Ebrahimi, "3D Video Quality Assessment," in Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering, p. 378-393, John Wiley & Sons Inc, 2013.