The Dissertation Committee for Zeina Sinno
certifies that this is the approved version of the following dissertation:

Statistical and Perceptual Properties of Images and Videos with Applications

Committee:

Alan C. Bovik, Supervisor

Joydeep Ghosh

Constantine Caramanis

Jonathan Valvano

Wilson S. Geisler

Statistical and Perceptual Properties of Images and Videos with Applications

by

Zeina Sinno

DISSERTATION

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN
May 2019

Dedicated to mom & dad.

Acknowledgments

Pursuing PhD has been an instructive journey academically, professionally and personally. I would like to thank some of the people that have supported me all along this time and without whom this work would not have been possible.

I would like to start by expressing my heartiest gratitude to my supervisor Dr. Bovik. Early on, his exceptional teaching style has inspired me to pursue my research in this area. His enthusiasm, guidance, dedication, and support all along this journey have shaped me to become a better researcher. His patience and confidence in my abilities gave me a lot of motivation to continue pushing to achieve my goals. The long discussions we had, detailed feedback, and the ease of communication with him have enriched my research, and helped me grow academically and professionally. His kindness, being a listening ear when I needed it, and getting advice from him about countless topics in life have helped me so much on a personal level. I admire a lot his deep knowledge in many areas which allowed me to receive the best advice. I am also very thankful to his positive energy that has brightened many of my days. I am forever grateful for the opportunity of knowing such a nice person, working with him, and for having him as a **SUPER**visor.

I also want to thank my committee members Dr. Geisler, Dr. Ghosh,

Dr. Valvano, and Dr. Caramanis for the help I received from them in my research, their comments, as well their diverse classes which broadened my knowledge. I enjoyed a lot interacting with Dr. Valvano over the years and learning about the latest embedded-systems gadgets in the making. I find it inspirational and amusing that he brings the concepts that he teaches to real-life by building very cool systems such as his own musical light show, his own Halloween costume with a micro-controller and a bunch of LEDs... Besides the awesome faculty members I interacted with, I am very thankful for the support I received from the ECE and WNCG staff: Melody, Melanie, Barry, Karen, Jaymie, and Apipol. All of them were very kind to me when I had any questions or needed any help. Navigating administrative tasks can be complex and confusing. Their hard-work, dedication and knowledge made such tasks way easier.

I would also like to thank the past and present members of LIVE for their continuous support and all the fun moments we had together: Lark, Deepti, Leo, Christos, Todd, Janice, Praful, Jerry, Meixu, Yize, Sahar, Dae, Pavan, Haoran, Somdyupti, Sungsoo, Li-Heng, Zhenqiang, and Zhengzhong. Working in such a diverse lab was an enriching personal experience as it broadened my perspective. I enjoyed learning about different cultures and trying out food coming from across the world during LIVE outings and potlucks. I am very thankful to the friendships I made in the lab. A very special thank you to Lark and his upbeat attitude and always smiling, the lab foodie Janice who was very supportive and with whom I enjoyed discovering a lot of places

and tastes around Austin, Deepti who felt like the older wise sister and her many attempts to make me acquire tolerance for spicy food, Meixu who felt like a little sister and with whom I had deep conversations, Praful who was a great source of support and encouragement, Christos and the many laughs we had together, Todd and his cool character and for being the best travel buddy during conferences, Leo for being a very caring friend always here to help, Sahar for enjoying conversations with her, and finally Jerry who was a very sweet friend. I am also grateful to the support I got from Sabine and Sara, who were part of this journey since day 1. I am glad to have an eclectic support network in town.

I would like to dedicate this dissertation to my beloved parents, Hala and Zaki for their infinite sacrifices and for their support in my educational pursuit. Their endless love and encouragement have inspired me to become who I am today. I also would like to convey my appreciation to my lovely siblings, Farah and Mounir, and my brother-in-law Maher for their affection, unfailing support, and big faith in me. And last but not least, I would like to thank my nephew Zein, born halfway throught this journey, and who brought so much happiness since then. The large amount of love and curiosity this little boy carries always amazes me.

Statistical and Perceptual Properties of Images and Videos with Applications

Publication No. _____

Zeina Sinno, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Alan C. Bovik

Abstract

The visual brain is optimally designed to process images from the natural environment that we perceive. Describing the natural environment statistically helps in understanding how the brain encodes those images efficiently. The Natural Scene Statistics (NSS) of the luminance component of images is the basis of several univariate statistical models. Such models were the fundamental building blocks of multiple visual applications, ranging from the design of faithful image and video quality models to the development of perceptually optimized image enhancing techniques. Towards advancing this area, I studied the bivariate statistical properties of images and developed the first of its kind closed-form model that describes the correlation of spatially separated bandpass image samples. I found that the model was useful in tackling different problems such as blindly assessing the quality of images and assessing 3D visual discomfort of stereo images.

