Visual Saliency in Image Quality Assessment

Wei Zhang

School of Computer Science and Informatics

Cardiff University

A thesis submitted in partial fulfilment

of the requirement for the degree of

Doctor of Philosophy

March 2017

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed . . . . (candidate)

Date . . . .

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed . . . . (candidate)

Date . . . .

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed . . . . (candidate)

Date . . . .

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed . . . . (candidate)

Date . . . .

To my family
for their love and support

Abstract

Advances in image quality assessment have shown the benefits of modelling functional components of the human visual system in image quality metrics. Visual saliency, a crucial aspect of the human visual system, is increasingly investigated recently. Current applications of visual saliency in image quality metrics are limited by our knowledge on the relation between visual saliency and quality perception. Issues regarding how to simulate and integrate visual saliency in image quality metrics remain. This thesis presents psychophysical experiments and computational models relevant to the perceptually-optimised use of visual saliency in image quality metrics. We first systematically validated the capability of computational saliency in improving image quality metrics. Practical guidance regarding how to select suitable saliency models, which image quality metrics can benefit from saliency integration, and how the added value of saliency depends on image distortion type were provided. To better understand the relation between saliency and image quality, an eye-tracking experiment with a reliable experimental methodology was first designed to obtain ground truth fixation data. Significant findings on the interactions between saliency and visual distortion were then discussed. Based on these findings, a saliency integration approach taking into account the impact of distortion on the saliency deployment was proposed. We also devised an algorithm which adaptively incorporate saliency in image quality metrics based on saliency dispersion. Moreover, we further investigated the plausibility of measuring image quality based on the deviation of saliency induced by distortion. An image quality metric based on measuring saliency deviation was devised. This thesis demonstrates that the added value of saliency in image quality metrics can be optimised by taking into account the interactions between saliency and visual distortion. This thesis also demonstrates that the deviation of fixation deployment due to distortion can be used as a proxy for the prediction of image quality.

Acknowledgements

The work in this thesis would not be possible without the support and help from so many people.

I would firstly like to express my sincere appreciation to my supervisors, Dr. Hantao Liu and Prof. Ralph R. Marin. Their advices on both my research and career have been invaluable. I would also like to thank Prof. Zhou Wang, Prof. Patrick Le Callet, Dr. Xianfang Sun and Dr. Steven Schockaert for their insightful comments on my research. Your priceless advices enable me to examine my research from various perspectives.

My thanks also go to my colleagues Lucie Lévěque and Juan Vicente Talens-Noguera, for the fun we had and the support they offered. I will never forget the inspiring discussion we had in our lab. I will always cherish the memory of our trip to Canada, American and Netherlands. As a part of my work is related to subjective eye-tracking experiment, I would like to thank all the participants for their time and efforts. Particularly, my thanks go to my colleague Juan Vicente Talens-Noguera for conducting the experiment with me.

I would especially like to thank my PhD funding institutions, China Scholarship Council (CSC) and the School of Computer Science & Informatics, Cardiff University. I am grateful for the scholarship that allowed me to pursue my study.

Last but not the least, my deep gratitude goes to my family and my friends. All of them have been there supporting me during the past three years and I dedicate this thesis to them.

Contents

List of Publications

The work introduced in this thesis is based on the following peer-reviewed publications. More specifically,

Chapter 3 is based on:

W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: a statistical evaluation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1266-1278, June 2016.

W. Zhang, Y. Tian, X. Zha and H. Liu, "Benchmarking state-of-the-art visual saliency models for image quality assessment," *in Proc. of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China, March, 2016, pp. 1090-1094.

W. Zhang, A. Borji, F. Yang, P. Jiang, and H. Liu, "Studying the added value of computational saliency in objective image quality assessment," *in Proc. of the IEEE International Conference on Visual Communication and Image Processing*, Valletta, Malta, Dec. 2014, pp. 21-24.

Chapter 4 is based on:

W. Zhang and H. Liu, "Towards a reliable collection of eye-tracking data for image quality research: challenges, solutions and applications," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2424-2437, May 2017.

W. Zhang and H. Liu, "SIQ288: a saliency dataset for image quality research," *in Proc. of the 18th International Workshop on Multimedia Signal Processing*, Montreal, CA, Sept. 2016, pp. 1-6.

W. Zhang and H. Liu, "Saliency in objective video quality assessment: what is the ground truth?," *in Proc. of the 18th International Workshop on Multimedia Signal Processing*, Montreal, CA, Sept. 2016, pp. 1-5.

Chapter 5 is based on:

W. Zhang and H. Liu, “Study of saliency in objective video quality assessment,” *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1275-1288, March 2017.

W. Zhang, J. V. Talens-Noguera and H. Liu, “The quest for the integration of visual saliency models in objective image quality assessment: a distraction power compensated combination strategy,” *in Proc. of the 22nd IEEE International Conference on Image Processing*, Quebec City, CA, Sept. 2015, pp. 1250-1254.

Chapter 6 is based on:

W. Zhang, R. R. Martin and H. Liu, “A saliency dispersion measure for improving saliency-based image quality metrics,” *IEEE Transactions on Circuits and Systems for Video Technology*, in press, DOI: 10.1109/TCSVT.2017.2650910

Chapter 7 is based on:

W. Zhang and H. Liu, “Learning Picture Quality from Visual Distraction: Psychophysical Studies and Computational Models,” *Neurocomputing*, vol. 247, pp. 183-191, July 2017.

In addition, I co-authored the following two papers which are closely related to visual quality assessment, but are not integrated in this thesis:

J. V. Talens-Noguera, **W. Zhang** and H. Liu, “Studying human behavioural responses to time-varying distortions for video quality assessment,” *in Proc. of the 22nd IEEE International Conference on Image Processing*, Quebec City, QC, 2015, pp. 651-655.

U. Engelke, **W. Zhang**, P. Le Callet and H. Liu, “Perceived interest versus overt visual attention in image quality assessment,” *in Proc. SPIE, Human Vision and Electronic Imaging XX*, March 2015, pp. 93941H-93941H-9.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The past decades have witnessed a significant growth in using digital image stimuli as a means of information representation and communication. In current digital image processing and communication systems, image signals are subject to various distortions due to causes such as acquisition errors, lossy data compression, noisy transmission channels and limitations in image rendering devices. The ultimate image content received by the human visual system (HVS) differs in image quality depending on the system and its underlying implementation. The undesired image quality degradation may affect the visual experiences of the end user or lead to interpretation errors in visual inspection tasks [1]. Finding ways to effectively control and improve image quality has become a focal concern in both academia and industry [2]. Therefore, considerable efforts were made on appropriately tuning the parameters of image processing systems in order to enhance the image quality. While controlling the parameters of image processing systems is important for achieving high image quality, it is more crucial to evaluate image quality from the users' perspective which is known as the subjective quality of experience (QoE) [3].

The subjective QoE can be directly measured by conducting subjective user study. Standardised subjective experimental methodologies have been proposed by the Radiocommunication Sector of the International Telecommunication Union (ITU-R) [4]. Though subjective test is regarded as the most accurate way of measuring QoE, it naturally has several disadvantages. First, the subjective test is expensive in terms of time and money. In addition, the results of the subjective QoE experiment collected in the laboratory environment may be inapplicable to the image quality assessment in real-world applications [3]. Moreover, the subjective test is impractical for any real-time applications.

To reduce the cost of subjective experiment and to facilitate the image quality assessment in real-world applications, image quality metrics (IQMs) — computational models for automatic assessment of perceived quality — have emerged as an important tool for the optimisation of modern imaging systems [5]. The performance of these IQMs is evaluated against the results of subjective test in order to check how well they can predict human scores. Nowadays, various IQMs are widely available in many imaging systems in a broad range of applications, e.g., for fine tuning image and video processing pipelines, evaluating image and video enhancement algorithms, and quality monitoring and control of displays. Substantial progress has been made on the development of IQMs over the last several decades, and many successful models have been devised. However, recent research shows that they demonstrate a lack of sophistication when it comes to dealing with real world complexity [6, 7, 8]. This makes image quality assessment a continued problem of interest. The fundamental challenge intrinsically lies in the fact that our knowledge about how the HVS assesses image quality and how to express that in an efficient mathematical model remains rather limited. Being able to reliably predict image quality as perceived by humans requires a better understanding of functional aspects of the HVS relevant to image quality perception, and optimal use of that to improve existing IQMs or devise more rigorous algorithms for IQMs.

Advanced IQMs benefit from embedding models of the HVS, such as contrast sensitivity function [9] and visual masking [10]. Recently, a growing trend in image quality research is to investigate how visual attention — a mechanism that allows the HVS to effectively select the most relevant information in a visual scene — plays a role in judging image quality. More specifically, the bottom-up stimulus-driven part of this attentional mechanism, i.e., visual saliency, is increasingly studied in relation to image quality. It is inferred that distortion in the salient areas is more annoying than that in the non-salient areas [11]. To understand whether this idea can be used to improve IQMs, initial effort has been made in the literature to investigate the added value of visual attention in IQMs by incorporating visual saliency models. Depending on the choice of saliency models and IQMs, some research findings revealed that the benefits of adding saliency to IQMs are marginal, whilst some research findings reported that saliency could significantly improve IQMs. Many saliency models and IQMs are available, the following issues such as how the benefits of inclusion of computational saliency in IQMs vary and what are the causes of this variation remain, which are worth further investigation.

Due to our limited understanding of the relation between visual attention and image quality, state of the art IQMs mainly focus on weighting local distortions (calculated by an IQM) with local saliency (resulted from a computational saliency model), yielding a more sophisticated means of image quality prediction. This concept, however, strongly relies on the simplification of the HVS that the visual attention aspects and the perception of local distortions are first treated separately and they are then combined artificially to determine the overall quality. The actual interactions between visual attention and image quality, however, are not considered. This simple combination of saliency and an IQM may downplay the importance of saliency in IQMs. It is highly desirable to investigate a perceptually optimised combination approach of adding saliency information to IQMs.

However, determining optimal use of visual attention aspects in IQMs is not straightforward [11]. The main challenge lies in the fact that how human attention affects image quality perception, and how to precisely simulate relevant functional aspects of the HVS in IQMs are not fully understood. To gain more knowledge of human vision, psychophysical studies have been undertaken to better understand visual attention aspects in relation to image quality assessment via eye-tracking [12, 13, 14, 15]. In general, these eye-tracking studies have shown that distortion occurring in an image alters gaze patterns relative to that of the image without distortion, and that the extent of the alteration tends to depend on several factors. These studies, unfortunately, are heavily limited by the choices made in their experimental design such as the use of a limited stimulus variability [13], an insufficient number of subjects [12], and the involvement of massive stimulus repetition [14]. Therefore, the conclusions of these studies are either biased or hardly reveal statistically sound findings. To ensure the validity and generalisability of empirical evidence, it is desirable to investigate a more reliable methodology for collecting eye-tracking data for the purpose of image quality study.

In addition, a previous eye-tracking study [13] has shown that the deployment of fixations changes as a result of the appearance of visual distortions, and that the extent of the changes seems to be related to the strength of distortion. From this, it can be inferred that the changes of gaze patterns driven by distortion might be correlated with the variation in perceived quality of natural images. Therefore, it is worth investigating the plausibility of directly using the deviation of saliency as the proxy for image quality prediction.

1.2 Hypotheses and Objectives

This thesis is based on the following hypotheses:

- IQMs benefit from the addition of computational saliency, and the benefits of adding a saliency term to an IQM can be further improved by taking into account the interactions between saliency and local distortions.
- Gaze is affected by distortion, and the deployment of fixations of a distorted image differs from that of the original image without distortion. The deviation of fixation deployment can be used as a proxy for the prediction of image quality.

To validate the first hypothesis, following objectives are set in this thesis:

- To statistically assess the added value of various computational saliency models in IQMs.

- To investigate the interactions between visual saliency and image distortions via eye-tracking.
- To improve saliency-based IQMs by taking into account the interactions between visual saliency and local distortions.

To validate the second hypothesis, following objectives are set in this thesis:

- To investigate the relationship between the deviation of fixation patterns driven by distortion and the perceived quality of natural images.
- To devise an IQM that is based on the measure of visual saliency deviation.

1.3 Contributions

This thesis presents the following contributions:

- We have conducted an exhaustive statistical evaluation to investigate the added value of incorporating computational saliency in IQMs and how that depends on various factors. Knowledge as the outcome of this evaluation is highly beneficial for the image quality research community to have a better understanding of saliency inclusion in IQMs. The evaluation also provides useful guidance for saliency incorporation in terms of the effect of saliency model dependency, IQM dependency and image distortion type dependency.
- We have built a reliable eye-tracking database for image quality research. We implemented dedicated control mechanisms in the experimental methodology to effectively eliminate potential bias due to the involvement of stimulus repetition. The resulting eye-tracking data provide insights into how visual attention behaviour is affected by visual distortions and how to optimise the inclusion of saliency in IQMs.
- We have proposed a new algorithm for the combination of saliency and IQMs by taking into account the distraction power of local distortions. The proposed algorithm explicitly includes the interactions of visual saliency and distortion, outperforming the conventionally used combination approach in terms of improving the performance of IQMs.
- We have proposed a new algorithm for reliably measuring the degree of saliency dispersion and used it to adaptively incorporate computational saliency in IQMs. We demonstrated that adaptive use of saliency according to saliency dispersion significantly outperforms fixed use of saliency in improving the performance of IQMs.

- We have conducted a dedicated eye-tracking experiment to investigate the relationship between the deviation of fixation patterns driven by distortion and the perceived quality of natural images. We demonstrated that the two variables mentioned above are highly correlated, which provides an empirical foundation for predicting image quality directly by the measurement of saliency deviation.

- We have devised a new IQM which is based on measuring saliency deviation between a distorted image and its reference. Experimental results show that the proposed IQM is among the best performing IQMs while at relatively low computational cost in the literature.

1.4 Thesis Organization

- Chapter 2 introduces the background knowledge regarding to image quality assessment and visual attention. The state of the art and challenges in the application of saliency information in image quality assessment are also presented.

- Chapter 3 details the statistical evaluation that investigates the capability of various computational saliency in improving the performance of IQMs. The relationship between how well a saliency model can predict human fixations and to what extent an IQM can profit from adding this saliency model is also explored. This chapter also assesses dependencies of the performance gain that can be achieved by including saliency in IQMs. Practical issues regarding the application of saliency models in IQMs are discussed.

- Chapter 4 describes the conduct of a large-scale eye-tracking experiment which aims to better understand visual saliency in relation to image quality assessment. A new experimental methodology is proposed and used in order to improve the reliability of eye-tracking data. Based on the resulting eye-tracking data, the impact of image distortions on human fixations is assessed. This chapter also discusses the optimal use of saliency in IQMs.

- Chapter 5 follows up the research conducted in Chapter 4 and describes a new algorithm that combines saliency and local distortions by taking into account the interactions between them.

- Chapter 6 investigates the content-dependent nature of the benefits of saliency inclusion in IQMs and presents a saliency dispersion measure which can be used to adaptively incorporate saliency models in IQMs.

- Chapter 7 explores the relation between the deviation of fixation patterns driven by distortion and the perceived quality of natural images, via an eye-tracking experiment. This chapter also discusses the case of replacing eye-tracking data with computational saliency.
- Chapter 8 presents a new IQM that is based on measuring the deviations of visual saliency features.
- Chapter 9 summarises the main conclusions of the thesis and discusses potential directions for future research.

Chapter 2

Background

2.1 Image Quality Assessment

Digital images usually undergo various phases of signal processing for the purpose of storage, transmission, rendering, printing or reproduction [1]. As a consequence, images are often subject to distortions at every stage of the processing chain, resulting in various types of artifacts or transmission errors. To prevent the appearance of visual distortions and to optimise the digital imaging chain, modelling image quality is essential.

Traditionally, image quality is assessed subjectively by human observers. In the subjective image quality assessment, a number of human subjects are requested to rate the perceived quality of images in a carefully controlled environment. This methodology is considered as the most reliable way of assessing image quality, since human beings are the ultimate receivers of most visual information. However, subjective assessment is expensive, time-consuming and most importantly, unrealistic for practical applications. The increasing demand for digital visual media has pushed to the forefront the need for computational algorithms that can predict image quality as perceived by humans. These algorithms are referred to as objective image quality metrics (IQMs). In the past few decades, many IQMs have been proposed and they are now serving as an important tool in digital imaging systems to benchmark the performance of image processing algorithms off-line, to monitor image quality in real-time and to improve the design and testing phases of image processing products.

2.1.1 Subjective image quality assessment

Subjective image quality assessment is important as it provides ground truth on how human visual system (HVS) judges image quality. After quality scoring by human subjects, a single score — mean opinion score (MOS) — representing the perceived quality of an image is obtained by pooling the individual subjective ratings. Alternatively, the final score can also be interpreted as a differential mean opinion scores (DMOS), which represents the difference in

MOS between the distorted image and its corresponding reference. An image of higher perceived quality corresponds to a greater value of MOS or a smaller value of DMOS. Standardised methodologies for the subjective assessment of the quality of natural images do exist, such as the Radiocommunication Sector of International Telecommunication Union (ITU-R) BT.500-13 [4]. This document establishes methodologies including viewing conditions (e.g., viewing environment, monitor set-up and selection of test material), rating methods (e.g., experimental procedure), and raw data processing (e.g., outlier screening and data pooling).

Representative rating methods in (ITU-R) BT.500-13 contain Double Stimulus Continuous Scale (DSCQS) and Single Stimulus Continuous Quality Evaluation (SSCQE). In DSCQS, both the source stimulus and its distorted stimulus are shown to the observers for rating their quality. The difference between two ratings is used to represent the quality of the distorted stimulus. This method is often adopted to measure the quality of a visual signal processing system relative to a pre-defined reference. In SSCQE, only the distorted stimuli are shown to the observers for quality rating as an attempt to reproduce the real-world viewing conditions where reference is normally unavailable. It should, however, be noted that each method documented so far contains advantages and disadvantages, and therefore, users should choose an appropriate method based on their own application environments. For example, double stimulus method is found to be more stable than single stimulus method for assessing small impairments due to that observers are easier to detect the impairment in the presence of reference images. In contrast, single stimulus method is of practical relevance in the circumstance where no reference is available.

In the meanwhile, research in image quality assessment has also lead to the emergence of various publicly available image quality databases. These databases can be used to benchmark the performance of IQMs. A typical image quality database usually contains a number of reference images, and for each reference there exist several distorted versions including various distortion types and distortion levels. The database also gives a MOS/DMOS for each stimulus. There are about twenty image quality databases in the literature, among which LIVE [16], CSIQ [17], TID2013 [18], IVC [19] and MICT [20] are the most widely used databases. The reliability of the above-mentioned databases is widely recognised in the image quality community since they were collected using standardised methodologies in controlled experimental environment [21]. In this thesis, we also use these five image quality databases for assessing the performance of IQMs to ensure an unbiased performance evaluation. Moreover, these databases are among the largest image quality databases in the literature in terms of stimulus variability especially for TID2013, CSIQ and LIVE. Additionally, as most of the IQMs are benchmarked on these databases, a direct comparison between the proposed IQMs in this thesis and other IQMs in the literature can be made immediately if we also use these databases. A detailed comparative study of some well-established databases was conducted in [21] regarding e.g., the composition of stimuli, experimental design and subjective rating. We summarize the details of these

Table 2.1: A comparison of five widely used image quality databases.

	No. of ref. images	No. of dist. images	No. of dis. types	No. of subjects
LIVE	29	799	5	29
TID2013	25	3000	24	971
CSIQ	30	866	6	35
IVC	10	185	5	15
MICT	14	168	2	27

databases as below and list their main features in Table. 2.1.

The LIVE database consists of 779 images distorted with five distortion types, i.e., JPEG compression (i.e., JPEG), JPEG2000 compression (i.e., JP2K), white noise (i.e., WN), Gaussian blur (i.e., GBLUR) and simulated fast-fading Rayleigh occurring in (wireless) channels (i.e., FF). Per image the database also gives a differential mean opinion score (DMOS) with a scale of zero to one hundred. The resolution of the images ranges from 634 $\times$ 438 to 768 $\times$ 512 pixels. The subjective ratings were obtained from 29 participants.

The TID2013 database is currently the largest database in the literature. It consists of 3000 distorted images derived from 25 reference images. There are 24 distortion types in the database, namely additive Gaussian noise (AGN), additive noise in color components (ANC), spatially correlated noise (SCN), masked noise (MN), high frequency noise (HFN), impulse noise (IN), quantization noise (QN), Gaussian blur (GB), image denoising (DEN), JPEG compression (JPEG), JPEG2000 compression (JP2K), JPEG transmission errors (JGTE), JPEG2000 transmission errors (J2TE), non eccentricity pattern noise (NEPN), local block-wise distortions (Block), mean shift (MS), contrast change (CTC), change of color saturation (CCS), multiplicative Gaussian noise (MGN), comfort noise (CN), lossy compression of noisy images (LCNI), image color quantization with dither (ICQD), chromatic aberrations (CHA) and sparse sampling and reconstruction (SSR). Per reference image, there are five distorted versions for each distortion type. All the stimuli in the TID2013 database are at a resolution of 512 $\times$ 384. The subjective ratings were obtained from 971 participants.

The CSIQ database consists of 866 distorted images derived from 30 reference images with each at a resolution of 512 $\times$ 512. It contains 5 distortion types, namely additive Gaussian white noise (AGWN), JPEG compression (JPEG), JPEG2000 compression (JP2K), additive Gaussian pink noise (AGPN), Gaussian blurring (GB) and global contrast decrements (GCD). The rating scores were obtained from 35 participants.

The IVC database consists of 185 distorted images derived from 10 reference images with each at a resolution of 512 $\times$ 512. More specifically, there are 20 images distorted with Gaussian blur, 50 images distorted with JPEG compression, 25 images distorted with JPEG compression of the luminance channel only, 50 images distorted with JPEG2000 compression and 40 images distorted with locally adaptive-resolution coding. The subjective ratings were obtained from 15 participants.

The MICT database consists of 168 distorted images derived from 14 reference images with each at a resolution of 768 $\times$ 512. It contains two distortion types: JPEG compression artifacts and JPEG2000 compression artifacts with each corresponding to 84 distorted images. The rating scores were obtained from 27 participants.

2.1.2 Objective image quality assessment

In the field of signal processing, signal fidelity metrics, e.g. mean square error (MSE) and peak signal-to-noise ratio (PSNR) are commonly used to objectively assess the signal quality. They remain widely used due to their simplicity and generalizability for implementation. However, these metrics usually show unsatisfactory performance when handling visual signals such as images and videos, and they have been long criticized for their inconsistency with how humans judge image quality [22]. The main reason account for this poor correlation between the objective measurements and human judgements is that these signal fidelity metrics are based on several implicit assumptions which may not be true for visual signals. For example, PSNR assumes that the image signals and distortions are independent, and the perceptual quality is purely determined by distortions independent of image content. Another assumption is that the perceived quality is independent of the spatial locations of distortion.

To improve the performance of objective image quality assessment, a lot of effort has been made on designing IQMs that take into account the way humans perceive image quality. The IQMs available in the literature differ in their application, ranging from metrics that assess a specific type of visual distortion to those that evaluate the overall image quality. These IQMs can be generally classified into two categories, namely the perception-driven metrics and the signal-driven metrics. The former attempts to simulate relevant functional components of the HVS, while the latter focuses on visual signal analysis.

The goal of the perception-driven IQMs is to come close to the behaviour of the HVS. Advances in human vision research have increased our understanding of the mechanisms in the HVS, and thus allowed integrating these psychophysical findings in designing IQMs [23, 24, 25]. Some well-established models that address the low-level aspects of early vision, such as contrast sensitivity [9], visual masking [10], luminance adaptation [26] and foveated vision [5] have been implemented in IQMs. Popular IQMs include Visual Signal-to-Noise Ratio (VSNR) [27], Most Apparent Distortion (MAD) [28] and the Noise Quality Measure (NQM) [29]. We hereby briefly introduce these IQMs as below:

- VSNR is inspired by the psychophysical experiments related to the detectability of distortions. A contrast threshold is modelled to determine the visibility of distortions in natural images. If the distortions are below the threshold, the quality of the distorted image is considered to be perfect. If the distortions are detectable by the HVS, the strength of distortions is quantified by the Euclidean distance between two image features of the reference and distorted images.
- MAD measures the image quality with two separate strategies based on the characteristics of the HVS. For images with high quality, MAD mimics how the HVS perceives visual artifacts in the presence of the image content whereas for images with low quality, MAD simulates how the HVS recognises image content in the presence of distortions.
- NQM is inspired by the psychophysical findings that frequency distortions and additive noise have independent effects on the visual quality perception. Thus, NQM decouples all distortions into these two forms and quantifies their impact on the HVS separately. Then, the final quality prediction is computed by integrating the two measures.

These HVS-based IQMs have been proven more reliable than the traditional signal fidelity metrics. Nevertheless, the perception-driven modelling approach remains limited in its sophistication and thus in its performance, mainly due to the fact that our knowledge of the HVS is limited and that it is impossible to precisely simulate all perception-related aspects in the HVS.

Instead of imitating the functional operations of the HVS, the signal-driven approach treats the HVS as a black box. This approach is usually concerned with the overall functionality of the HVS, and concentrates on image statistics as well as analysis of distortions. Many IQMs based on this philosophy have been devised and demonstrated rather effective in predicting image quality. Representative IQMs in this category include the universal quality index (UQI) [30], the structural similarity index (SSIM) [31], the multi-scale SSIM (MS-SSIM) [32], the information content weighting PSNR (IWPSNR) [33], the information content weighting SSIM (IWSSIM) [33], the visual information fidelity (VIF) [34], and the feature similarity index [35], the generalized block-edge impairment metric (GBIM) [36], the no-reference perceptual blur metric (NPBM) [37], the just noticeable blur metric [38], and the no-reference blocking artifact measure [39]. We hereby briefly introduce these IQMs as below:

- UQI measures the image quality degradation as a combination of the loss of pixel correlation, luminance and contrast.

- SSIM is based on the observation that the HVS is highly adapted to extract structural information from a visual scene. Thus, SSIM attempts to measure image quality by quantifying the structural similarity between a distorted image and its original version.

- MS-SSIM represents a refined and flexible version of the single-scale SSIM, incorporating the variations of viewing conditions.

- Based on the hypothesis that the importance of the locally measured distortion is proportional to the local information content, IWPSNR was proposed by extending PSNR with an extra weighting process to refine the relative importance of local distortions.

- Similarly, IWSSIM was also devised by refining the local distortion measured by SSIM.

- VIF aims to assess image quality using natural scene statistics. The shared information between an original image and its distorted version is used to measure the quality of the distorted image.

- Based on the assumption that phase congruency and gradient magnitude play complementary roles in characterising local image quality, FSIM predicts image quality by measuring the deviations of these two features between an original image and its distorted version.

- GBIM measures the quality of images that are distorted with blocking artifacts as an inter-pixel difference across block boundaries.

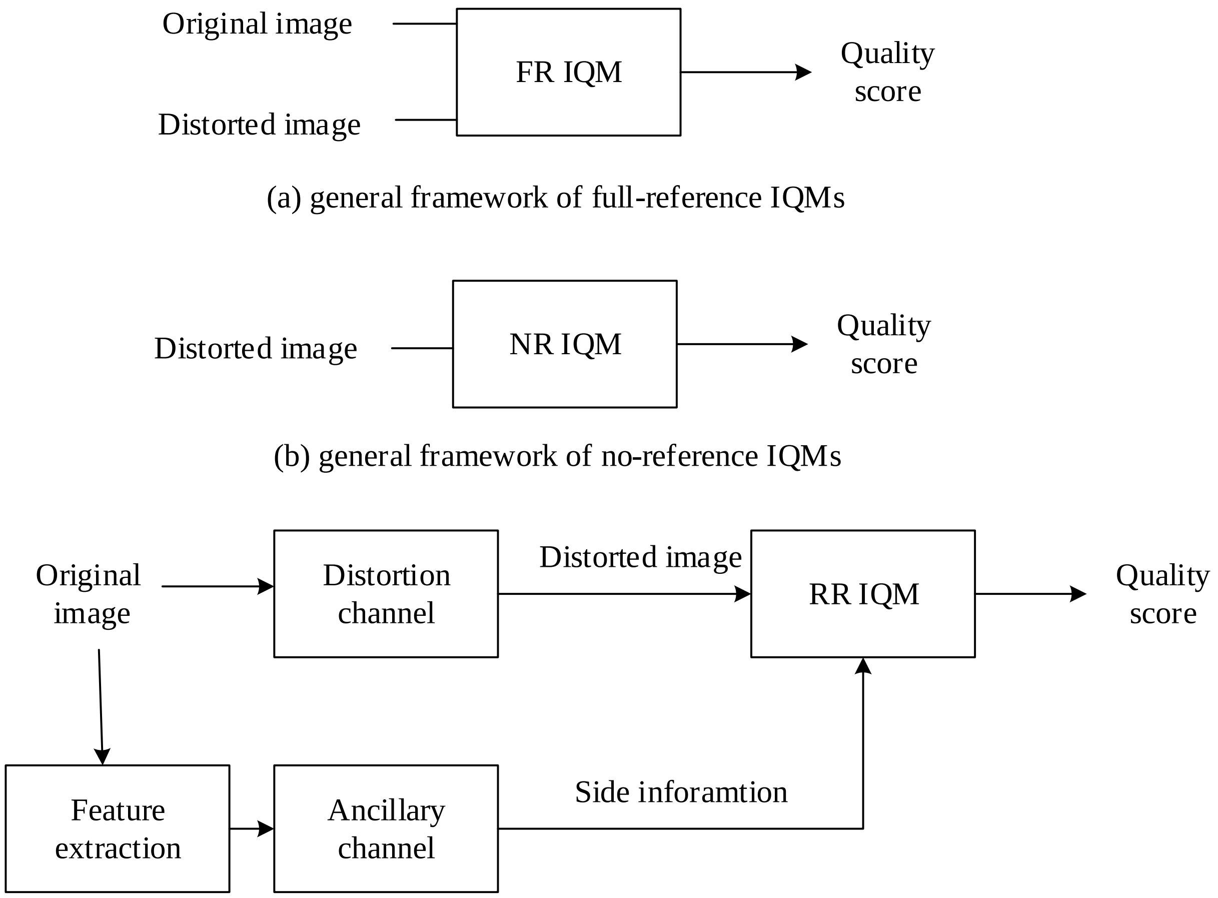
- NBAM considers the visibility of the blocking artifacts by computing the local contrast in gradient.

- NPBM measures the quality of blurring images based on extracting sharp edges in an image and measuring the width of these edges.

- JNBM refines the measurement of the spread of the edges by integrating the concept of just noticeable blur.

In general, compared to the perception-driven IQMs, signal-driven IQMs provide simplified solutions which can be easily embedded in real-time applications. Additionally, signal-driven IQMs do not rely on the success of modelling the rather complex HVS. However, it should be noted that the effectiveness of the signal-driven IQMs depends on the relevance of prior knowledge of image statistics.

It should be noted that there have been a variety of IQMs that are based on machine learning techniques. They have become an emerging category of IQM apart from the signal-driven IQMs



(a) general framework of full-reference IQMs

(b) general framework of no-reference IQMs

(c) general framework of reduced-reference IQMs

Figure 2.1: General frameworks of full-reference (FR), reduced-reference (RR) and no-reference (NR) metrics.

and perceptual-driven IQMs. Generally, these learning-based IQMs extract features from image first, and then use machine learning methods to map the image features to a single quality score. These learning-based IQMs are not considered in this thesis. How to effectively apply visual saliency as an image feature in learning-based IQMs is worth investigating and may be considered in future work.

IQMs can also be classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) metrics, depending on to what extent the quality assessment algorithms rely on the undistorted reference. Figure 2.1 illustrates the general frameworks for IQMs in each category. FR metrics require the full access to the reference and are generally implemented using the framework as shown in Fig. 2.1(a). They assume that the undistorted reference image exists and is fully available. These IQMs are also called image similarity or fidelity measurement since the quality scores predicted by these IQMs are based on quantifying the similarity or difference between the reference image and the distorted image. In contrast, NR metrics attempt to predict the perceived quality solely based on the distorted image. The general framework of the IQMs in this category is illustrated in Fig. 2.1(b). NR metrics can be further classified into general-purpose NR IQMs and distortion-specific NR IQMs. General-purpose NR metrics aim to measure the quality of images without any information from the distortion. Most of these metrics are based on feature extraction and training on subjective quality scores. On the other hand, distortion-specific NR metrics focus on a specific type of distortion, e.g., JPEG/JPEG2000 compression artifacts, ringing or blurring, and characteristics of specific distortions are utilized to increase the performance of NR IQMs. In the scenarios where the reference is partially available (e.g., in complex communication networks), RR metrics are meant to assess image quality with partial information extracted from the reference (e.g., some image features). Figure 2.1(c) shows the processing pipeline for RR image quality assessment systems. At the sender's side, some image features are extracted from the original undistorted images. These extracted features are then transmitted to the receiver's side through an ancillary channel as side information. This information is later used to assist the quality assessment of the image transmitted through the distorted channel. Generally, FR metrics achieve higher performance than RR and NR metrics due to the availability of extra information extracted from reference images [40]. However, the requirement for the access of reference images may limit the deployment of FR metrics in certain applications.

It is worth noting that the IQA framework can be easily extended to a video quality assessment (VQA) framework since one straightforward design of video quality metrics (VQMs) may be applying current IQMs on a frame-by-frame basis. The overall video quality can then be derived by pooling the frame level quality scores with other video features [41]. Therefore, designing IQMs of high performance is of fundamental importance to the visual quality research community.

2.2 Visual Saliency

It is estimated that the visual data travelling into our eyes are approximately 10^8 to 10^9 bits per second [42]. Dealing with this data flow in real-time is an incredibly heavy mission for the HVS. Fortunately, only a portion of the data is selected and processed further in detail by the HVS. This selective mechanism in the HVS is called *visual attention*. Such an attentional behaviour is believed to be guided by two types of mechanisms, namely the stimulus-driven, bottom-up mechanism and the expectation-driven, top-down mechanism [43]. The bottom-up attention is mainly driven by the attributes of visual scenes including orientation, contrast, colour, motion and etc. The top-down attention is associated with cognitive aspects including experience, memory and cultural background and etc. In the area of computer vision, visual attention is mainly concerned with the former attentional mechanism due to its simplicity, which is often interchangeably referred to as *visual saliency* [44].

2.2.1 Eye-Tracking

The most straightforward way to study human visual attention is through the use of eye-tracking [45, 46, 47]. In an eye-tracking experiment, the eye movements of observers are recorded when viewing images. Neuroscientists, psychologists and computer vision engineers are using eye-tracking in a broad range of applications including medicine [48], engineering [49], psychology [50], education [51], robotics [52], marketing [53] and gaming [54].

The devices to obtain the eye-tracking data are called eye-trackers. They can be generally classified into three categories, namely the optical tracking system, the eye attached tracking system and the electrooculography (EoG) tracking system. The optical tracking is the most commonly used method which captures the infrared light reflected from the eye. In eye attached tracking systems, eye movements are recorded by measuring the movements of an attachment (e.g., contact lenses with a magnetic sensor embedded in) to human eyes. In EoG tracking systems, eye movements are measured by quantifying the change of electric signals around human eyes. Among these tracking systems, EOG tracking systems are less accurate due to noise in the electric signal. Eye attached systems suffer from potential slips of the attachment. Therefore, the optical tracking method is the most widely used approach in the literature. The eye-tracker used in this thesis (i.e., SensoMotoric Instrument (SMI) RED-m) belongs to the optical tracking category. Moreover, it features a contact-free property that allows free head movement. Therefore, it enables a collection of eye-tracking data over long duration without causing discomfort to subjects.

Abundant information is contained in the eye-tracking data, including fixations, saccades, pupil dilation and scanpaths [55]. Among these variables, researchers in the field of computer vision are concerned with fixation as they provide important information for bottom-up saliency. Eye fixation is defined as a spatially stable gaze lasting for several hundreds of milliseconds [56]. A *fixation map*, also known as *gaze map* of an image is often derived by accumulating all fixations of all observers recorded for that image. The reason to combine fixations from all observers instead of using an individual's fixations is to minimise the bias due to personal preference. A fixation map can be simply visualised as a binary map with "1" representing fixated locations and "0" representing unfixated locations or further post-processed into a grayscale map which is constructed by convolving a Gaussian kernel with each of the fixations. The binary map gives exact pixel locations of fixations in an image, while the grayscale map reflects attentive regions of the visual field. Both types of fixation maps are being used in different applications.

A number of eye-tracking databases have been created for computer vision researchers to better understand visual attention behaviour and to benchmark saliency prediction algorithms [57]. Unfortunately, standardised methodologies for eye-tracking data collection do not exist. Eye-tracking experiments are usually conducted in different laboratories and under different conditions and the raw data are processed in slightly different ways. Therefore, the usefulness of these databases may differ for different applications.

2.2.2 Visual saliency models

Computational models of visual saliency (i.e., bottom-up aspects of visual attention) aim to predict where people look in images. Note top-down aspects of visual attention are complex and are therefore rarely included in a saliency model. So far, many saliency models have been proposed in the literature and they have proven useful to various applications, including computer vision (e.g., object detection [58] and object recognition [59]), robotics (e.g., human-machine interaction [60]) and visual signal processing (e.g., region-of-interest-based compression [61] and image resizing [62]).

Pioneering work in saliency modelling was conducted in 1980s when Tresiman and Gelade proposed the *Feature-Integration Theory (FIT)* [63].The FIT states that when the HVS perceives a visual stimulus, different categories of image features are first parallelly selected and then combined in a later stage in order to identify objects. Following this framework, Koch and Ullman [43] proposed a mathematical descriptor for the FIT, resulting in a so-called *saliency map* that represents conspicuousness of a visual scene. This mathematical descriptor was further implemented by Itti et.al [64] for the construction of a saliency model, which has become one of the best-known models in the literature. Nowadays, a large number of saliency models are available in the literature, among which a majority of them are based on the FIT framework. In general, these saliency models take a three-step approach. First, multi-scale image pyramids of the input image are created, mimicking the multi-channel and multi-scale nature of the HVS [6]. In the second step, various image features are extracted from the pyramids, resulting in a set of *feature maps*. Finally, these feature maps are normalized and combined to form the final saliency map.

Saliency models can be categorized into pixel-based models and object-based models. The pixel-based models aim to highlight pixel locations where fixations are likely to occur. The object-based models focus on detecting salient objects in a visual scene. The majority of saliency models in the literature are pixel-based saliency models, such as ITTI [64], STB [65], AIM [66], SUN [67], CovSal [68] DVA [69], GBVS [70], Torralba [71], SR [72], PQFT [73], EDS [74], AWS [75], Gazit [76], SDSR [77], SDSF [78] and SDCD [79]. Representative object-based saliency models include CBS [80], FTS [81], salLiu [82], SVO [83] and CA [84]. In general, there are less object-based saliency models in the literature if compared with the number of pixel-based models. One of the reasons may be that the object segmentation process involved in a typical object-based model remains as an issue of computer vision. All the saliency models mentioned above are used in this thesis and are briefly summarized as below:

- ITTI is perhaps the first notable work in the field of computational saliency modelling, which combines multiscale image features into a single topographical saliency map.
- STB is meant to improve the output of ITTI for extracting the region of interest (ROI) — a binary mask that highlights the portion of an image where observers pay their attention to.
- AIM computes visual saliency using Shannon's self-information measure of visual features.
- SUN compares the features observed at each pixel location to the statistics of natural images and calculate the probability of each pixel to be salient using Bayes' rule.
- CovSal employs a local definition of saliency and measures the saliency of a pixel as how much it differs from its surroundings.
- DVA measures saliency with an attempt to maximize the entropy of the sampled visual features.
- GBVS is based on graph theory and is achieved by concentrating mass on activation maps, which are formed from certain raw features.
- Torralba measures saliency by incorporating several low-level features including contrast, colour, edge and orientation and two high-level features including objectness and context.
- SR is a simple model based on Fourier transform, where both the amplitude spectrum and phase spectrum are obtained.
- PQFT combines the phase spectrum information and the motion information to form a spatiotemporal saliency models.
- EDS relies on multi-scale edge detection and produces a simple and non-parametric method for detecting salient regions.
- AWS computes saliency by taking into account the decorrelation and distinctiveness of multi-scale low level features.

- Gazit employs a local-regional multi-level approach to detect edges of salient objects.
- SDSR computes the saliency using local descriptors from a given image which measure likeness of a pixel to its surroundings.
- SDFS measures saliency by combining global image features from frequency domain and local image features from spatial domain.
- SDCD works in the compressed domain and adopts intensity, colour and texture features for saliency detection.
- salLiu focuses on the salient object detection problem for images, using a conditional random field to learn ROI from a set of pre-defined features.
- CA employs multiple principles: local low-level features, visual organisation, global features and high-level features to separate the salient object from the background.
- FTS aims for the detection of well-defined boundaries of salient objects, which is achieved by retaining more frequency content from the image.
- CBS is formalized as an iterative energy minimization framework, which results in a binary segmentation of the salient object.
- SVO detects salient objects by fusing the cognitive-based objectness together with the image-based saliency.

Alternatively, saliency models can also be classified into spatial models and spatiotemporal models. Spatial models predict visual saliency according to the spatial cues only whereas spatiotemporal models estimate saliency based on both the spatial and temporal features of video sequences. Most saliency models in the literature fall into the former category, since simulating the effect of temporal saliency cues on the fixation deployment remains an academic challenge. Current spatiotemporal saliency models usually compensate spatial saliency models with motion features. For example, SDSR and GBVS add additional dynamic features (e.g., motion and flicker) in their design for video saliency estimation. However, this artificial temporal compensation for detecting video saliency is often inconsistent with ground truth [85].

Recently, Borji et al. [86] divided saliency models into eight categories on the basis of the modelling approach used, including information theoretic models, cognitive models, graphical models, spectral analysis models, pattern classification models, Bayesian models, decision theoretic models and other models. We hereby briefly introduce each category as below:

- Information theoretic models treat human eyes as information selectors and manage to select the most informative regions from a visual scene.

- Cognitive models are concerned with the biological plausibility of attention behaviour which are usually inspired by psychological findings of the HVS. Most models in this category follow the FIT framework and aim to simulate visual features related to selective attention.
- Graphical models consider eye fixations as a time series and use a graph-based representation for expressing the conditional dependence structure between random variables.
- Spectral analysis models operate in the frequency domain rather than in the spatial domain.
- Pattern classification models resort to machine learning approaches by training a saliency predictor with eye fixations or labelled salient areas. These models are usually not considered purely bottom-up since top-down image features (e.g., faces) are used during learning.
- Bayesian models combine prior knowledge of visual scenes (e.g., scene context) using Bayes' rule.
- Decision theoretic models treat the attention deployment as a decision making process in which the attention is determined by optimality.

It should be noted that clearly classifying saliency models according to the modelling approach is difficult, as some saliency models may fall into more than one category.

To evaluate the performance of saliency models, the modelled saliency maps are compared with ground truth human data. More specifically, saliency models that predict the bottom-up aspects of visual attention are validated against the eye-tracking data created under free-viewing conditions. Generally, state of the art saliency models can achieve promising accuracy when predicting the saliency of simple scenes or scenes with obvious regions of interest. However, there still exists a large gap between the current performance of saliency models and human performance especially when dealing with complex scenes [44].

2.3 Visual Saliency in Image Quality Assessment

Notwithstanding the tremendous progress made in the development of IQMs, recent research shows that the current performance of these IQMs remains limited when it comes to deal with the real-world complexity (e.g., a mixture of multiple types of distortion in images) [6, 7, 8]. To further improve the performance of IQMs, a significant research trend is to incorporate visual saliency information in IQMs. The rationale is that visual distortions perceived in the salient regions are considered to have relatively higher impact on image quality perception than those in the non-salient regions. Based on this, psychophysical studies and computational modelling have been conducted to investigate the added value of visual saliency in IQMs.