Provided the success of NSS in tackling image processing problems, I decided to use them as a tool to tackle the blind video quality assessment (VQA) problem. First, I constructed a video quality database, the LIVE Video Quality Challenge Database (LIVE-VQC). This database is the largest across several key dimensions: number of unique contents, distortions, devices, resolutions, and videographers. For collecting the subjective scores, I constructed a new framework in Amazon Mechanical Turk. A massive number of subjects from across the globe participated in my study. Those efforts resulted in a VQA database that serves as a great benchmark for real-world videos. Next, I studied the spatio-temporal statistics of a wide variety of natural videos and created a space-time completely blind VQA model that deploys a directional temporal NSS model to predict quality. My newly created model outperforms all previous completely blind VQA models on the LIVE-VQC.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Natural Scene Statistics (NSS) models are useful probes of the visual brain, and of how it has evolved to efficiently process gigantic amounts of visual data [1]. Previous work on NSS based image models has focused primarily on characterizing the univariate bandpass statistics of single pixels. The parameters of univariate NSS models samples of bandpass images have been used as fundamental low-level picture descriptors to successfully solve image and video processing and analysis tasks such as image interpolation [2], texture modeling [3, 4], full reference and blind image quality prediction [5–9], and color depth modeling [10]. Extending NSS models to characterize the bivariate behavior of images could help advance improved solutions to a wide variety of applications. However to date, little effort has been applied towards modeling the bivariate NSS of bandpass image samples. Towards addressing this problem, I developed a closed form bivariate spatial correlation model of bandpass and normalized image samples that completes an existing two-dimensional joint generalized gaussian distribution model [11] of adjacent bandpass pixels. I also studied the behavior of the model in presence of distortions and I was able to demonstrate that the parameters of my model vary systematically as a function of the type and the amount of distortions introduced to an image.

This observation was the ground for building a distortion classification tool, and a no-reference image quality predictor that outperformed the performance of other state-of-the-art models. I took this model one step further by studying the statistics of 3D images, and I was able to see that my model holds in this case too. Then, I created a no-reference 3D visual discomfort predictor based on the parameters of my model that outperformed all the other perception-based and deep neural network-based predictors.

Given the success of NSS to tackle multiple image quality related problems, I decided to NSS as a tool to tackle the blind video quality assesment (VQA) problem. Digital video has become ubiquitous and now accounts for the largest portion of Internet traffic. By 2021, it is expected that 82% of all transmitted bits will contain video content [12]. The number of videos that are being streamed is skyrocketing, with much of the traffic being driven by "cord-cutter" streaming video services, such as Netflix, Hulu, and Amazon Prime Video. However, the number of videos that are being captured and shared by casual users is also dramatically growing. The significantly improved quality of cameras that are found in smartphones is one of the important factors driving this popularity, along with a wide range of camera-related apps that facilitate the sharing, editing or aesthetic modification of images and videos. Sharing "in-the-moment" experiences in the form of video has become quite popular using applications such as Instagram, Facebook, Twitter via Periscope, Snapchat, and so on. Online videos have also revolutionized modern journalism as they enable online news stories to unfold live, and allow the viewing

audience to comment on or otherwise interact with it. Over the past year, Facebook alone generated more than 100 million of video watch hours each day [13]. On YouTube, the overall durations of the videos that are uploaded daily exceeds 65 years, and more than 1 billion hours of their video content is watched each day [14]. These numbers are continuing to rise and to reshape digital marketing, entertainment, journalism, and amateur videography. The volume of streaming video viewed online has become so large that more than 139 million people are now Netflix subscribers [15]. Streaming videos now comprise the majority of Internet traffic today. It is no surprise that videos account for the largest portion of Internet traffic, which is expected to eclipse 82% of all transmitted bits by 2021 [12]. These videos are captured using a very wide variety of camera devices by users having very diverse goals and expertise, sometimes under difficult lighting conditions. This, coupled with the many currently available camera devices built with different characteristics leads to complex and often commingled distortions that degrade the quality of the videos. These videos are captured and often uploaded to cloud services such as iCloud and Google Photos and might be shared on platforms such as YouTube, Instagram, and Facebook. Being able to predict the quality of these videos is an important goal for a variety of invested practitioners, such as camera designers, cloud engineers, and users who could be directed to recapture videos of poor quality. In nearly every instance, a high-quality reference video is not available, hence no-reference video quality predictors are of the greatest interest. Current no-reference video quality models are unable to handle the

diversity of distortions. This is true in part because available VQA databases which are benchmarks containing videos and associated quality scores present very limited content of fixed resolutions, captured using a small number of camera devices by only a few videographers. This content is then subjected to only a modest number of synthetic distortions. As such, these databases fail to adequately represent real world videos, which contain very different kinds of content obtained under various imaging conditions and are subject to authentic, complex and often commingled distortions that are difficult or impossible to simulate.