2.3.1 Relevance of saliency for image quality

To validate the relevance of visual saliency for image quality assessment, Alers et al. [87] conducted dedicated eye-tracking and quality scoring experiments. In that study, observers were first asked to score a set of images with different levels of distortion. The eye-tracking data were also recorded during the quality rating task. The images were then divided into regions of interest (ROI) and background (BG) according to the eye-tracking data. Observers were then asked to rate the quality of a series of new images, which were created by combining the BG and ROI at different quality levels. Experimental results showed that the quality of the combined images is dominated by the quality of their ROI, demonstrating that visual distortions present in salient regions are more important than those in non-salient regions.

2.3.2 Adding ground truth saliency to IQMs

To investigate the intrinsic added value of visual saliency to the performance of IQMs, some researchers used eye-tracking data [88, 89, 90, 91]. By integrating the “ground-truth” saliency into state of the art IQMs, one could identify whether and to what extent the addition of saliency is beneficial for IQMs in a genuine manner. However, while some researchers reported that integrating ground truth saliency improves the performance of IQMs, others reported that marginal or no performance gain can be obtained from saliency integration.

Pioneering study was conducted by Larson et al. in [88], where five widely-cited IQMs were augmented with eye-tracking data. Experimental results showed that the performance of most IQMs was improved by adding saliency. In addition, Larson et al. [89] attempted to optimise the performance of IQMs with manually labelled regions-of-interest (ROI) information. More specifically, the input image was first segmented into primary ROI, secondary ROI and non ROI based on the eye-tracking data obtained in [88]. Then, the local distortions were measured separately within individual levels of ROI (i.e., primary ROI, secondary ROI and non ROI). Finally, the overall quality was obtained by a linear combination of the locally measured distortions based on three ROI regions. Experimental results demonstrated that the performance of IQMs could be improved. However, no improvements were found to be significant.

A more comprehensive study with statistical evaluation was carried out by Liu et al. [90]. In that work, two eye-tracking experiments were conducted with both undistorted images and their corresponding JPEG compressed versions. Statistically significant improvements were reported when both types of eye-tracking data were integrated in four widely-cited IQMs. The deviation observed between the two types of eye-tracking data also yielded differences in the performance gain in IQMs. Including eye-tracking data obtained from undistorted images gives relatively larger performance gain for IQMs than using eye-tracking data of distorted images. It also concluded that the added value of visual saliency in IQMs might be related to the characteristics of image content [92].

In contrast to the above findings, the study in [91] tends to suggest that integrating visual saliency information to IQMs is of no benefit. The eye-tracking data collected under the quality scoring task rather than task-free conditions was integrated in two IQMs when assessing JPEG and JPEG2000 distorted images. Experimental results showed that no improvement was found for both IQMs, even though various saliency integration approaches were applied.

2.3.3 Adding computational saliency to IQMs

In any real-world quality assessment systems, it is impractical to involve human users in order to acquire the attentional information for the saliency integration process. Instead, fully automatic computational saliency models should be used. To this end, researchers investigate whether a saliency model, at least with the current performance of visual saliency modelling, is also able to improve the performance of IQMs, and if so, to what extent.

Literature on studying the added value of computational saliency in IQMs mainly focuses on the extension of a specific IQM with a specific saliency model. For example, to enhance the performance of an IQM concerning image sharpness [93], a saliency model proposed in [65] was applied by multiplying the local distortion estimated by the IQM with local saliency value. The performance of the IQM, in terms of the Pearson linear correlation coefficient between the IQM's predictions and human judgements, was significantly improved from 0.58 to 0.69. Similarly, Moorthy et al. [94] integrated an existing saliency model [95] in the IQM proposed in [31], achieving an improvement of 1% to 4% in terms of correlation across all distortion types assessed. In [96], an NR metric for assessing the JPEG2000 compressed artifacts was designed using the saliency model proposed in [97]. Experimental results demonstrated that the saliency information yielded significant improvements in the performance of the IQM without saliency. Ma et al. [98] incorporated the saliency model proposed in [72] in two state of the art IQMs, i.e., MSSIM [32] and VIF [34]. The experimental results showed that the MSSIM significantly benefited from the addition of saliency. However, no performance gain was observed for VIF.

In [99], the authors investigated the added value of four saliency models in three popular IQMs, resulting in twelve saliency-based IQMs. Experimental results showed that all saliency models were able to improve the performance of all IQMs with the performance gain ranging from 1.1% to 1.9%.

As shown above, employing a specific saliency model to specifically optimise a target IQM is often effective. However, these research findings also revealed that the performance gain that can be achieved for an IQM by the inclusion of a saliency model tends to depend on the saliency model, the IQM and the distortion type to be assessed. Some saliency models may not be designed to fit the IQA purposes at all, so blindly applying them to IQMs may not work well. On the other hand, some IQMs already consider elements on saliency. A double inclusion of saliency may cause saturation effect in the saliency-based optimisation. In addition, the method used to combine saliency and an IQM may also affects the actual gain to some extent.

2.3.4 Existing Issues

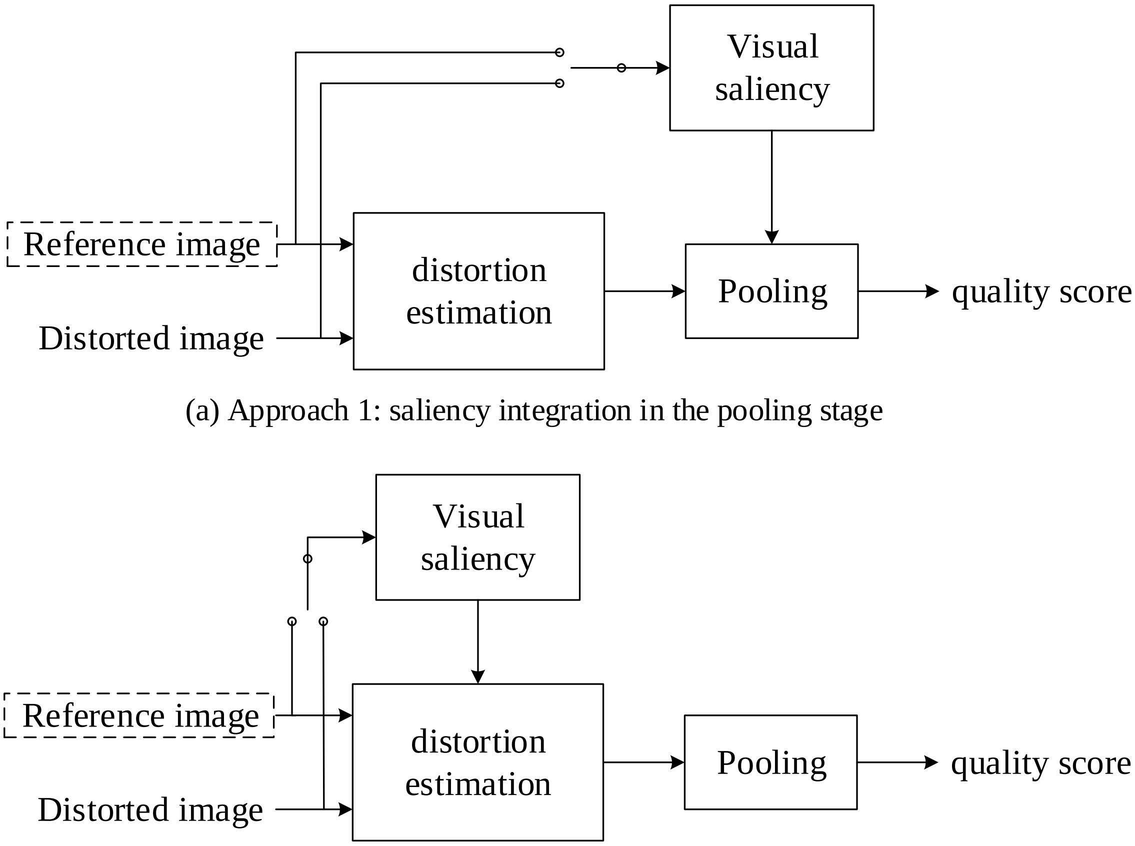
Integration approach

Research on modelling visual saliency in IQMs remains very limited. This is primarily due to the fact that how human attention affects the perception of image quality is largely unknown, and also due to the difficulties of precisely simulating visual attention aspects in IQMs. Existing saliency-based IQMs are generally based on a simple approach — weighting local distortions with local saliency. Most IQMs in the literature are implemented using a two-stage framework, as shown in Fig. 2.2. In the first stage, visual distortions are estimated locally, yielding a distortion map (or quality map) that represents the distortion level (or quality level) at each pixel location in an image. In the second stage, a single quality score for the whole image is calculated by pooling the local distortions. A saliency term can be added to this framework in two different approaches:

Approach 1 aims to locally adjust the relative importance of the estimated distortions in the pooling stage. The most widely used method is to multiply the distortion map with the saliency map. This approach is often called *visual saliency pooling* and can be formulated as:

$$WIQM = \sum_{x=1}^{M}\sum_{y=1}^{N} D(x,y)S(x,y) / \sum_{x=1}^{M}\sum_{y=1}^{N} S(x,y) \tag{2.1}$$

where D represents the local distortion, S denotes the saliency map, $WIQM$ denotes the final output of the saliency-weight IQM, and (x, y) denotes the pixel location. So far, this approach is the most widely used method for incorporating saliency information in IQMs.



(a) Approach 1: saliency integration in the pooling stage

(b) Approach 2: saliency integration in the local distortion estimation stage

Figure 2.2: Two saliency Integration Approaches.

Instead of using saliency in the pooling stage, **Approach 2** utilizes saliency information to optimise the local distortion estimation. For example, Lu et al. [100] used saliency to modulate the visual sensitivity of the HVS to distortions. A detectability threshold model was created by taking into account the modulation effect of visual saliency. The distortion below a certain threshold is considered imperceptible and is thus excluded from the following calculations of the IQM. Larson et al. [88] and Engelke et al. [101] segmented an image into primary regions of interest, secondary regions of interest and background regions. These regions were treated differently when estimating local distortions.

It should be noted that the two approaches mentioned above can be combined to form a more complicated method. One IQM using this combination is the visual saliency-induced index (VSI) [102] where saliency information was not only used to refine the importance of local distortions in the pooling stage (i.e., Approach 1), but also used as an image feature to estimate the local distortions (i.e., Approach 2). The success of VSI implies that the added value of visual saliency in IQMs can be optimised by refining the saliency integration method. It should also be noted that the approaches mentioned above strongly rely on the assumption of the HVS that the visual saliency and image distortions are treated separately and the estimations are combined artificially to determine the overall quality. The actual interactions between saliency and distortions may be more complex.

Task-free eye-tracking v. s. quality-scoring eye-tracking

Apart from the selection of saliency integration approaches, how to properly conduct the eye-tracking experiment is another issue in the research community. To investigate the intrinsic added value of visual saliency in IQMs, some researchers incorporate eye-tracking data obtained under task-free conditions (e.g., [90]) while some researchers use the eye-tracking data recorded during image quality scoring (e.g. [91]). It is worth noting that when saliency is added to IQMs, it should be the saliency that reflects bottom-up aspects of visual attention rather than the saliency including top-down aspects, e.g., the effects of a quality scoring task. This means that ground truth saliency should be obtained from eye-tracking under task-free conditions. It is well known that human fixation behaviours are significantly affected by viewing tasks [103]. Asking the observers to rate the image quality often results in fixations that are spread out into the background, simply because observers have learnt where to search for distortions [13]. As such, incorporating the scoring eye-tracking data in IQMs gives more weight (in relative term) to the distortions in the background, which overestimates their annoyance level. During free looking, observers view images as they would normally do, thus they tend to focus more on the regions of interest instead of the background [11]. Liu and Heynderickx [90] found that incorporating task-free eye-tracking data results in larger performance gain than using eye-tracking data of quality scoring.

Saliency of original scenes v. s. saliency of distorted scenes

Research has shown that visual distortions impact the fixation deployment [12, 13, 14, 15]. In the literature, there is a debate about when adding saliency in IQMs, whether saliency of the original scene or saliency of the distorted scene should be included in IQMs. It is not known yet whether the difference between both types of saliency is sufficiently large to actually affect the performance gain for the IQMs. To better understand this problem, ground truth saliency should be collected via task-free eye-tracking on both the original and distorted images. Without such empirical evidence, current approaches add either saliency computed from the undistorted scene (e.g. [94, 99]) or that calculated from the distorted scene (e.g. [104, 91]) to IQMs. Assuming the former is more appropriate, existing saliency models would be immediately useful since they have been designed and validated against undistorted images. However, assuming the latter is more appropriate, issues arise whether existing saliency models which are originally designed for undistorted images would be useful for detecting saliency of distorted images, and hence for improving IQMs. If not, further effort is needed, e.g., to develop a dedicated saliency model for distorted scenes or compensate for the reduction of benefits of using existing saliency models.

As the issues discussed above, challenges to optimising the performance of saliency-based IQM remain, and determining optimal use of saliency in IQMs requires further investigation.

2.4 Performance Evaluation Criteria

Previous sections have introduced the background knowledge on image quality and visual attention. The relationship between these two concepts has also been discussed. In this section, we briefly introduce how to measure the performance of an IQM or a saliency model and how to validate the performance difference between different IQMs or saliency models with statistical hypothesis testings.

2.4.1 Image quality metric evaluation

To evaluate the performance of an IQM, objective scores for all distorted images in an image quality database should be computed first. Then, a non-linear mapping is performed to map the objective scores to the scale of subjective ratings. We list below three widely used regression functions for the nonlinear mapping:

$$x' = ax^3 + bx^2 + cx + d \tag{2.2}$$

$$x' = \frac{b_1}{1 + e^{-b_2 \cdot (x - b_3)}} \tag{2.3}$$

$$x' = b_1 \cdot (\frac{1}{2} - \frac{1}{1 + e^{b_2 (x - b_3)}}) + b_4 \cdot x + b_5 \tag{2.4}$$

where x denotes an IQM's output, x' denotes the mapped score, b_i denotes a fitting parameter. After the non-linear regression, several performance evaluation criteria can be applied to measure the relationship between the predicted quality scores and the subjective quality scores (i.e., MOSs/DMOSs). The Video Quality Expert Group (VQEG) [105] recommends four criteria to quantify the performance of IQMs, including Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), root mean square error (RMSE) and outlier ratio (OR). Additional criteria such as Kendall's rank order correlation coefficient (KROCC) and mean absolute error (MAE) are also commonly used in the literature. PLCC, MAE and

RMSE measure the prediction accuracy of an IQM — the ability to reproduce the ground truth quality scores; SROCC and KROCC measure the prediction monotonicity of an IQM — the agreement between the rank order of ground truth ratings and that of IQM's predictions; OR measures the prediction consistency of an IQM — the degree to which an IQM remains accurate over a variety of distorted images. An IQM that obtains higher PLCC, SROCC and KROCC values (or lower MAE, RMSE and OR values) is considered to perform better than an IQM with lower PLCC, SROCC and KROCC values (or higher MAE, RMSE and OR values). Among these measures, PLCC, SROCC and RMSE are the most widely used criteria in the literature. We now briefly introduce these three measures as follows:

- Pearson linear correlation coefficient (PLCC)
  PLCC measures the linear relation between the subjective ratings and objective scores as:

  $$PLCC = \frac{\sum_{i=1}^{N}(x'_i - \bar{x}')(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x'_i - \bar{x}')}^2 \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})}^2} \tag{2.5}$$

  where x'_i denotes the mapped objective score, $\bar{x}'$ denotes the mean value of x', y_i denotes the corresponding subjective rating and $\bar{y}$ denotes the mean value of y. The result of PLCC is between -1 to $+1$, where a value approaching ± 1 reveals strong correlation.

- Spearman rank order correlation coefficient (SROCC)
  SROCC measures the agreement between the rank order of subjective ratings and that of IQM's outputs as:

  $$SROCC = 1 - \frac{6\sum\limits_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{2.6}$$

  where d_i^2 denotes the difference between the $i-th$ rank in subjective ratings and an IQM's output, N denotes the number of distorted images in the database.

- root mean square error (RMSE)
  RMSE measures the distance between the subjective ratings and an IQM's outputs as:

  $$RMSE = \sqrt{\frac{1}{N - df}\sum_{i=1}^{N}(y_i - \bar{x}')} \tag{2.7}$$

  where y is the subjective rating, x' denotes the mapped IQM's output, N is the number of images in the database and df denotes the degree of freedom of the mapping function.

It should be noted that the image quality community is increasingly accustomed to evaluating an IQM with several different databases. It may, e.g., account for the innate limitations of a typical subjective experiment in terms of the diversity in stimuli, and therefore provide more implications on the robustness of an IQM. It should also be noted that when reporting the performance gain between a saliency-based IQM and its original version, we followed the common practice (e.g., methods in [90] and [106]) where the nonlinear fitting is avoided. Adding a nonlinear mapping process can bias the added value of saliency term since it is unknown whether the performance gain obtained is due to saliency integration or due to the fitting process. In this thesis, Chapters 3 to 6 focus on investigating the added value of saliency in IQMs and the nonlinear regression is therefore not used. In contrast, Chapters 7 to 8 focus on devising a new IQM and the nonlinear regression is adopted for benchmarking its performance with other IQMs in the literature.

2.4.2 Saliency model evaluation

The prediction accuracy of a saliency model is measured by calculating the similarity between the ground truth fixation maps (FM) provided by an eye-tracking database and their corresponding modelled saliency maps (SM). There are many evaluation criteria in the literature, among which Area Under the Receiver Operating Characteristic Curve (AUC) [44], Normalized Scanpath Saliency (NSS) [107], Linear Correlation Coefficient (CC) and Kullback-Lebler Divergence (KLD) [108] are the most widely used measures. Some measures use the binary fixation maps as input whilst others use the grayscale fixation maps. We here denote the binary fixation map as FM^B and the grayscale fixation map as FM^G. We briefly introduce some widely used criteria as follows:

- Kullback-Lebler Divergence (KLD)
  KLD is originally designed as a measure for the difference between two probability distributions. Tatler et al. [108] apply this measure to quantify the difference between human fixations and computational saliency. It is formulated as:

$$KLD(SM, FM^G) = \sum_i FM_i^G * \log(\varepsilon + \frac{FM_i^G}{\varepsilon + SM_i}) \tag{2.8}$$

  where ε is a small constant to increase the stability of the measure (i.e., in case the denominator is approaching zero) and i represents a pixel in the map. KLD results in a positive score with a larger value indicating lower saliency prediction accuracy.

- Normalized Scanpath Saliency (NSS)
  NSS was proposed in [107] with a focus on measuring the responses of saliency models

at the locations of human fixations. This measure can be formulated as:

$$NSS(SM, FM^B) = \frac{1}{N} \cdot \sum_{f=1}^{N} NSS(f) \tag{2.9}$$

$$NSS(f) = \frac{SM(f) - \mu_{SM}}{\sigma_{SM}} \tag{2.10}$$

where $NSS(f)$ is the normalized NSS value of a fixation location f in FM^B; μ_{SM} and σ_{SM} are mean and standard deviation of SM respectively. When $NSS > 0$, the higher the value of the measure the more similar modelled saliency and ground truth are, whereas $NSS < 0$ indicates that a saliency model is able to predict human fixations is likely due to chance.

- Area Under the Receiver Operating Characteristic Curve (AUC)
  AUC is widely used to measure the performance of a binary classifier at various thresholds. In the context of measuring the performance of saliency models, there exist many different implementations for calculating the AUC score. In a conventional AUC measurement, human fixations of an image constitute a positive set, whereas a set of negative points is randomly selected. Saliency maps under different thresholds are then treated as binary classifiers to separate the positives from the negatives. After sweeping over all thresholds, a ROC curve is drawn by plotting the true positive rate against the false positive rate. The area under the ROC is considered as the performance of the saliency model. Shuffled AUC (SAUC) [44] is a refined version of AUC that is now widely used in the literature. Because of the more or less centred distribution of the human fixations (e.g., human eye tends to look at the central area of an image and/or photographers often place salient objects in the image centre [44]) in a typical image database, a saliency model could take advantage of such so-called center-bias by weighting its saliency map with a central Gaussian blob. This usually yields a dramatic increase in the AUC score. SAUC is proposed to normalise the effect of center-bias and as a consequence, to ensure a fair comparison of saliency models. Instead of selecting negative points randomly, all fixations over other images in the same database are used to form the negative set. By doing so, SAUC gives more credit to the off centre information. In this sense, SAUC is considered as a more rigorous measure; the bad performance of a saliency model cannot be masked by simply adding a central Gaussian filter. A perfect prediction of human fixations corresponds to a AUC score of 1 whereas a score of 0.5 indicates a random guess.

- Linear Correlation Coefficient (CC)
  CC represents the strength of the linear relationship between the saliency map SM and the grayscale fixation map FM^G. The definition is the same as the PLCC mentioned

above for evaluating image quality metrics. When CC is close to ±1 there is almost a perfect linear relationship between the two maps.

In [109], these measures are categorized into distribution-based criteria (e.g., KLD), location-based criteria (e.g., AUC) and value-based criteria (e.g., NSS). These measures focus on different aspects when evaluating a saliency model, suggesting that a comprehensive saliency model evaluation should include all criteria.

2.4.3 Statistical testings

The performance evaluation criteria mentioned above are designed to measure the absolute performance of individual IQMs and saliency models. In order to verify whether there exist significant differences in performance between different IQMs and different saliency models, statistical testings are usually performed. A statistical testing determines whether there is enough evidence to infer that a statistically significant difference exists between variables. A p-value is calculated according to the sample data of variables and is then compared to a pre-defined cut-off threshold which is known as the significance level. If the p-value is less than or equal to the significance level, a conclusion can be drawn that there exists statistically significantly difference between these variables. We list different types of statistical testings used in this thesis as below:

- Paired samples t-test
  A paired samples t-test is used to verify the difference between two related variables. It assumes that the two variables being compared follow the normal distribution. The normality of distributions is commonly tested by calculating the Kurtosis value of that distribution. Generally, a normal distribution tends to have a Kurtosis of 3. In practical implementations, a distribution with a Kurtosis value between 2 and 4 is considered to be normally distributed. In the context of this thesis, the paired samples t-test is used to compare the performance between an original IQM and its saliency-based version.

- Wilcoxon signed rank test
  The Wilcoxon signed rank test is the non-parametric version of a paired samples t-test. It is used when the two variables being compared do not follow the normal distribution.

- One-way ANOVA
  A one-way analysis of variance (ANOVA) is used to measure the effect of one independent variable on a dependent variable. It assumes that the dependent variable is normally distributed. In the context of this thesis, the one-way ANOVA is used to check whether

different saliency models have statistically significant difference in predicting human fixations. The saliency model is the independent variable and the performance of different saliency models is the dependent variable.

- Factorial ANOVA
  A factorial ANOVA measures the effect of two or more independent variables on the dependent variable. It also assumes that the dependent variable is normally distributed. In the context of this thesis, the factorial ANOVA is used to measure the effect of various influential factors (e.g., the type of distortions, the type of IQMs and the type of saliency models) on the performance gain of IQMs obtained from saliency integration. The influential factors are the independent variables and the performance gain of IQMs is the dependent variable.

Chapter 3

Computational Saliency in IQMs: A Statistical Evaluation

3.1 Introduction

Employing a specific saliency model to specifically optimise a target IQM is often effective, and these saliency-based IQMs outperform their original versions without saliency. However, the added value of saliency in IQMs reported in the literature heavily varies. The variation of the benefits causes several concerns. Firstly, a variety of saliency models are available in the literature. They are either specifically designed or chosen for a specific domain, but the general applicability of these models in the context of image quality assessment is so far not completely investigated. A rather random selection of a particular saliency model runs the risk of compromising the possibly optimal performance gain for IQMs. It is, e.g., not known yet whether the gain in performance (if existing) when adding a randomly chosen saliency model is comparable to the gain when "ground-truth" saliency is used. Secondly, questions still arise whether a saliency model successfully embedded in one particular IQM is also able to enhance the performance of other IQMs, and whether a dedicated combination of a saliency model and an IQM that can improve the assessment of one particular type of image distortion would also improve the assessment of other distortion types. If so, it remains questionable whether the gain obtained by adding this pre-selected saliency model to a specific IQM (or to IQMs to assess a specific distortion) is comparable to the gain that can be obtained with alternative IQMs (or when assessing other distortion types). Finally, it has been taken for granted in the literature that a saliency model that better predicts human fixations is expected to be more advantageous in improving the performance of IQMs. This speculation, however, has not been statistically validated yet. The various concerns discussed above imply that before implementing saliency models in IQMs, it is desirable to have a comprehensive understanding on whether and to what extent the addition of computational saliency can improve IQMs, in the context of existing saliency models and IQMs available in the literature.

In this chapter, we examined the capability and capacity of computational saliency in improving an IQM's performance in predicting perceived image quality. A statistical evaluation was conducted by integrating state of the art saliency models in several IQMs well-known in the literature. We investigated whether there is a significant difference in predicting human fixations between saliency models, and whether and to what extent such difference can affect the actual gain in performance that can be obtained by including saliency in IQMs. The statistics also allowed us to explore whether or not there is a direct relation between how well a saliency model can predict human fixations and to what extent an IQM can profit from adding this saliency. Furthermore, This work explicitly evaluated to what extent the amount of performance gain when adding computational saliency depends on the saliency model, IQM and type of image distortion. This work intends to, based on in-depth statistical analysis, provide recommendations and practical solutions with respect to the application of saliency models in IQMs.

3.2 Evaluation Environment

To evaluate the added value of computational saliency in IQMs, the saliency map derived from a saliency model was integrated into an IQM, and the resulting IQM's performance was compared to the performance of the same IQM without saliency. To ensure a study of sufficient statistical power, the validation was carried out with twenty saliency models, twelve IQMs, and three image quality assessment databases, which are all so far widely recognised in the research community.

3.2.1 Visual saliency models

Twenty state of the art models of visual saliency, namely AIM, AWS, CBS, EDS, FTS, Gazit, GBVS, CA, SR, DVA, ITTI, SDFS, PQFT, salLiu, SDCD, SDSR, STB, SUN, SVO and Torralba were implemented. These models have been summarized in Section 2.2. These models cover a wide range of modelling approaches and application environments, thus enable a study of sufficient diversity. Figure 3.1 illustrates the saliency maps generated by the models mentioned above for one of the source images in the LIVE image quality assessment database [16]. They can be generally classified into two categories: pixel-based and object-based models. Pixel-based saliency models focus on mimicking the behaviour and neuronal architecture of the early primate visual system, aiming to predict human fixations (see, e.g., ITTI, AIM and GBVS). Object-based models are driven by the practical need of object detection for machine vision applications, attempting to identify explicit salient regions/objects (see, e.g., FTS, CBS and SVO).

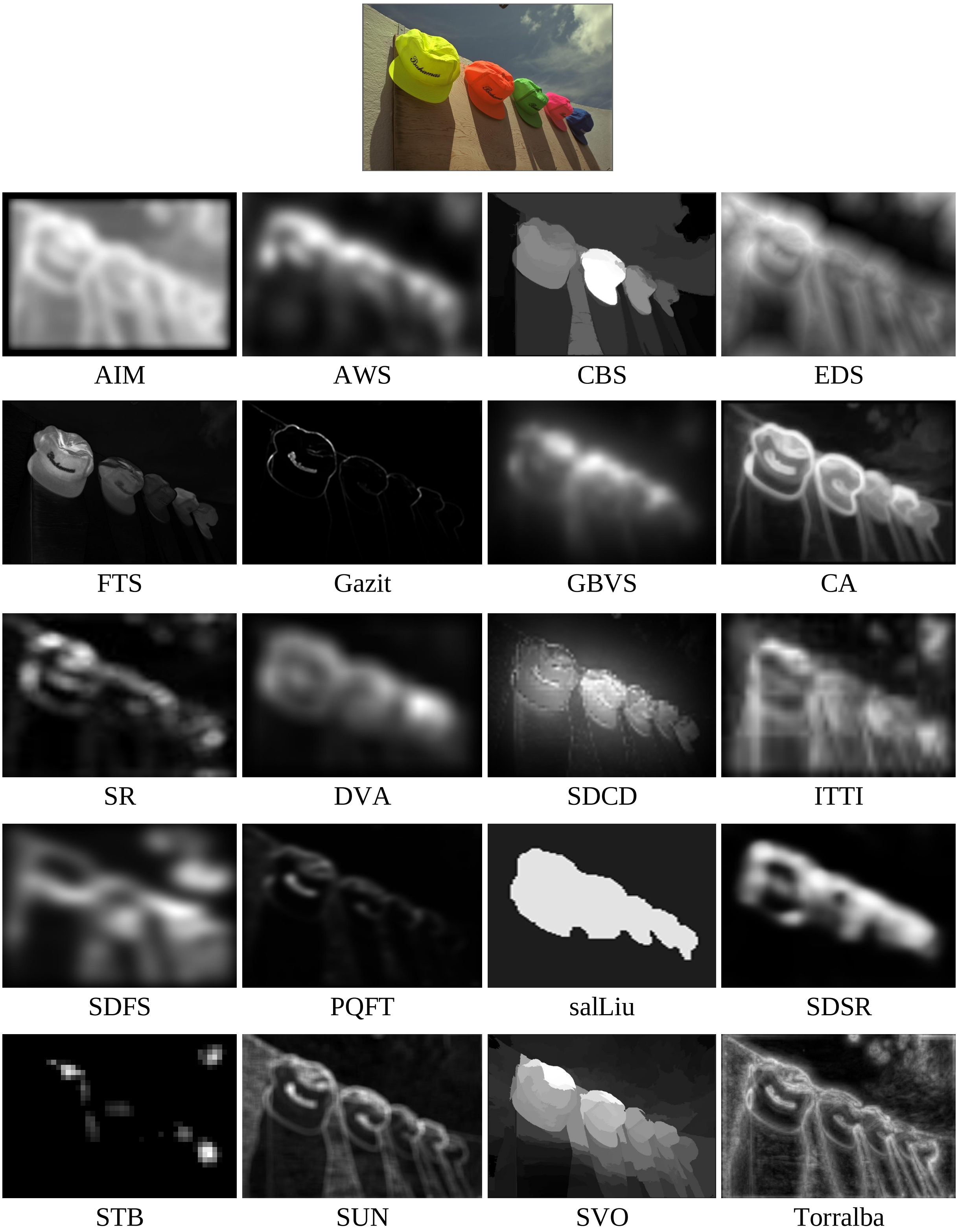


Figure 3.1: Illustration of saliency maps generated by twenty state-of-the-art saliency models for one of the source images in the LIVE database.

3.2.2 Image quality metrics

Twelve widely recognised IQMs, namely PSNR, UQI, SSIM, MSSIM, VIF, FSIM, IWPSNR, IWSSIM, GBIM, NBAM, NPBM and JNBM, were applied in our evaluation. They estimate image quality locally, resulting in a quantitative distortion map which represents a spatially varying quality degradation profile. We have already summarized these IQMs in Section 2.1. These IQMs include eight FR and four NR metrics, and range from the purely pixel-based IQMs without characteristics of the HVS to IQMs that contain complex HVS modelling. The FR metrics are PSNR, UQI, SSIM, MS-SSIM, IWPSNR, IWSSIM, VIF and FSIM. The NR metrics are GBIM, NPBM, JNBM and NBAM. These IQMs are implemented in the spatial domain. It is noted that other well-known IQMs formulated in the transform domain, such as VSNR, MAD and NQM were not included in our study. Integrating a saliency map in a rather complex IQM calculated in the frequency domain is not straightforward, and is therefore, outside the scope of this chapter.

3.2.3 Integration approach

We followed the most widely used integration approach in the literature where the distortion map (DM) of an IQM is weighted by the saliency map (SM) using the following formula:

$$SW - IQM = \sum_{x=1}^{M}\sum_{y=1}^{N} DM(x,y) \cdot SM(x,y) / \sum_{x=1}^{M}\sum_{y=1}^{N} SM(x,y) \tag{3.1}$$

where $SW - IQM$ is the output of the saliency-weighted IQM, SM is generated from the reference undistorted image. In the case of an NR IQM, the SM was either assumed to be available, which is analogous to a RR framework in practice, or considered to be possibly calculated from the distorted image by separating natural scene and distortion apart. This combination is simple and parameter-free, and consequently, fulfils a generic implementation. It allows us to conduct a large scale statistical evaluation. A more sophisticated combination strategy may further improve an IQM's performance, e.g., in assessing a specific type of distortion. However, the increase in the effectiveness is often achieved at the expense of the generality of the combination strategy. As such, this simplified approach seems to be a viable and probably so far the most acceptable way of including visual saliency aspects in IQMs.

3.2.4 Evaluation databases

The evaluation of the performance of an IQM was conducted on the LIVE database [16]. The reliability of the LIVE database is widely recognized in the image quality community. Indeed, the image quality community is more and more accustomed to the evaluation of IQMs with different databases that are made publicly available. It may, e.g., account for the innate limitations of a typical subjective experiment in terms of the diversity in image content and distortion type, etc., and therefore, provide more implications on the robustness of an IQM. With this in mind, a cross-database evaluation was carried out by repeating our evaluation protocol on other two existing image quality databases, i.e., IVC [19] and MICT [20], which are customarily used in the literature. It should, however, be noted that the meaningfulness of a cross database validation heavily depends on, e.g., the consistency between different databases. The measured difference in the performance of an IQM can be attributed to the difference between the designs of different subjective experiments.

The evaluation of the performance of a saliency model was conducted on the TUD eye-tracking database [90], which is obtained from 20 human observers looking freely to all the 29 reference undistorted images of the LIVE database. The TUD database has been validated as a reliable ground truth in [110]. We benchmarked the saliency models against the TUD database due to that it shares the same set of stimuli as used in LIVE database, which enabled us to compare the performance of a saliency model to its added value to IQMs.

3.2.5 Performance measures

To quantify the similarity between a "ground-truth" fixation map (FM) obtained from eye-tracking and the modelled saliency map (SM) derived from a saliency model, we used three measures, namely CC, NSS and AUC.

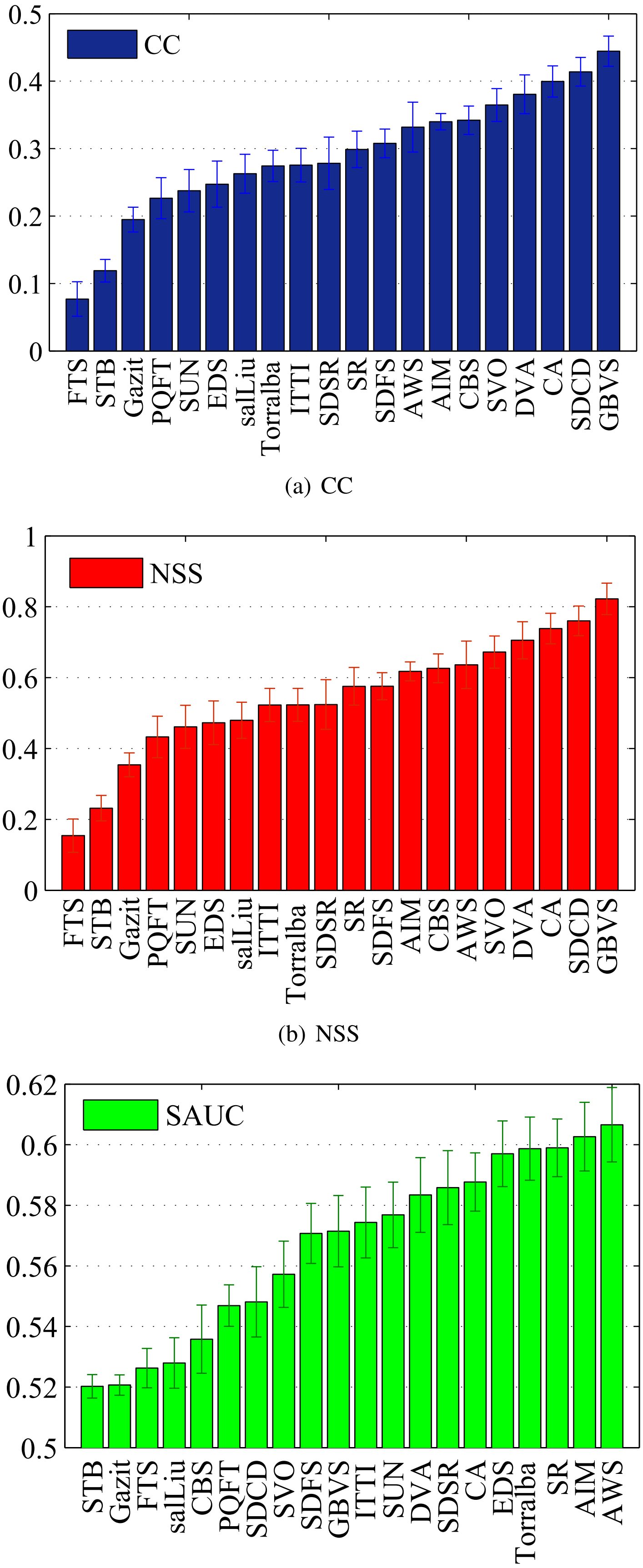
To quantify the performance gain of an IQM, we used the PLCC and SROCC. It should be noted that the image quality community is accustomed to fitting the predictions of an IQM to the subjective scores [105]. A nonlinear mapping may, e.g., account for a possible saturation effect in the quality scores at high quality. It usually yields higher PLCCs in absolute terms, while generally keeping the relative differences between IQMs [111]. As also explained in [90], without a sophisticated nonlinear fitting the PLCCs cannot mask a bad performance of the IQM itself. To better visualize differences in performance, we avoided any nonlinear fitting and directly calculated correlations between an IQM's predictions and the DMOS scores.

3.3 Overall Effect of Computational Saliency in IQMs

In this section, the overall effect of including computational saliency in IQMs is evaluated. The evaluation protocol breaks down into three steps: first, the difference in predictability between saliency models is checked; second, by applying these saliency models to individual IQMs, the gain in performance for the IQMs is evaluated regarding its meaningfulness; finally, the relation between the predictability of saliency models and the profitability of including different saliency models in IQMs is investigated.

3.3.1 Prediction accuracy of saliency models

Per saliency model, CC, NSS and SAUC were calculated between the FM and SM, and averaged over the 29 stimuli. Figure 3.2 illustrates the rankings of saliency models in terms of CC, NSS and SAUC respectively. It shows that the saliency models vary over a broad range of predictability independent of the measure used. Notwithstanding a slight variation in the ranking order across three measures, there is a strong consistency between different ranking results. Based on SAUC, hypothesis testing was performed in order to check whether the numerical difference in predictability between saliency models is statistically significant. Before being able to decide on an appropriate statistical test, we evaluated the assumption of normality of the SAUC scores. A simple kurtosis-based criterion (as used in [112]) was used for normality: if the SAUC scores have a kurtosis between 2 and 4, they were assumed to be normally distributed, and the difference between saliency models could be tested with a parametric test, otherwise a non-parametric alternative could be used. Since the variable SAUC was tested to be normally distributed, an ANOVA (analysis of variance) was conducted by selecting SAUC as the dependent variable, and the categorical saliency model as the independent variable. The ANOVA results showed that the categorical saliency model had a statistically significant effect (p-value = 1.47e-17, $p < 0.05$ at 95% confidence level) on SAUC. Pairwise comparisons were further performed with a t-test between two consecutive models in the SAUC rankings. The results (i.e., with p-value ranging from 0.073 to 0.831, $p > 0.05$ at 95% confidence level) indicated that the difference between any pair of consecutive models was not significant. This, however, does not necessarily mean that two models that are not immediately close to each other are not significantly different. This can be easily revealed by running all pairwise comparisons. For example, AWS was tested to be better than SVO and manifested itself significantly better than all other models on the left-hand side of SVO. In general, we may conclude that there is a significant variation in predictability among saliency models. Based on this finding, we set out to investigate whether adding these saliency models to IQMs can produce a meaningful gain, and whether the existence and/or status of such gain is affected by the predictability of a saliency



(a) CC

(b) NSS

(c) SAUC

Figure 3.2: Illustration of the rankings of saliency models in terms of CC, NSS and SAUC, respectively. The error bars indicate the 95% confidence interval.

model.

3.3.2 Added value of saliency models in IQMs

Integrating saliency models into IQMs results in a set of new saliency-based IQMs. FR metrics and their saliency-based derivatives are intended to assess image quality independent of distortion type, and therefore, were applied to the entire LIVE database. This resulted in 800 combinations, i.e. 8 FR metrics × 20 saliency models × 5 distortion types. The NR blockiness metrics (i.e., GBIM and NBAM) and their derivatives were applied to the JPEG subsets of the LIVE database. The NR blur metrics (i.e., NPBM and JNBM) and their derivatives were applied to the GBLUR subset of the LIVE database. This resulted in 80 combinations, i.e. 4 NR metrics × 20 saliency models × 1 distortion type. PLCC and SROCC are calculated between the subjective DMOS scores and the objective predictions of an IQM. In total, 880 PLCCs and SROCCs were calculated for all the possible combinations in our evaluation framework. Table 3.1 and Table 3.2 summarize the overall performance gain (averaged over all distortion types where appropriate) of a saliency-based IQM over its original version in terms of PLCC and SROCC respectively. The gain in performance that can be obtained by adding ground truth fixation map (FM) in IQMs is also included as a reference. It should be noted that the analysis conducted in this chapter is based on the ΔPLCC values. ΔSROCC exhibits the same trend of changes as ΔPLCC and we do not expect a major change in the conclusions made. In general, these two tables demonstrate that there is indeed a gain in performance when including computational saliency in IQMs, being most of the ΔPLCC and ΔSROCC values are positive.

It is noticeable in Table 3.1 that some ΔPLCC values are relatively marginal, but not necessarily meaningless. In order to verify whether the performance gain as obtained in Table 3.1 is statistically significant, hypothesis testing was conducted. As suggested in [105], the test was based on the residuals between DMOS and the quality predicted by an IQM (hereafter, referred to as M-DMOS residuals). Before being able to run an appropriate statistical significance test, we evaluated the assumption of normality of the M-DMOS residuals. The results of the test for normality are summarized in Table 3.3. For the vast majority of cases, in which paired M-DMOS residuals (i.e., two sets of residuals being compared: one is from the original IQM and one is from its saliency-based derivative) were both normally distributed, a paired samples t-test was performed (as used in [90]). Otherwise, in the case of non-normality, a non-parametric version (i.e., Wilcoxon signed rank sum [113]) analogue to a paired samples t-test was conducted. The test results are given in Table 3.4 for all combinations of IQMs and saliency models. It illustrates that in most cases the difference in performance between an IQM and its saliency-based derivative is statistically significant. In general terms, this suggests that the addition of computational saliency in IQMs makes a meaningful impact on their prediction performance.

Table 3.1: Performance gain (i.e., △PLCC) between a metric and its saliency-based version over all distortion types for LIVE database. Each entry in the last row represents the △PLCC averaged over all saliency models excluding the FM. The standard deviations of the mean values range from 0.001 to 0.019.

	PSNR	UQI	SSIM	MSSIM	VIF	FSIM	IWPSNR	IWSSIM	GBIM	NPBM	JNBM	NBAM	MEAN
FM	0.017	0.038	0.025	0.015	0.022	0.004	0.001	-0.013	0.039	0.046	0.02	0.023	0.02
AIM	0.005	0.033	0.013	0.009	0.009	0.001	0.001	-0.004	0.039	0.03	0.004	0.005	0.012
AWS	0.021	0.042	0.028	0.017	-0.13	0.002	0.004	-0.008	0.036	0.027	0.028	0.025	0.008
CA	0.019	0.043	0.031	0.018	0.021	0.001	0.003	-0.011	0.048	0.037	0.027	0.02	0.021
CBS	-0.023	0.026	0.013	0.007	0.018	0.002	-0.001	-0.006	0.009	0.022	0.04	0	0.009
DVA	0.015	0.041	0.025	0.015	0.018	0.001	0.003	-0.011	0.043	0.029	0.02	0.016	0.018
EDS	0.015	0.043	0.027	0.018	0.01	0.001	0.002	-0.006	0.014	0.021	0.022	0.021	0.016
FTS	0.015	0.035	0.009	0.005	0.004	0.01	0.002	-0.005	-0.025	0.003	0.007	0.015	0.006
Gazit	0.038	0.029	0.016	-0.001	-0.001	-0.034	0.006	-0.036	0.02	0.044	0.007	0.035	0.01
GBVS	0.017	0.038	0.026	0.016	0.022	0.003	0.002	-0.01	0.049	0.032	0.026	0.025	0.021
ITTI	0.011	0.044	0.026	0.017	-0.129	0.003	0.002	-0.006	0.045	0.031	0.016	0.021	0.007
PQFT	0.027	0.043	0.041	0.025	0.015	-0.007	0.003	-0.02	0.034	0.036	0.033	0.026	0.021
salLiu	0.014	0.017	0.015	0.008	0.024	0.003	0.002	-0.011	0.034	0.029	0.015	0.001	0.013
SDCD	0.017	0.038	0.023	0.015	0.02	0.003	0.002	-0.007	0.038	0.037	0.039	0.027	0.021
SDFS	0.005	0.036	0.019	0.011	0.022	-0.002	0.001	-0.012	0.033	0.036	0.023	0.001	0.014
SDSR	0.027	0.047	0.03	0.017	0.018	0.002	0.005	-0.014	0.03	0.041	0.023	0.037	0.022
SR	0.027	0.047	0.036	0.021	0.02	0.002	0.004	-0.018	0.049	0.041	0.029	0.039	0.025
STB	0.004	0.015	0.001	-0.008	0.017	-0.007	-0.005	-0.025	0.004	0.006	-0.107	-0.001	**-0.009**
SUN	0.015	0.032	0.029	0.019	0.008	-0.001	0.002	-0.011	0.024	0.025	0.022	0.012	0.015
SVO	0.005	0.025	0.012	0.008	0.021	0.001	0.001	-0.004	0.024	0.019	0.012	0.004	0.011
Torralba	0.014	0.041	0.025	0.016	0.009	0.001	0.001	-0.005	0.026	0.022	0.009	0.017	0.015
MEAN	0.014	0.036	0.022	0.013	0.001	**-0.001**	0.002	**-0.011**	0.029	0.028	0.015	0.017	0.014

Table 3.2: Performance gain (i.e., ∆SROCC) between a metric and its saliency-based version over all distortion types for LIVE database. Each entry in the last row represents the ∆SROCC averaged over all saliency models excluding the FM. The standard deviations of the mean values range from 0.002 to 0.017.

	PSNR	UQI	SSIM	MSSIM	VIF	IWPSNR	IWSSIM	GBIM	NPBM	JNBM	NBAM	MEAN
FM	0.024	0.037	0.026	0.013	0.010	0.003	-0.005	0.028	0.030	0.021	0.014	0.018
AIM	0.007	0.016	0.009	0.007	-0.007	0.002	0.001	0.033	0.026	0.029	0.011	0.012
AWS	0.024	0.028	0.019	0.014	-0.024	0.004	0.004	0.029	0.020	0.021	0.013	0.014
CA	0.023	0.031	0.020	0.014	-0.003	0.004	0.002	0.031	0.025	0.044	0.013	0.018
CBS	-0.009	0.007	0.007	0.005	-0.008	0.000	-0.001	0.027	0.011	0.056	0.012	0.010
DVA	0.022	0.027	0.017	0.012	-0.007	0.004	0.002	0.031	0.024	0.043	0.011	0.017
EDS	0.016	0.029	0.019	0.014	-0.005	0.003	0.003	0.000	0.020	0.022	0.013	0.012
FTS	0.012	0.016	0.006	0.004	-0.010	0.003	0.001	-0.005	-0.015	-0.014	0.008	0.001
Gazit	0.049	0.063	0.028	0.019	-0.005	0.009	0.005	0.002	0.033	-0.015	0.011	0.018
GBVS	0.028	0.022	0.018	0.012	-0.003	0.004	0.002	0.031	0.018	0.039	0.014	0.017
ITTI	0.016	0.029	0.015	0.011	-0.031	0.002	0.002	0.031	0.024	0.031	0.015	0.013
PQFT	0.034	0.044	0.029	0.019	-0.006	0.005	0.003	0.024	0.040	0.046	0.015	0.023
salLiu	0.019	-0.004	0.009	0.003	0.000	0.003	-0.002	0.028	0.026	0.045	0.010	0.012
SDCD	0.021	0.020	0.014	0.010	-0.004	0.003	0.002	-0.001	0.023	0.050	0.014	0.014
SDSF	0.006	0.024	0.014	0.009	-0.003	0.001	0.001	0.031	0.023	0.033	0.009	0.014
SDSR	0.032	0.036	0.021	0.015	-0.006	0.007	0.002	0.024	0.027	-0.014	0.009	0.014
SR	0.034	0.040	0.025	0.019	-0.004	0.007	0.004	0.033	0.017	-0.004	0.016	0.017
STB	0.010	0.001	-0.008	-0.012	-0.001	-0.002	-0.010	-0.013	-0.045	-0.126	-0.014	**-0.020**
SUN	0.020	0.019	0.018	0.013	-0.007	0.003	0.001	0.028	0.022	0.030	0.010	0.014
SVO	0.008	0.006	0.007	0.005	-0.003	0.001	0.001	0.030	0.024	0.020	0.011	0.010
Torralba	0.015	0.027	0.016	0.012	-0.007	0.002	0.002	0.030	0.022	0.013	0.013	0.013
MEAN	0.019	0.024	0.015	0.010	**-0.007**	0.003	0.001	0.021	0.018	0.017	0.011	0.012

Table 3.3: Normality of the M-DMOS residuals. Each entry in the last column is a codeword consisting of 21 digits. The position of the digit in the codeword represents the following saliency models (from left to right): FM, AIM, AWS, CBS, EDS, FTS, Gazit, GBVS, CA, SR, DVA, SDCD, ITTI, SDFS, PQFT, salLiu, SDSR, STB, SUN, SVO, and Torralba. "1" represents the normal distribution and "0" represents the non-normal distribution.

IQM	Normality	saliency-based IQM	Normality
PSNR	1	SW-PSNR	111111111111111111111
UQI	1	SW-UQI	111111111111111111111
SSIM	1	SW-SSIM	111111111111111111111
MSSIM	1	SW-MSSIM	111101111111111101111
VIF	1	SW-VIF	111111111111111111111
FSIM	1	SW-FSIM	111111111111111111111
IWPSNR	1	SW-IWPSNR	111111111111110111111
IWSSIM	1	SW-IWSSIM	111111111111111111111
GBIM	1	SW-GBIM	111111111011111111111
NBAM	1	SW-NBAM	111111111111111111111
NPBM	1	SW-NPBM	101010101101101110111
JNBM	1	SW-JNBM	111111111111111111111

Table 3.4: Results of statistical significance testing based on M-DMOS residuals. Each entry is a codeword consisting of 21 symbols refers to the significance test of an IQM versus its saliency based version. The position of the symbol in the codeword represents the following saliency models (from left to right): FM, AIM, AWS, CBS, EDS, FTS, Gazit, GBVS, CA, SR, DVA, SDCD, ITTI, SDFS, PQFT, salLiu, SDSR, STB, SUN, SVO, and Torralba. "1" (parametric test) and "*" (non-parametric test) means that the difference in performance is statistically significant; "0" (parametric test) and "#" (non-parametric test) means that the difference is not statistically significant.

IQM vs. saliency-weighted IQM	Significance
PSNR vs. SW-PSNR	111111111111111111111
UQI vs. SW-UQI	111111111111111111111
SSIM vs. SW-SSIM	011011111111111110111
MSSIM vs. SW-MSSIM	0100*11010101111#0111
VIF vs. SW-VIF	111101111111011111011
FSIM vs. SW-FSIM	110111111000111111110
IWPSNR vs. SW-IWPSNR	10101101000110*101111
IWSSIM vs. SW-IWSSIM	111111111111111111111
GBIM vs. SW-GBIM	111011011*11100111011
NBAM vs. SW-NBAM	111001101100100010101
NPBM vs. SW-NPBM	1*1*1*1*11*01*111#111
JNBM vs. SW-JNBM	111111111111111111111

In accordance with custom, we also evaluated the potential impact of different image quality databases on the performance gain that can be obtained by adding computational saliency to IQMs. We repeated the aforementioned evaluation protocol once for the IVC database, and once for the MICT database. In terms of the performance gain for IQMs (expressed by ΔPLCC), the Pearson correlation coefficient is 0.84 between LIVE and IVC, and 0.82 between LIVE and MICT. The cross database validation indicated that the same trend of changes in performance gain is consistently found for the three image quality databases.

3.3.3 Predictability versus profitability

Having identified the overall benefits of including computational saliency in IQMs, one could intuitively hypothesize that the better a saliency model can predict human fixations, the more an IQM may profit from adding this saliency model in the prediction of image quality. To check this hypothesis, we calculated the correlation between the predictability of saliency models (based on the SAUC scores as shown in Fig. 3.2(c)) and the average performance gain achieved by using these models (based on ΔPLCC averaged over all IQMs as shown in the last column of Table 3.1). The resulting Pearson correlation coefficient is equal to 0.44, suggesting that the relation between the predictability of a saliency model and the actual added value of this model for IQMs is rather weak. Saliency models that are ranked relatively highly in terms of predictability do not necessarily correspond to a larger amount of performance gain when they are added to IQMs. For example, AWS ranks the 1st (out of 20) in predictability. However, the rank of AWS in terms of the added value for IQMs is the 17th (out of 20). On the contrary, PQFT is ranked comparatively low in terms of predictability, but it produces higher added value for IQMs compared to other saliency models. In view of the statistical power, which is grounded on all combinations of 20 saliency models and 12 IQMs, this finding is fairly dependable but indeed surprising, and it suggests that our common belief in the selection of appropriate saliency models for inclusion in IQMs is being challenged. However, it may be still far from being conclusive whether or not the predictability has direct relevance to the performance gain, e.g., it is arguable that the measured predictability might be still limited in its sophistication. But we may conclude that the measure of predictability should not be used as the only criterion to determine the extent to which a specific saliency model is beneficial for its application in IQMs, at least, not with the current performance of visual saliency modelling.

Table 3.5: Results of the ANOVA to evaluate the impact of the IQM, saliency model and image distortion type on the added value of computational saliency in IQMs.

ANOVA

Source	df	F	Sig.
saliency model	19	4.036	.000
distortion	4	32.944	.000
IQM	11	56.651	.000
distortion * IQM	28	9.414	.000
saliency model * IQM	209	4.111	.000
saliency model * distortion	76	1.107	.262

3.4 Dependencies of Performance Gain

The aforementioned section provided a grounding in the general view of the added value of including computational saliency in IQMs. Granted that a meaningful impact on the performance gain is in evidence, the actual amount of gain, however, tends to be different for different IQMs, for different saliency models and for different distortion types. Such dependencies of the performance gain have highly practical relevance to the application of computational saliency in IQMs, e.g., in a circumstance where a trade-off between the increase in performance and the expense needed for saliency modelling is in active demand. To this effect, the observed tendencies in the changes of the performance gain were further statistically analysed in order to comprehend the impact of individual categorical variables being the kind of IQM, the kind of saliency model and the distortion type. The statistical test was based on the 880 data points of performance gain in terms of ΔPLCC resulted from the entire LIVE database. The test for the assumption of normality indicated that the variable performance gain is normally distributed and consequently, a factorial ANOVA was conducted with the performance gain as the dependent variable, the kind of IQM, kind of saliency model and distortion type as independent variables. The results are summarized in Table 3.5, and show that all main effects are highly statistically significant. The significant interaction between IQM and distortion (excluding NR cases due to data points being incomplete for irrelevant combinations) is caused by the fact that the way the performance gain changes among IQMs depends on the distortion type. The interaction between saliency model and IQM is significant since the impact the different saliency models have on performance gain also depends on the IQM.

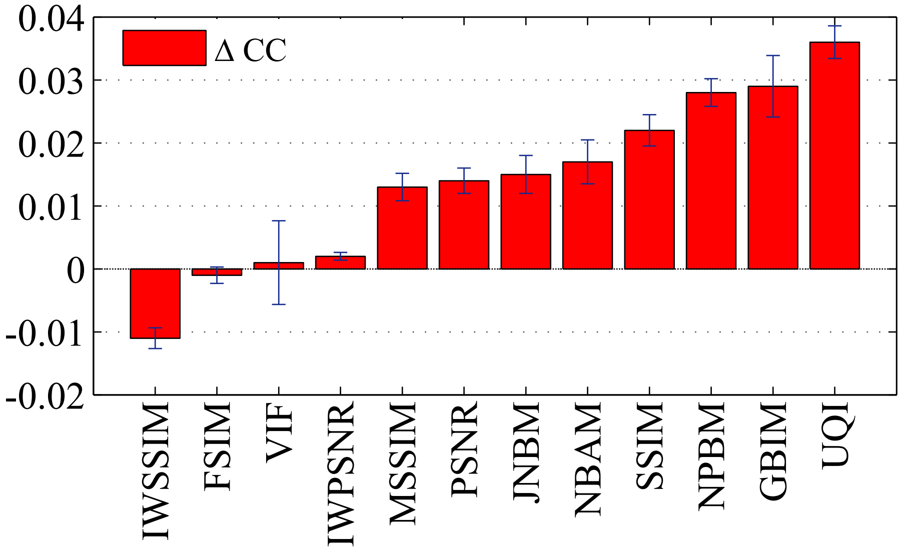
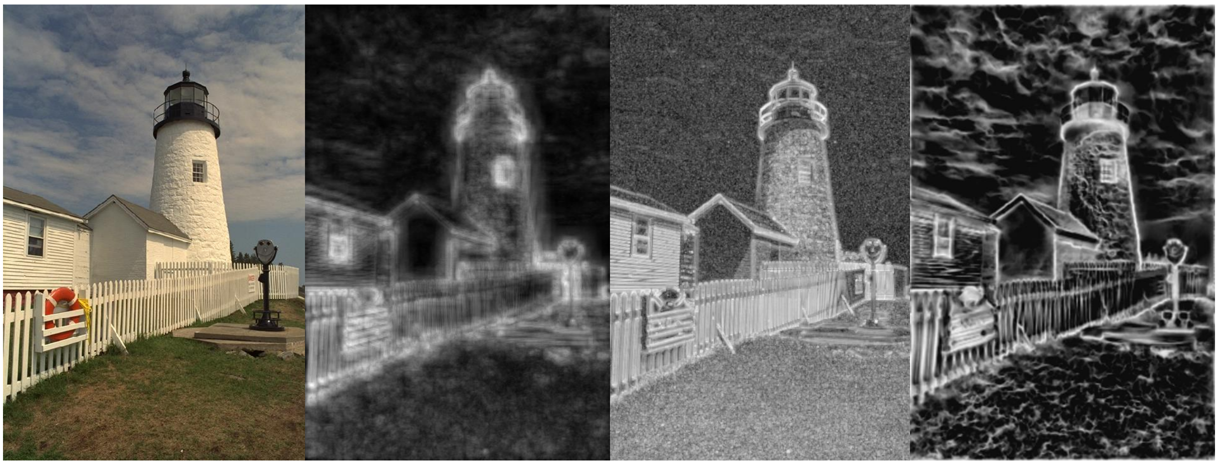


Figure 3.3: Illustration of the rankings of IQMs in terms of the overall performance gain (expressed by ΔPLCC, averaged over all distortion types, and over all saliency models where appropriate) between an IQM and its saliency-based version. The error bars indicate the 95% confidence interval.

3.4.1 Effect of IQM dependency

Obviously, the kind of IQM has a statistically significant effect on the performance gain. Figure 3.3 illustrates the order of IQMs in terms of the overall performance gain. It shows that adding computational saliency results in a marginal gain for IWSSIM, FSIM, VIF and IWPSNR; the performance gain is either non-existent or even negative (i.e., the averaged ΔPLCC is -1.1% for IWSSIM, -0.1% for FSIM, 0.1% for VIF and 0.2% for IWPSNR). Compared to such a marginal gain, adding computational saliency to other IQMs, such as UQI, yields a larger amount of performance gain (e.g., the averaged ΔPLCC is 3.6% for UQI). The difference in performance gain between IQMs may be attributed to the fact that some IQMs already contain saliency aspects in their metric design but others do not. For example, IWSSIM, VIF and IWPSNR incorporate the estimate of local information content, which is often applied as a relevant cue in saliency modelling [66]. Phase congruency, which is implemented in FSIM, manifests itself as a meaningful feature of visual saliency [114]. Figure 3.4 contrasts the so-called “information content map (ICM)” (i.e., extracted from IWSSIM, VIF or IWPSNR) and the “phase congruency map (PCM)” (i.e., extracted from FSIM) to a representative saliency map (i.e., Torralba). It clearly visualizes the similarity between ICM/PCM and the real saliency map: the Pearson correlation coefficient is 0.72 between ICM and Torralba, and 0.79 between PCM and Torralba. Similarly, JNBM and NBAM intrinsically bear saliency characteristics (e.g., contrast). As such, the relatively small gain obtained for the aforementioned IQMs is probably caused by the saturation effect in saliency-based optimisation (i.e., the double inclusion of saliency).



(a) original image (b) saliency map (c) ICM (d) PCM

Figure 3.4: Illustration of the comparison of the "information content map (ICM)" (c) extracted from IWSSIM, VIF or IWPSNR, the "phase congruency map (PCM)" (d) extracted from FSIM and a representative saliency map(i.e., Torralba (b))for one of the source images in the LIVE database (a).

Based on the observed trend, one may hypothesize that adding computational saliency produces a larger improvement for IQMs without built-in saliency than for IQMs that intrinsically include saliency aspects. To validate this hypothesis, we performed a straightforward statistical test. On account of a normally distributed dependent variable performance gain, a t-test was performed with two levels of the variable being the IQMs with built-in saliency (i.e., IWSSIM, VIF, FSIM, IWPSNR, NBAM and JNBM) and the IQMs without (i.e., PSNR, UQI, SSIM, MSSIM, NPBM and GBIM). The t-test results (p-value = 3.78e-24, $p < 0.05$ at 95% confidence level) showed that IQMs without built-in saliency (<gain>=2.3%) receive on average statistically significantly higher performance gain than IQMs with built-in saliency (<gain>=0.18%).

Since IQMs can also be characterized at a different aggregation level, using FR/NR as the classification variable, a practical question arises whether FR/NR has an impact on the performance gain, and if so, to what extent. To check such effect with a statistical analysis, a t-test was performed again in a similar way as described above, but with two new independent variables to substitute the variable with/without built-in saliency: i.e., FR and NR. The t-test results (p-value = 0.021, $p < 0.05$ at 95% confidence level) showed that overall NR IQMs (<gain>=2.5%) obtain a statistically significantly larger amount of performance gain than FR IQMs (<gain>=0.9%). This implies that applying computational saliency to an NR IQM has potential to significantly boost its performance in an effective way.

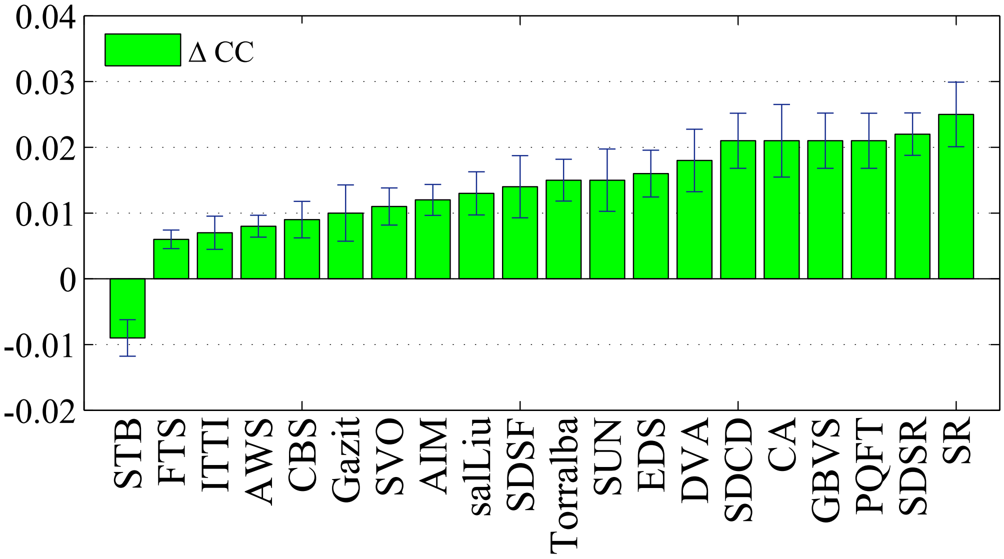


Figure 3.5: Illustration of the rankings of the saliency models in terms of the overall performance gain (expressed by ΔPLCC, averaged over all distortion types, and over all IQMs where appropriate) between an IQM and its saliency based version. The error bars indicate the 95% confidence interval.

3.4.2 Effect of saliency model dependency

There is a significant difference in performance gain between saliency models. Figure 3.5 illustrates the order of saliency models in terms of the average performance gain that can be obtained by adding individual models to IQMs. A promising gain is found when adding SR (<gain>=2.5%), SDSR (<gain>=2.2%), PQFT (<gain>=2.1%), GBVS(<gain>=2.1%), CA (<gain>=2.1%) and SDCD(<gain>= 2.1%) to IQMs. The gain achieved for these models is fairly comparable to (but not necessarily statistically significantly better than) the gain of adding “ground truth” FM (<gain>=2.0%) to IQMs. At the other extreme, STB(<gain>=-0.9%) tends to deteriorate the performance of IQMs, and saliency models, such as FTS(<gain>=0.6%), do not yield an evident profit for IQMs. Figure 3.6 illustrates the saliency models sitting at the two extremes of performance gain: the most profitable models (i.e., SR, SDSR, PQFT and GBVS) versus the least profitable models (i.e., STB and FTS). The comparison indicates that SR, SDSR, PQFT and GBVS make a sufficiently clear distinction between the salient and non-salient regions, which aligns with the appearance of FM as shown in Fig. 3.6. STB, which predicts the order in which the eyes move, often highlights the fixation locations (e.g., a certain portion of a cap) rather than salient regions (e.g., the entire cap). Adding such saliency to IQMs may result in an overestimation of localized distortions. The relatively lower performance gain obtained with FTS is possibly caused by the fact that it segments objects, which are sequentially labelled in a random order. As such, adding saliency in an IQM could randomly give more weight to artifacts in one object (e.g., the yellow cap) than that in another object (e.g., the red cap).

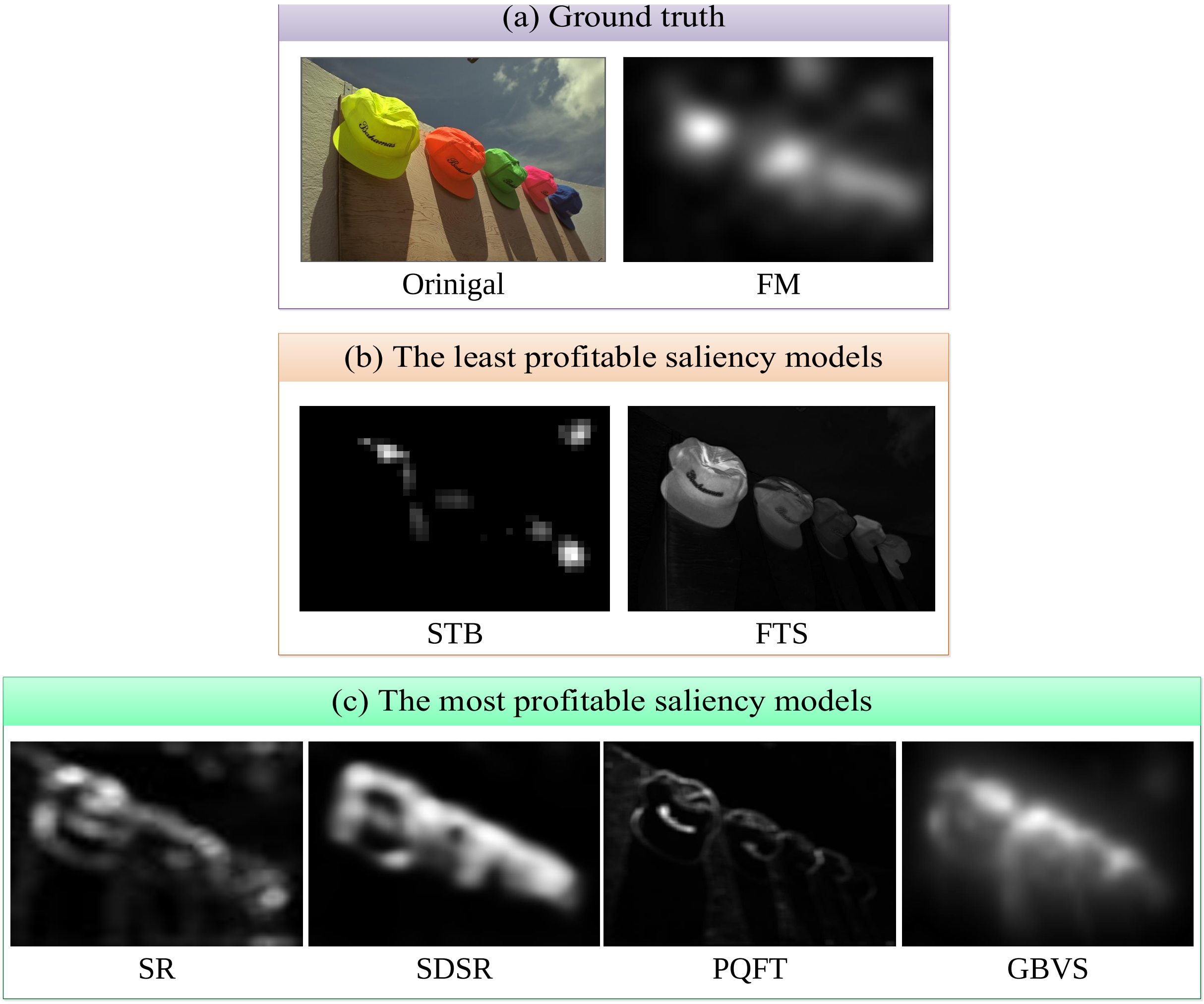

(a) Ground truth
Orinigal
FM
(b) The least profitable saliency models
STB
FTS
(c) The most profitable saliency models
SR
SDSR
PQFT
GBVS

Figure 3.6: Illustration of the saliency maps as the output of the least profitable saliency models and of the most profitable saliency models for IQMs. The original image is taken from the LIVE database.

Since it is customary to classify saliency models into two categories, which are referred to as salient object detection (SOD) and fixation prediction (FP), we checked whether and to what extent this categorical variable affects the performance gain. Based on the classification criteria defined in [115], CBS, FTS, salLiu, SVO, CA are categorized as SOD and the rest models belong to FP. A t-test was conducted with the performance gain as the dependent variable (note that it was tested to be normally distributed), and SOD and FP as independent variables. The results (p-value = 0.56, p $>$ 0.05 at 95% confidence level) revealed that there is no significant difference in performance gain between these two categories. This suggests that the classification of saliency models to SOD and FP does not have direct implications for the trend of changes in performance gain of IQMs.

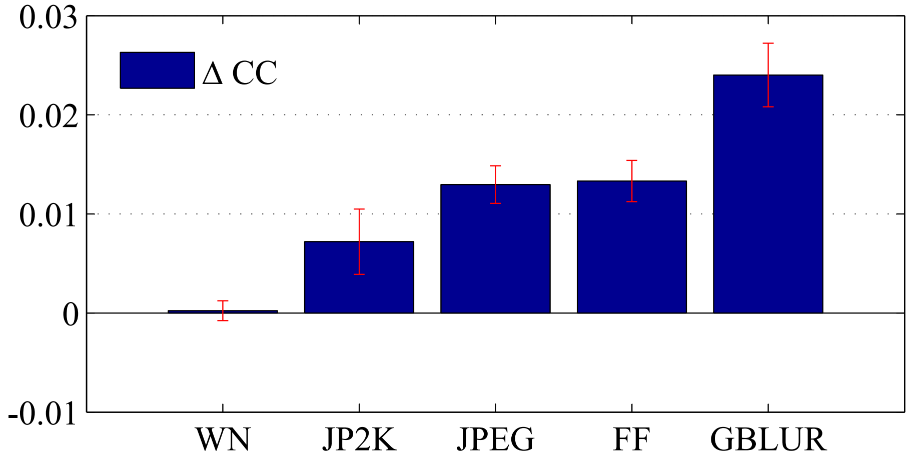


Figure 3.7: Illustration of the ranking in terms of the overall performance gain (expressed by ΔPLCC, averaged over all IQMs, and over all saliency models where appropriate) between an IQM and its saliency based version, when assessing WN, JP2K, JPEG, FF, and GBLUR. The error bars indicate the 95% confidence interval.

3.4.3 Effect of distortion type dependency

On average, the distortion type has a statistically significant effect on the performance gain, with the order as illustrated in Fig. 3.7. It shows that GBLUR (<gain>=2.4%) profits most from adding computational saliency in IQMs, followed by FF (<gain>=1.4%), JPEG (<gain>=1.3%), JP2K (<gain>=0.7%) and finally WN (<gain>=0%). Such variation in performance gain may be attributed to the intrinsic differences in perceptual characteristics between individual distortion types. In the case of an image degraded with WN as shown in Fig. 3.8(a), artifacts tend to be uniformly distributed over the entire image. At low quality, the distraction power of the (uniformly distributed) annoying artifacts is so strong that it may mask the effect of the natural scene saliency. As such, directly weighting the distortion map with saliency intrinsically underestimates the annoyance of the artifacts in the background, and their impact on the quality judgement. This case may eventually offset any possible increase in performance and, as a consequence, may explain the overall non-existing performance gain.

The promising performance gain obtained for GBLUR may be attributed to two possible causes. First, in the particular case of images distorted with both unintended blur (e.g., on a high-quality foreground object) and intended blur (e.g., in the intentionally blurred background to increase the field of depth), IQMs often confuse these two types of blur and process them in the same way. Adding saliency happens to circumvent such confusion by reducing the importance of blur in the background, and as such might improve the overall prediction performance of an IQM. Second, blur is predominantly perceived around strong edges in an image [37]. The addition of saliency effectively accounts for this perception by eliminating regions (e.g., the background) that are perceptually irrelevant to blur, and consequently may enhance the performance of an

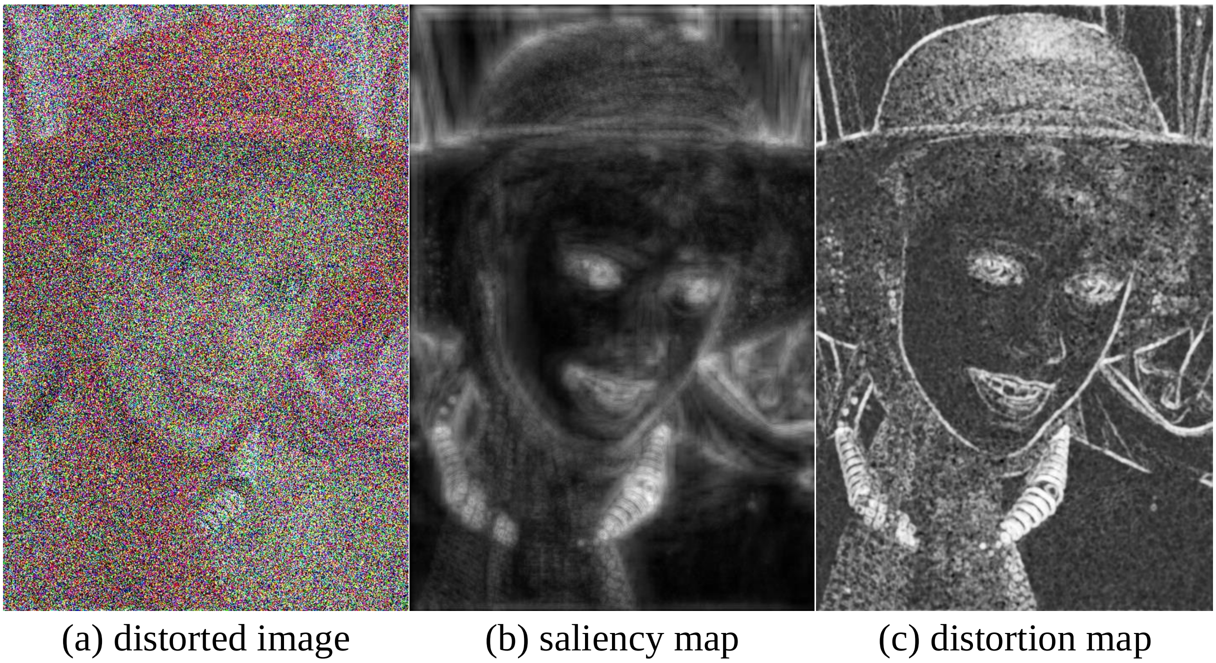


Figure 3.8: Illustration of an image distorted with white noise (WN) and its measured natural scene saliency and local distortions. (a) A WN distorted image extracted from LIVE database. (b) The saliency map (i.e., Torralba) based on the original image of (a) in the LIVE database. (c) The distortion_map of (a) calculated by an IQM (i.e., SSIM).

IQM for blur assessment. To further confirm whether adding saliency indeed preserves the perceptually relevant regions for blur, we first partitioned an image into blur-relevant (i.e., strong-edge positions) and blur-irrelevant (i.e., non-strong-edge positions) regions, and then compared the saliency residing in the relevant regions to that in the irrelevant regions. Figure 3.9 illustrates the comparison of the average saliency in the blur-relevant and blur-irrelevant regions, for the 29 source images of the LIVE database. It demonstrates that including saliency intrinsically retains the regions that are perceptually more relevant to perceived blur, and this explains the improvement of an IQM in assessing GBLUR.

In JPEG, JP2K and FF, the perceived artifacts tend to be randomly distributed over the entire image due to the luminance and texture masking of the HVS [5]. This could further confuse the issue of assessing artifacts with the addition of saliency, despite the general effectiveness as found in Fig. 3.7. Figure 3.10 illustrates a JPEG compressed image (bit rate = 0.4bit/pixel), and its corresponding saliency (i.e., generated by Torralba [71]). Due to HVS masking, this image exhibits imperceptible artifacts in the salient regions (e.g., the lighthouse and rocks in the foreground), but relatively annoying artifacts in the non-salient regions (e.g., the sky in the background). In such a demanding condition, directly combining the measured distortions with saliency to a large extent overlooks the impact of the background artifacts on the overall quality. In view of this, we may speculate such type of images may not profit from adding saliency in

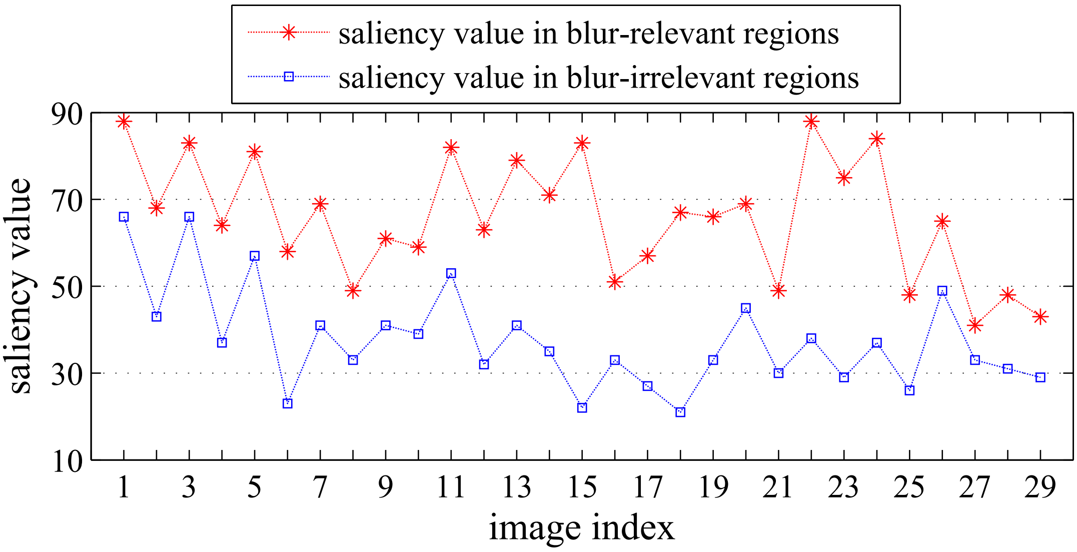


Figure 3.9: Illustration of the comparison of the averaged saliency residing in the blur-relevant regions (i.e., positions of the strong edges based on the Sobel edge detection) and blur-irrelevant regions (i.e., positions of the rest of the image) for the 29 source images of the LIVE database. The vertical axis indicates the averaged saliency value (based on the saliency model called Torralba), and the horizontal axis indicates the twenty-nine test images (the content and ordering of the images can be found in the LIVE database.

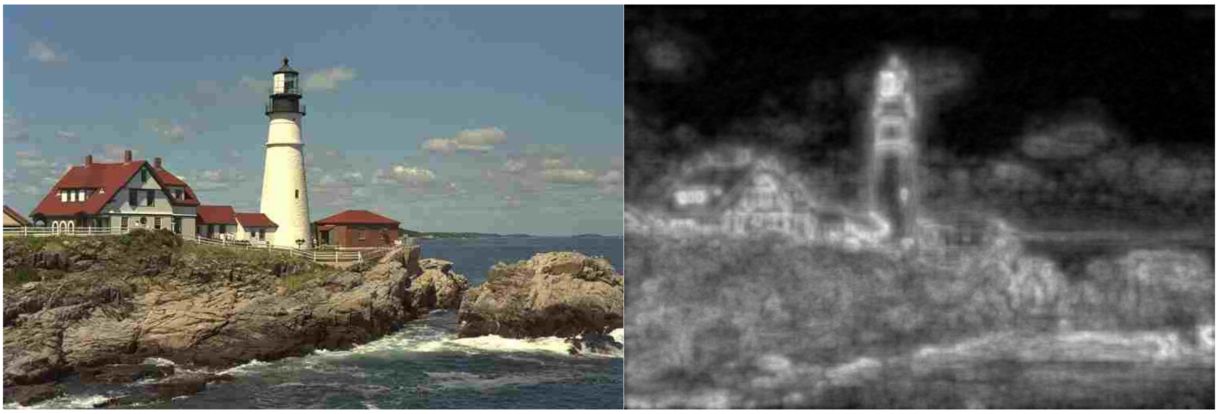

(a) JPEG compressed image (b) saliency map

Figure 3.10: Illustration of a JPEG compressed image at a bit rate of 0.4b/p, and its corresponding natural scene saliency as the output of a saliecny model (i.e., Torralba).

IQMs, which also implies that the performance gain obtained so far for JPEG, JP2K and FF may not be optimal amount. The overall positive gain as illustrated in Fig.7, however, can be explained by the fact that most of the images in the LIVE database exist of one of the following types: first, images having visible artifacts uniformly distributed over the entire image; second, images having the artifacts masked by the content in the less salient regions, but showing visible artifacts in the more salient regions. Obviously, for these two types of images, adding saliency is reasonably safe.

Also, as the speculation mentioned in [13] and [90], the observed trend that the amount of performance gain varies depending on the type of distortion may be associated with the performance of IQMs without saliency. For example, it may be more difficult to obtain a significant increase in performance by adding saliency when IQMs (without saliency) already achieve a high prediction performance for a given type of distortion. This phenomenon can be further revealed by checking the correlation between the original performance (without saliency) of IQMs and the performance gain (with the integration of saliency) of IQMs. The Pearson correlation coefficient between these two variables is -0.71 which indicates that the higher the original performance of an IQM, the more the gain is limited by adding saliency. Future study may focus on investigating the relative improvements as same amount of improvement can mean different for different IQMs.

3.5 Summary

In this chapter, a statistical evaluation was conducted to investigate the added value of including computational saliency in objective image quality assessment. The testbed comprised twenty best-known saliency models, twelve state of the art FR and NR IQMs, and five image distortion types. It resulted in 880 possible combinations: each represented a case of performance gain of a saliency-based IQM over its original version, when assessing the quality of images degraded with a given distortion type.

Based on the experimental results, we found that the current performance of visual saliency modelling is sufficient for IQMs to yield a statistically meaningful gain in their performance. On average, such improvement is fairly comparable to the gain that can be obtained by adding "ground-truth" eye-tracking data into IQMs. However, the actual amount of performance gain varies among individual combinations of the two variables: saliency models and IQMs. This variation directs the real-world applications of saliency-based IQMs, in which implementation choices are often confronted with a trade-off between performance and computational efficiency. The measured gain for a given combination can be used as a reference to assist in making decisions about how to balance the performance gain of a saliency-based IQM against the additional costs needed for the saliency modelling and inclusion.

To decide upon whether a saliency model is in a position to deliver an optimized performance gain for IQMs, it is essential to check the overall gain that can be actually obtained by adding this saliency model in state-of-the-art IQMs. We found a threshold value in the overall gain, i.e. 2%, above which the effectiveness of a saliency model, such as SR, SDSR, PQFT, GBVS, CA and SDCD, is comparable to that of the eye-tracking data and thus is considered to be an optimized amount. Such profit achieved by a saliency model, surprisingly, has no direct relevance to its measured prediction accuracy of human fixations. Moreover, the customary classification of saliency models (i.e., salient object detection and fixation prediction) is not informative on the trend of changes in performance gain. The most profitable models and the least profitable models can be found in both classes.

When it comes to the issues relating to the IQM dependency of the performance improvement, care should be taken to make a distinction between the IQMs with and without built-in saliency aspects. Adding computational saliency to the former category intrinsically confuses the workings of saliency inclusion, and often produces a smattering of profit. The performance of the latter category of IQMs, however, can be boosted to a large degree with the addition of computational saliency. In terms of a different aggregation level, NR IQMs significantly profit more from including computational saliency than FR IQMs.

The effectiveness of applying saliency-based IQMs in the assessment of different distortion types is subject to the perceptual characteristics of the distortions. The appearance of the perceived artifacts, such as their spatial distribution due to HVS masking, tends to influence the extent to which a certain image may profit from adding saliency to IQMs. Overall, we found that images degraded with Gaussian blur respond positively to the addition of saliency in IQMs, whereas saliency inclusion does not deliver added value when assessing the quality of images degraded with white noise. In practice, it should, however, be mindful of the images distorted with localized artifacts, e.g., JPEG, JP2K and FF, which may further confuse the operations of adding saliency in IQMs.

Knowledge as the outcome of this study is highly beneficial for the image quality community to have a better understanding of saliency modelling and inclusion in IQMs. Our findings are valuable to guide developers or users of IQMs to decide on appropriate saliency model for their specific application environments. The statistical evaluation also provides a grounding for the quest of a more reliable saliency modelling in the context of image quality assessment.

Chapter 4

A Reliable Eye-tracking Database for Image Quality Research

4.1 Introduction

In Chapter 3, we have statistically assessed the benefits of integrating computational saliency in IQMs. However, finding ways to achieve such integration in a perceptually optimised way remains largely unexplored. The challenge lies in the fact that our knowledge about how saliency is actually affected by the concurrence of visual signals and their distortions as well as the associated implications for image quality judgements is very limited. To advance the research, dedicated eye-tracking experiments are essential.

Psychophysical studies have been undertaken to better understand visual saliency in relation to image quality assessment [12, 13, 14, 15]. For example, an eye-tracking study was performed in [12] to investigate (via visual inspection of fixation patterns) how task-free fixations (i.e., saliency) of undistorted images may be affected by two variables, i.e., quality rating task and visual distortion. Based on the visualisations of eye-tracking data, white noise and blurring (under quality rating conditions) were not observed to impact the fixation patterns (relative to the task-free conditions), whereas the impact tends to be more obvious in the case of compression artifacts. On the contrary, the similar eye-tracking experiment conducted in [15] revealed that white noise and blur do lead to changes in gaze patterns and that this impact is predominately driven by the intensity of distortion. In [14], task-free eye-tracking experiments were conducted to investigate how JPEG compression affects fixations. It showed that the impact of JPEG artifacts on fixations is more disruptive at low image quality than the high quality. The eye-tracking data in [13] indicated that fixations change as visual distortion occurs, and that the extent of the change seems to be more related to the strength of artifacts rather than the type of artifacts. In general, psychophysical studies revealed that visual distortions may lead to a deviation from the natural scene saliency, and that such deviation tends to depend on the type of distortion, the level of distortion and the visual content.

Notwithstanding the above effort, it should be noted that the generalisability of the findings reported in these studies remains limited by the choices made in their experimental design. For example, some experiments used a limited number of human subjects (i.e., 5 subjects were used in [12]). Some experiments were restricted to a small degree of stimulus variability in terms of scene content (i.e., 6 original images were used in [13], 1 distortion type was used in [14] and 2 degradation levels were used in [12]). Some eye-tracking studies involved top-down aspects of visual attention (e.g., the involvement of a quality rating task) rather than studying free-viewing bottom-up saliency [12, 13, 15]).

Apart from the above drawbacks, existing studies by their nature potentially suffer from an inherent bias due to the involvement of stimulus repetition. Typical eye-tracking data collection for the purpose of image quality assessment often involves each observer viewing the same scene repeatedly several times (with multiple variations of distortion) throughout a session. This repetition (i.e., repeated versions of the same scene) becomes massive as the number of distortion types and/or levels increases and would potentially skew the intended eye-tracking data. In [116], eye-tracking data were collected where participants first viewed 12 short videos and then after a 2-minute break they viewed the same 12 videos again. The results showed that there was a notable difference in the locations of the participants' gaze for the first and second viewings of the same video. The eye-tracking experiments in [117] included 10 original videos and their 50 impaired versions (i.e., five levels of degradation per original). The results showed evidence for a memory or learning effect for several viewings of the same video content, and that the observers' gaze behaviour tended to be affected by the involvement of stimulus repetition. Both studies suggest that to ensure the consistency of oculomotor behaviour throughout the experiment (i.e., observing stimuli naturally rather than being forced to learn where to look for visual artifacts) and as such to guarantee the reliability of fixation data collection, there is a need for reducing the impact of stimulus repetition.

In this chapter, a new experimental methodology with carefully control mechanisms was proposed. This methodology allowed reliably obtaining a substantial eye-tracking data with a large degree of stimulus variability in terms of scene content, distortion type as well as degradation level. Unlike previous eye-tracking studies that have focused more on a limited dataset and rather qualitative analysis, the resulting eye-tracking data enabled us to thoroughly evaluate the relation between saliency and distortion. In particular, a statistical analysis was performed to provide a comprehensive view of the extent to which different types of distortion with each represented at different levels of degradation can actually affect fixation deployment. Moreover, an important question has arisen whether saliency derived from an original scene or that from the same scene affected by distortions should be included in IQMs. Based on our eye-tracking data, we assessed whether the difference between both types of saliency is sufficiently large to affect the performance gain for existing IQMs.

4.2 A Refined Experimental Methodology

Unlike previous studies, our experiment contains a large degree of stimulus variability in terms of scene content, distortion type as well as distortion level. In addition, a dedicated protocol was devised to eliminate potential bias due to the involvement of massive stimulus repetition, which inherently occurs in a typical image quality study. An eye-tracking database was collected with 160 human observers and 288 test stimuli. Each stimulus was viewed by 20 observers, resulting in 5760 eye movement trials (i.e., 288 × 20 = 5760).

4.2.1 Stimuli

A set of test stimuli was constructed by systematically selecting images from the LIVE image quality database [16]. The construction of reference images and distorted images is detailed below.

From the fixation deployment perspective, natural scenes can be classified based on the degree of saliency dispersion [92]. As the observation revealed from eye-tracking studies in [118, 119], if an image contains highly salient objects, then most viewers will concentrate their fixations around them, whereas if there is no obvious object-of-interest viewers' fixations will appear as a more evenly distributed pattern. Thus, images with salient objects tend to have less variation in fixations between viewers than images without salient objects. By the use of eye-tracking data in [92], the degree of saliency dispersion — the degree of agreement between observers for human fixations — was determined and used to categorise all the 29 reference images in the LIVE database. The results showed that 6 images clustered around the range of small degree of saliency dispersion, 19 images clustered around the range of medium degree of saliency dispersion and 4 images clustered around the range of large degree of saliency dispersion. To maximum the stimulus variability of the database, all the images in the small saliency dispersion category and in the large saliency dispersion category were included. Also, to mitigate the unbalanced distribution of reference images, we decided to remove some images having a medium degree of saliency dispersion (i.e., to keep 8 images out of 19). This yielded a rather balanced set of 18 reference images as illustrated in Fig. 4.1. The new make-up consists of 6 images of a small degree of saliency dispersion (e.g., images with distinct foreground/background configurations); 4 images of a greater saliency dispersion (e.g., images without any specific object-of-interest); and 8 images that fall into the range of medium degree of saliency dispersion.

Distorted stimuli used in our experiment cover the full range of distortion types available in the LIVE database, including white noise (WN), JPEG compression (JPEG), Gaussian blur (GBLUR), JPEG2000 compression (JP2K) and simulated fast-fading in wireless channels (FF).

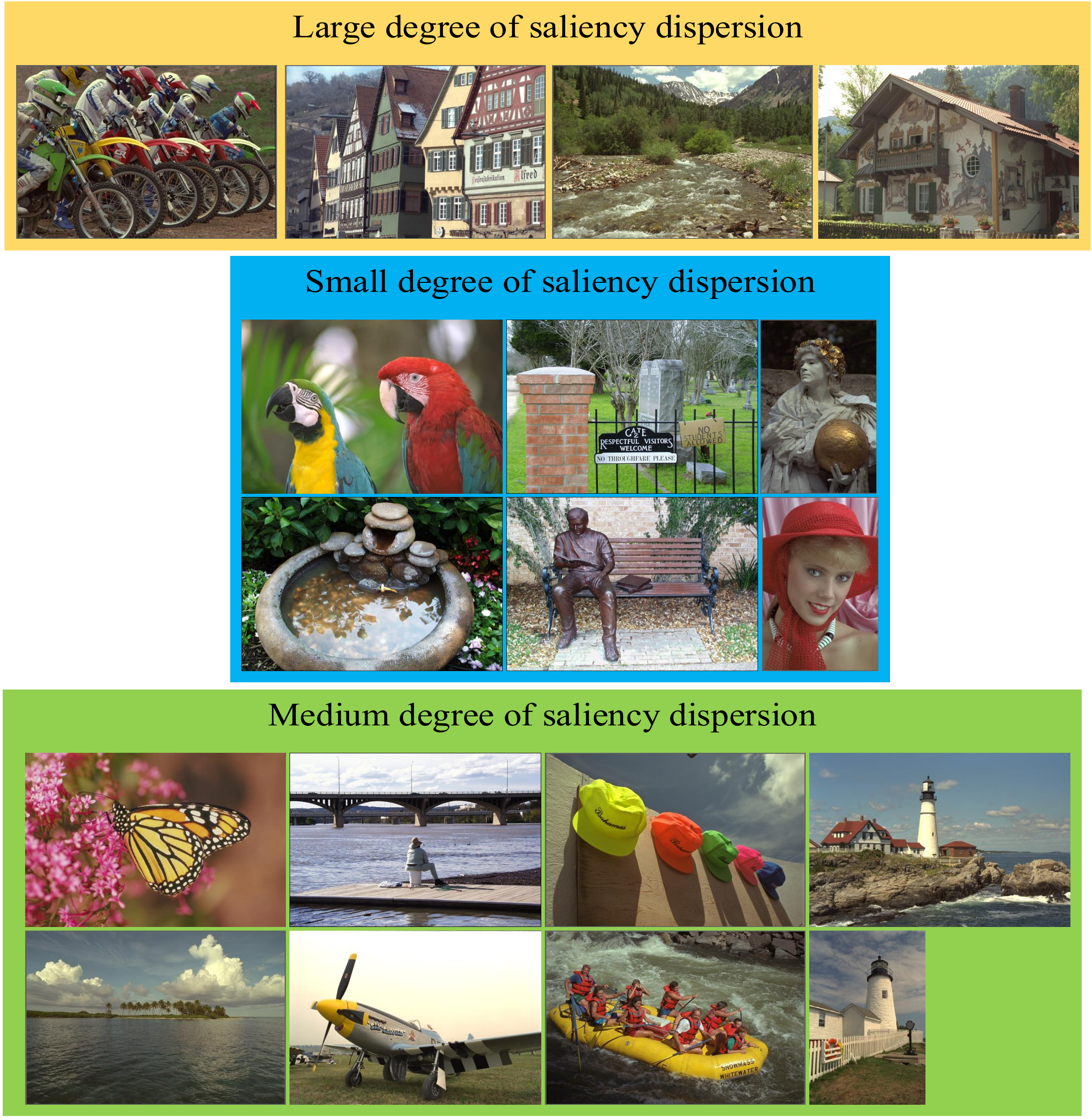

Large degree of saliency dispersion
Small degree of saliency dispersion
Medium degree of saliency dispersion

Figure 4.1: Illustration of reference images with different degrees of saliency dispersion used in our experiment, which yield 288 test images.

For each distortion type, three distorted versions per reference image were systematically selected, which were intended to reflect three distinct levels of perceived quality: “High” (i.e., with perceptible but not annoying artifacts), “Medium” (i.e., with noticeable and annoying artifacts) and “Low” (i.e., with very annoying artifacts). Taking advantage of the LIVE database that contains per image a “ground truth” quality score (i.e., DMOS), distortion strengths/levels were adjusted perceptually by using the following mapping: DMOS = [10, 40] to “High” quality, DMOS = [40, 70] to “Medium” quality and DMOS = [70, 100] to “Low” quality. By doing so, for a specific distortion type, the selected 18 “High” quality versions of reference images

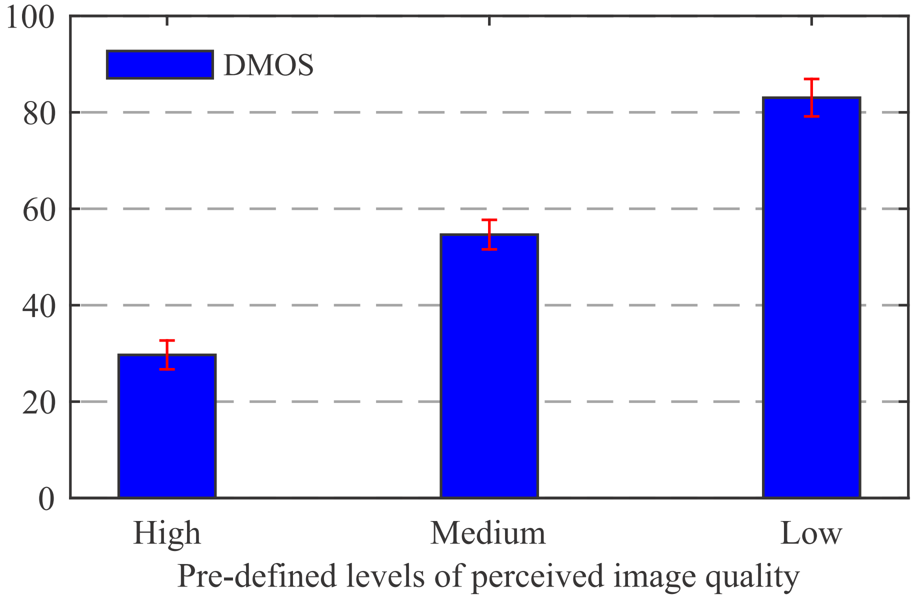


Figure 4.2: Illustration of average DMOS of images assigned to a pre-defined level of distortion. The distortion levels are meant to reflect three perceptually distinguishable levels of image quality (i.e., denoted as "High", "Medium" and "Low"). The error bars indicate a 95% confidence interval.

are meant to have approximately the same perceived quality; and similarly for other distortion levels (i.e., "Medium" and "Low"). In addition, a "High" quality version of any reference image chosen under a specific distortion type is meant to have approximately the same perceived quality as the "High" quality version of the same reference image chosen under any other distortion type; and similarly for other distortion levels (i.e., "Medium" and "Low"). The selection procedure resulted in a set of **288** test stimuli (including the reference images) from the LIVE database. Figure 4.2 illustrates the average DMOS of images (i.e., 90 images based on 18 reference images × 5 distortion types) assigned to individual distortion levels. It clearly shows three distinct means of DMOS (i.e., 30, 55 and 83 within the score range [0, 100]); and hypothesis testing (i.e., based on t-test preceded by a test for the assumption of normality) reveals that the difference between these three pre-defined categories is statistically significant (i.e., p-value = 5.88e-11 between high and medium, p-value = 2.37e-12 between medium and low) with $p < 0.05$ at the 95% confidence level).

4.2.2 Proposed experimental protocol

There is little consensus on which method is the most appropriate for the conduct of an eye-tracking experiment for the purpose of image quality study. A within-subjects method, in which the same group of subjects views all test stimuli, is commonly used in relevant studies [90, 12, 15, 14, 13]. This experimental methodology, however, potentially contaminates the results due to carry-over effects, which refer to any effect that carries over from one experimental condition to another [120]. Such effects become more pronounced as the number of test stimuli and/or the rate of stimulus repetition increase in eye-tracking. In our case, the test dataset contains a total of 288 stimuli representing 16 repeated versions (i.e., 15 distorted + 1 original) per reference image, which makes the use of a within-subjects method prone to undesirable effects such as fatigue, boredom and learning from practice and experience, and thus increases the chances of skewing the results. To overcome these problems, an alternative method, namely between-subjects [121] was employed in our experiment. In a between-subjects method, multiple groups of subjects are randomly assigned to partitions of test stimuli, each contains little or no stimulus repetition. We decided to divide the test dataset into 8 partitions of 36 stimuli each; and to allow only 2 repeated versions of the same scene in each partition. To further reduce the carry-over effects, each session per subject was divided into two sub-sessions with a "washout" period between sub-sessions; and by doing so, each subject effectively had to view 18 stimuli with no stimulus repetition in a separate session. Mechanisms were further applied to control the order in which participants per group perform their tasks: (1) half of the participants view the first half partition of stimuli first, and half of the participants view the second half partition first; (2) the stimuli in each sub-session are presented to each subject in a random order. A dedicated control mechanism was also adopted in each sub-session to deliberately include a mixture of all distortion types and the full range of distortion levels. We recruited 160 participants in our experiment, consisting of 80 male and 80 female university students and staff members (between 19 to 42 years of age), all inexperienced with image quality assessment and eye-tracking. The experiment went through the required ethics review process and all the participants volunteered for this eye-tracking experiment (i.e., no payment was made to the participants). The participants were not tested for vision defects, and we considered their verbal expression of the soundness of vision was adequate. The participants were first randomly divided into 8 groups of equal size, each with 10 males and 10 females; and the 8 groups of subjects were then randomly assigned to 8 partitions of stimuli. Based on the rule of thumb for determining sample size in relevant studies (i.e., 5-15 subjects per test stimulus), we assume 20 per stimulus is an adequate sample size (note that the validity of sample size will be further quantitatively tested in Section. 4.3).

4.2.3 Experimental procedure

We set up a standard office environment as to the recommendations of [4] for the conduct of our experiment. The test stimuli were displayed on a 19-inch LCD monitor (native resolution is 1024 × 768 pixels). The viewing distance was set to be approximately 60cm. Eye movements were recorded using an image processing based contact-free tracking system with sufficient

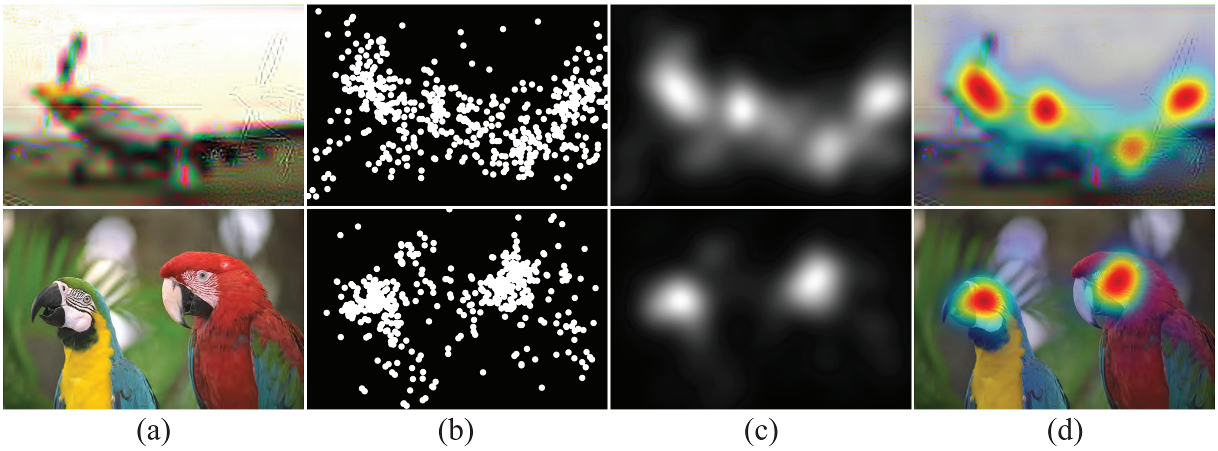


Figure 4.3: (a) Two sample stimuli of distinct perceived quality (DMOS = 95.96 (top image) and DMOS = 32.26 (bottom image)). (b) The collection of human eye fixations over 20 subjects. (c) Grayscale fixation maps (the darker the regions are, the lower the saliency is). (d) Saliency superimposed on the sample stimuli.

head movement compensation (SensoMotoric Instrument (SMI) RED-m). The eye tracking system features a sampling rate of 120Hz, a spatial resolution of 0.1 degree and a gaze position accuracy of 0.5 degree. Each subject was provided with instructions on the purpose and general procedure of the experiment before the start of the actual experiment. Each session per subject contained two successive sub-sessions with a break of 60 minutes between sub-sessions. Since each subject had only two viewings of the same scene, the 60-minute "washout" period was considered sufficient to balance between further reducing the carry-over effects and completing the entire data collection within a reasonable time-scale. Each individual sub-session was preceded by a 9-points calibration of the eye-tracking equipment. The participants were instructed to look at the stimuli in a natural way ("view it as you normally would"). Each stimulus was shown for 10 seconds followed by a mid-gray screen of 3 seconds.

4.3 Experimental Results

4.3.1 Fixation map

A binary fixation map representative for stimulus-driven, bottom-up visual attention was derived from the recorded fixations. Fixations were extracted from the raw eye-tracking data using the SMI BeGaze Analysis Software with minimum fixation duration threshold set to 100ms. A fixation was defined by SMI's Software using the dispersal and duration based algorithm established in [122]. Figure 4.3(b) illustrates the collection of fixations over all subjects (i.e., 20) for each of the two sample stimuli. To construct a grayscale map for an average human

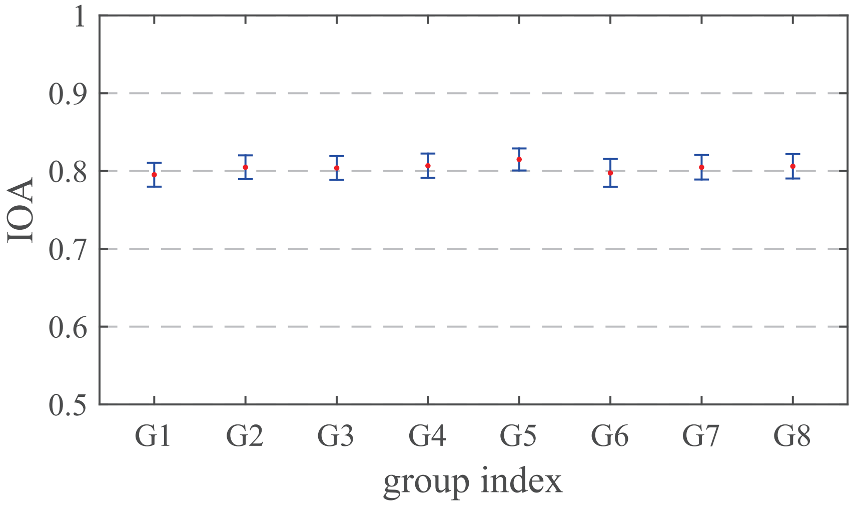


Figure 4.4: Illustration of inter-observer agreement (IOA) value averaged over all stimuli assigned for each subject group in our experiment. The error bars indicate a 95% confidence interval.

observer, each fixation location (contained in the aggregated data as shown in Fig. 4.3(b)) gave rise to a gray-scale patch that simulates the foveal vision of the HVS. The activity of the patch was modelled as a Gaussian distribution of which the width approximates the size of the fovea (2 degree of visual angle). As treated similarly in relevant literature (see e.g., [13, 14, 90]) the duration of fixation was not included when creating a grayscale fixation map.

4.3.2 Validation: reliability testing

Since standardised methodology for the collection of eye-tracking data does not exist, researchers often follow best practice guidelines for the design of their own experiments. The resulting data, however, differ in their reliability depending on the choices made in the experimental methodology, such as the sample size and the ways of presenting stimuli [57]. To make use of eye-tracking data as a solid "ground truth", it is crucial to validate the reliability of the collected data. We, therefore, proposed and performed systematic reliability testing to assess: (1) whether the variances in the eye-tracking data obtained from different subject groups (in a between-subjects method) are similar; (2) whether the sample size (number of participants) per stimulus is sufficient to create a stable fixation map; and (3) whether the eye-tracking data collected in our study are comparable to similar data obtained from other independent studies. Note, hereafter, when performing a statistical significance test, if the assumption of normality is tested to be satisfied a parametric test (e.g., t-test) is used; otherwise a non-parametric alternative (e.g., Wilcoxon signed rank test) is used.

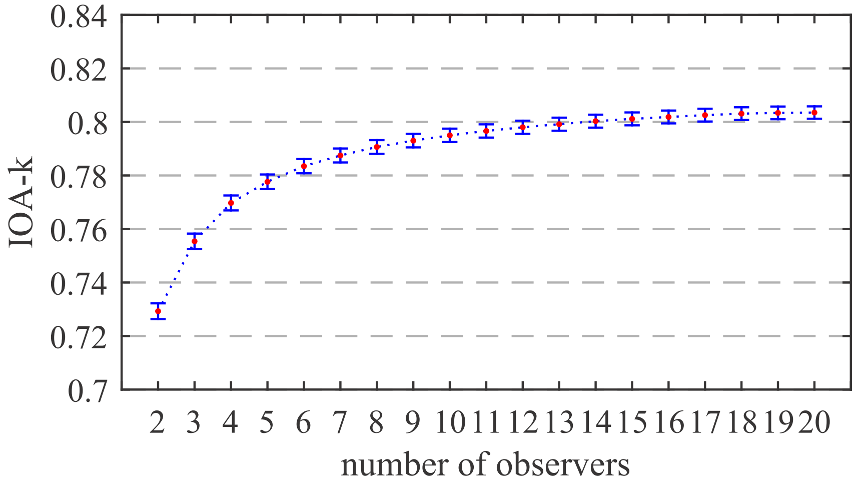


Figure 4.5: Illustration of inter-k-observer agreement (IOA-k) value averaged over all stimuli contained in our entire dataset. The error bars indicate a 95% confidence interval.

Homogeneity of variances between groups

Since a between-subjects method was adopted, assuming the representativeness of participants in each group is satisfied, we tested whether variances of eye-tracking data across all groups are homogeneous. To identify such homogeneity, we measured the inter-observer agreement (IOA), which refers to the degree of agreement in saliency among observers viewing the same stimulus [123, 124]. In our implementation, per stimulus and per subject group, IOA was quantified by comparing the fixation map generated from the fixations over all-except-one observers to the fixation map built upon on the fixations of the excluded observer; and by repeating this operation so that each observer serves as the excluded subject once. The similarity between two fixation maps is commonly measured by AUC (i.e., area under the receiver operating characteristic curve) [44]. Figure 4.4 illustrates the IOA value averaged over all stimuli assigned to each subject group in our experiment. It shows that the IOA remains similar across eight groups. A statistical significance test (i.e., analysis of variance (ANOVA)) was performed and the results showed that there is no statistically significant difference between groups (p-value = 0.41, $p > 0.05$ at the 95% confidence level). The above evaluation indicates that a high degree of consistency across groups was found in our data collection.

Data (saliency) saturation

There is, unfortunately, no general agreement on how many participants are adequate to achieve reliable eye-tracking data. Researchers often use “data saturation” as a guiding principle to check whether a given/chosen sample size is sufficient to cause a “saturated” fixation map. This means a fixation map reaches the point at which no new information is observed. We tested the adequacy of sample size required to reach saliency “saturation” (i.e., a proxy of sufficient degree of reliability) in our experimental data. The validation was again based on the principle of IOA, which is extended to an inter-*k*-observer agreement measure (i.e., referred to as IOA-*k*, and k=2, 3...20). More specifically, for a given stimulus, IOA-*k* was calculated by randomly selecting k participants among all observers. Figure 4.5 illustrates the IOA-*k* value averaged over all stimuli contained in our entire dataset. It shows that “saturation” occurs with 16 participants, although a reasonably high degree of consistency in fixation deployment is already reached with 12 participants. It demonstrates that our chosen number of 20 observers for each subject group is fairly sufficient to yield a stable/saturated fixation map.

Cross-database similarity

To further evaluate the reliability of our eye-tracking data as a “ground truth”, we compared our data to other relevant databases that are publicly available and obtained from independent laboratories. In terms of free-viewing eye movement recordings related to the LIVE database, there exist three widely cited eye-tracking databases (with stimuli being only the 29 reference images of the LIVE database), namely TUD [90], UN [110] and UWS [11]. A comparative study was already conducted in [110], and showed a high degree of similarity between these databases, despite the fact that they were independently collected under different experimental conditions. As a reference provided in [110], for the same image, when comparing its two independently generated fixation maps by means of Pearson correlation, the result that falls into the range [0.8, 0.9] indicates a high degree of similarity. Since we only selected 18 reference images from the LIVE database, the comparison had to be based on these 18 images only. The Pearson correlation averaged over all images between our data and TUD is 0.87; and is 0.87 and 0.89 with respect to UN and UWS, respectively. This suggests that our eye-tracking data should be considered as reliable “ground truth”.

4.3.3 Validation: impact of stimulus repetition

The above testings have validated the reliability of the collected eye-tracking data. We now justify the necessity of the proposed methodology by proving stimulus repetition would bias the gaze data. We hereby conducted a dedicated eye-tracking experiment in a within-subjects fashion, combining the ideas of both [116] and [117] as mentioned in Section 4.1. Note our main purpose here is to raise awareness of the need for eliminating stimulus repetition in the scenario where subjects have to view the same scene repeatedly, e.g., 16 times, rather than compare the general usage of different subjective testing methodologies. Our experiment aimed to investigate two aspects: 1) how stimulus repetition affects fixation behaviour when viewing several distorted versions of the same scene (as also similarly studied for videos in [117]); 2) how stimulus repetition affects fixation behaviour when viewing several times the same undistorted scene (as also similarly studied for videos in [116]).

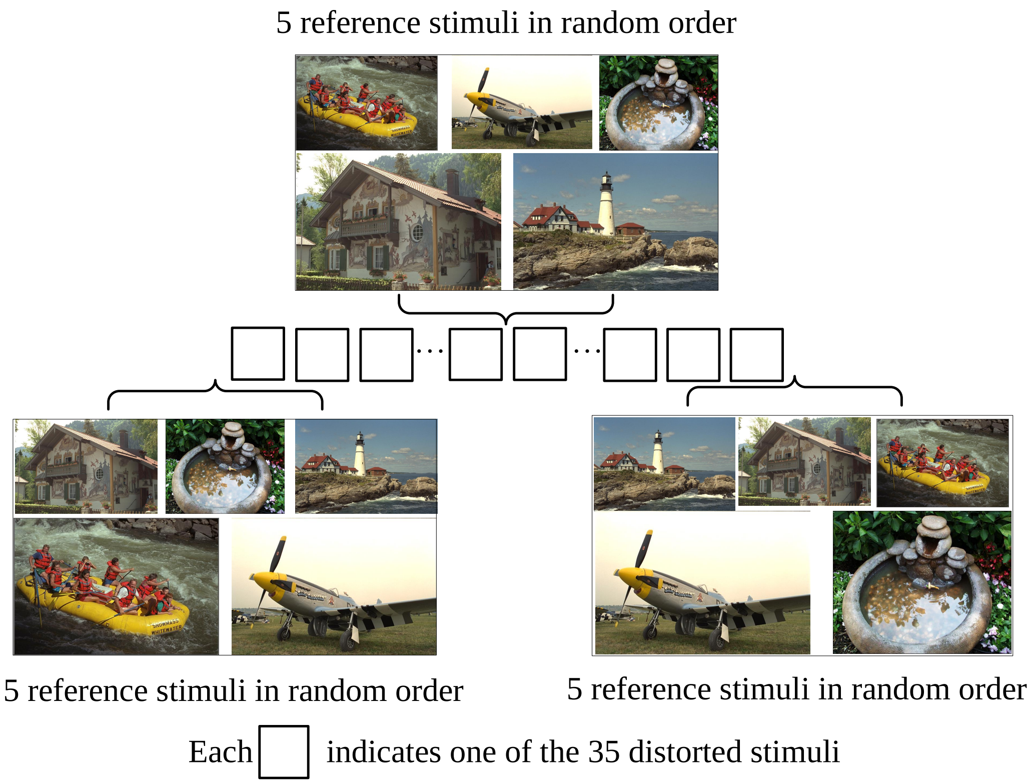

5 reference stimuli in random order
5 reference stimuli in random order
5 reference stimuli in random order
Each
indicates one of the 35 distorted stimuli

Figure 4.6: The construction of stimuli in a single trail. The boxes indicate 35 stimuli in random order. The 5 original images, as a group, are inserted in the front end, middle and back end of each trail in random order.

We chose five reference images to construct our test stimuli. In creating distorted stimuli, we selected 7 distorted images (covering all available distortion types and the full range of DMOS) per content from the LIVE database, resulting in 35 distorted images. In creating undistorted stimuli, we just used the 5 reference images three times. This gave a total of 50 test stimuli. As illustrated in Fig. 4.6, the 35 distorted stimuli were presented in a random order to each participant. The three groups of the same reference images (presented in a random order within group) were positioned in the beginning, middle and end of the presentation. Therefore, in terms of the distorted stimuli, there are 7 repetitions per content; and in terms of the undistorted stimuli, there are 3 repetitions per content. We recruited 20 participants (10 females and 10 males) in our experiment. Each participant viewed freely all stimuli. Each stimulus was shown for 10 seconds followed by a mid-grey screen for 3 seconds. We followed the same experimental set-up as described in Section 4.2.3.

The effects for distorted stimuli (7 repetitions)

For each participant, first the similarity in fixations between each distorted image and the corresponding reference image (presented in the beginning) was measured by AUC. Then, the 7 AUC values per content were ranked in the order of viewing, averaged over all contents and all participants as shown in Fig. 4.7. It clearly shows the general trend that the similarity decreases as the viewing order increases, independent of the image content, distortion type and distortion level. The results of t-test showed that there is a statistically significant difference between the 1st viewing and the Nth viewing (N=3 to 7) with the p-value ranging from 0.028 to 0.046 for each pair, all smaller than 0.05 at the 95% confidence level. This suggests that stimulus repetition can significantly impact the fixation behaviour, and consequently bias the intended fixation data.

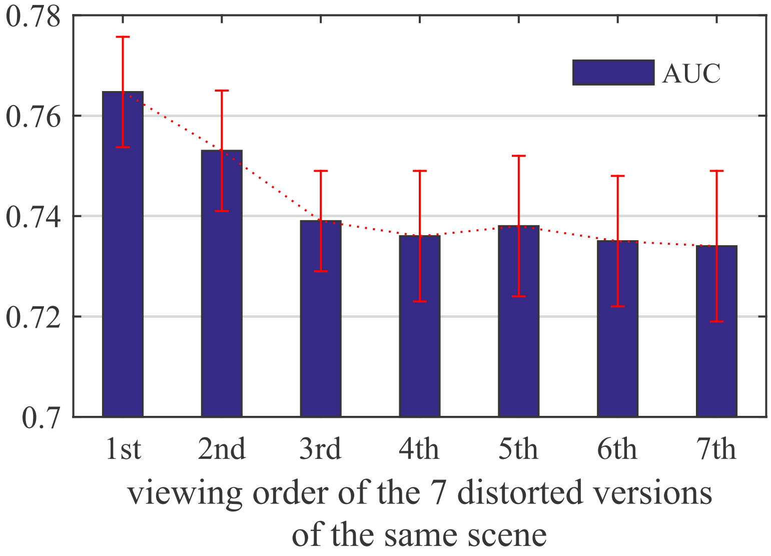


Figure 4.7: Illustration of the impact of stimulus repetition on fixation behaviour. When viewing 7 distorted versions of the same scene, the similarity in fixations (measured by AUC) relative to its original decreases as the viewing order increases. The error bars indicate a 95% confidence interval.

The effects for undistorted stimuli (3 repetitions)

A mean fixation map (over all subjects) was produced for each undistorted stimulus, and was compared by AUC to the corresponding baseline fixation map taken from the TUD database [90]. The fixation maps contained in the TUD database were collected under task-free, no distortion, no stimulus repetition conditions, using the reference images of the LIVE database. Figure 4.8 illustrates the AUC values in viewing order, averaged over all 5 reference images. It shows that the similarity dramatically drops after the first viewing of a scene, independent of image

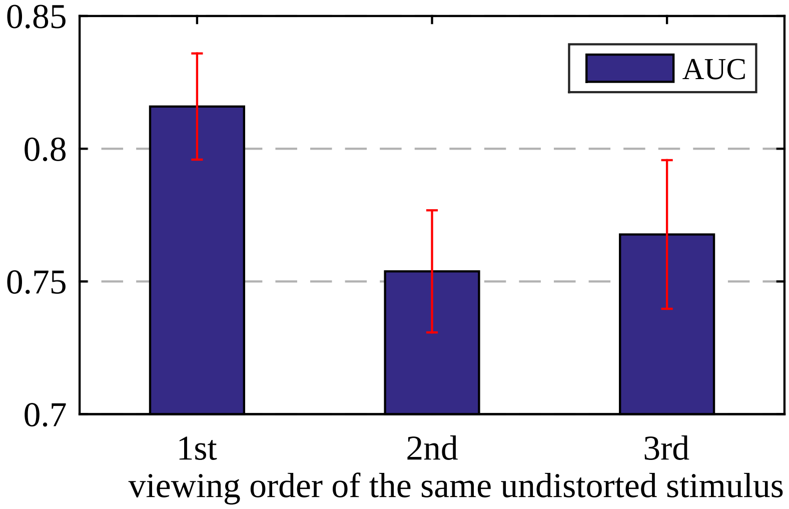


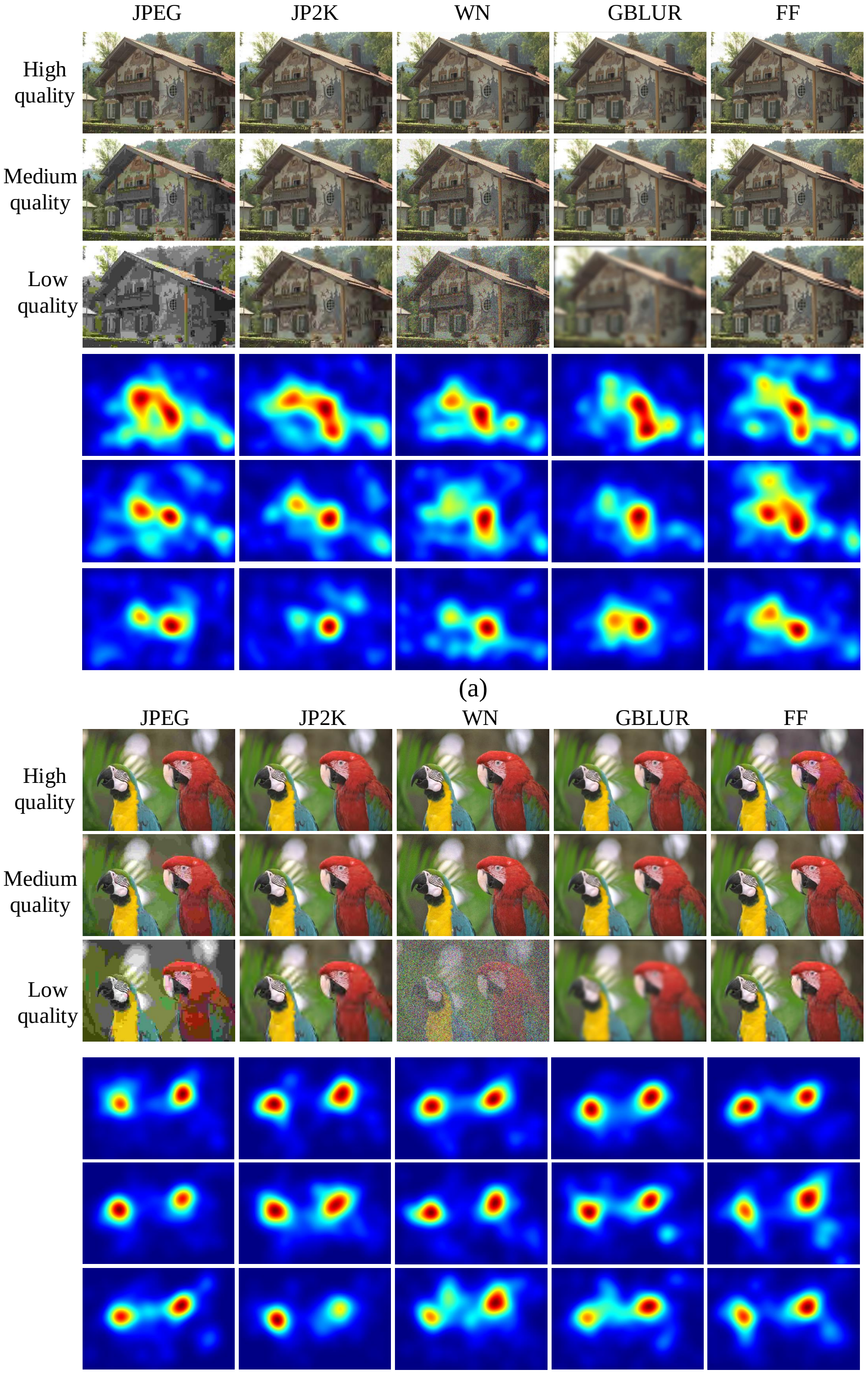
Figure 4.8: Illustration of the impact of stimulus repletion on fixation behaviour. When viewing 3 times the same undistorted scene, the similarity in fixations (measured by AUC) relative to its baseline taken from the TUD database decreases as the viewing order increases. The error bars indicate a 95% confidence interval.

content. A Wilcoxon signed rank test showed that there is a statistically significant difference between the first and the second viewing (p-value = 0.018) and between the first and the third viewing (p-value = 0.043), with $p < 0.05$ at the 95% confidence level.

The above study provides evidence that when subjects view the same stimuli repeatedly the fixation data are likely to be biased, and care should be taken to eliminate the effect of stimulus repetition in such a scenario.

4.3.4 Fixation deployment

Figure 4.9(a) illustrates an overview of all distorted versions (5 distortion types × 3 distortion levels) of a reference image (of a large degree of saliency dispersion) and their corresponding fixation maps (i.e., referred to as distorted scene saliency (DSS)). The same layout of distorted images and DSS for a different reference image (of a small degree of saliency dispersion) is illustrated in Fig. 4.9(b). The grids visualise typical correspondences and differences between DSS rooted from the same reference image. In general, there exist consistent patterns among the relevant DSS, e.g., the highly salient regions tend to cluster around the same positions. However, there are some deviations, which are seemingly caused by either the distortion type or distortion level. It is observed in Fig. 4.9(a) that as the quality degrades (i.e., the strength of distortion increases) the saliency patterns become more convergent (i.e., less amount of heated areas in DSS); and that at the same distortion level how saliency dispersion tends to depend on the distortion type, e.g., at “High” quality saliency is more spread out for JPEG, JP2K and FF



(a)

(b)

Figure 4.9: (a) Illustration of all distorted versions of a reference image (of a large degree of saliency dispersion) and their corresponding fixation maps. The same layout of distorted images and fixation maps for a different reference image (of a small degree of saliency dispersion) is illustrated in (b).

than for WN and GBLUR. In addition, the two examples (rooted from two different reference images) exhibit different trends in terms of the variation in the array of DSS. For example, the change in quality seems to cause a more obvious rate of convergence in saliency in Fig. 4.9(a) than in Fig. 4.9(b). This may be due to the fact that the two reference images fall into distinct categories of visual content in terms of saliency dispersion (see Fig. 4.1). It implies that image content also has an impact on the deployment of DSS, as already mentioned in [92].

4.4 Interaction Between Saliency and Distortion

The resulting eye-tracking data represent sufficient statistical power, which allows further statistical analysis on the observed tendencies in the changes of saliency induced by the changes of image quality aspects. More specifically, we evaluated the impact of three individual categorical variables (i.e., distortion type, distortion level and image content) on the deployment of fixation.

4.4.1 Evaluation criteria

We used saliency derived from the original undistorted scene (i.e., referred to as scene saliency (SS)) as the reference, and quantified the deviation of DSS from its corresponding reference SS. The deviation between two fixation maps was quantified by three similarity measures, namely CC, NSS and AUC.

4.4.2 Evaluation results

The statistical evaluation was based on 270 data points (i.e., 270 distorted stimuli rooted from 18 originals) of SS-DSS similarity (i.e., the similarity calculated by CC, NSS and AUC between a given DSS and its corresponding SS). A full factorial ANOVA was conducted with the SS-DSS similarity as the dependent variable (the test for the assumption of normality indicated that the dependent variable was normally distributed); and the distortion type, distortion level and image content as independent variables. The results are summarized in Table. 4.1, and show that all main effects (except for the case of distortion type when AUC and NSS are used for SS-DSS similarity) are statistically significant.

Impact of distortion type on SS-DSS similarity

As shown in Table 4.1, “distortion type” has a statistically significant effect on SS-DSS similarity measured by CC. The same effect, however, is not found when the SS-DSS similarity is

Table 4.1: Results of the ANOVA to evaluate the impact of distortion type, distortion level and image content on the measured similarity between SS and DSS. df denotes degree of freedom, F denotes F-ratio and Sig denotes the significance level.

ANOVA		CC		NSS		AUC	
Source	df	F	Sig	F	Sig	F	Sig
Distortion type	4	2.89	**.02**	1.48	.21	0.92	.45
Distortion level	2	46.7	**.00**	23.44	**.00**	27.89	**.00**
Image content	2	124.33	**.00**	439.96	**.00**	483.7	**.00**
Distortion type * Distortion level	8	2.03	**.04**	1.15	.33	0.95	.48
Distortion type * Image content	8	1.92	**.05**	0.74	.66	0.96	.47
Distortion level * Image content	4	2.82	**.03**	0.1	.98	1.02	.39
Distortion type * Distortion level * Image content	16	0.71	.79	0.32	.99	0.4	.98

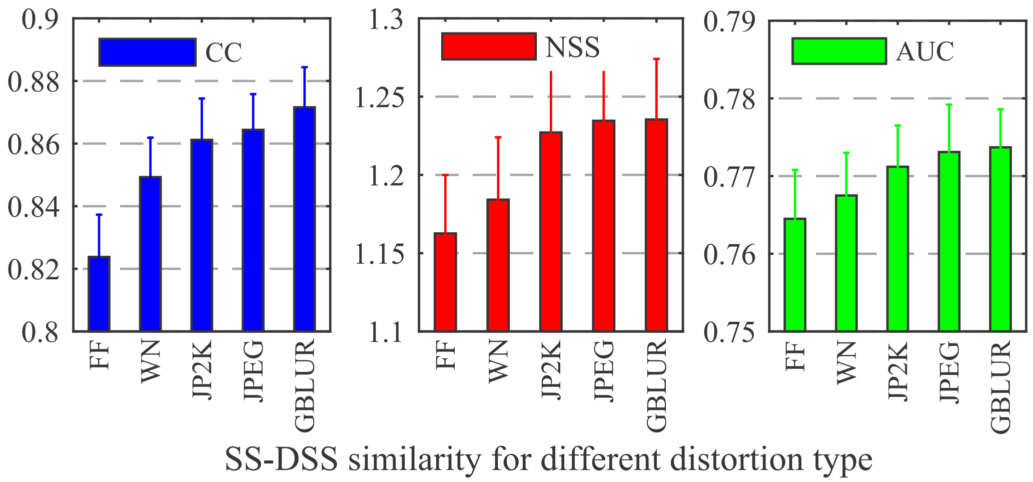


Figure 4.10: Illustration of rankings of five distortion types contained in our database in terms of the SS-DSS similarity measured by CC, NSS and AUC, respectively. The error bars indicate a 95% confidence interval.

calculated based on NSS or AUC. The inconsistency in the results is attributed to the fact that different similarity measures capture different characteristics of saliency changes while being consistent in measuring SS-DSS similarity, as already mentioned in [125]. CC focuses on the similarity in terms of the spatial distribution of fixation, whereas NSS and AUC are based on the estimation of similarity in terms of the locality and density of fixations. Figure 4.10 illustrates the rankings of the five available distortion types in terms of the SS-DSS similarity measured by CC, NSS and AUC, respectively. They consistently produce the same rank order for the five distortion types. For each subplot, the results of hypothesis testing (i.e., Wilcoxon signed

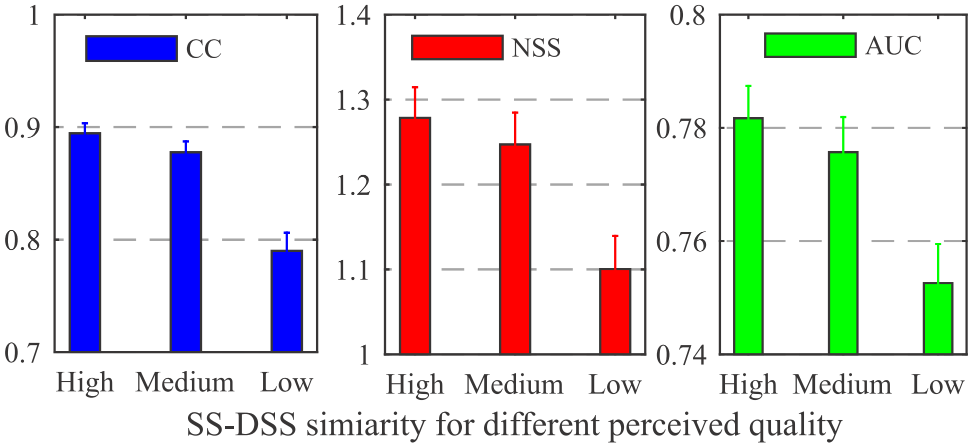


Figure 4.11: The measured SS-DSS similarity in terms of CC, NSS and AUC for images of different perceived quality. The error bars indicate a 95% confidence interval.

rank test) showed that the impact of distinct distortion types (e.g., FF and GBLUR) on SS-DSS similarity is statistically different with p-value = 0.022, $p < 0.05$ at the 95% confidence level. The distortions contained in FF (i.e., high-frequency, localised artifacts) produce a large extent of saliency deviation, whereas the GBLUR distortions (i.e., low-contrast, uniformly distributed artifacts) cause only slight changes in saliency.

Impact of distortion level on SS-DSS similarity

Table 4.1 shows that "distortion level" has a statistically significant effect on SS-DSS similarity, independent of the similarity measure used. The degree of saliency deviation increases as the perceived quality decreases (or strength of distortion increases). Figure 4.11 illustrates the measured SS-DSS similarity (again in terms of CC, NSS and AUC) for three levels of perceived quality. It reveals a statistically significant (i.e., based on t-test with p-value = 4.19e-26 between low and high and p-value = 1.66e-22 between low and medium, $p < 0.05$ at the 95% confidence level) drop in SS-DSS similarity at low quality relatively to the other two cases, which means that the distraction power of the annoying artifacts (or strong distortions) present in an image comes into impact the perception of the natural scene.

Impact of image content on SS-DSS similarity

Table 4.1 also shows that SS-DSS similarity is strongly affected by "image content" (i.e., classified by the degree of saliency dispersion). Figure 4.12 illustrates the measured SS-DSS similarity (again in terms of CC, NSS and AUC) for images having different degrees of saliency dispersion. In the case of images that do not contain highly salient objects (i.e., a large degree of

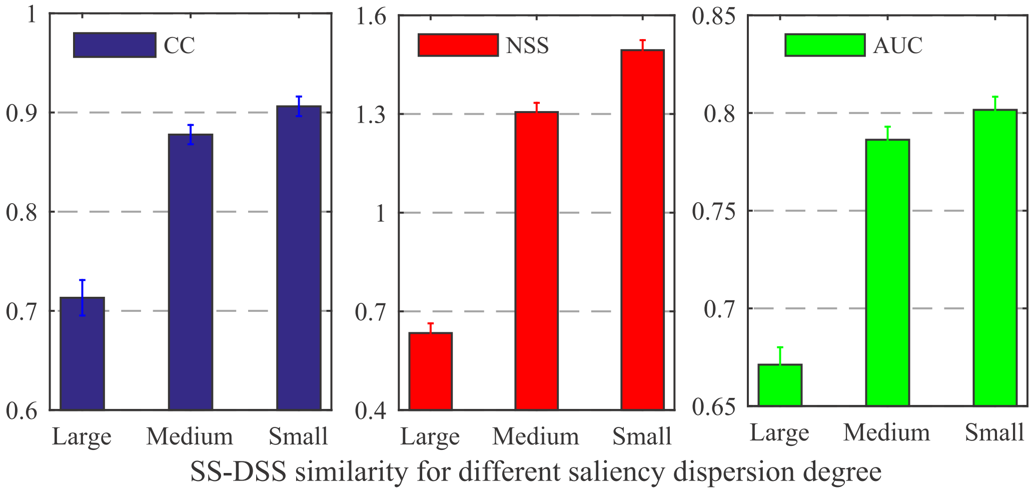


Figure 4.12: The measured SS-DSS similarity in terms of CC, NSS and AUC for images of different visual content (i.e., classified by the degree of saliency dispersion). The error bars indicate a 95% confidence interval.

saliency dispersion), adding artifacts to these images results in substantial changes between SS and DSS, as indicated by the statistically significant (i.e, based on t-test with p-value = 3.97e-26 between large and medium and p-value = 4.77e-29 between large and small, $p < 0.05$ at the 95% confidence level) drop in SS-DSS similarity relatively to the other two cases. On the other hand, images with highly salient objects (i.e., a small degree of saliency dispersion) are less sensitive to the distortions since human fixations are allocated to these highly salient objects though the distortions are perceived.

4.5 SS versus DSS on the Performance Gain

Previous research [90] has demonstrated that adding "ground truth" SS does improve the performance of IQMs in predicting perceived image quality. The findings, however, also showed that the performance gain could be potentially optimised by taking into account the interactions between SS and distortion. DSS, to some extent, represents the interactive effect of the concurrence of natural scene and unnatural artifacts. The added value of DSS as opposed to SS in IQMs, however, has not been investigated. To provide insights into this matter, both types of saliency are added to several IQMs in this section.

4.5.1 Evaluation criteria

We followed the general framework established in Chapter 3 for assessing the added value of computational saliency in IQMs. The basic idea is to quantify the performance gain of an

IQM by comparing its predictive power with and without saliency. The predictive power of an IQM can be simply measured by the PLCC between the output of the IQM and the subjective quality ratings [105]; and the performance gain can be effectively expressed by the increase in PLCC (i.e., ΔPLCC). The IQMs used in our evaluation consisted of six full-reference (FR) IQMs, namely PSNR, UQI, SSIM, MS-SSIM, VIF and FSIM, and four no-reference (NR) IQMs, namely GBIM, NBAM, NPBM and JNBM.

4.5.2 Evaluation results

Original versus saliency-based IQMs

Per IQM, adding SS and DSS results in two new saliency-based IQMs. The performance (i.e., PLCC) of an IQM was calculated based on the subjective quality scores contained in our database, which is summarised in Table 4.2. In general, it shows that the performance of IQMs is improved by using both SS and DSS. The gain (i.e., ΔPLCC) ranges from 0.002 (FSIM extended with SS) to 0.058 (GBIM extended with DSS). Note VIF and FSIM obtain relatively small gain by adding saliency, due to the fact that some well-established saliency aspects (i.e., information content feature in VIF [33] and phase congruency feature in FSIM [114]) are already embedded in these metrics, which consequently causes a saturation effect in saliency optimisation [106].

The observed effects were statistically analysed with hypothesis testing, selecting the metric strategy (SS-based v.s. original or DSS-based v.s. original) as the independent variable and the performance gain as the dependent variable. A Wilcoxon signed rank test was performed using the data points contained in Table 4.2. The results (p-value = 7.53e-4 between original and SS-based, and p-value = 6.05e-3 between original and DSS-based, $p < 0.05$ at the 95% confidence level) revealed that both SS and DSS statistically significantly improve the original IQMs. To further check the effectiveness of adding saliency for individual IQMs, the differences were statistically analysed per IQM as the implementation detailed in Section 3.3.2 (i.e., based on the residuals between DMOS and the quality predicted by an IQM). In the case of normality, t-test was performed; otherwise a Wilcoxon signed rank test was conducted, as the results summarised in Table 4.3.

SS-based versus DSS-based IQMs

As can be seen in Table 4.2, on average (over all IQMs), the gain achieved by use of SS is similar to that of using DSS. To check the effects with a statistical analysis, a Wilcoxon signed rank test was performed, selecting the type of saliency as the independent variable and the performance as the dependent variable. The test results (i.e., p-value = 0.484, $p > 0.05$ at the 95% confidence level) showed that there is no statistically significant difference between the inclusion of both types of saliency.

In response to the evaluation framework identified in Section 4.4, we further assessed how the performance gain between SS-based and DSS-based IQMs is affected by the observed main effects, i.e., the distortion type, distortion level and image content. More specifically, our database was again characterised at three individual aggregation levels, using “distortion type”, “distortion level” and “image content” as the classification variables, respectively.

Table 4.2: Performance for 10 IQMs (PLCC without non-linear fitting) and their corresponding saliency-based versions on our database with 270 distorted stimuli.

	PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	GBIM	NPBM	JNBM	NBAM	average ΔPLCC
original	0.784	0.891	0.773	0.791	0.917	0.855	0.809	0.842	0.854	0.828	-
SS-based	0.800	0.910	0.817	0.824	0.922	0.857	0.834	0.862	0.871	0.871	0.022
DSS-based	0.801	0.912	0.818	0.824	0.920	0.858	0.867	0.849	0.855	0.866	0.023

Table 4.3: Results of statistical significance testing for individual IQMs. "1" means that the difference in performance is statistically significant with P<0.05 at the 95% confidence level. "0" means that the difference is not significant.

	PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	GBIM	NPBM	JNBM	NBAM
original vs. SS-based	1	1	1	0	1	0	1	1	1	1
original vs. DSS-based	1	1	1	1	1	1	1	1	1	1

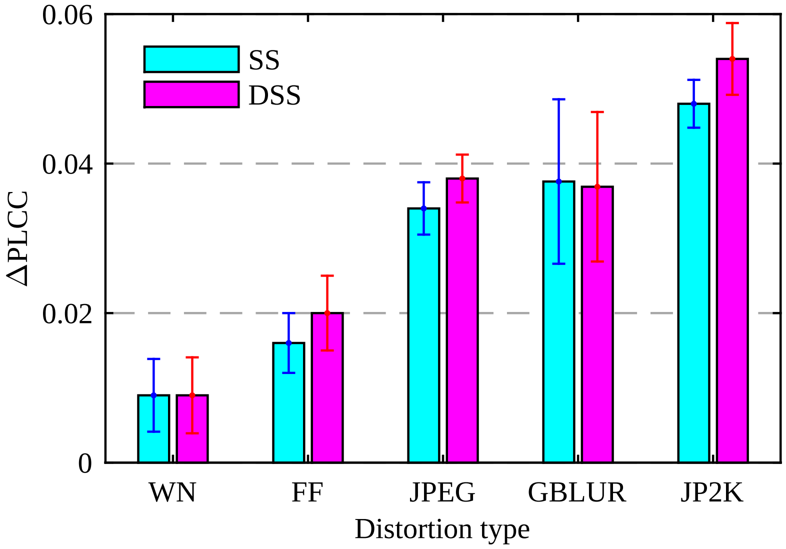


Figure 4.13: Comparison of performance gain between SS-based and DSS-based IQMs, with the effect of distortion type dependency. The error bars indicate a 95% confidence interval.

Figure 4.13 illustrates the performance gain (i.e., ΔPLCC) averaged once over all SS-based IQMs and once over all DSS-based IQMs, when assessing WN, JPEG, GBLUR, JP2K and FF, respectively. It shows that both types of saliency are beneficial for IQMs (i.e., ΔPLCC values are positive in all cases). Results of a Wilcoxon signed rank test showed that the difference in performance gain between the use of SS and DSS is not statistically significant different (i.e., p-value ranges from 0.132 to 0.895, $p > 0.05$ at the 95% confidence level) for all distortion types except for JP2K. For JP2K, using DSS improves the IQMs' performance more, which is in line with the conclusions drawn in [106] that when saliency is added in IQMs for accessing localised distortion, such as JP2K, taking into account the interactions between saliency and distortion can be used to optimise the performance gain. Note the same trend can also be observed for the localised JPEG and FF distortion, although the results were not significant in our current samples. This indicates that the use of saliency in IQMs potentially benefits from taking into account the interactions between saliency and distortion, especially for assessing localised distortion types.

Figure 4.14 shows the comparison of ΔPLCC between SS-based and DSS-based IQMs, when accessing images with three distinct levels of perceived quality. At low quality, IQMs do not benefit from the use of saliency (i.e., marginal values of ΔPLCC). At high quality, there is no statistically significant difference (i.e., based on t-test with p-value = 0.792, $p > 0.05$ at the 95% confidence level) between the added value of SS and DSS, which is attributed to the fact that SS and DSS is very similar (i.e., a small degree of SS-DSS deviation as shown in Fig. 4.11). In terms of the medium level of quality, the results of a t-test (p-value = 0.041, $p < 0.05$ at the

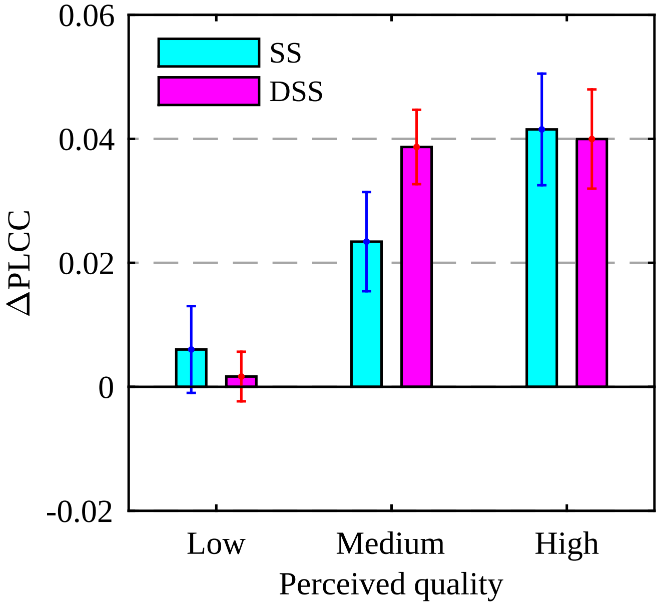


Figure 4.14: Comparison of performance gain between SS-based and DSS-based IQMs with the effect of distortion level dependency. The error bars indicate a 95% confidence interval.

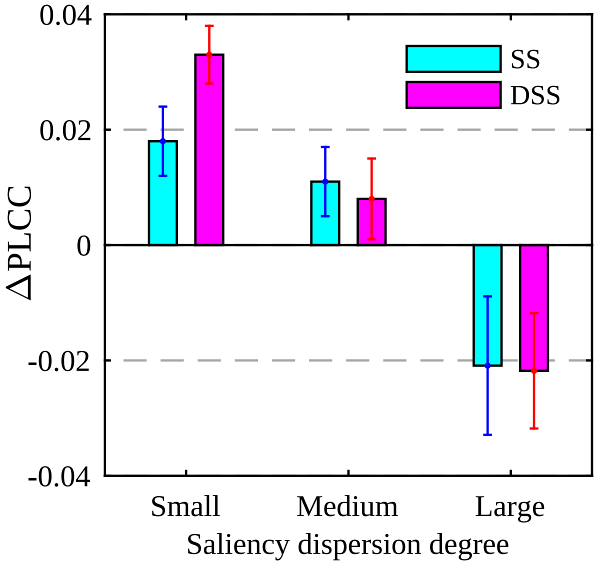


Figure 4.15: Comparison of performance gain between SS-based and DSS-based IQMs with the effect of saliency dispersion degree dependency. The error bars indicate a 95% confidence interval.

95% confidence level) demonstrated that adding DSS to IQMs yields statistically significantly higher performance gain than adding SS, suggesting that the use of saliency in IQMs potentially benefits from taking into account the interactions between saliency and distortion.

Figure 4.15 illustrates the difference in ΔPLCC between SS-based and DSS-based IQMs when accessing images with three distinct degrees of saliency dispersion. Adding saliency deteriorates the performance of IQMs for assessing images with a large degree of saliency dispersion, which should be avoided in saliency optimisation. This is mainly due to the uncertainty of a dispersed fixation map, which confuses the workings of IQMs by e.g., unhelpfully downplaying the importance of high distortion in certain regions [92]. Images with a medium range of saliency dispersion do not profit from adding saliency to an IQM (i.e., marginal ΔPLCC). For images having a small degree of saliency dispersion, the use of DSS produces statistically significantly (i.e., based on t-test with p-value = 0.035, $p < 0.05$ at the 95% confidence level) larger ΔPLCC than that of using SS. Again, this suggests the interactions between saliency and distortion play a significant role in optimising the increase in the performance of IQMs.

4.6 Summary

In this chapter, we investigated a more reliable methodology for collecting eye-tracking data for the purpose of image quality study. We proposed dedicated control mechanisms to effectively eliminate potential bias due to the involvement of massive stimulus repetition. The refined methodology resulted in a new eye-tracking database with a large degree of stimulus variability, including 288 test images distorted with different types of artifacts at various levels of degradation. The database contains 5760 eye movement trials recorded with 160 human observers.

Based on the "ground truth" data, we assessed the interactions between saliency and distortion. A statistical evaluation was conducted to provide insights into the tendencies in the changes of saliency induced by distortion. We found that the occurrence of distortion in an image tends to deviate fixation deployment. We also quantified the extent of such deviation as a function of distortion type, degradation level and image content, respectively. In terms of optimal use of saliency in IQMs, we investigated whether it is saliency of the undistorted scene or that represents the same scene affected by distortion would deliver the best performance gain for IQMs. The results showed that both types of saliency are beneficial for IQMs, but the latter which reflects the interactions between saliency and distortion tends to further boost the effectiveness of the integration of saliency in IQMs.

Chapter 5

A Distraction Compensated Approach for Saliency Integration

5.1 Introduction

In Chapter 4, eye-tracking study has shown that including saliency of distorted scene provides more benefits for IQMs than including saliency of undistorted scene. This is due to the former saliency better addresses the interactions between saliency and distortion. A realistic IQM would use a saliency model instead of eye-tracking data. This means to implement the idea mentioned above, saliency of distorted images needs to be automatically detected. Unfortunately, existing computational saliency models are designed to detect saliency of undistorted natural images. Their ability to capture saliency of distorted images is unknown. Instead of investigating saliency detection for distorted images, we focus on improving the effectiveness of saliency inclusion in IQMs.

In this chapter, we proposed a more sophisticated saliency integration strategy that better takes into account the attentional power of distortion. The proposed method compensated the attentional power of the visual distortions on the basis of modelled saliency of undistorted scenes instead of using the modelled saliency of distorted scenes.

5.2 Proposed Integration Approach

In the conventional approach, the distortion map computed by an IQM is simply multiplied by the modelled scene saliency. This process may run the risk of underestimating or neglecting the distraction power of e.g., strong artifacts in non-salient areas. To compensate for such deficiency, in the proposed approach, for the distortion map computed by an IQM, instead of using modelled scene saliency as a weighting factor, we now use two components: the modelled attraction power of a scene (i.e., denoted as α) and the distraction power of the visual artifacts (i.e., denoted as β) to produce a local weighting factor ω. Given a pixel location (i, j), ω is defined as:

$$\omega(i, j) = f(\alpha, \beta) \tag{5.1}$$

where $f()$ denotes a combination operator. In this chapter, α can be the modelled scene saliency calculated from any saliency model. The measurement of β was derived using an information theory based approach. This approach treated the HVS as an optimal information extractor [33]; and the distraction power β was considered to be proportional to the perceived information of distortion.

Based on the principle in [126], the perceived information I of a stimulus can be modelled as the number of bits transmitted from this stimulus (with the stimulus power S) through the visual channel of the HVS (with the noise power C); and can be computed as:

$$I = \frac{1}{2}\log(1 + \frac{S}{C}) \tag{5.2}$$

If we simply consider the distortion as the input stimulus, the perceived inforamtion of distortion can now be measured by the above formula. In such a scenario, the component S/C is analogous to the power of the locally measured perceived distortion using the distortion map. Due to the fact the HVS is not sensitive to pixel-level variations [5], the implementation of the algorithm was thus performed on the basis of a local patch of 45 × 45 pixels (to approximate a 2° visual angle when viewing images from a distance of 150 cm with a screen resolution of 1024×768 pixels). Thus Equation (5.2) can be further defined as:

$$I_P = \frac{1}{2}\log(1 + \sigma_p^2) \tag{5.3}$$

where σ_P^2 estimates the power of local distortion within the patch P centred at a given pixel (i, j) in the distortion map; and σ_P corresponds to the standard deviation of P.

Moreover, our algorithm is motivated by the significant findings in [127] that each perceptible artifacts suppresses each other artifact's effect especially for those with close proximity. This so-called surround suppression effect (SSE) is used to approximate the proportional relationship between β and I, where the effect of I is suppressed by its local neighbourhood. The β can be defined as:

$$\beta_P = \frac{I_P}{\bar{I}} \tag{5.4}$$

where $\bar{I}$ represents the averaged distraction power surrounding the local patch P. In this chapter, the vicinity was defined as the Moore neighbourhood or 8-neighbours of the local patch P (i.e., the set of eight patchesP_k (k=1 to 8) of the same size which share a vertex or edge with P).

Finally, we combined α and β using a simply multiplication operator, resulting in a specific form of the local weighting factor:

$$\omega(i, j) = \alpha^m \cdot \beta^n \tag{5.5}$$

where $m > 0$ and $n > 0$ are parameters to adjust the relative importance of different components. We set $m = n = 1$ in our experiment for simplification. Tuning the parameters may improve the algorithm; however it goes beyond the merits of this chapter. The final form of the weighting factor is:

$$\omega(i, j) = \alpha \cdot \frac{\log(1 + \sigma_P^2)}{\frac{1}{8}\sum_{k=1}^{8} \log(1 + \sigma_{P_k}^2)} \tag{5.6}$$

It should be noted that the estimation of the distraction power takes advantage of the distortion map that have already been generated in each IQM. Therefore, the computational cost significantly drops if compared with measuring the distraction power based on other image features. It should also be noted that the proposed combination strategy keeps the generalizability as of the conventional strategy and hence can be easily implemented independent of the saliency model, IQM and image distortion type.

5.3 Performance Evaluation

To evaluate the performance of the saliency-based IQMs using the proposed approach, we repeated the performance evaluation (i.e., with the same 12 IQMs and 20 saliency models) in Chapter 3 on the LIVE database, but by replacing the conventional weighting approach with the proposed distraction power compensated weighting approach. We then compared the performance of the saliency-based IQMs using either integration approach.

Figure 5.1 illustrates the comparison in performance gain (in terms of PLCC) when adding computational saliency in IQMs using either the conventional approach or the proposed one. In general, it shows a consistent trend that the proposed combination strategy results in a larger amount of performance gain independent of the distortion type assessed. The performance gain has arisen from 0 to 0.003 for WN, from 0.007 to 0.026 for JP2K, from 0.013 to 0.027 for JPEG, from 0.014 to 0.023 for FF, and from 0.024 to 0.029 for GBLUR, respectively. A t-test was performed per distortion type to check whether the numerical difference in performance

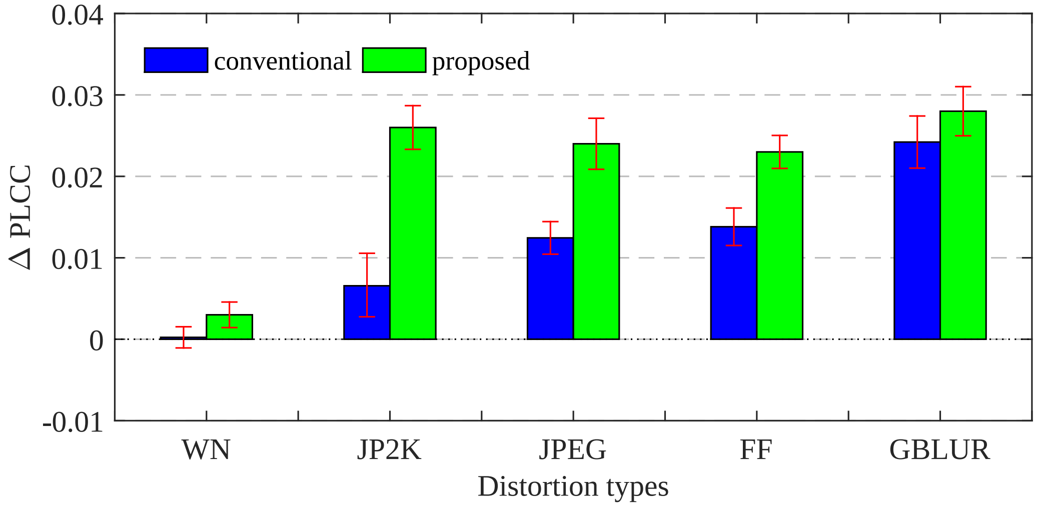


Figure 5.1: Comparison in performance gain using two combination strategies. The error bars indicate the 95% confidence interval.

gain between two combination strategies is statistically significant. In each case, the combination approach (conventional *v.s.* proposed) was selected as the independent variable and the performance gain as the dependent variable. The t-test results showed that the proposed strategy is statistically significantly better than the conventional strategy (i.e., p-value = 0.012 for JP2K, p-value = 0.024 for JPEG and p-value = 0.038 for FF) with $p < 0.05$ at 95% confidence level for the three localised distortion types including JP2K, JPEG and FF. The significant improvement for these three distortion types is consistent with the conclusion in Section 4.5 that the interaction between saliency and distortion should be taken into account for assessing localised distortion.

To also check the effectiveness of the proposed combination strategy on individual IQMs, we reformed the results of Fig. 5.1 and illustrated them again in Table 5.1. In general, the proposed combination strategy outperforms the conventional strategy in terms of improving the performance of an IQM. The difference was further statistically analysed with a t-test per IQM. All differences, except for the cases of FSIM and VIF, were statistically significant (p-value ranges from 4.55e-12 to 6.57e-10, $p < 0.05$ at 95% confidence level). Table 5.1 also shows a trend of a larger improvement in predictability of PSNR, UQI, SSIM, MS-SSIM, GBIM, NPBM, JNBM and NBAM when using either combination strategy. The relatively small amount of performance gain for VIF, FSIM, IWPSNR and IWSSIM when adding saliency may be attributed to the fact that some saliency-driven aspects (e.g., the phase congruency used in FSIM [114], the information of content used in VIF, IWSSIM, IWPSNR [33]) are already integrated in these metrics. As such, it is more difficult to obtain a significant increase in performance by adding a dedicated saliency model.

To further validate the robustness of the proposed combination approach, we also evaluated the performance of saliency-based IQMs on another two image quality databases, namely the TID2013 [18] and the CSIQ [17]. As can be seen from Table 5.2, the proposed saliency combination approach outperforms the conventional approach in all cases. A paired samples t-test analysis (preceded by a test for the assumption of normality) was performed, selecting the combination strategy as the independent variable and the performance as the dependent variable. Experimental results showed that the proposed approach is statistically better (p-value =1.43e-6 for TID2013 and p-value = 1.66e-5 for CSIQ, $p < 0.05$ at 95% confidence level) than the conventional approach on both the TID2013 and CSIQ databases.

Table 5.1: Performance (in terms of PLCC, without nonlinear regression) of 12 IQMs and their corresponding saliency-based versions (using either the conventional approach or the proposed approach) of the LIVE database. Note that PLCC is averaged over all saliency models and over all distortion types where appropriate.

	PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	IWPSNR	IWSSIM	GBIM	NPBM	JNBM	NBAM
original	0.880	0.895	0.909	0.919	0.952	**0.915**	0.931	0.897	0.773	0.843	0.833	0.836
conventional combination	0.894	0.928	0.930	0.931	0.953	0.914	0.932	0.885	0.802	0.871	0.848	0.853
proposed combination	**0.910**	**0.941**	**0.957**	**0.958**	**0.956**	0.913	**0.938**	**0.907**	**0.818**	**0.890**	**0.855**	**0.871**

Table 5.2: Performance (in terms of PLCC, without nonlinear regression) of 12 IQMs and their corresponding saliency-based versions (using either the conventional approach or the proposed approach) of the TID2013 and CSIQ database.

		PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	IWPSNR	IWSSIM	GBIM	NPBM	JNBM	NBAM
TID2013	original	0.478	0.615	0.653	0.707	**0.608**	0.820	0.538	**0.764**	0.841	0.787	0.794	0.765
	conventional	0.482	0.666	0.695	0.733	0.587	0.823	0.540	0.754	0.855	0.774	0.786	0.797
	proposed	**0.490**	**0.686**	**0.711**	**0.755**	0.604	**0.830**	**0.552**	0.761	**0.877**	**0.792**	**0.805**	**0.810**
CSIQ	original	0.751	0.829	0.765	0.815	0.883	**0.805**	0.802	0.795	0.731	0.814	0.793	0.625
	conventional	0.770	0.847	0.832	0.841	0.858	0.796	0.806	0.788	0.741	0.814	0.799	0.650
	proposed	**0.782**	**0.870**	**0.849**	**0.855**	**0.888**	0.804	**0.810**	**0.805**	**0.758**	**0.829**	**0.805**	**0.667**

5.4 Summary

This chapter considered how to enhance the added value of visual saliency in IQMs by taking into account the interaction of saliency and distortion. A new saliency integration strategy was proposed by compensating the distraction power of local distortions. Experimental results showed that the proposed combination strategy significantly and consistently yields a larger amount of performance gain for an IQM than the conventionally used saliency combination strategy. Moreover, the proposed algorithm is based on the use of saliency of undistorted scenes, and therefore existing saliency models can be directly applied to improve IQMs.

Chapter 6

A Saliency Dispersion Measure for Improving Saliency-Based IQMs

6.1 Introduction

In Chapter 4, the eye-tracking study has demonstrated an image content-dependent nature of the improvement to IQMs by incorporating saliency. It showed that incorporating saliency in IQMs when assessing images with a clear region-of-interest results in a promising gain in these IQMs' performance, while integrating saliency in IQMs for the assessment of images with spread-out saliency deteriorates their original performance. This observation may be used to improve saliency-based IQMs by adaptively applying saliency in IQMs, depending on image content.

Previous study [92] based on eye-tracking revealed that the inter-observer agreement (IOA) for human fixations — the degree of agreement between observers freely viewing the same visual stimulus — is strongly image content dependent. Furthermore, this measure predicts the extent to which a certain image may profit from adding saliency information to an IQM. As the observation also revealed from eye-tracking studies in [118] and [119], if an image has highly salient objects, then most viewers will concentrate their fixations around them, whereas if there is no obvious object of interest, viewers' fixations will appear as a more evenly distributed pattern. Thus, images with salient objects tend to have less variation in fixations between viewers (i.e. higher IOA) than images without salient objects. When saliency is spread throughout the scene, incorporating saliency in an IQM is less likely to benefit image quality prediction [92], as different observers tend to look at different parts of the image. Incorporating saliency into an IQM may give a low weight to some region with high distortion, and therefore weighting the IQM might unhelpfully downplay the importance of distortion in this region. To make better use of saliency in IQMs, a sophisticated integration strategy taking into account the dispersion of saliency is needed.

However, calculating the IOA values for saliency from eye-tracking data is unrealistic in any practical application context. A saliency dispersion measure is needed as a proxy for the variation in human fixation (i.e., IOA). Meanwhile, it should be noted that the observed content dependency of the performance gain is validated with eye-tracking data. A realistic IQM, however, will use a computational model of saliency rather than eye tracking data. Therefore, the feasibility of such content-adaptive integration approach is based on that the content-dependency of performance gain still holds when computational saliency is used. To determine whether content dependency still remains significant, and potentially useful, the effect of content-dependency should be investigated with computational saliency models in the first instance.

In this chapter, the effect of content-dependency was first validated by conducting a statistical evaluation using 15 state of the art saliency models and 10 of the best-known IQMs. A saliency dispersion measure that provides a reliable proxy for IOA was then proposed. The saliency dispersion measure was used to devise an adaptive saliency integration approach for IQMs.

6.2 Effect of Image Content Dependency

In [92], ground truth eye-tracking data and IOA (calculated as the average correlation coefficient between the mean fixation map and each observer's fixation map) were measured for the LIVE database [16]; based on IOA for scene content, the entire database was divided into 3 subsets: images with low, intermediate and high IOA. To determine whether content dependency still remains significant, and potentially useful, we conducted a statistical evaluation using 15 state of the art saliency models and 10 of the best-known IQMs. In this evaluation, saliency was incorporated by weighting the distortion map calculated by an IQM using the saliency map computed from the original scene. For each subset of images, we quantified the performance gain of a saliency-based IQM over its original form without saliency, using the same evaluation framework established in Section 3.2.3.

The IQMs in this evaluation include PSNR, UQI, SSIM, MS-SSIM, VIF, FSIM, GBIM, NBAM, NPBM and JNBM. As suggested in [90], for all NR metrics saliency was computed from the original scene rather than the distorted scene. Such saliency was either assumed to be practically available (e.g., as a side information, in which case the framework is analogous to the reduced-reference (RR) case), or considered to be plausibly approximated from the distorted image (e.g., by filtering out distortion). The 15 saliency models were AIM, AWS, CA, CBS, DVA, GBVS, ITTI, PQFT, SDCD, SDFS, SDSR, SR, SUN, SVO and Torralba, representing the best performing saliency models in terms of the capability of improving the performance of IQMs.

The study thus resulted in 150 possible combinations (10 IQMs $\times$ 15 saliency models). The performance of an IQM was quantified by the PLCC and SROCC between the IQM's output and

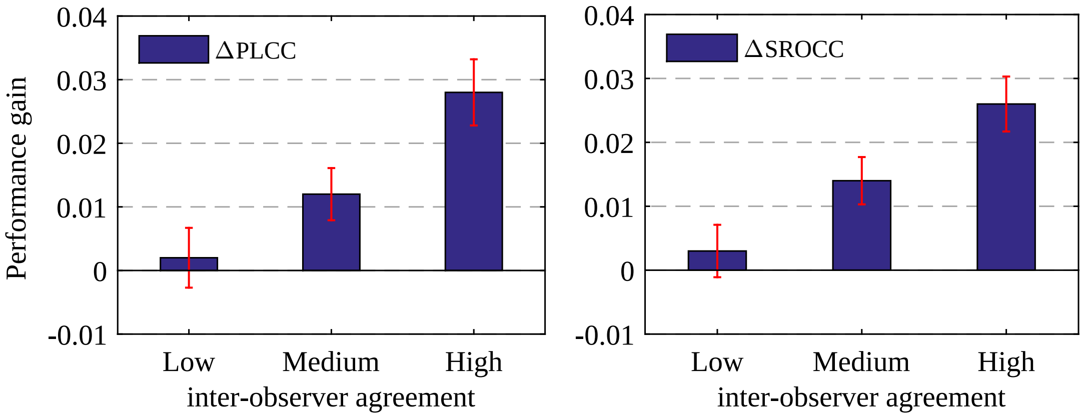


Figure 6.1: Performance gain (i.e., ΔPLCC and ΔSROCC) of saliency-augmented IQMs for three degrees of IOA. Error bars indicate a 95% confidence interval.

the subjective quality ratings [105]. Figure 6.1 illustrates the performance gain averaged over all 150 cases for different degrees of IOA. Results of t-test (preceded by a Kurtosis test for the assumption of normality [90]) showed that the difference in performance gain between each pair of subsets is statistically significant (i.e, p-value = 0.039 between low IOA and medium IQA groups, p-value = 0.012 between low IOA and high IQA groups and p-value = 0.025 between medium IOA and high IQA groups) all with $p < 0.05$ at the 95% confidence level. This confirms that the benefits of inclusion of computational saliency in IQMs depend on image content. For images with low IOA, incorporating saliency runs the risk of reducing IQM's performance (i.e., the performance gain can appear negative as shown in Fig. 6.1).

6.3 Saliency Dispersion Measure

To optimise the saliency integration by incorporating the above observation, we proposed an algorithm to measure the saliency dispersion and used that as a proxy for the variation in human fixation (i.e., IOA) on natural scenes. Reliably quantifying saliency dispersion in agreement with IOA is very challenging, despite research on the topic. Existing methods either have limited sophistication (e.g., the simple saliency coverage measure in [118]) or limited applicability to real-world systems (e.g., the complex approaches in [128] and [129]). In [118], the saliency dispersion was approximated as the amount of the stimulus covered by the fixated region. The fixated region was obtained by thresholding the saliency map with a specific value. This measure, however, fails to take into account the compactness of the fixation deployment. In addition, this measure can be strongly affected by the threshold selected. In [128] and [129], the IOA was measured on the basis of a set of visual features including both low-level (e.g., color and contour) and high-level (e.g., face and object) features. Applying such complex algorithms to

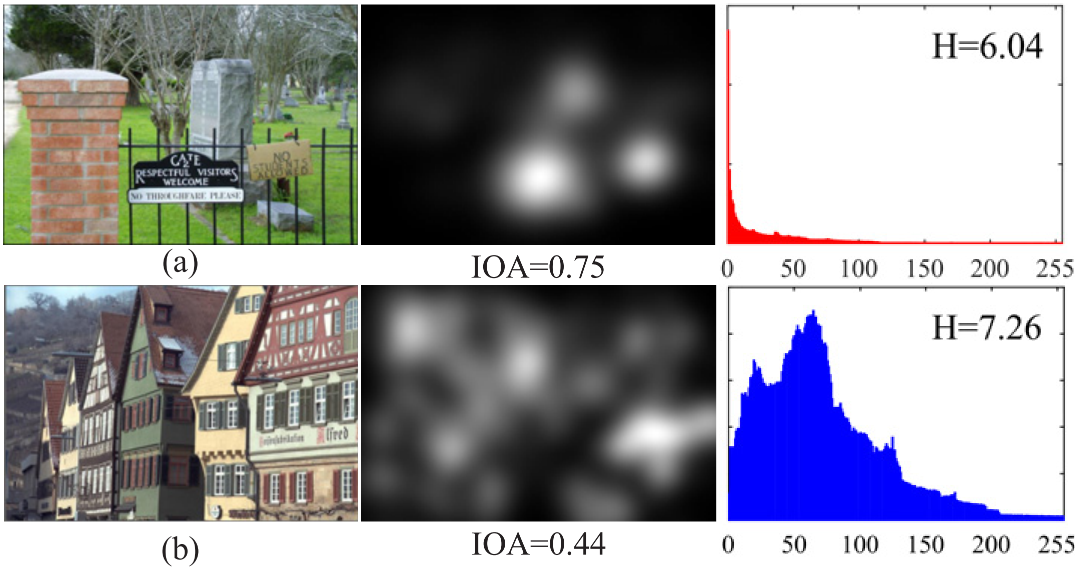


Figure 6.2: Natural scenes, their ground truth fixation maps, corresponding IOA scores, and entropy of scene saliency. (a): an image with a few highly salient objects; IOA is high. (b): an image lacking salient objects; IOA is low. IOA values and fixation maps were determined from human eye fixations in the TUD eye-traking database.

measure the IOA is impractical for IQMs due to the massive computational cost introduced. We have thus devised our own simple, but reliable, method.

Our method is based on Shannon entropy, which is a measure of the randomness or uncertainty of a variable [130]. We analysed saliency maps as realisations of random variables. Figure 6.2(a) shows a ground truth fixation map (grayscale values represent the intensity of saliency) of a natural scene derived by accumulating human fixations of 20 observers [90]. The normalized histogram of the saliency map represents an estimate of the underlying probabilities of pixel intensities: $p(i) = h(i)/K$, where $h(i)$ is the histogram entry for intensity value i in the saliency map S, and K is the total number of pixels in S. The entropy of the saliency map is given by:

$$H(S) = -\sum_{i} p(i) \log p(i) \tag{6.1}$$

For the saliency map in Fig. 6.2(a) it is 6.04 bits. The entropy calculated for the saliency map of a different natural scene shown in Fig. 6.2(b) is 7.26 bits. Saliency in Fig. 6.2(a) is more concentrated in fewer areas than in Fig. 6.2(b), which results in a smaller value of entropy.

Note, however, that even a single large salient object may also lead to a spread-out saliency map. For example, the saliency map in Fig. 6.3(b) is more concentrated than the saliency map in Fig. 6.3(a), but the entropy values are similar (i.e., $H = 7.26$ for Fig. 6.3(a) and $H = 6.99$ for Fig. 6.3(b)). This is because entropy is a single value summarising the whole image; it does not consider spatial characteristics and relations of fixation patterns [131]. To perform a more

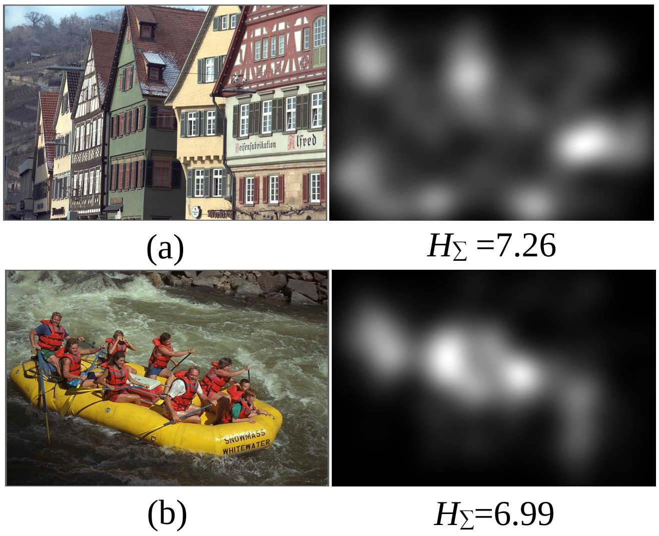


Figure 6.3: Illustration of two scenes with their corresponding ground truth saliency. (a) an image with spread-out saliency. (b) an image with a large salient objects. Saliency maps were determined from human eye fixations in TUD eye-tracking database.

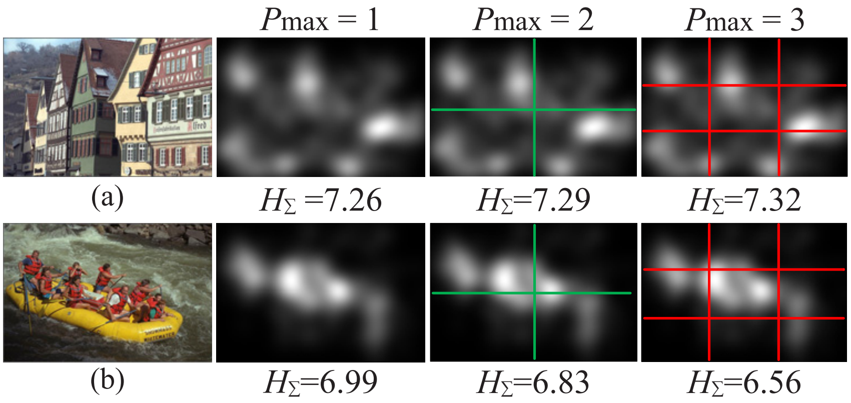


Figure 6.4: Calculation of multi-level entropy H_Σ. At each level the saliency map is partitioned into blocks of equal size. H_Σ is found by adding the entropies computed at each level of partition. $P_{\max}$ is the level with finest partitioning.

refined saliency dispersion analysis, we used a multi-level approach to entropy calculation. To do so, the saliency map was partitioned at level P into $P \times P$ non-overlapping blocks of equal size: see Fig. 6.4. At $P = 2$ the original map was subdivided into 4 equal quadrants, at $P = 3$, into 9 equal partitions, and so on. We defined the multi-level entropy of the saliency map to be:

$$H_\Sigma(S) = \frac{1}{P_{\max}} \sum_{P=1}^{P_{\max}} \sum_{B=1}^{N_{\max}} H(B) \tag{6.2}$$

where $P_{\max}$ is the finest level of division, and $N_{\max} = P_{\max}^2$; B runs over each block. In the case illustrated in Fig. 6.4, the disparity in entropy between saliency maps increases as the number of partitions increases, which allows the multi-level entropy to better distinguish the two saliency

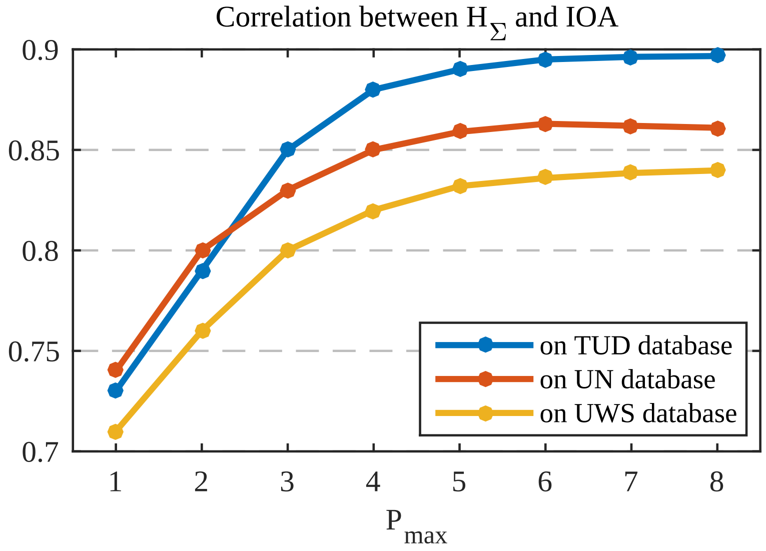


Figure 6.5: The absoulte value of the Pearson correlation (as shown for each data point) between estimated saliency dispersion, H_{Σ}, and its ground truth counterpart IOA, for difference choices of $P_{\max}$. IOA values were determined for the same set of images from three independent eye tracking databases.

maps than the whole-image entropy, giving the more compact saliency map a lower entropy.

To determine the number of levels to use, we used an empirical approach, based on quantifying the correlation between the estimated saliency dispersion and its ground truth counterpart (i.e. IOA). Figure 6.5 plots the absolute value of the Pearson correlation between H_{Σ} for different choices of $P_{\max}$, and ground truth IOA values determined for the same set of images from three independent eye-tracking databases, namely TUD, UN and UWS as listed in [92]. While correlation increases with $P_{\max}$, saturation starts to occur at about $P_{\max} = 4$. Hypothesis testing was performed to verify whether there is a significant difference between the use of $P_{\max} = 4$ and a higher level of $P_{\max}$. To do so, a Wilcoxon signed rank test (i.e., a non-parametric version of t-test in the case of non-normality) based on the residuals between H_{Σ} and IOA, as suggested in [90], was conducted. The results showed that there was no statistically significant difference between $P_{\max} = 4$ and $P_{\max} = 5$ (i.e., p-value = 0.15, p $>$ 0.05 at the 95$\%$ confidence level), and between $P_{\max} = 4$ and $P_{\max} = 6$ (i.e., p-value = 0.11, p $>$ 0.05 at the 95$\%$ confidence level). We therefore, used $P_{\max} = 4$ in our experiments.

6.4 Proposed Integration Approach

We now consider how to use the above formula for assessing saliency dispersion to improve saliency-based IQMs.

Suppose we are given a particular saliency model, and an IQM. For an input scene of size $M \times N$, we can compute a saliency map together with its degree of dispersion H_Σ. The key idea is to *only include saliency in the computation of image quality if the dispersion is not too large*, in line with the observation that using saliency in cases of low IOA may be of no benefit to or even reduce the IQM performance.

In principle, we wish to do the following. If H_Σ is below a threshold T, saliency is combined with the pre-existing IQM to provide a modified method of quality assessment, as follows:

$$I' = \sum_{x=1}^{M}\sum_{y=1}^{N} D(x,y)S(x,y) / \sum_{x=1}^{M}\sum_{y=1}^{N} S(x,y) \tag{6.3}$$

where D represents the distortion map measured by an IQM, and S indicates the saliency map generated by the saliency model. If the saliency dispersion is large, the saliency of the scene contains much uncertainty, and so is ignored: the pre-existing IQM is used directly without saliency.

However, using a hard threshold will lead to a discontinuous IQM, and two very similar scenes whose saliency dispersions are just above and below the threshold may end up with significantly different quality scores. To avoid such sudden changes, instead of using a step function to switch between using saliency, or not, a sigmoid function $\sigma(\cdot)$ was applied to smooth the IQM near the transition region. Our integrated image quality metric I'' is given by:

$$I'' = \sigma(H_\Sigma)I + (1 - \sigma(H_\Sigma))I' \tag{6.4}$$

where I is the original IQM value, and $\sigma(x)$ is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-\tau(x-T)}}, \tag{6.5}$$

where T is the threshold value and τ controls the steepness of the sigmoid function.

As different saliency models lead to intrinsically different scales of entropy measurements (i.e. different ranges of H_Σ values), T should be individually determined for each saliency model. To ensure generality of the technique and to perform a more rigorous procedure to determine reliable parameters, τ and T were empirically determined from a separate larger-scale saliency database to that used in our experiments: we used the MIT300 database [132] containing 300 natural scenes and a wide diversity of content. Figure 6.6 gives H_Σ for these scenes, ordered from lowest to highest value, for the 15 saliency models considered in Section 6.2. The median H_Σ value for each saliency model was used as the corresponding threshold T (e.g. $T = 4.38$ for AIM), while the slope of the envelope of the values between the 25th and 75th percentiles

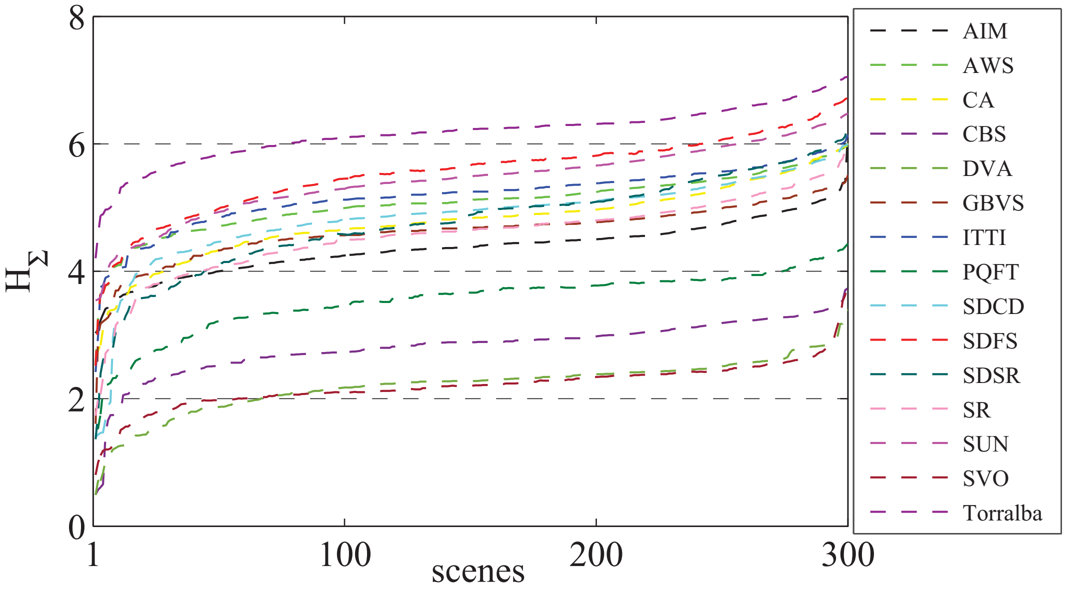


Figure 6.6: H_Σ calculated for 300 scenes from the MIT300 database, using saliency values generated by 15 state of the art saliency models. H_Σ values are ordered from lowest to highest for each model.

was used to determine an appropriate value of the steepness control τ; in practice, these were similar, so we used $\tau = 20$ for all saliency models. Note other saliency databases (e.g., [133] and [134]) may be used to estimate these parameters, but we do not expect it to change the results significantly.

6.5 Performance Evaluation

The performance of each IQM was evaluated against three recognised image quality databases: CSIQ [28], TID2013 [135] and LIVE. In each case we compared its performance between *no* use of saliency, *fixed* use of saliency and *adaptive* use of saliency according to saliency dispersion.

Table 6.1 shows the performance (in terms of PLCC) in each case, averaged over 15 saliency models (SROCC values exhibit similar trends and thus are not presented here). Following the approach taken in [106], PLCC values are reported without nonlinear fitting in order to better visualise differences in IQM performance. As can be seen in Table 6.1, the adaptive approach outperforms fixed use of saliency in all cases over all databases. On average, VIF and FSIM, do not benefit from fixed use of saliency, but are improved by using adaptive saliency. Note VIF and FSIM obtain relatively small gain by adding saliency. This is probably due to the fact that some well-established saliency aspects are already embedded in VIF and FSIM, which consequently causes a saturation effect in saliency optimisation.

Table 6.1: Performance for 10 IQMs (in terms of PLCC, without non-linear regression) on all images of three databases, using versions which did not use saliency, always used saliency, or adaptively used saliency according to saliency dispersion.

		PSNR	UQI	SSIM	MS-SSIM	VIF	FSIM	GBIM	NPBM	JNBM	NBAM
CSIQ	original metric	0.751	0.829	0.765	0.815	0.883	0.805	0.731	0.814	0.793	0.625
	saliency-based metric	0.769	0.851	0.834	0.846	0.862	0.794	0.740	0.815	0.796	0.654
	adaptive-saliency-based metric	**0.782**	**0.876**	**0.852**	**0.858**	**0.891**	**0.809**	**0.757**	**0.831**	**0.811**	**0.666**
TID2013	original metric	0.478	0.615	0.653	0.707	0.608	0.820	0.841	0.787	0.794	0.765
	saliency-based metric	0.485	0.668	0.694	0.740	0.562	0.822	0.859	0.773	0.788	0.799
	adaptive-saliency-based metric	**0.497**	**0.687**	**0.728**	**0.759**	**0.587**	**0.828**	**0.875**	**0.798**	**0.803**	**0.811**
LIVE	original metric	0.859	0.898	0.825	0.830	0.945	0.859	0.773	0.843	0.833	0.836
	saliency-based metric	0.872	0.915	0.867	0.865	0.935	0.851	0.802	0.872	0.852	0.855
	adaptive-saliency-based metric	**0.883**	**0.929**	**0.882**	**0.887**	**0.952**	**0.872**	**0.815**	**0.886**	**0.866**	**0.874**

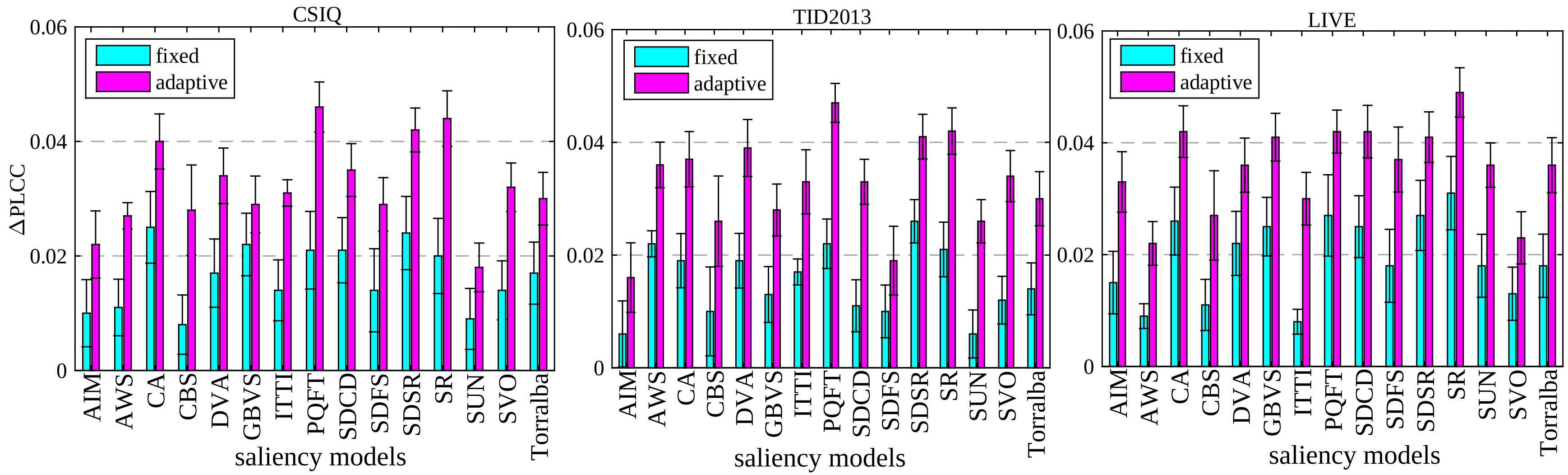


Figure 6.7: Comparison of performance gain (i.e., ΔPLCC) between saliency-augmented IQMs using fixed and adaptive use of saliency for each saliency models.

More detailed results are given in Fig. 6.7, which shows the performance gain (i.e., increase in correlation when using fixed or adaptive saliency approaches), averaged over all IQMs, for individual saliency models. On average, the gain achieved by adaptive use of saliency is more than double that of always using saliency. As well as the observed *relative* difference in gain, Figure 6.7 also gives the *absolute* gain of the adaptive approach for individual saliency models—this can be easily used to decide which of these models are more useful for IQMs. For example, by applying a threshold ΔPLCC = 0.04 to all databases picks out the good models to be PQFT, SDSR, SR. However, we again note that the purpose of this chapter is not to find the best IQM (or to target specific IQMs), but rather to compare *fixed* use of saliency to *adaptive* use of saliency according to saliency dispersion.

A paired samples t-test analysis (preceded by a test for the assumption of normality) was performed, selecting the integration strategy as the independent variable and the performance as the dependent variable. Using the $150 \times 2 \times 3$ data points contained in Table 6.1 demonstrated that an adaptive strategy is statistically significantly better (p-value = 2.18e-10, $p < 0.05$ at the 95% confidence level) than fixed inclusion of saliency.

6.6 Summary

Previous eye-tracking studies have shown that the performance gain of IQMs obtained from saliency integration is image content dependent. This chapter first demonstrated that the content dependent effect still holds when computational saliency is used, indicating that content dependency is useful in practice for optimising the saliency-augmented IQMs. This chapter then presented an efficient algorithm to reliably measure saliency dispersion in natural scenes, and considered how it can be used to adaptively incorporate computational saliency into image quality metrics. Experimental results showed that adaptive use of saliency according to saliency dispersion significantly outperforms fixed use of saliency in improving IQMs.

Chapter 7

Relation Between Visual Saliency Deviation and Image Quality

7.1 Introduction

Incorporating saliency potentially leads to improved ability of IQMs in predicting perceived quality. In Chapters 5 and 6, advanced methods have been proposed to further increase the added value of saliency in IQMs. It should, however, be noted that the use of visual saliency in IQMs is limited; saliency is mainly used to refine the importance of local distortions. Challenges to optimising the application of saliency in IQMs remain. Exploring perceptually optimised ways to use saliency information in IQMs is worth further investigating.

Previous psychophysical studies have shown that distortion occurring in an image causes visual distraction and consequently alters gaze patterns relative to that of the image without distortion [12, 13, 14, 15]. However, these studies remain limited by the choices made in their experimentation, such as the use of a limited number of human subjects [12], a small degree of stimulus variability [12, 13, 14] and the involvement of strong bias due to stimulus repetition [12, 13, 14, 15]. In Chapter 4, we systematically investigated the relation between distortion and saliency with a more reliable experimental methodology. We found that the degree of saliency deviation between an undistorted image and its distorted version increases as the perceived quality of the distorted image decreases. From this, it can be inferred that the measurable changes of gaze patterns driven by distortion may be used as a proxy for the likely variation in perceived quality of natural images. Note, in Chapter 4, the trend was observed between images using three significantly distinct levels of image quality. It is worthwhile to investigate whether the observed tendencies remain significant when more levels of perceived image quality are involved.

In this chapter, rather than using saliency as an add-on (i.e., a post-processing weighting factor) for IQMs, we investigate the plausibility of approximating image quality by means of measuring saliency deviation induced by distortion. We aim to clarify how visual distortions at different levels of intensity affect the deployment of visual fixations. We hypothesize that distortion would cause deviation in saliency, and that the extent of saliency deviation should be related to the strength of distortion and may be used as a proxy for the likely variation in image quality. To validate this hypothesis, we first conduct an eye-tracking experiment with sufficient levels of distortion. This psychophysical validation will provide empirical evidence on modelling image quality using saliency deviation. Second, from a practical point of view, we further evaluate the relation between saliency deviation and quality change using computational saliency models. This computational validation will help to select computational saliency models that best characterize saliency deviation.

7.2 Psychophysical Validation

To investigate the changes of saliency induced by distortion and their relation to image quality, we performed an eye-tracking experiment, where ground truth data of saliency were collected on natural scenes of varying quality. To be able to vary the perceived quality, each natural scene should be distorted with different types of distortion and at various levels of degradation. As evaluated in Chapter 4, asking a human subject to view multiple variations of the same scene is likely to result in biased eye-tracking data. To eliminate such potential bias, we followed the between-subjects experimental design as used in Chapter 4. Our experiment contains 149 test stimuli with a large degree of variability in terms of image content, distortion type as well as distortion level, and involves 100 human observers.

7.2.1 Experimental methodology

Stimuli

To leave out expensive image quality scoring experiments, we decided to construct our set of stimuli by systematically selecting images from the LIVE database [16], which already contains per image a ground truth quality score i.e., DMOS. The range of the DMOS in LIVE is between 0 to 100, with a higher DMOS indicating a lower perceived quality. The LIVE database consists of five subsets: JP2K, JPEG, WN, GBLUR and FF. For each subset, we define 8 distinguishable perceived quality levels (PQLs) (i.e., $DMOS = 10 * k, k = 1, 2...8$) by dividing the DMOS between 0 and 80 to 8 intervals. The images with a DMOS larger that 80 are excluded in our experiment since the original content are totally covered by distortions. For each PQL, we accommodated three images of different scenes. Note we wish to include more stimuli for each

Table 7.1: Configuration of test stimuli from LIVE image quality database.

	JPEG2000		JPEG		WN		GBLUR		FF	
PQL	image	DMOS	image	DMOS	image	DMOS	image	DMOS	image	DMOS
1	img191	78.16	img91	78.98	img106	77.63	img121	74.67	img21	78.40
	img79	79.17	img100	81.20	img61	78.47	img69	79.76	img18	81.02
	img107	79.85	img188	83.03	img50	79.44	img11	83.27	img92	81.44
2	img220	70.84	img207	70.02	img96	69.26	img125	68.02	img112	70.08
	img227	70.88	img175	70.50	img134	70.65	img118	70.26	img3	70.14
	img28	70.95	img134	72.36	img39	71.76	img120	71.84	img141	71.64
3	img122	58.56	img156	59.65	img32	60.28	img53	59.57	img88	59.17
	img160	58.75	img9	59.74	img124	61.46	img40	59.73	img66	59.22
	img137	60.61	img41	59.87	img26	61.63	img73	60.11	img98	60.48
4	img91	48.72	img69	48.87	img25	49.83	img130	49.54	img81	48.85
	img163	49.97	img21	50.12	img102	50.19	img38	49.57	img56	49.80
	img170	50.19	img128	52.32	img139	52.70	img30	50.78	img32	52.33
5	img8	39.38	img15	41.37	img132	38.09	img76	40.32	img93	38.39
	img133	40.05	img90	42.04	img90	39.71	img77	40.77	img135	38.98
	img120	40.34	img86	42.58	img70	41.32	img103	41.62	img123	41.42
6	img187	29.92	img163	28.82	img22	29.05	img29	29.71	img38	30.78
	img78	30.57	img63	30.00	img1	29.05	img35	29.95	img109	30.89
	img222	31.28	img56	32.01	img84	31.25	img13	30.51	img89	31.06
7	img61	21.46	img38	18.03	img54	20.53	img32	20.42	img14	19.72
	img21	22.46	img174	19.47	img138	22.15	img90	20.70	img44	20.70
	img80	20.57	img216	20.93	img103	22.88	img51	21.91	img98	20.90
8	img198	9.47	img101	10.25	img59	11.04	img95	17.16	img119	8.32
	img59	10.11	img58	10.55	img101	11.17	img137	17.75	img64	10.14
	img75	11.04	img130	11.02	img98	13.27	img63	18.64	img27	10.90

PQL. The stimulus variability of LIVE, however, limited us from including more. This yielded 120 distorted images (5 distortion type × 8 PQLs × 3 scenes) that covers all the 29 image content and 5 distortion types in LIVE database. We also included the 29 reference images for comparing the saliency deviation, resulting in 149 stimuli in total in our experimental design. Table 7.1 shows the configuration of the stimuli. Figure 7.1 further illustrates the average DMOS of all stimuli for each PQL. It clearly visualises that our dataset contains eight distinct levels of perceived picture quality. Note that making sure the quality levels are perceptually distinct is an important prerequisite for the following study. Pairwise comparisons were performed with

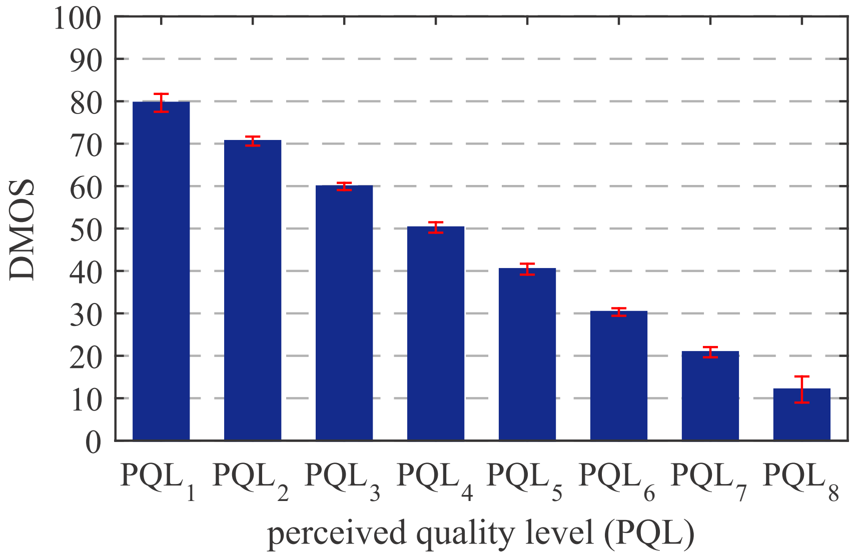


Figure 7.1: Illustration of the average DMOS of all stimuli for each perceived quality level (PQL) in our database. The error bars indicate the standard deviation.

a Wilcoxon signed rank test (i.e., an alternative for t-test for non-normal distributions) between two successive quality levels, selecting DMOS as the dependent variable, and the quality level as the independent variable. The results indicated that the difference between any pair of consecutive levels is statistically significant (p-value ranges from 6.11e-15 to 5.15e-13, $p < 0.05$ at 95% confidence level).

Protocol

Again, to reduce the undesirable effect due to stimulus repetition, we followed the "between-subjects" design established in Chapter 4. In particular, the test dataset was divided into 5 partitions (i.e., one contains 29 images and the other four contain 30 images each), and only up to two repeated versions of the same scene were allowed in each partition. Stimuli assigned to each partition covered all distortion types and the full range of quality levels.

Experimental procedure

We set up a standard office environment as to the guidelines specified in [4] for the conduct of our experiment. The test stimuli were displayed on a 19-inch LCD monitor (native resolution: 1024×768 pixels). Again, we used the SensoMotoric Instrument (SMI) RED-m eye-tracker to conduct the experiment. Each subject was provided with instructions on the purpose and general procedure of the experiment (e.g., the task, the format of stimuli and timing) before the start of the actual experiment. A training session was carried out to familiarise the participants with the experiment, using 10 images that were different from those used in the real experiment. Each session per subject was preceded by a 9-point calibration of the eye-tracking equipment. The participants were instructed to experience the stimuli in a natural way ("view it as you normally would"). Each participant saw all stimuli (in his/her assigned partition) in a random order. Each stimulus was shown for 10 seconds followed by a mid-gray screen of 2 seconds.

We recruited a total of 100 participants in our experiment. The subject pool consisted of 50 male and 50 female university students and staff members. They were aged between 22 to 47 years, and all inexperienced with image quality assessment and eye-tracking recordings. The experiment went through the required ethics review process and all the participants volunteered for this eye-tracking experiment (i.e., no payment was made to the participants). The subjects were not examined for vision defects, and their verbal expression of the soundness of their vision was considered sufficient. The participants were first randomly divided into 5 groups of equal size, each with 10 males and 10 females; and the 5 groups of subjects were then randomly assigned to 5 partitions of stimuli. This gave a sample size of 20 subjects per test stimulus.

7.2.2 Experimental results

Fixation map

A topographic fixation map that reflects the stimulus-driven, bottom-up aspects of visual attention was derived from free-viewing fixations. For a given stimulus, its fixation map was constructed by first accumulating fixations over all viewers (i.e., 20 subjects in our experiment) and then convolving the resulting fixation points with a Gaussian kernel. The width of the Gaussian kernel approximates the size of fovea (i.e., 2 degrees of visual angle [90, 13, 91], and 45 pixels in our experiment). The intensity of the resulting fixation map was linearly normalised to the range [0, 1]. Figure 7.2 illustrates the saliency maps for the original images (i.e., referred to as scene saliency (SS)) used in our experiment, and for samples of their distorted versions (i.e., referred to as distorted scene saliency (DSS)). In Fig. 7.2 (b), we intentionally selected distorted stimuli from our dataset and showed them in order of perceived quality level (i.e., DMOS increases from 8.31 for the left end image to 81.44 for the right end image). As can be seen, the difference between SS and DSS increases as the perceived quality decreases (or DMOS increases), which will be further investigated below.

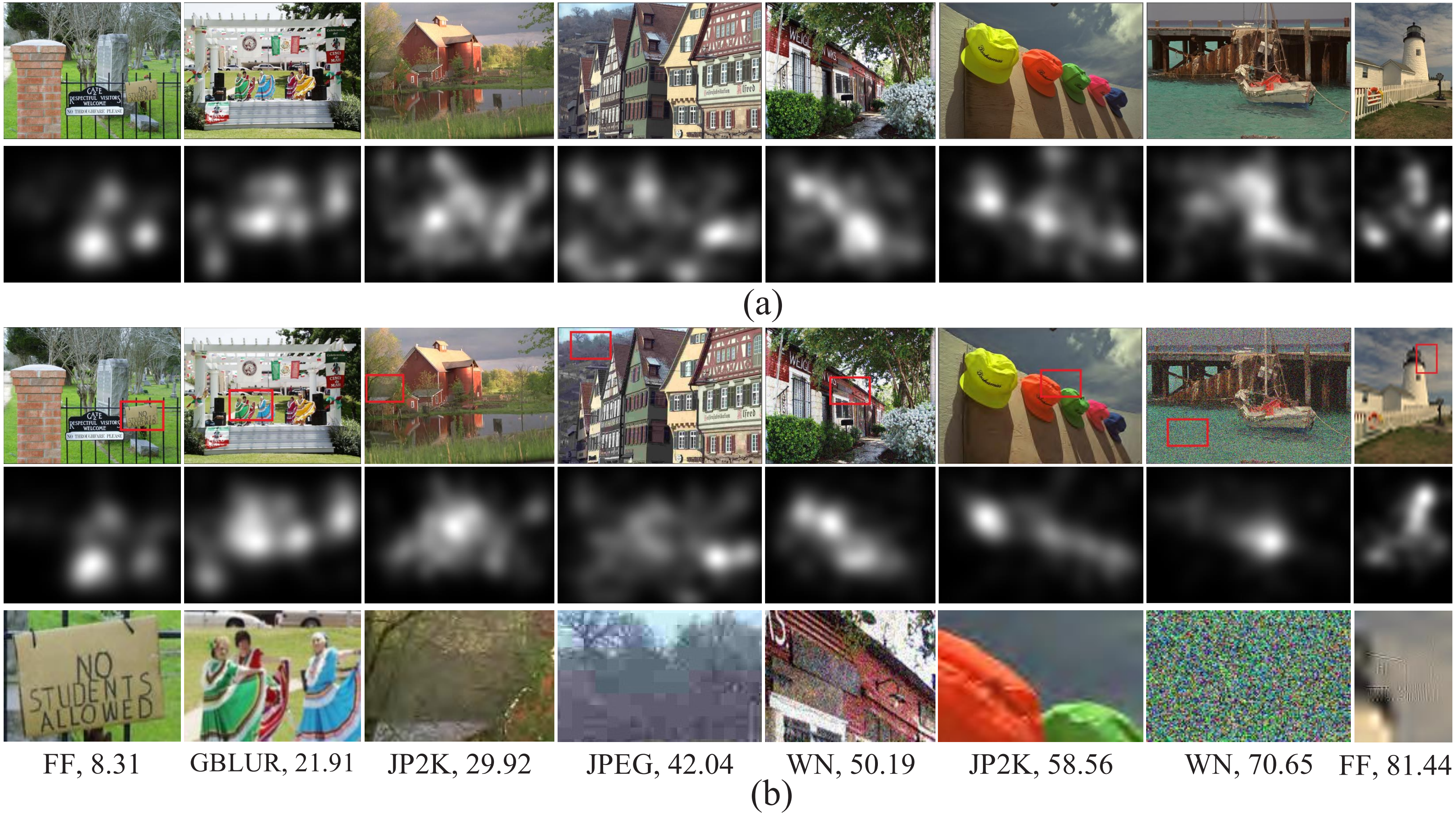


Figure 7.2: Illustration of the ground truth saliency maps for the original images used in our database (a) and for samples of their distorted versions (b). The stimuli in the first row of (b) are placed in order of perceived quality (with the corresponding DMOS values listed at the bottom of (b)), and the third row of (b) shows the image patches extracted from the stimuli (i.e., as indicated by the red boxes in the stimuli).

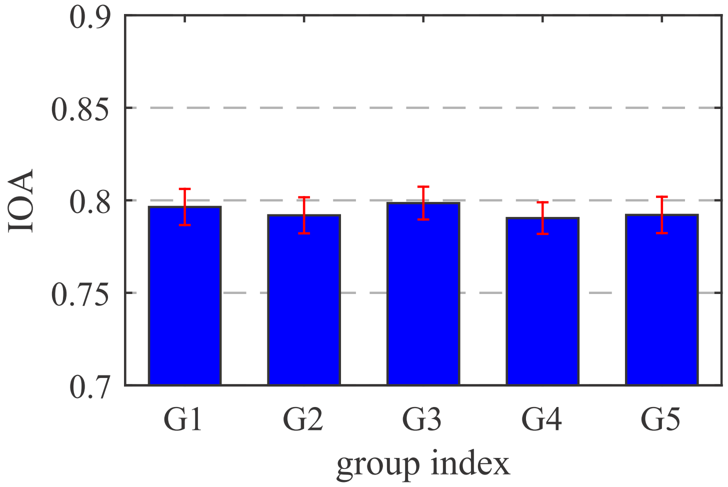


Figure 7.3: Illustration of the inter-observer agreement (IOA) averaged over all stimuli assigned to each group in our experiment. The error bars indicate a 95% confidence interval.

Reliability testing

Since the reliability of eye-tracking data strongly depends on, e.g., the number of viewers used and the strength of carry-over effects, it is crucial to validate the reliability of the collected data before using them as a "ground truth". We followed the proposed testings in Chapter 4 to assess (1) whether the variances of eye-tracking data are homogeneous across different subject groups (in a between-subjects method); (2) whether the sample size (number of participants) used per stimulus is sufficient to achieve a stable saliency map.

Homogeneity of variance across groups: To be able to assess the homogeneity, again, we measured the IOA among observers viewing the same stimulus as we did in Chapter 4. Figure 7.3 illustrates the IQA value averaged over all stimuli assigned to each subject group, showing that the IQA remains similar across five groups in our experiment. A statistical significance test (i.e., analysis of variance (ANOVA)) was performed and the results (i.e., p-value = 0.68, $p > 0.05$ at 95% confidence level) showed that there is no statistically significant difference between groups, suggesting a high degree of consistency in eye-tracking data across groups.

Data saturation: To determine the number of participants required for an eye-tracking experiment, we evaluated whether the sample size used in our experiment is adequate to reach saliency "saturation" (i.e., a proxy of sufficient degree of reliability). The validation was again based on the IOA-k as used in Chapter 4. Figure 7.4 illustrates the IOA-k value averaged over all stimuli contained in our entire dataset. It shows that "saturation" starts to occur with 18

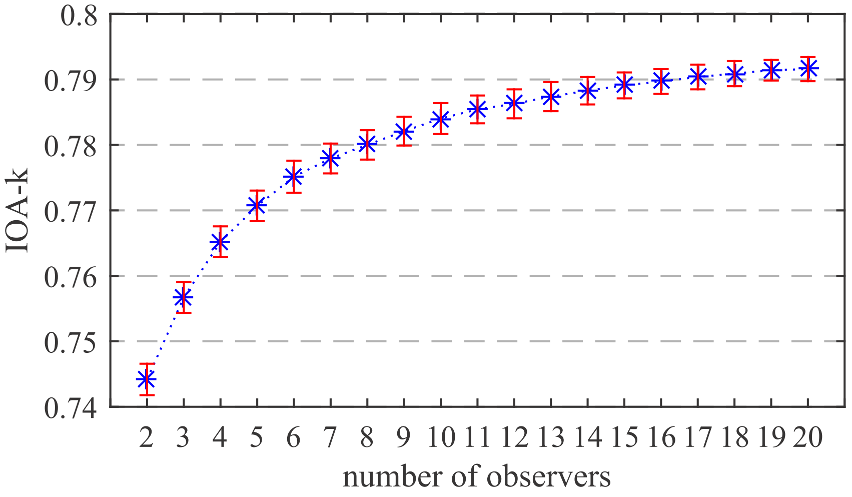


Figure 7.4: Illustration of the inter-k-observer agreement (IOA-k) averaged over all stimuli contained in our entire dataset. The error bars indicate a 95% confidence interval.

participants, suggesting that our sample size (i.e., 20 observers per stimulus) gives a stable and saturated saliency map.

Saliency deviation v.s. quality variation

As can be seen from Fig. 7.2, when comparing a DSS map to its corresponding SS map, there exist some consistent patterns, e.g., the highly salient regions tend to occur around the same places in both maps; the deviation of DSS from SS seems to be associated with the strength of distortion (or level of perceived quality). To verify this observation, we used the SS map derived from the original undistorted scene as the reference, and quantified the deviation of a DSS map from the reference (i.e., referred to as SS-DSS deviation), using three popular similarity measures: AUC, NSS and KLD.

Figure 7.5 illustrates the SS-DSS deviation (in terms of AUC, NSS and KLD) averaged over all distorted stimuli within each perceived quality level. In general, the figure shows that distortion strength has a strong effect on SS-DSS deviation, independent of the similarity measure used. The degree of SS-DSS deviation increases as the perceived quality decreases (or strength of distortion increases). An ANOVA test was further performed with the quality level as independent variable and the SS-DSS deviation as dependent variable. Experimental result showed that there exists statistically significant difference between the SS-DSS deviation of eight levels with p-value = 0.007, $p < 0.05$ at 95% confidence level. However, significant difference was merely detected between two successive quality levels due to the long error bars. Including more stimuli for each level can reduce the error bar thus resulting in a stronger significance level. Overall, we can conclude that the SS-DSS deviation driven by distortion correlated well

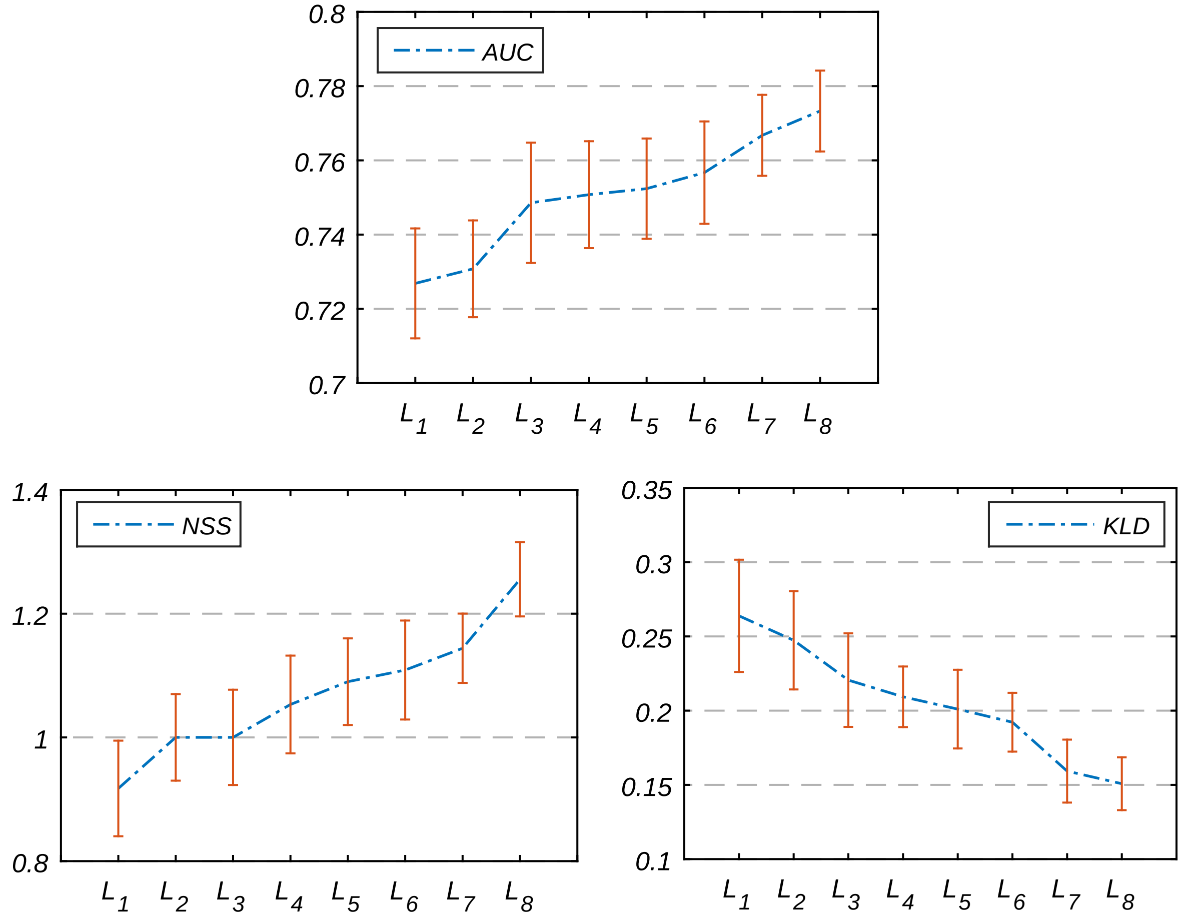


Figure 7.5: The measured SS-DSS deviation in terms of AUC, NSS and KLD for images of different perceived quality (or distortion strength). The error bars indicate a 95% confidence interval.

with the change of the perceived quality.

7.3 Computational Validation

Based on the findings of our psychophysical studies, it can be inferred that the changes of saliency induced by distortion are strongly associated with the changes of image quality. We further investigated the plausibility of modelling SS-DSS deviation as a proxy for the likely variation in perceived image quality. Eye-tracking is cumbersome and impractical in many real-world applications. A more realistic and practical system will use a computational model of saliency rather than eye-tracking. Fundamental problems such as how to gauge the effectiveness of saliency models and to what extent the SS-DSS deviation measured by these models is useful for image quality prediction remain unsolved, and are the topics to be investigated below. Note that, in this chapter, we focus on validating the plausibility of using computational saliency in place of eye-tracking data for SS-DSS deviation, rather than developing an IQM. The latter is done in the following chapter.

7.3.1 Evaluation criteria

To have sufficient statistical power, our evaluation was conducted using various state of the art saliency models and three widely recognised image quality databases: LIVE, CSIQ and TID2013. The saliency models considered are CA, CBS, CovSal, DVA, GBVS, ITTI, PQFT, salLiu, SDSF, SDSR, SR, SUN, Torralba, GR [136], LGS [137], RARE2012 [138] and SigSal [139]. Figure 7.6 shows the saliency maps generated by these models for one of the reference images and one of its distorted versions in our testbed.

To quantify the SS-DSS deviation of modelled saliency maps, KLD was used. Note since both AUC and NSS essentially require the access to the original fixation locations which are, however, not available for a computational saliency model, they become unrealistic for comparing two modelled saliency maps and producing SS-DSS deviation. To compensate for the lack of appropriate similarity measures, we devised a new measure, saliency deviation measure (SDM), as follows:

$$SDM_{SS-DSS} = mean(\frac{2 \cdot SS \cdot DSS + \varepsilon}{SS^2 + DSS^2 + \varepsilon}) \tag{7.1}$$

where SS denotes the grayscale modelled SS, DSS denotes the grayscale modelled DSS, and ε denotes a constant to avoid instability when $SS^2 + DSS^2$ is very close to zero (i.e., ε = 0.01 is used in our experiment). Note that the value is somewhat arbitrary, but we found that the performance of the SDM algorithm is fairly insensitive to variations of this value).

Finally, the strength of the relationship between the SS-DSS deviation and the perceived image quality (i.e., DMOS) was quantified by the PLCC, SROCC and KROCC. Note the PLCC is customarily calculated after performing a nonlinear regression.

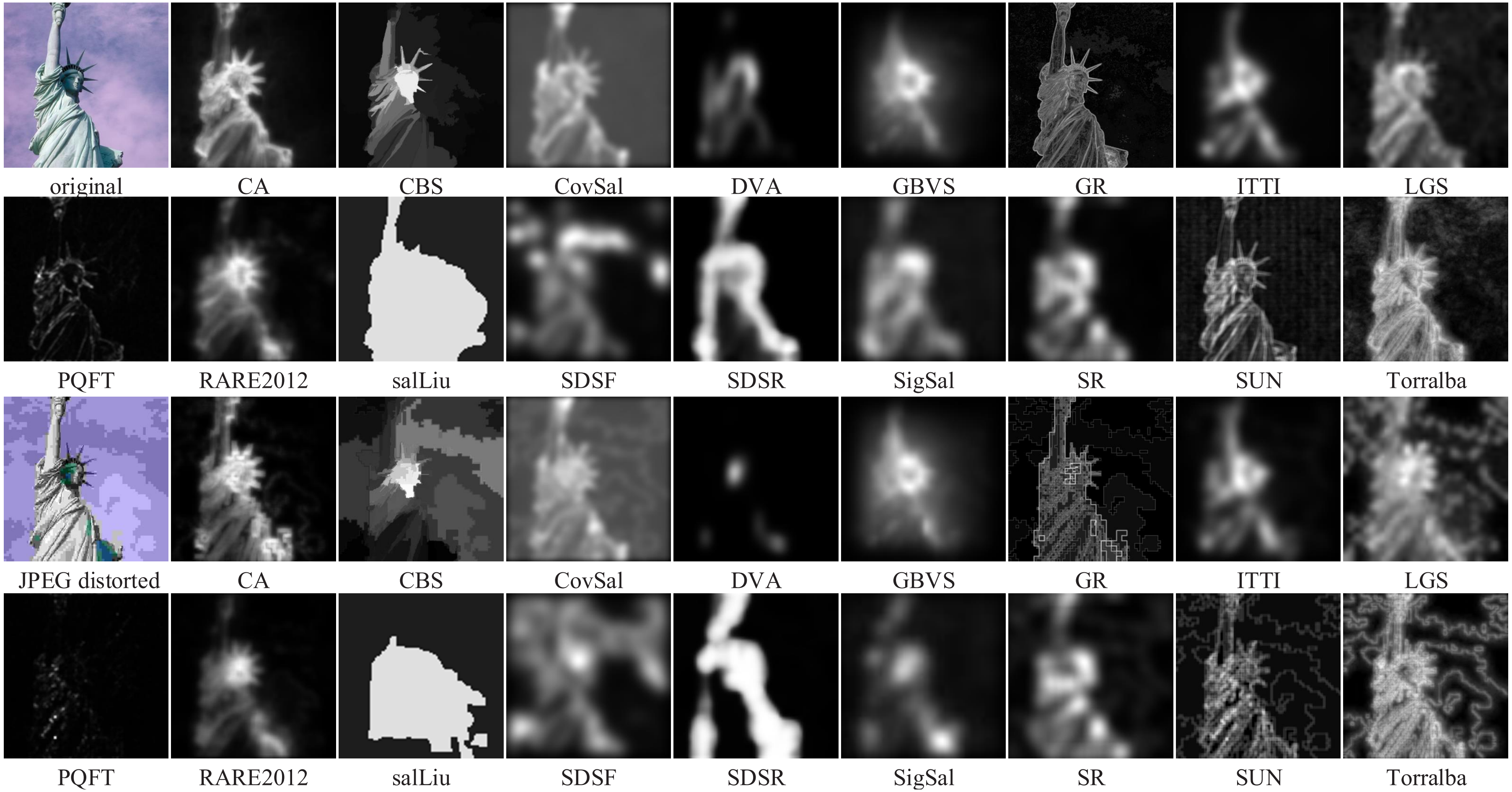


Figure 7.6: Illustration of saliency maps generated by 17 state of the art saliency models for one of the original images in CSIQ database and for one of its JPEG distorted versions.

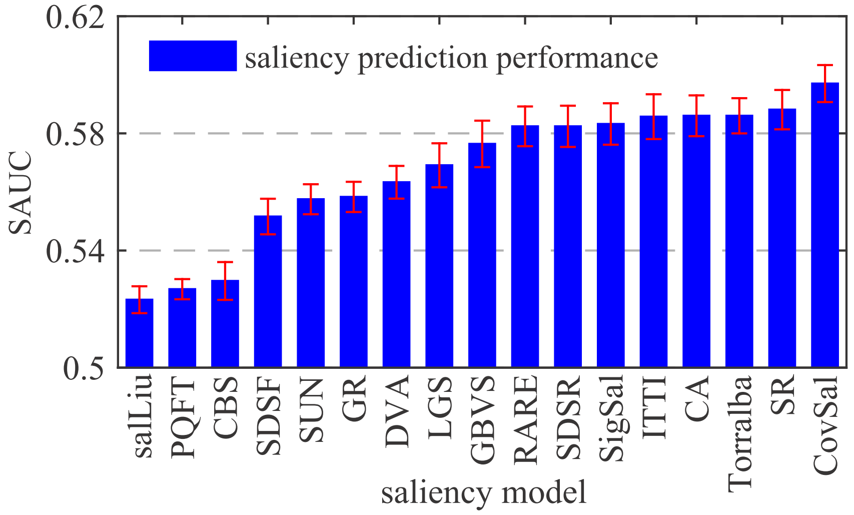


Figure 7.7: Illustration of the rankings of saliency models in terms of predictive power measured by SAUC. The error bars indicate a 95% confidence interval.

7.3.2 Experimental results

Performance of saliency models on ground truth SS and DSS

Benchmarking saliency models against ground truth SS has been extensively attempted [44, 115], however, little is known about the performance of existing saliency models in terms of detecting DSS. Being able to detect both kinds of saliency would justify the applicability of a saliency model in the specific area of image quality, where stimuli are distorted. We used the obtained eye-tracking data that contain both SS and DSS to evaluate the performance of saliency models. To quantify the predictive power of a saliency model against ground truth fixations, we used SAUC to account for the centre-bias issue in the modelled saliency maps, thus precisely comparing the performance of different saliency models. Each model ran over each stimulus in the database to calculate an SAUC score, which was then averaged over all stimuli. Figure 7.7 illustrates the rankings of saliency models in terms of the average SAUC. It shows that there is variation in performance among saliency models. The observed variation was further statistically analysed with an ANOVA, and the results (i.e., p-value = 8.22e-16, $p < 0.05$ at 95% confidence level) showed that saliency model has a statistically significant effect on SAUC. This suggests the ability of predicting the ground truth saliency in terms of both SS and DSS is different for different saliency models.

Relationship between SS-DSS deviation and image quality

For each stimulus, the SS-DSS deviation can be computed using a saliency model combined with a similarity measure. Table 7.2 and Table 7.3 illustrate the correlation between the SS-DSS

Table 7.2: The correlation between SS-DSS deviation and image quality, using KLD as the similarity measure.

	LIVE			CSIQ			TID2013		
model	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
CA	0.7690	0.5734	0.7920	0.7468	0.5416	0.7639	0.5435	0.3747	0.5733
CBS	0.4664	0.3396	0.4795	0.5680	0.3940	0.5765	0.3793	0.2588	0.3910
CovSal	0.8052	0.6025	0.8099	0.7995	0.5947	0.7840	0.7343	0.5512	0.7512
DVA	0.6889	0.4862	0.7169	0.7068	0.500	0.7370	0.5242	0.3591	0.5189
GBVS	0.7895	0.5921	0.8177	0.7813	0.586	0.8003	0.5747	0.4027	0.5909
GR	0.5592	0.3848	0.5613	0.5251	0.3593	0.5438	0.3815	0.2609	0.4042
ITTI	0.8366	0.6394	0.8455	0.6602	0.4806	0.6965	0.6470	0.4639	0.6678
LGS	0.7986	0.5970	0.8078	0.6072	0.4334	0.6325	0.5298	0.3696	0.5505
PQFT	0.7900	0.5762	0.7924	0.6896	0.4922	0.7609	0.4659	0.3202	0.4949
RARE	0.6665	0.4722	0.6874	0.6124	0.4242	0.6746	0.4333	0.2965	0.4682
salLiu	0.4708	0.3258	0.4912	0.5008	0.3500	0.5329	0.2830	0.1934	0.3091
SDSF	0.7227	0.5313	0.7286	0.6280	0.4498	0.6327	0.4959	0.3445	0.4914
SDSR	0.7219	0.5256	0.7386	0.7117	0.5115	0.7417	0.6363	0.4526	0.6490
SigSal	0.6900	0.4948	0.7541	0.7065	0.5213	0.7758	0.6172	0.4414	0.6666
SR	0.7863	0.5952	0.7845	0.7803	0.5914	0.8033	0.6336	0.4552	0.6782
SUN	0.6728	0.4858	0.6818	0.5334	0.3656	0.5706	0.4060	0.2742	0.4064
Torralba	0.8193	0.635	0.8159	0.6759	0.4937	0.7367	0.6540	0.4750	0.6814

deviations (i.e., measured by KLD or SDM) and image quality scores (i.e., DMOS) for three databases (i.e., LIVE, CSIQ and TID2013), using different saliency models. Both table show that the correlation varies significantly depending on the saliency model used. For example, in Table 7.2, the SROCC values range from 0.4664 to 0.8366 on LIVE, from 0.5008 to 0.7995 on CSIQ and from 0.2830 to 0.7343 on TID2013. However, the results are insensitive to the variations of similarity measure: both KLD and SDM yield consistent correlation coefficients. For example, as shown in Table 7.2 and Table 7.3, Torralba and CovSal give the strongest correlation and salLiu and CBS give the weakest correlation, independent of the similarity measure used.

In order to provide the overall rankings of saliency models, we used the following formula:

$$Overall_{CC}(model) = \beta_1 \cdot CC_{LIVE} + \beta_2 \cdot CC_{TID2013} + \beta_3 \cdot CC_{CSIQ} \tag{7.2}$$

where β_1, β_2 and β_3 indicate a weighting factor that is proportional to the number of distorted images in a dataset as shown in Table 2.1. In particular, β_1 equals to 0.168 (i.e., $779/(779{+}866{+}$

Table 7.3: The correlation between SS-DSS deviation and image quality, using SDM as the similarity measure.

	LIVE			CSIQ			TID2013		
model	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
CA	0.7699	0.5714	0.7813	0.7758	0.5617	0.7866	0.5732	0.3866	0.5797
CBS	0.6451	0.4584	0.6580	0.5633	0.3878	0.5647	0.3856	0.2616	0.3954
CovSal	0.7086	0.5116	0.7158	0.6605	0.4679	0.6508	0.6697	0.4855	0.6881
DVA	0.7317	0.5208	0.7438	0.6909	0.4804	0.7150	0.5751	0.3997	0.5762
GBVS	0.7876	0.5912	0.8031	0.7827	0.5833	0.7933	0.5500	0.3831	0.5527
GR	0.8300	0.6306	0.8337	0.7373	0.5335	0.7752	0.5078	0.3563	0.5598
ITTI	0.8305	0.6325	0.8340	0.6447	0.4665	0.6890	0.6449	0.4608	0.6582
LGS	0.7348	0.5357	0.7440	0.6130	0.4353	0.6299	0.4875	0.3377	0.5092
PQFT	0.6230	0.4333	0.6177	0.4933	0.3154	0.5169	0.2682	0.1775	0.3092
RARE	0.6718	0.4779	0.6949	0.6081	0.4214	0.6792	0.4270	0.2921	0.4570
salLiu	0.4678	0.3247	0.4867	0.4535	0.3160	0.4906	0.2593	0.1771	0.2844
SDSF	0.7251	0.5326	0.7304	0.6272	0.4430	0.6212	0.4875	0.3355	0.4836
SDSR	0.7507	0.5530	0.7596	0.7388	0.5345	0.7590	0.6537	0.4662	0.6639
SigSal	0.6903	0.4903	0.7399	0.6858	0.4933	0.7297	0.6077	0.4294	0.6311
SR	0.8846	0.7153	0.8766	0.8551	0.6706	0.8887	0.6649	0.4883	0.7250
SUN	0.5533	0.3808	0.5797	0.5747	0.3871	0.6003	0.3250	0.2174	0.3429
Torralba	0.8854	0.6937	0.8904	0.7204	0.5321	0.8007	0.6905	0.5029	0.7423

$3000)$), β_2 equals to 0.646 (i.e., i.e., $3000/(779+866+3000)$) and β_3 = 0.186 (i.e., $866/(779+866+3000)$). CC indicates the averaged score over a certain correlation coefficient(i.e., either SROCC or KROCC or PLCC) and over two cases (i.e., based on KLD and SDM).

Figure 7.8 illustrates the saliency models in order of overall correlation for SROCC (the results for KROCC and PLCC exhibit similar trends and thus not presented here). It shows that by use of saliency models, such as CovSal, SR and Torralba, the resulting SS-DSS deviation can reasonably predict the perceived image quality. To further justify our idea of using SS-DSS deviation as a proxy for image quality prediction, we compared our best performing models (i.e., SS-DSS based on CovSal, SR and Torralba) to four state of the art image quality metrics in the literature, including SSIM, VIF, VSNR and MAD. Table 7.4 shows the comparison of performance of these models in terms of predicting image quality. It shows that the proposed SS-DSS models are fairly comparable to some of the traditional image quality metrics, suggesting that measuring saliency deviation induced by distortion is a plausible method for assessing image quality.

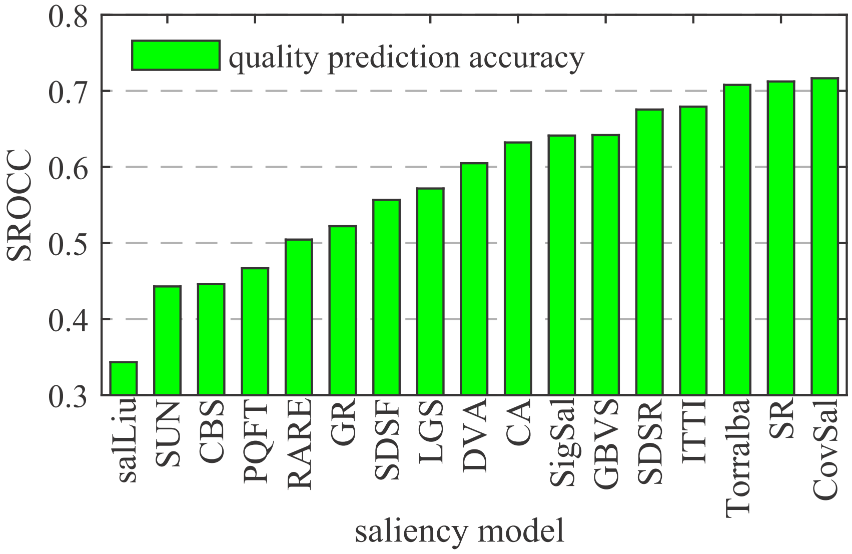


Figure 7.8: Illustration of the overall ability of saliency models in producing the correlation (in terms of SROCC) between SS-DSS deviation and image quality.

Table 7.4: The comparison of performance (in terms of SROCC) of two groups of models in predicting image quality: one refers to the proposed SS-DSS model and one refers to state of the art image quality metrics.

	proposed SS-DSS deviation models			image quality metrics			
	$SS-DSS_{Torralba}$	$SS-DSS_{SR}$	$SS-DSS_{CovSal}$	VSNR	VIF	SSIM	MAD
SROCC	0.7079	0.7124	0.7166	0.7466	0.7708	0.8012	0.8428

Impact of saliency models on the effectiveness of SS-DSS deviation measure

Having identified the benefits of SS-DSS deviation for image quality prediction, one could intuitively hypothesize that the better a saliency model can predict ground truth SS and DSS, the better the resulting SS-DSS model can predict image quality. To validate this hypothesis, we scatter plotted in Fig. 7.9 the following two variables: the saliency predictive power of a saliency model (i.e., SAUC, based on the results of Fig. 7.7) and the quality predictive power of the corresponding SS-DSS model (i.e., SROCC, based on the results of Fig. 7.8). The PLCC is equal to 0.87, suggesting a fairly strong positive relationship. In general, saliency models (e.g., CovSal, SR, Torralba and ITTI) that perform well in detecting ground truth SS and DSS lead to competitive performance of quality prediction when using these models for SS-DSS deviation measure, and vice versa (e.g., salLiu, Sun, CBS and PQFT). The findings above suggest that the performance of a saliency model in detecting both SS and DSS should be used as a criterion to determine whether or not a specific saliency model is suitable for SS-DSS deviation measure and for image quality prediction. It should be noted that this conclusion is drawn based on the

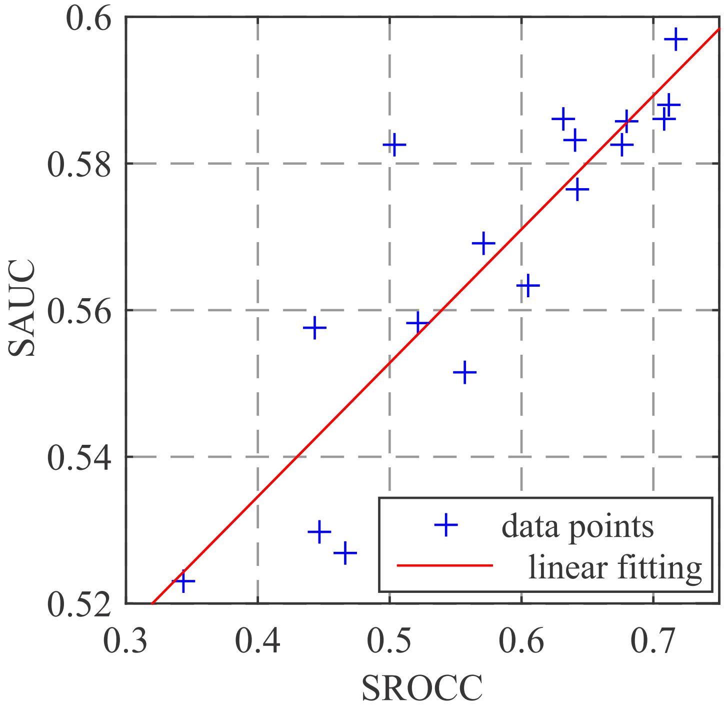


Figure 7.9: Scatter plot of two variables: the saliency predictive power of a saliency model (i.e., SAUC, based on the results of Fig. 7.7) and the quality predictive power of the corresponding SS-DSS model (i.e., SROCC, based on the results of Fig. 7.8).

saliency models used in this chapter. Including other saliency models may impact the PLCC value between the saliency model's prediction accuracy and its ability to be used as a quality indicator.

7.4 Summary

In this chapter, we investigated the relationship between the changes of gaze patterns driven by distortion and the likely variability of perceived image quality. Preliminary eye-tracking experiments have been conducted in the literature for exploring their relationship. These psychophysical studies are, however, either biased in their data collection or limited by the generalisability of their results. To provide substantial and statistically sound empirical evidence, we started from designing and conducting a large-scale eye-tracking experiment by means of a reliable methodology. This allowed us to clarify the knowledge on the relationship between the distortion-driven saliency variation and perceived image quality. We used a refined "between-subjects" experimental methodology with an aim to eliminate bias induced by each subject having to view multiple variations of the same scene in a conventional experiment. This methodology allowed reliably collecting eye-tracking data with a large degree of stimulus variability in terms of scene content, distortion type as well as degradation level. Based on the statistical evaluation on the obtained fixation data, we found that the occurrence of distortion in an image causes the deviation of fixations from their original places in the image without distortion, and that the extent of distortion determines the amount of saliency deviation. We also considered how the findings can be used to devise an algorithm for image quality prediction. To do so, we investigated the ability of several saliency models to be used as image quality indicators. Experimental results showed that it is highly plausible to approximate image quality by means of saliency deviation.

Chapter 8

A Saliency Deviation Index (SDI) for Image Quality Assessment

8.1 Introduction

Recent study in image quality assessment has shown the benefits of incorporating visual saliency in IQMs. Research so far mainly focuses on the extension of a specific IQM with a specific visual saliency model. Previous chapters have revealed potential disadvantages of this approach. First, the added value of visual saliency in IQMs strongly depends on the saliency model, the IQM, and the characteristics of the test image. An unsuitable saliency model may give a low weight to some regions of high distortion and consequently may unhelpfully downplay the importance of distortion in those regions. Also, some IQMs have already intrinsically incorporated saliency features in their design. Directly weighting the local distortion with an extra saliency model duplicates saliency inclusion and may result in marginal or even negative performance gain of the IQM. Second, existing saliency incorporation approaches, though improved in Chapters 5 and 6, implicitly assume that the attentional mechanism of the HVS functions in a post-processing manner when assessing image quality. This assumption may limit the potential use of visual saliency in IQMs. Finally, additive computational cost is added to IQMs by generating saliency maps and using them to refine the importance of local distortions.

In Chapter 7, we demonstrated that the saliency deviation between an original image and its distorted version is highly correlated with the change of image quality. Hence, it is highly plausible to approximate image quality by means of saliency deviation. In this chapter, we propose an IQM called Saliency Deviation Index (SDI), which can predict image quality by detecting and measuring saliency deviation. SDI avoids using saliency as an add-on to existing IQMs and it is also computationally less expensive. One key question arises what saliency features are suitable for SDI. In Chapter 7, the ability of various saliency models to be used as image quality assessors has been thoroughly evaluated. A straightforward solution maybe to use the best performing saliency models as found in Chapter 7, including SR, CovSal and

Torralba. Among these three saliency models, recent research in [78, 140] has proven that the amplitude spectral residual term of SR is of little significance and it is the phase spectrum term that corresponds to saliency. Therefore phase spectrum can serve as a useful feature for measuring the saliency deviation. Torralba is proved to be accurate in detecting global scene saliency. However, taking into account its high computational cost, it is impractical to use it in any IQMs which are expected to work in real-time. In contrast to SR and Torralba which highlight the salient regions from a global point of view, CovSal measures saliency in a local definition. This implies that local saliency features can also be considered in calculating the saliency deviation.

Since a single visual cue is far from complete to deal with complex natural scenes [44], we therefore decided to combine both the local saliency features and the global saliency features in designing the SDI. In our implementation, phase spectrum (PS) was selected as the global feature due to its simplicity. We did not use Torralba as the global feature due to the high computational cost involved. For the local saliency cue, we did not directly use CovSal also due to its high complexity. Instead, we simply followed how CovSal defines saliency and proposed a simple saliency feature called local detail (LD). PS and LD were combined to form the SDI. Note, both PS and LD were extracted from the luminance channel of images, we further extracted saliency features from chrominance channels of images and incorporated them in the SDI to improve its robustness.

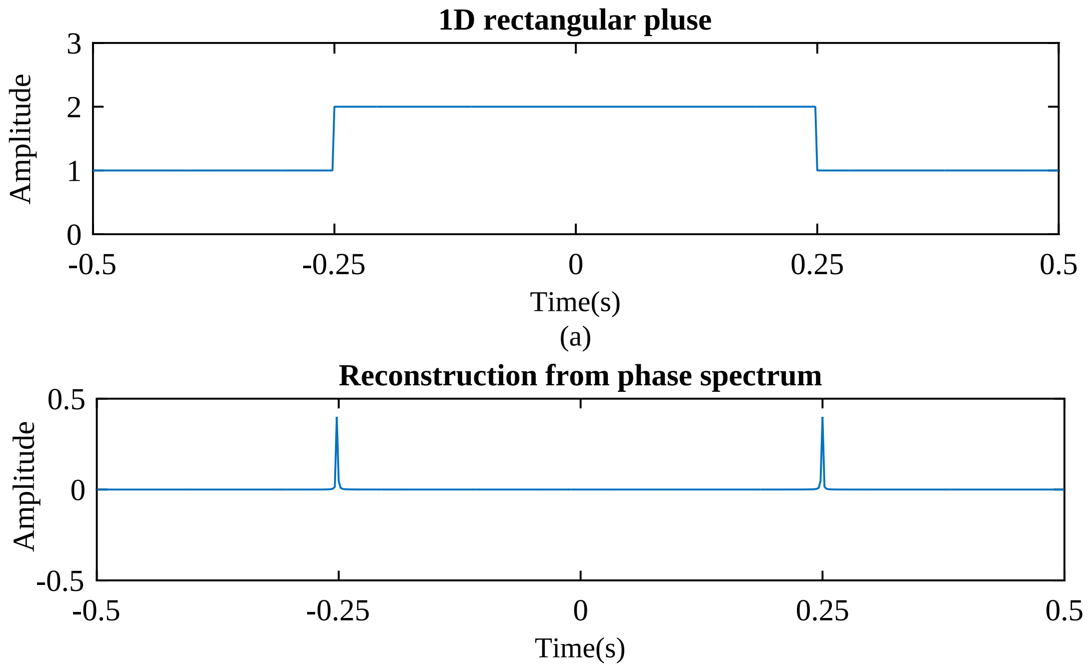
8.2 Saliency Feature Extraction

This section introduces the saliency features used in the SDI as well as how the deviations of these features due to distortion correspond to image quality.

8.2.1 Phase spectrum

It is well known that the phase spectrum specifies the location where less homogeneity is in a signal in comparison with the entire waveform. Figure 8.1(a) shows a one-dimensional rectangular pulse signal. Its reconstruction solely based on the phase spectrum is shown in Fig. 8.1(b). The two spikes in the reconstruction clearly show where the sudden changes are located in the pulse signal. In the context of a two-dimensional image, the phase spectrum indicates the locations where pop-out objects are placed. The phase spectrum saliency modelling for images is introduced in [140] and the calculation steps are as follows:

$$f(x,y) = \mathscr{F}\{I(x,y)\} \tag{8.1}$$



(a)

(b)

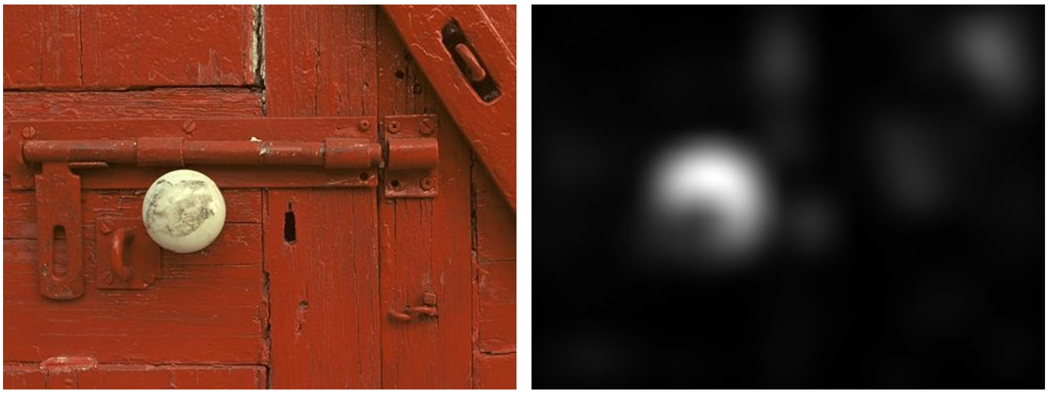
Figure 8.1: Illustration of (a) a one-dimensional rectangular pulse signal and (b) its reconstruction using phase spectrum.

$$p(x, y) = \mathscr{P}\{f(x, y)\} \tag{8.2}$$

$$PS(x, y) = g(x, y) * \left\| \mathscr{F}^{-1}\left[e^{i \cdot p(x,y)}\right] \right\|^2 \tag{8.3}$$

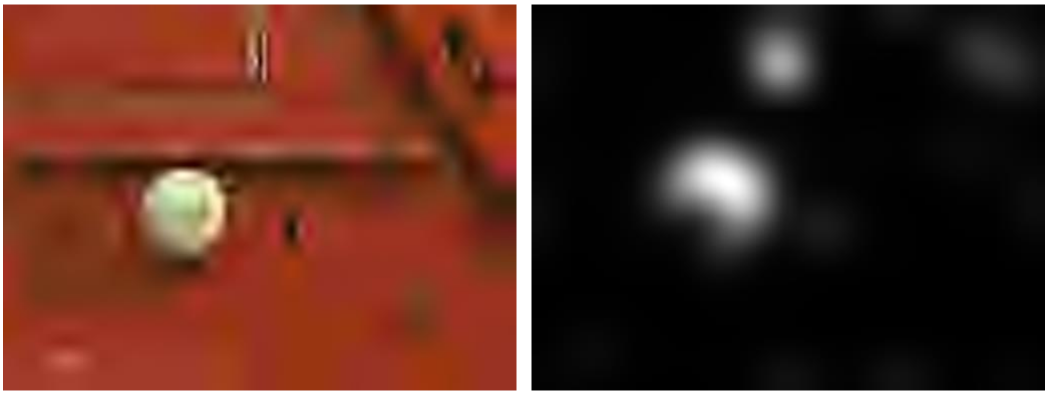
where $\mathscr{F}$ and $\mathscr{F}^{-1}$ denote the Fourier transform and inverse Fourier transform, I denotes the luminance channel of the image, $p(x, y)$ denotes the phase spectrum of the image, and $g(x, y)$ is a Gaussian filter applied on the result to smooth out the salient regions. We followed the same implementation in [140] where images are first down-sampled to a smaller resolution (i.e., 64×64) in order to avoid highlighting the unnecessary image details (e.g., textures). Figure 8.2 illustrates one original undistorted image and its saliency modelling result using PS. It shows that PS effectively highlights the door handle in the scene whilst suppresses the homogeneous background in the scene.

To calculate the deviation of the PS features between a reference image and its distorted version, the PS saliency feature of the distorted image was also extracted for comparison. Figure 8.3(a) illustrates one distorted version (i.e., JPEG2000 compression artifacts) of the reference image shown in Fig. 8.2(a), together with its PS saliency. By comparing the two PS feature maps, it can be seen that the PS features clearly show the saliency deviations due to global visual artifacts. To quantify the deviation of PS features between a reference image x and its distorted



(a) original input (b) PS

Figure 8.2: Illustration of (a) a reference image and (b) its reconstruction using phase spectrum.



(a) distorted input (b) PS

Figure 8.3: Illustration of (a) a distorted image and (b) its reconstruction using phase spectrum.

version y, a similarity measure was defined as:

$$S_{PS} = \frac{2 \cdot PS_x \cdot PS_y + T_1}{{PS_x}^2 + {PS_y}^2 + T_1} \tag{8.4}$$

where PS_x and PS_y denote the phase spectrum saliency of x and y respectively, S_{PS} is a two dimensional map representing the PS feature deviation for each pixel, and T_1 is a constant. The main role of T_1 is to improve the numerical stability (i.e., in case ${PS_x}^2+{PS_y}^2$ is close to zero) of the measure. In addition, it also serves as the added parameter of a generalised model which can be tuned to improve the performance of the proposed metric. This similarity measure is widely applied in image quality community to measure the similarity of two features. It provides a symmetric measure making $S_{PS}(x, y) = S_{PS}(y, x)$. This definition also has a fixed range (0, 1] for each local pixel with the maximum value 1 indicating PS_x and PS_y are exactly the same at that location and a value approaching 0 indicating PS_x and PS_y are significantly different at

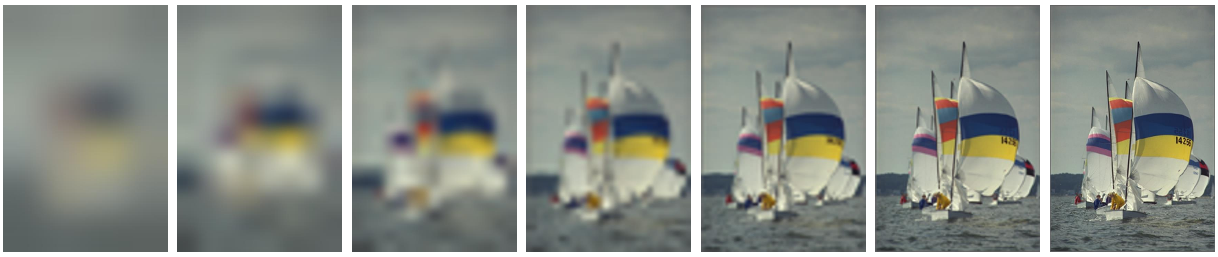


Figure 8.4: Illustration of the coarse-to-fine mechanism of the HVS. Note all lower scales are upsampled to the original resolution o the image.

that location.

8.2.2 Local detail

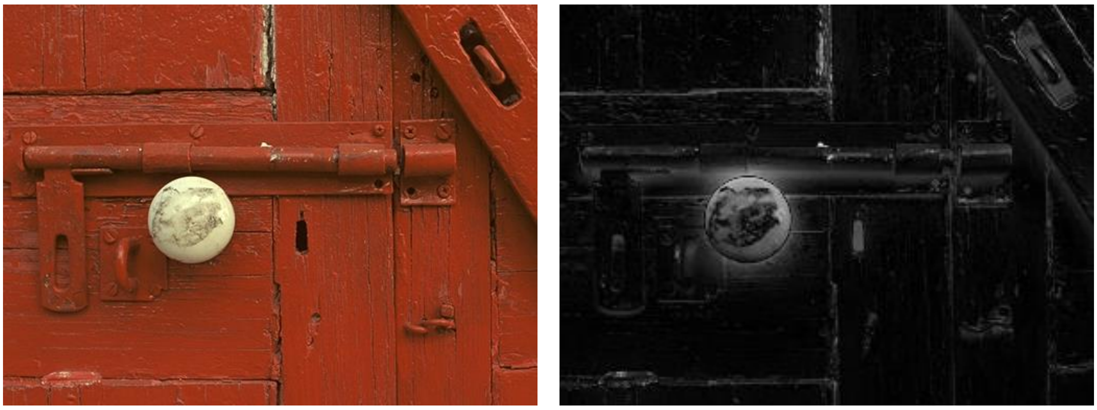
The CovSal saliency model measures the saliency locally as how much a pixel differs from its surroundings. Inspired by this idea, a saliency feature called local detail (LD) was proposed to represent saliency in a local manner. Different from the approach taken in CovSal where saliency is defined as the dissimilarity between a center patch and all its surrounding patches, the proposed LD measures local saliency as the differences between different scales of an image.

Psychophysical studies have shown that the HVS generally operates at multi scales in a coarse-to-fine manner [141]. This means that the HVS first perceives a scene in a low resolution, followed by several enhanced resolutions. The HVS treats the regions where significant deviations exist between a lower resolution and a higher resolution as important. To mimic this multi-scale mechanism of the HVS, in our implementation, a low-pass filter was iteratively applied to the image which down-samples the image by a factor of 2, resulting in a N scales image pyramid. The relation between two successive scales was formulated as:

$$S_{k+1} = f_{lowpass} * S_k \tag{8.5}$$

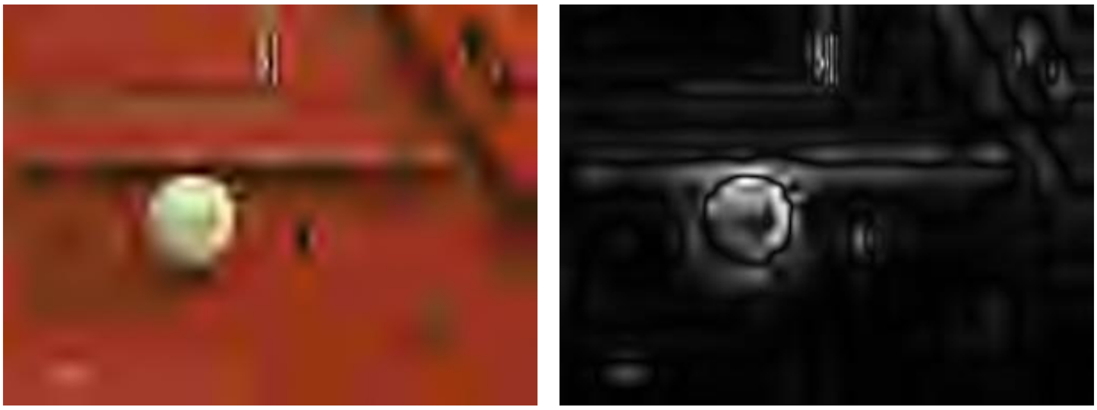
where S_k denotes the kth scale and $f_{lowpass}$ denotes the low-pass filter applied. In this implementation, the Gaussian filter was used as it is simple while being able to sufficiently preserve the edge information. Figure 8.4 illustrates one image with N (e.g., $N = 7$) scales (with each scale up-sampled to the original resolution). As can be seen in Fig. 8.4, the HVS first perceives the boats that stand out of the ocean; then perceives the details of the boats (e.g., the sails) and the persons in the boats. In order to highlight the salient regions that pop-out into our visual field at different scales, LD was simply modelled as the difference between two consecutive scales:

$$LD = \frac{1}{N} \sum_{k=0}^{N} (S_k - S_{k+1}) \tag{8.6}$$



(a) original input (b) LD

Figure 8.5: Illustration of (a) an original input and its corresponding (a) LD.



(a) distorted input (b) LD

Figure 8.6: Illustration of (a) a distorted input and its corresponding (a) LD.

It should be noted that the LD feature extraction is also performed on the luminance channel of the input image. To determine the number of scales used, our implementation related N to the original resolution, in order to adaptively determine a proper N for images of different size. We chose N to be:

$$N = floor(\log_2[\min(width, height)]) \tag{8.7}$$

where $floor()$ is the round down function. Figure 8.5 illustrates LD feature of the same image used in Fig. 8.2. It shows that LD feature successfully highlights the local details at different level from coarse (e.g., the white handle) to fine (e.g., textures of the red door). To check whether LD is able to detect the local saliency deviations due to distortion, we also extracted the LD feature on the same distorted image as used in Fig. 8.3 and visualize it in Fig. 8.6. By comparing Fig. 8.6(b) and Fig. 8.5(b), we can see that the LD feature deviates due to distortion.

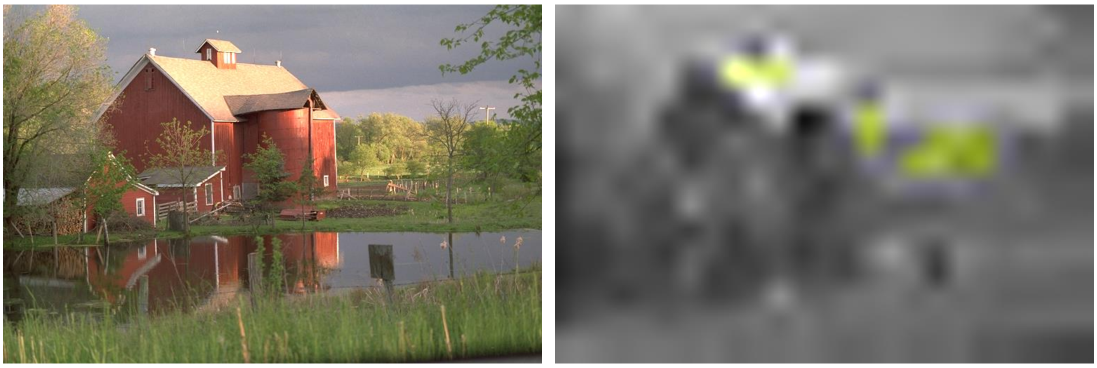

(a) original image (b) a FF distorted version

Figure 8.7: Example of the effect of the transmission error on chromatic channels.

The deviation can again be quantified as:

$$S_{LD} = \frac{2 \cdot LD_x \cdot LD_y + T_2}{{LD_x}^2 + {LD_y}^2 + T_2} \tag{8.8}$$

where LD_x and LD_y denote the local detail saliency of the reference image x and its distorted version y respectively, S_{LD} is a two dimensional map representing the LD feature deviation for each pixel, and T_2 is a constant that plays the same role of T_1.

8.2.3 Colour feature

The PS and LD saliency features are extracted from the luminance channel of images whereas color information is not taken into account. However, some types of visual distortions will also affect the chromatic channels, thus deteriorate the image quality. For example, Fig. 8.7 shows an image with strong FF transmission errors. The visual distortion significantly influences the chromatic channels of the original image, resulting in a change in color perception. The salient regions in the distorted image may shift from the red house as shown in Fig. 8.7(a) to the yellow regions in Fig. 8.7(b). In this case, the saliency features purely based on the luminance channel may not be able to fully characterise the impact of distortion on the HVS. Moreover, some types of artifacts, e.g., color saturation change during image printing and JPEG-based image compression [18] only impact the chromatic channels. In this case, the luminance-based saliency features for the reference image and distorted image are totally the same, leading to a false conclusion that there is no quality difference in between. Figure 8.8 illustrates an image (i.e., taken from TID2013 image quality database [18]) that suffers from different levels of color saturation change. The salient level of the red flower in the bottom right corner may significantly drop since the decrease of saturation impacts the salient level of that flower.

To compensate for such deficiency, the saliency deviation measure in chromatic channels should

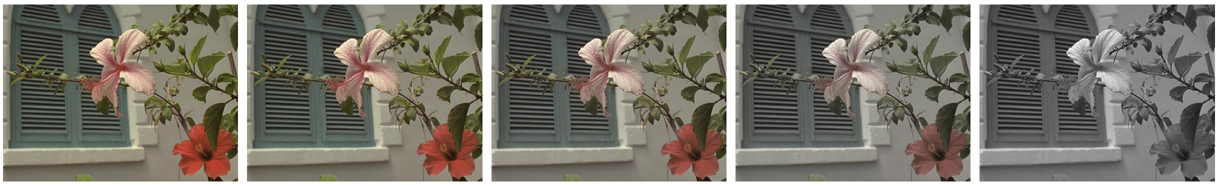


Figure 8.8: Example of color saturation distortion. The distortion level increases from left to right.

also be included in the SDI. The HVS is proved to be less sensitive to color change than the luminance change [142, 143]. Taking advantage of this characteristics of the HVS, we chose not to extract a specific color feature, but simply measured the deviation of color feature as the change of chrominance channels. Prior to the color deviation quantification, we first converted the images in RGB color space to a color opponent space [144] where the color perception of the HVS is better reflected. Following the conversion method in [145], two opponent color channels were simply modelled as:

$$RG = R - G \tag{8.9}$$

$$BY = 2B - (R + G) \tag{8.10}$$

where RG indicates the red-green color opponency and BY indicates the blue-yellow color opponency. Figure 8.9 illustrates the RG and BY channels for both the original and distorted images in Fig. 8.7. It can be seen that the both RG and BY channels change due to the visual distortions and the change of chromatic channels reflects the deviation of saliency deployment. In our implementation, we defined the chrominance deviation as:

$$S_C = \frac{2 \cdot RG_x \cdot RG_y + T_3}{{RG_x}^2 + {RG_y}^2 + T_3} \cdot \frac{2 \cdot BY_x \cdot BY_y + T_3}{{BY_x}^2 + {BY_y}^2 + T_3} \tag{8.11}$$

where RG_x, BY_x, RG_y, and BY_y denotes the RG and BY opponent color channels of the original image x and its distorted version y respectively, S_C is a two dimensional map representing the color change for each pixel, and T_3 is a constant that plays the same role as T_1.

8.3 Saliency Deviation Index

Previous process provided us three saliency deviation maps: the global saliency deviation, the local saliency deviation and the saliency deviation due to chrominance errors. We combined these maps to form a unique quality map (QM) with each pixel value indicating the quality for that location as follow:

$$QM = {S_{LD}}^{\alpha} \cdot {S_{PS}}^{\beta} \cdot {S_C}^{\gamma} \tag{8.12}$$

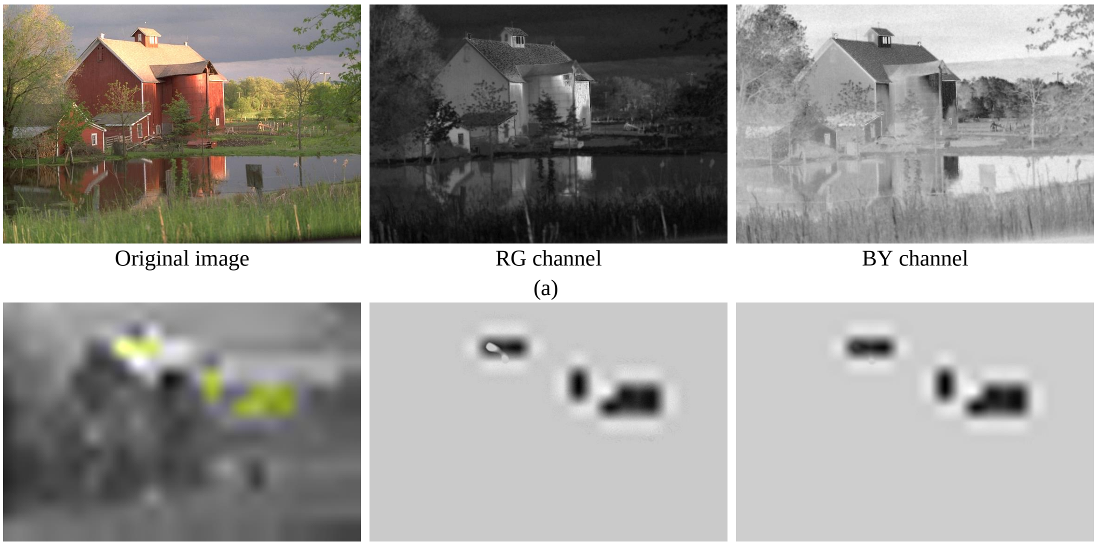


Figure 8.9: Illustration of the RG and BY channels for (a) an original image and (b) a distorted image.

where $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ are parameters to adjust the relative importance of different components. In this map, a larger pixel value corresponds to a higher quality (i.e., higher similarity between the reference and the distorted) of that location. We set $\alpha = \beta = 1$ in our implementation for the saliency features of luminance channel for simplification. Since the HVS is less sensitive to the change in chrominance channels, we set γ to be a positive constant less than 1 to limit its impact on the final quality map.

Once we have the overall quality map, we should consider how to pool the local quality values for all the pixels into a single score representing the overall quality of the distorted image. Researchers in National Telecommunications and Information Administration (NTIA) found that the quality judgement towards a visual stimulus is predominated by the worst part of the picture whereas the low level distortions have less impact on the quality perception [146]. We thus pooled the quality map into a single score as the output of SDI as:

$$SDI = \frac{\sum_i S_{LD} \cdot S_{PS} \cdot S_C^{\gamma} \cdot P(i)}{\sum_i P(i)} \tag{8.13}$$

where i indicates each pixel in the quality map and P indicates a penalty function to assign more weight to the pixels with high degree of distortion (i.e., low value in our quality map). Therefore we followed the definition of P as a monotonically decreasing function [147]:

$$P(i) = \left(\frac{1}{QM(i)}\right)^k \tag{8.14}$$

where $k > 0$ is a parameter to adjust the level of penalty for pixel i in the QM. By using this definition, a larger pixel value in the QM (i.e., a smaller degree of distortion) is given a smaller weight while a lower pixel value in the QM (i.e., a larger degree of distortion) is given a larger weight.

Before we can deploy the SDI in any real world application, we need to determine all the parameters involved. The SDI metric consists of five parameters: T_1 in the S_{PS}, T_2 in the S_{LD}, T_3 in the S_C, γ for the importance of chrominance channels and k in the penalty function. To perform a parameter selection, we conducted a 5-fold cross-validation on the TID2013 image quality database [18]. The TID2013 database includes 25 image contents with each content having 120 distorted images. We randomly partitioned the 25 content into 5 subsets of equal size. We retained a single subset as the test set and used the remaining 4 subsets as training set. We then repeated the process 5 times with each of the 5 subsets being used as the test set once. Experimental results showed that the 5 groups of parameters were similar to each other and the performance of SDI with each group of parameters were also similar to each other. We thus chose one group of parameters and fixed them as: $T_1 = 0.01$, $T_2 = 0.04$, $T_3 = 400$, $\gamma = 0.04$ and $k = 0.4$.

8.4 Performance Evaluation

In this section, the performance of SDI was evaluated on three widely used image quality assessment databases. We also compared the performance of SDI with several state of the art IQMs in the literature regarding their performance and complexity. The databases are LIVE [16], CSIQ [17] and TID2013 [18]. The IQMs used for comparison are the state of the arts in the literature, namely PSNR, SSIM [31], MS-SSIM [32], VIF [34], VSNR [27], MAD [28], FSIM [35] and GSM [148]. The performance evaluation criteria are PLCC, SROCC and RMSE. We followed the suggestion in [105] where a non-linear mapping is performed before calculating the PLCC and RMSE. We used the following nonlinear mapping function:

$$x' = b_1 \cdot (\frac{1}{2} - \frac{1}{1 + e^{b_2(x-b_3)}}) + b_4 \cdot x + b_5 \tag{8.15}$$

where x denotes the SDI's output, x' denotes the mapped score, b_i denotes a fitting parameter.

8.4.1 Prediction accuracy

Table 8.1 shows the performance of different IQMs on three databases. The IQM that ranks the highest on individual database is highlighted in bold. It can be seen that SDI performs

Table 8.1: Performance comparison of eight state of the art IQMs on three image quality datasets.

		PSNR	SSIM	MS-SSIM	VIF	MAD	VSNR	FSIM	GSM	SDI
TID2013	SROCC	0.640	0.742	0.786	0.677	0.781	0.681	0.802	0.795	**0.839**
	PLCC	0.580	0.790	0.833	0.772	0.827	0.740	0.859	0.846	**0.862**
	RMSE	1.010	0.761	0.686	0.788	0.698	0.839	0.634	0.660	**0.629**
LIVE	SROCC	0.876	0.948	0.951	0.964	**0.967**	0.927	0.963	0.956	0.951
	PLCC	0.872	0.945	0.949	0.960	**0.968**	0.923	0.960	0.951	0.946
	RMSE	13.36	8.946	8.618	7.614	**6.907**	10.51	7.678	8.433	8.86
CSIQ	SROCC	0.806	0.876	0.913	0.920	0.947	0.811	0.924	0.910	**0.949**
	PLCC	0.800	0.861	0.899	0.928	0.950	0.800	0.912	0.896	**0.951**
	RMSE	0.158	0.133	0.115	0.098	0.082	0.158	0.108	0.116	**0.081**

Table 8.2: Overall rankings of IQMs based on SROCC.

IQMs	SROCC	ranking
SDI	**0.878**	**1**
FSIM	0.852	2
GSM	0.851	3
MAD	0.843	4
MS-SSIM	0.837	5
SSIM	0.801	6
VIF	0.771	7
VSNR	0.747	8
PSNR	0.709	9

well on each database under various criteria. It ranks the highest on TID2013 and CSIQ with the SROCC scores being 0.839 and 0.943 respectively, demonstrating a high prediction monotonicity. Also, a high performance is evidenced by the highest PLCC score obtained on both TID2013 and CSIQ database. In order to provide the overall rankings (i.e., based on SROCC) of IQMs over three databases, we used the following formula:

$$Overall_{SROCC} = \beta_1 \cdot SROCC_{LIVE} + \beta_2 \cdot SROCC_{TID2013} + \beta_3 \cdot SROCC_{CSIQ} \tag{8.16}$$

where β_1, β_2 and β_3 indicate a weighting factor that is proportional to the number of distorted images in a dataset as shown in Table 2.1. In particular, β_1 equals to 0.168, β_2 equals to 0.646 and β_3 = 0.186. The overall rankings based on SROCC are shown in Table. 8.2. It shows that SDI outperforms all other counterparts.

Table 8.3: Performance comparison in terms of SROCC for individual distortion types on CSIQ dataset.

	Dis. Type	PSNR	SSIM	MS-SSIM	VIF	MAD	VSNR	FSIM	GSM	SDI
CSIQ	AGWN	0.936	0.897	0.947	**0.958**	0.951	0.924	0.926	0.944	**0.960**
	JPEG	0.888	0.955	0.963	**0.971**	0.962	0.904	0.965	0.963	**0.966**
	JP2K	0.936	0.961	0.968	0.967	**0.975**	0.948	0.968	0.965	**0.972**
	AGPN	0.934	0.892	0.933	0.951	**0.957**	0.908	0.923	0.939	**0.954**
	GB	0.929	0.961	0.971	**0.975**	0.960	0.945	**0.972**	0.959	0.969
	GCD	0.862	0.792	**0.953**	0.935	0.921	0.870	0.942	0.935	**0.943**

Table 8.4: Performance comparison in terms of SROCC for individual distortion types on TID2013 dataset.

	Dis. Type	PSNR	SSIM	MS-SSIM	VIF	MAD	VSNR	FSIM	GSM	SDI
TID2013	AGN	**0.929**	0.867	0.865	0.899	0.884	0.827	0.897	0.906	**0.916**
	ANC	**0.898**	0.773	0.773	0.830	0.802	0.731	0.821	0.818	**0.861**
	SCN	**0.919**	0.852	0.854	0.884	0.891	0.801	0.875	0.816	**0.903**
	MN	0.831	0.777	0.807	**0.845**	0.738	0.707	0.794	0.729	**0.839**
	HFN	**0.914**	0.863	0.860	0.897	0.888	0.846	0.898	0.887	**0.911**
	IN	**0.896**	0.750	0.763	0.854	0.277	0.736	0.807	0.797	**0.882**
	QN	0.878	0.866	0.871	0.785	0.851	0.836	0.872	**0.884**	**0.880**
	GB	0.914	0.967	**0.967**	0.965	0.932	0.947	0.955	**0.969**	0.953
	DEN	**0.947**	0.925	0.927	0.891	0.925	0.908	0.930	0.943	**0.951**
	JPEG	0.919	0.920	0.927	0.919	0.922	0.901	**0.932**	0.928	**0.953**
	JP2K	0.884	0.947	0.950	0.952	0.951	0.927	0.958	0.960	0.962
	JGTE	0.768	0.849	0.848	0.841	0.828	0.791	0.846	**0.851**	**0.852**
	J2TE	0.888	0.883	0.889	0.876	0.879	0.841	0.891	**0.918**	**0.910**
	NEPN	0.686	0.782	0.797	0.772	**0.832**	0.665	0.792	**0.813**	0.753
	Block	0.154	**0.572**	0.480	0.531	0.281	0.177	0.549	**0.642**	0.319
	MS	0.765	0.775	**0.791**	0.628	0.645	0.487	0.753	**0.788**	0.610
	CTC	0.441	0.378	0.463	**0.839**	0.197	0.332	0.469	**0.486**	0.180
	CCS	0.359	**0.414**	0.410	0.310	0.058	0.368	0.275	0.358	**0.831**
	MGN	**0.890**	0.780	0.779	0.847	0.841	0.764	0.847	0.835	**0.865**
	CN	0.841	0.857	0.853	0.895	0.906	0.868	**0.912**	0.912	**0.914**
	LCNI	0.914	0.906	0.907	0.920	0.944	0.882	**0.947**	**0.956**	0.924
	ICQD	0.826	0.854	0.856	0.841	0.875	0.867	0.876	**0.897**	**0.879**
	CHA	**0.887**	0.878	0.878	0.885	0.831	0.865	0.872	0.882	**0.898**
	SSR	0.904	0.946	0.848	0.935	0.957	0.934	0.957	**0.967**	**0.959**

Table 8.5: Performance comparison in terms of SROCC for individual distortion types on LIVE dataset.

	Dis. Type	PSNR	SSIM	MS-SSIM	VIF	MAD	VSNR	FSIM	GSM	SDI
LIVE	JP2K	0.895	0.961	0.963	**0.970**	0.968	0.955	**0.972**	0.970	0.948
	JPEG	0.881	0.976	0.982	**0.985**	0.976	0.966	**0.984**	0.978	0.971
	WN	0.985	0.969	0.973	**0.986**	0.984	0.979	0.972	0.977	**0.986**
	GBLUR	0.782	0.952	0.954	**0.973**	0.947	0.941	**0.971**	0.952	0.940
	FF	0.891	0.956	0.947	**0.965**	**0.957**	0.903	0.950	0.940	0.943

We further evaluated the performance of IQMs when assessing the image quality for individual distortion types. Note that similar tendencies can be obtained when either the SROCC, PLCC or RMSE is applied. Table 8.3, Table 8.4 and Table 8.5 show the SROCC scores for assessing individual distortion types in CSIQ, TID2013 and LIVE respectively. In total, there are 35 subsets (as categorized by distortion type) in all the three databases. We highlighted the IQMs that rank among the top two places in bold. As can be seen from these tables, SDI are among the top two places 23 (out of 35) times, indicating that SDI performs consistently independent of the distortion type assessed.

Table 8.6: The average processing time of each IQM per image (milliseconds per image).

IQM	PSNR	SSIM	MS-SSIM	VIF	MAD	VSNR	FSIM	GSM	SDI
time	5.3	12.4	36.0	66.3	704.8	25.5	141.8	15.1	34.7

To show the impact of parameter selection on the performance of SDI, we tuned all the five parameters around their determined values and then measured the performance of SDI as a function of the parameter values. Figure 8.10 plots the relation between all the five parameters included in the SDI and the performance of SDI on all the databases in terms of SROCC. As can be seen from the figure, the performance of SDI is insensitive to the change of parameter values. Additionally, SDI shows similar preference to the value of these parameters for different databases.

8.4.2 Computational complexity

A useful IQM should not only feature a high performance in terms of predicting subjective quality scores, but also maintain a low computational complexity in order to deal with real-

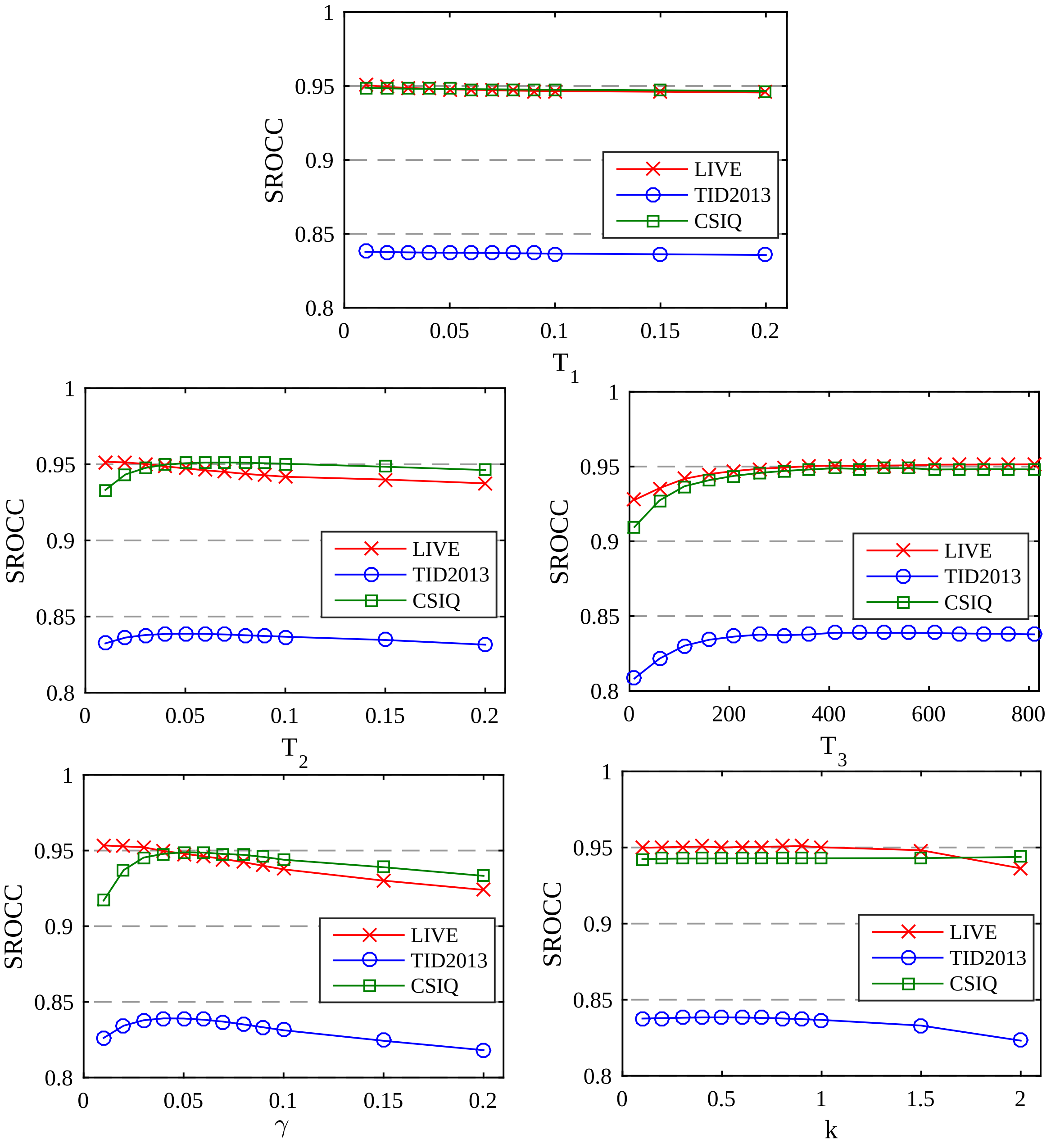


Figure 8.10: Plots of SROCC as a function of different parameters used in SDI for LIVE, TID2013 and CSIQ databases.

time applications. In this section, we tested the running time for each IQM as the proxy for their computational complexity. The test was conducted on an office PC with an Intel Core i7-4790 CPU and 32GB RAM. All the codes tested were released by the authors and implemented in Matlab. Each IQM was evaluated on TID2013 database with the resolution of each test stimulus by 512×384 pixels. The processing time per image (i.e., milliseconds/image) is listed in Table. 8.6. It shows that SDI exhibits a relatively low computational cost among all the IQMs tested.

8.5 Summary

In this chapter, we proposed an SDI metric based on measuring the saliency deviation driven by visual distortions. The modelling principle for SDI is that the saliency deviation due to visual distortion is well-correlated with the variation of image quality. Experimental results showed that the proposed SDI metric features a high performance in terms of predicting subjective quality scores. Meanwhile, the SDI metric was tested to have a low computational cost, thus can be deployed in real-time applications. It should be noted that the proposed SDI metric falls into the full-reference IQM category, which means the SDI metric needs the reference image to make quality judgement.

Chapter 9

Conclusions and Future Work

This thesis aims to optimise the application of visual saliency in image quality assessment algorithms. Previous chapters have identified existing issues in current usage of visual saliency in IQMs. A statistical evaluation was performed to clarify the effectiveness of integrating computational saliency in IQMs. An eye-tracking experiment was conducted to investigate the relation between visual saliency and image quality. Based on the empirical evidence obtained from the eye-tracking data, two perceptually-optimised saliency integration approaches were proposed. Furthermore, a new IQM based on measuring saliency deviation was devised and demonstrated to be effective. Overall, our research hypotheses are validated and the research objectives are fulfilled. In this chapter, we first summarise the main conclusions drawn in previous chapters. Then, possible future directions as well as their relation to the work in this thesis are discussed.

9.1 Conclusions

Incorporating features of the HVS in IQMs has shown its effectiveness for improving the performance of IQMs. Visual saliency, a feature closely related to how humans perceive visual information, has been recently demonstrated to have impact on image quality perception. However, due to a lack of understanding on how exactly visual saliency affects the perception of image quality, researchers usually incorporate saliency in IQMs in an *ad hoc* way, by weighting local saliency with local distortion. This method often leads to marginal or even non-existent performance gain for IQMs. To clarify our knowledge, we conducted a statistical evaluation to justify the added value of computational saliency in objective image quality assessment. Quantitative results show that the difference in predicting human fixations between saliency models is sufficient to yield a significant difference in the performance gain when adding these saliency models to IQMs. However, surprisingly, the extend to which an IQM can profit from adding a saliency model does not appear to have direct relevance to how well this saliency model can predict human fixations. In addition, we found that the added value of saliency in IQMs depends on various factors including the distortion type, IQM and saliency model.

An eye-tracking experiment with a new experimental methodology was conducted to investigate the relation between visual saliency and image quality. We found that the occurrence of distortion in an image tends to deviate fixation deployment. We quantified the extent of such deviation as a function of distortion type, degradation level and image content, respectively. In terms of the optimal use of saliency in IQMs, we investigated whether it is the saliency of the undistorted scene or that represents the same scene affected by distortion would deliver the best performance gain for IQMs. We concluded that both types of saliency are beneficial for IQMs, but the latter which reflects the interactions between saliency and distortion tends to further boost the effectiveness of visual saliency integration in IQMs. We also demonstrated that adding saliency deteriorates the performance of IQMs for assessing image contents with a large degree of saliency dispersion.

Based on the above findings, we proposed a new generic saliency integration strategy taking into account the interactions between saliency of the natural scene and the distraction power of the image distortions. We found that the proposed integration strategy consistently outperforms the conventionally used integration strategy in the literature. Also, according to the image content-dependent nature of the added value of visual saliency in IQMs, we proposed an algorithm that can provide a reliable proxy for the degree of saliency dispersion. We then used this algorithm to adaptively incorporate computational saliency in IQMs. We found that the adaptive use of saliency according to saliency dispersion can significantly improve the added value of visual saliency in IQMs.

Inspired by the psychophysical studies, we explored the plausibility of approximating image quality based on measuring the deviation of saliency induced by distortion. A large-scale eye-tracking experiment was conducted and a statistical evaluation was performed on the resulting data. We found that the extent of distortion determines the amount of saliency deviation. We also showed that it is highly plausible to approximate image quality by measuring saliency deviation with computational saliency models. We then proposed a new IQM which can quantify the saliency deviation and used it as a proxy for image quality. We showed that the proposed metric are among the best performing IQMs in the literature while exhibiting a relatively low computational cost.

9.2 Future work

Following up the work discussed in this thesis, several possible research directions could be considered in the future:

- **A saliency model for visual quality research:**

Existing computational saliency models in the literature are mainly designed for predicting the saliency of undistorted images. The performance of these saliency models are usually validated against eye-tracking data obtained from distortion-free stimuli. These models are applied for object detection, recognition, tracking and human-machine interaction etc.; and are even expected to be distortion-resistant in terms of handling the real-world distortions such as blurriness due to focus failure, image sensor noise due to low illumination and blocking artifacts due to lossy compression. However, for visual quality research, a desirable saliency model should response to distortions in consistent with human behaviour. Therefore, future work should focus on investigating a dedicated saliency model for image quality research. Chapter 8 has already identified several saliency features which are able to characterize the deviation of saliency between an original scene and its distorted version. We may create a saliency model by combining these saliency features. Moreover, the observations in Chapter 4 regarding to the interactions between saliency and distortion is useful for creating such a model. Finally, temporal/-motion information should also be taken into account in the saliency model when dealing with dynamic scenes. By doing so, the proposed saliency model can be used for video quality assessment.

- **A saliency integration approach for video quality assessment:**
  Chapter 5 and Chapter 6 present two saliency integration approaches for IQMs, in which the interactions between visual saliency aspects and visual quality are explicitly taken into account. In video quality metrics (VQMs), the relevance of visual saliency has also been confirmed [149]. Researchers have integrated saliency information to VQMs in a similar way to the conventional approach used in IQMs, i.e., by multiplying local saliency with local distortion on a frame-by-frame basis [150, 151]. Relevant studies on IQMs can be extended to VQMs, e.g., investigating whether integrating eye-tracking data obtained from distorted scenes results in a higher performance gain than integrating eye-tracking data obtained from original scenes and the necessity of taking into account the interaction between visual saliency and visual distortion for video saliency integration. Future work may also include investigating how the visual distortions alter the fixation deployment. The outcome of this investigation can be used to derive a perceptually-optimised saliency integration method for VQMs.

- **An extension of SDI to a video quality metric:**
  Chapter 8 presents the SDI metric for assessing image quality based on saliency deviation measurement. Future work may focus on extending the SDI to a VQM which can deal with dynamic scenes. To do so, one straightforward method is to apply the SDI directly to videos on a frame-by-frame basis. The final quality for the whole video sequence is then calculated as the average of the frame-level SDI scores. It, however, should be

noted that the SDI metric only makes use of spatial saliency features. Directly using the SDI metric for assessing video quality may be problematic since the temporal visual cues in video sequences, e.g., motion, can significantly affect quality perception in the HVS [152, 153, 154]. To offset this drawback, temporal saliency features should be selected and incorporated in the SDI metric.

Bibliography

[1] Z. Wang, “Applications of objective image quality assessment methods,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 137–142, 2011.

[2] S. Winkler, “Video quality measurement standards: Current status and trends,” in *Proc. of the 7th International Conference on Information, Communications and Signal Processing*, (Macau, China), pp. 848–852, 2009.

[3] Y. Chen, K. Wu, and Q. Zhang, “From qos to qoe: A tutorial on video quality assessment,” *IEEE Communications Surveys Tutorials*, vol. 17, pp. 1126–1165, Secondquarter 2015.

[4] ITU, “Methodology for the subjective assessment of the quality of television pictures, Recommendation ITU-R BT-500.13,” Jan. 2012.

[5] Z. Wang and A. C. Bovik, *Modern image quality assessment*. San Rafael, CA: Morgan & Claypool, 2006.

[6] D. M. Chandler, “Seven challenges in image quality assessment: past, present, and future research,” *ISRN Signal Processing*, vol. 2013, 2013.

[7] D. Ghadiyaram and A. C. Bovik, “Massive online crowdsourced study of subjective and objective picture quality,” *IEEE Transactions on Image Processing*, vol. 25, pp. 372–387, Jan 2016.

[8] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, “Group MAD competition - a new methodology to compare objective image quality models,” in *Proc. of the IEEE Conferece on Computer Vision and Pattern Recognition*, 2016.

[9] S. J. Daly, “Application of a noise-adaptive contrast sensitivity function to image data compression,” *Optical Engineering*, vol. 29, no. 8, pp. 977–987, 1990.

[10] G. E. Legge and J. M. Foley, “Contrast masking in human vision,” *Journal of the Optical Society of America*, vol. 70, pp. 1458–1471, Dec 1980.

[11] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Processing Magazine*, vol. 28, pp. 50–59, Nov. 2011.

[12] E. Vu and D. M. Chandler, "Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience," in *Proc. of the IEEE Southwest Symposium Image Analysis and Interpretation*, pp. 73–76, 2008.

[13] J. Redi, H. Liu, R. Zunino, and I. Heynderickx, "Interactions of visual attention and quality perception," in *Proc. of SPIE, Human Vision and Electronic Imaging*, (San Francisco, USA), pp. 78650S–78650S–11, Jan. 2011.

[14] X. Min, G. Zhai, Z. Gao, and C. Hu, "Influence of compression artifacts on visual attention," in *Proc. of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2014.

[15] F. Röhrbein, P. Goddard, M. Schneider, G. James, and K. Guo, "How does image noise affect actual and predicted human gaze allocation in assessing image quality?," *Vision Research*, vol. 112, pp. 11 – 25, 2015.

[16] L. C. H.R. Sheikh, Z.Wang and A. Bovik, "LIVE image quality assessment database release 2."

[17] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006–011006–21, 2010.

[18] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57 – 77, 2015.

[19] P. Le Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database," 2005.

[20] Z. M. P. Sazzad, Y. kawayoke, and Y. Horita, "Mict image quality evaluation database," 2000.

[21] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, pp. 616–625, Oct 2012.

[22] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, pp. 98–117, Jan 2009.

[23] A. B. Watson, *Digital images and human vision.* Cambridge, MA: The MIT Press, 1997.

[24] B. A. Wandell, *Foundations of vision.* Sunderland, MA, US: Sinauer Associates, 1995.

[25] W. S. Geisler and M. S. Banks, "Visual performance," in *Handbook of Optics*, New York: McGraw-Hill, 1995.

[26] H. R. Blackwell, "Contrast thresholds of the human eye," *Journal of the Optical Society of America*, vol. 36, pp. 624–643, Nov 1946.

[27] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, pp. 2284–2298, Sep. 2007.

[28] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011006–011006, 2010.

[29] 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, pp. 636–650, Apr. 2000.

[30] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, pp. 81–84, Mar. 2002.

[31] 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, pp. 600–612, Apr. 2004.

[32] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. of the 37th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1398–1402, Nov. 2003.

[33] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, pp. 1185–1198, May 2011.

[34] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, pp. 430–444, Feb. 2006.

[35] L. Zhang, D. Zhang, and X. Mou, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, pp. 2378–2386, Aug. 2011.

[36] H. Wu and M. Yuen, “A generalized block-edge impairment metric for video coding,” *IEEE Signal Processing Letters*, vol. 4, pp. 317–320, Nov. 1997.

[37] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, “A no-reference perceptual blur metric,” in *Proc. of the 9th IEEE International Conference on Image Processing*, (Rochester, USA), pp. 57–60, Sep. 2002.

[38] R. Ferzli and L. J. Karam, “A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB),” *IEEE Transactions on Image Processing*, vol. 18, pp. 717–728, Apr. 2009.

[39] R. Muijs and I. Kirenko, “A no-reference blocking artifact measure for adaptive video processing,” in *Proc. of 13th European Signal Processing Conference*, (Antalya, TR), Sep. 2005.

[40] Y. Fang, W. Lin, and S. Winkler, “Review of existing objective qoe methodologies,” *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*, vol. 29, 2015.

[41] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, “Objective video quality assessment methods: A classification, review, and performance comparison,” *IEEE Trans. on Broadcast.*, vol. 57, pp. 165–182, June 2011.

[42] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. B. II, V. Balasubramanian, and P. Sterling, “How much the eye tells the brain,” *Current Biology*, vol. 16, no. 14, pp. 1428 – 1434, 2006.

[43] C. Koch and S. Ullman, *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry*, pp. 115–141. Dordrecht: Springer Netherlands, 1987.

[44] A. Borji, D. N. Sihite, and L. Itti, “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study,” *IEEE Transactions on Image Processing*, vol. 22, pp. 55–69, Jan. 2013.

[45] B. Fischer and B. Breitmeyer, “Mechanisms of visual attention revealed by saccadic eye movements,” *Neuropsychologia*, vol. 25, no. 1, pp. 73–83, 1987.

[46] J. E. Hoffman and B. Subramaniam, “The role of visual attention in saccadic eye movements,” *Perception & psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[47] K. Rayner, “Eye movements and attention in reading, scene perception, and visual search,” *The quarterly journal of experimental psychology*, vol. 62, no. 8, pp. 1457–1506, 2009.

[48] P. S. Holzman, L. R. Proctor, and D. W. Hughes, "Eye-tracking patterns in schizophrenia," *Science*, vol. 181, no. 4095, pp. 179–181, 1973.

[49] P. L. Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proceedings of the IEEE*, vol. 101, pp. 2058–2067, Sept 2013.

[50] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 4, pp. 455–470, 2002.

[51] D. A. Slykhuis, E. N. Wiebe, and L. A. Annetta, "Eye-tracking students' attention to powerpoint photographs in a science education setting," *Journal of Science Education and Technology*, vol. 14, no. 5, pp. 509–520, 2005.

[52] R. Jacob and K. S. Karn, "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises," *Mind*, vol. 2, no. 3, p. 4, 2003.

[53] M. Wedel and R. Pieters, *A Review of Eye-Tracking Research in Marketing*, pp. 123–147. 2008.

[54] P. M. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi, "Real-time eye gaze tracking for gaming design and consumer electronics systems," *IEEE Transactions on Consumer Electronics*, vol. 58, pp. 347–355, May 2012.

[55] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, (New York, NY, USA), pp. 71–78, ACM, 2000.

[56] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception & Psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[57] S. Winkler and R. Subramanian, "Overview of eye tracking datasets," in *Proc. of the 5th International Workshop on Quality of Multimedia Experience*, pp. 212–217, July 2013.

[58] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2049–2056, 2006.

[59] D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, vol. 100, no. 1âĂŞ2, pp. 41 – 63, 2005.

[60] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proc. of the 16th International Joint Conference on Artificial Intelligence*, IJCAI '99, (San Francisco, CA, USA), pp. 1146–1153, Morgan Kaufmann Publishers Inc., 1999.

[61] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The jpeg2000 still image coding system: an overview," *IEEE Transactions on Consumer Electronics*, vol. 46, pp. 1103–1127, Nov 2000.

[62] Y. Fang, J. Wang, Y. Yuan, J. Lei, W. Lin, and P. L. Callet, "Saliency-based stereoscopic image retargeting," *Information Sciences*, vol. 372, pp. 347 – 358, 2016.

[63] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97 – 136, 1980.

[64] 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 Intellegence*, vol. 20, pp. 1254–1259, Nov. 1998.

[65] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[66] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, p. 5, 2009.

[67] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.

[68] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, p. 11, 2013.

[69] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. of the 22nd Conference on Advances in Neural Information Processing Systems*, (Vancouver,CA), pp. 681–688, Dec. 2008.

[70] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. of the 20th Conference on Advances in Neural Information Processing Systems*, (Vancouver, CA), pp. 545–552, Dec. 2006.

[71] A. Torralba, "Modeling global scene factors in attention," *Journal of the Optical Society of America*, vol. 20, no. 7, pp. 1407–1418, 2003.

[72] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. of the 20th IEEE Conference on Computer Vision and Pattern Recognition*, (Minneapolis, MN), pp. 1–8, Jun. 2007.

[73] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, pp. 185–198, Jan. 2010.

[74] P. L. Rosin, "A simple method for detecting salient regions," *Pattern Recognit.*, vol. 42, no. 11, pp. 2363–2371, 2009.

[75] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[76] M. Holtzman-Gazit, L. Zelnik-Manor, and I. Yavneh, "Salient edges: A multi scale approach," in *Proc. of the 11th European Conference on Computer Vision*, (Crete, Greece), Sep. 2010.

[77] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, p. 15, 2009.

[78] J. Li, M. D. Levine, X. An, H. He, *et al.*, "Saliency detection based on frequency and spatial domain analyses," in *Proc. of 22th British Machine Vision Conference*, (Dundee, UK), Sep. 2011.

[79] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Transactions on Image Processing*, vol. 21, pp. 3888–3901, Sep. 2012.

[80] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proc. of 22nd British Machine Vission Conference*, (Dundee, UK), p. 7, Sep. 2011.

[81] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. of the 19th IEEE Conference on Computer Vision and Pattern Recognition*, (Ghaziabad), pp. 1597–1604, Feb. 2009.

[82] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 353–367, Feb. 2011.

[83] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. of the 14th IEEE International Conference on Computer Vision*, (Barcelona, SP), pp. 914–921, Nov. 2011.

[84] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. Oct., pp. 1915–1926, 2012.

[85] N. Riche and M. Mancas, *Bottom-Up Saliency Models for Videos: A Practical Review*, pp. 177–190. New York, NY: Springer New York, 2016.

[86] A. Borji and L. Itti, “State-of-the-Art in visual attention modeling,” *IEEE Transactions on Pattern Analysis and Machine Intellegence*, vol. 35, pp. 185–207, Jan 2013.

[87] H. Alers, J. Redi, H. Liu, and I. Heynderickx, “Studying the effect of optimizing image quality in salient regions at the expense of background content,” *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 043012–043012, 2013.

[88] E. C. Larson, C. Vu, and D. M. Chandler, “Can visual fixation patterns improve image fidelity assessment?,” in *Proc. of the 15th IEEE International Conference on Image Processing*, pp. 2572–2575, Oct 2008.

[89] E. C. Larson and D. M. Chandler, “Unveiling relationships between regions of interest and image fidelity metrics,” in *Proc. of SPIE*, vol. 6822, pp. 68222A–68222A–16, 2008.

[90] 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, pp. 971–982, Jul. 2011.

[91] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, “Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric,” in *Proc. of the 14th IEEE International Conference on Image Processing*, vol. 2, pp. II – 169–II – 172, Sept 2007.

[92] H. Liu, U. Engelke, J. Wang, P. Le Callet, I. Heynderickx, *et al.*, “How does image content affect the added value of visual attention in objective image quality assessment?,” *IEEE Signal Processing Letter*, vol. 20, Apr. 2013.

[93] N. Sadaka, L. Karam, R. Ferzli, and G. Abousleman, “A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling,” in *Proc. of the 15th IEEE Int. Conf. Image Process.*, (San Diego, CA), pp. 369–372, Oct. 2008.

[94] A. K. Moorthy and A. C. Bovik, “Visual importance pooling for image quality assessment,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, pp. 193–201, April 2009.

[95] U. Rajashekar, I. Van Der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Transactions on Image Processing*, vol. 17, pp. 564–573, Apr. 2008.

[96] R. Barland and A. Saadane, "Blind quality metric using a perceptual importance map for jpeg-20000 compressed images," in *Proc. of the 13th IEEE International Conference on Image Processing*, (Atlanta, GA), pp. 2941–2944, Oct. 2006.

[97] W. M. Osberger and A. M. Rohaly, "Automatic detection of regions of interest in complex video sequences," in *Proc. of SPIE*, vol. 4299, pp. 361–372, 2001.

[98] Q. Ma and L. Zhang, "Image quality assessment with visual attention," in *Proc. of the 15th International Conference on Pattern Recognition*, (Tampa, FL), pp. 1–4, Dec. 2008.

[99] M. C. Q. Farias and W. Y. L. Akamine, "On performance of image quality metrics enhanced with visual attention computational models," *Electronics Letters*, vol. 48, pp. 631–633, May 2012.

[100] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Transactions on Image Processing*, vol. 14, pp. 1928–1942, Nov 2005.

[101] U. Engelke and H.-J. Zepernick, "Framework for optimal region of interest based quality assessment in wireless imaging," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011005–011005–13, 2010.

[102] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 23, pp. 4270–4281, Oct 2014.

[103] H. Alers, J. Redi, H. Liu, and I. Heynderickx, "Effects of task and image properties on visual-attention deployment in image-quality assessment," *Journal of Electronic Imaging*, vol. 24, no. 2, p. 023030, 2015.

[104] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency based objective quality assessment of decoded video affected by packet losses," in *Proc. of the 15th IEEE International Conference on Image Processing*, pp. 2560–2563, Oct 2008.

[105] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II (FR_TV2)," tech. rep., Video Quality Experts Group, 2003.

[106] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1266–1278, 2016.

[107] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397 – 2416, 2005.

[108] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643 – 659, 2005.

[109] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proc. of the IEEE International Conference on Computer Vision*, December 2013.

[110] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H.-J. Zepernick, and A. Maeder, "Comparative study of fixation density maps," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1121–1133, 2013.

[111] S. Winkler, *Vision models and quality metrics for image processing applications*. PhD thesis, Univ. of Lausanne, Switzerland, 2000.

[112] 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, pp. 3440–3451, Nov 2006.

[113] D. C. Montgomery, *Applied Statistics and Probability for Engineers 6th edition*. Wiley, 2013.

[114] L. Ma, J. Tian, and W. Yu, "Visual saliency detection in image using ant colony optimisation and local phase coherence," *Electronic Letters*, vol. 46, no. 15, pp. 1066–1068, 2010.

[115] A. Borji, M. M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, pp. 5706–5722, Dec 2015.

[116] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic, "Eye-tracking database for a set of standard video sequences," *IEEE Transactions on Image Processing*, vol. 21, pp. 898–903, Feb 2012.

[117] O. L. Meur, A. Ninassi, P. L. 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, pp. 547 – 558, 2010.

[118] D. S. Wooding, “Eye movements of large populations: II. deriving regions of interest, coverage, and similarity using fixation maps,” *Behavior Research Methods, Instruments & Computers*, vol. 34, no. 4, pp. 518–28, 2002.

[119] M. Mancas and O. L. Meur, “Memorability of natural scenes: The role of attention,” in *Proc. of the IEEE International Conference on Image Processing*, pp. 196 – 200, 2013.

[120] A. G. Greenwald, “Within-subjects designs: To use or not to use?,” *Psychological Bulletin*, vol. 83, no. 2, p. 314, 1976.

[121] G. Keren, “Between-or within-subjects design: A methodological dilemma,” *A Handbook for Data Analysis in the Behaviorial Sciences*, p. 257, 2014.

[122] D. D. Salvucci and J. H. Goldberg, “Identifying fixations and saccades in eye-tracking protocols,” in *Proc. of the 2000 Symposium on Eye Tracking Research & Applications*, (Florida, USA), pp. 71–78, 2000.

[123] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, “Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.,” *Psychological review*, vol. 113, no. 4, p. 766, 2006.

[124] T. Judd, F. Durand, and A. Torralba, “Fixations on low-resolution images,” *Journal of Vision*, vol. 11, no. 4, pp. 14–14, 2011.

[125] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, “Saliency and human fixations: state-of-the-art and study of comparison metrics,” in *Proc. of the IEEE International Conference on Computer Vision*, pp. 1153–1160, 2013.

[126] Z. Wang and X. Shang, “Spatial pooling strategies for perceptual image quality assessment,” in *Proc. of the 13th IEEE International Conference on Image Processing*, pp. 2945–2948, IEEE, 2006.

[127] L. K. Chan and W. G. Hayward, “Dimension-specific signal modulation in visual search: evidence from inter-stimulus surround suppression,” *Journal of vision*, vol. 12, no. 4, p. 10, 2012.

[128] O. Le Meur, T. Baccino, and A. Roumy, “Prediction of the inter-observer visual congruency (iovc) and application to image ranking,” in *Proc. of the 19th ACM International Conference on Multimedia*, pp. 373–382, 2011.

[129] S. Rahman and N. D. B. Bruce, “Factors underlying inter-observer agreement in gaze patterns: Predictive modelling and analysis,” in *Proc. of the 9th Biennial ACM Symposium on Eye Tracking Research & Applications*, pp. 155–162, 2016.

[130] R. M. Gray, *Entropy and Information Theory*. New York: Springer, 1990.

[131] M. R. Sabuncu, *Entropy-based image registration*. PhD thesis, Princeton University, USA, 2006.

[132] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, “MIT saliency benchmark.” http://saliency.mit.edu/.

[133] G. Kootstra and L. Schomaker, “Prediction of human eye fixations using symmetry,” *Proceedings of CogSci*, 2009.

[134] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *Proc. of the 12th IEEE International Conference on Computer Vision*, pp. 2106–2113, Sept 2009.

[135] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, *et al.*, “Color image database TID2013: Peculiarities and preliminary results,” in *Proc. of the 4th Eur.Workshop Vis. Inf. Process.*, pp. 106–111, Jun. 2013.

[136] M. Mancas, C. Mancas-Thillou, B. Gosselin, B. M. Macq, *et al.*, “A rarity-based visual attention map-application to texture description.,” in *Proc. of the 13th IEEE International Conference on Image Processing*, pp. 445–448, 2006.

[137] A. Borji and L. Itti, “Exploiting local and global patch rarities for saliency detection,” in *Proc. of the 22th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 478–485, IEEE, 2012.

[138] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, “Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis,” *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013.

[139] X. Hou, J. Harel, and C. Koch, “Image signature: Highlighting sparse salient regions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[140] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[141] A. Oliva and P. G. Schyns, “Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli,” *Cognitive psychology*, vol. 34, pp. 72–107, 1997.

[142] S. Winkler, M. Kunt, and C. J. van den Branden Lambrecht, *Vision and Video: Models and Applications*, pp. 201–229. Boston, MA: Springer US, 2001.

[143] M. Rabbani and P. W. Jones, *Digital image compression techniques*, vol. 7. SPIE Press, 1991.

[144] L. M. Hurvich and D. Jameson, “An opponent-process theory of color vision.,” *Psychological review*, vol. 64, no. 6p1, p. 384, 1957.

[145] K. van de Sande, T. Gevers, and C. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1582–1596, Sept 2010.

[146] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, vol. 50, pp. 312–322, Sept 2004.

[147] Z. Wang and X. Shang, “Spatial pooling strategies for perceptual image quality assessment,” in *Proc. of the IEEE International Conference on Image Processing*, pp. 2945–2948, Oct 2006.

[148] A. Liu, W. Lin, and M. Narwaria, “Image quality assessment based on gradient similarity,” *IEEE Transactions on Image Processing*, vol. 21, pp. 1500–1512, April 2012.

[149] H. Alers, J. A. Redi, and I. Heynderickx, “Quantifying the importance of preserving video quality in visually important regions at the expense of background content,” *Signal Processing: Image Communication*, vol. 32, pp. 69 – 80, 2015.

[150] X. Feng, T. Liu, D. Yang, and Y. Wang, “Saliency inspired full-reference quality metrics for packet-loss-impaired video,” *IEEE Transactions on Broadcasting*, vol. 57, pp. 81–88, March 2011.

[151] W. Y. L. Akamine and M. C. Q. Farias, “Video quality assessment using visual attention computational models,” *Journal of Electronic Imaging*, vol. 23, no. 6, p. 061107, 2014.

[152] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, “Considering temporal variations of spatial visual distortions in video quality assessment,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, pp. 253–265, April 2009.

[153] 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, pp. 266–279, April 2009.

[154] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE Transactions on Image Processing*, vol. 19, pp. 335–350, Feb 2010.