Towards advancing this area, I constructed a real-world, "in the wild", VQA database which consists of 585 videos, sourced from 80 different inexpert videographers, and captured using 101 unique devices. The new database represents unprecedented degrees of realism, data authenticity, and relevance. To collect the quality scores, I constructed a new framework in Amazon Mechanical Turk to crowdsource the data from participants from across the world. About 5000 participants, from 45 different countries took part of the subjective study resulting in over 200000 opinion scores. The significant diversity of the subject pool raised many technical challenges owing to widely differing viewing conditions and resources. However, I demonstrated that the framework I built is robust against the many variables affecting the video rating process. This effort resulted in a VQA database that is the largest along several key dimensions: number of unique contents, capture devices, distortion types and combinations of distortions, study participants, and recorded

subjective scores.

Next, I studied the spatio-temporal statistics of a wide variety of natural videos and I created a space-time blind VQA model that deploys directional temporal statistic models. The constituent statistical features of the new predictor show excellent consistency between its perceived quality. It also utilizes a different temporal pooling strategy. Instead of creating a trained model, I devised a completely blind video quality predictor that outperforms other existing completely blind video quality models on the largest available subject quality dataset of authentically distorted videos. The new model is simple and computationally efficient as compared to other models.

The remainder of the dissertation is organized as follows; in Chapter 2, in the background section, I review relevant work in the literature to my bivariate natural scene statistics model as well as to the blind VQA problem. In Chapter 3, I summarize the contributions of my dissertation. In Chapter 4, I present the details of the closed-form bivariate NSS model and study the model in the context of image distortions. Then, in Chapter 5, I breifly summarize two applications of the bivariate NSS model: blind image quality assesment and 3D visual discomfort prediction. In Chapter 6, I present a new database that I built for tackling the blind VQA problem, the LIVE-VQC database as well as a new framework for scaling up the collection of video quality scores. And in Chapter 7, a new completely blind VQA model based on spatio-temporal naturalness of videos is presented. The dissertation concludes with future thoughts about the VQA problem in Chapter 8.

Chapter 2

Background

I begin this chapter by describing relevant background to the second order statistics modeling of natural scenes. Then, I overview relevant background for the blind VQA problem related to the construction of VQA databases and models that can predict quality closely to humans' judgment.

2.1 Background Related to Second Order Natural Scene Statistics Modeling

In this section, I describe existing previous relevant models of the second order statistics of natural scenes, followed by a few concepts necessary to understand my closed-form model.

2.1.1 Second Order Natural Scene Statistics Models

Early on, Simoncelli [16] and Liu *et al.* [17] observed that the coefficients of orthonormal wavelet (i.e; bandpass) decompositions of natural images tend to be much less spatially correlated than the source images, yet they exhibit strong intra and inter scale dependencies between bands [18]. These observations formed the basis of an image texture model [3], where a set of parametric constraints imposed on pairs of complex wavelet coefficients occu-

pying adjacent spatial locations, orientations and scales are used to represent and synthesize textures. Po *et al.* [19] developed a natural image model using a hidden Markov tree, a Gaussian mixture model and two dimensional contourlet features that capture interlocation, interscale and interdirection dependencies. Mumford *et al.* [20] proposed an infinitely divisible statistical bandpass image model that assumes natural segmentations of images into high-information objects, cast against an ergodic field of low-information regions. However, their model does not capture the two-dimensional dependencies that occur within bandpass images. Lee *et al.* [21] found that the power law dominates the short spatial covariance function of pairs of bandpass image samples obeying a reciprocal power law over short distances.

Prior efforts on the bivariate NSS have not produced closed form representations. The first attempt to do so was reported in Su *et al.* [11]; but their model was incomplete. The authors found it to be useful for tasks such as color depth and range modeling [22] and stereopair quality evaluation [23]. In [24] and [25], I extended their work by studying the bivariate distributions of the responses of horizontally related, oriented bandpass image samples separated by distance of up to 10 pixels. In this dissertation, I generalize my findings even further by diversifying the model across spatial orientations and by extending the studied distances to 25 pixels and more (up to to 35 pixels for some of the spatial orientations). I demonstrate that for any image, the bivariate NSS model correlation can be expressed using 6 parameters, per spatial orientation. I also study the bivariate NSS of distorted images and I find

that my model is capable of representing the correlations between distorted image samples. The observed changes in the bivariate NSS model parameters when distortions are introduced are found to be systematic, suggesting their usefulness in image distortion analysis and future image quality models.

2.1.2 *1/f* Processes

In order to understand better my model, it is necessary to also review the concept of *1/f* processes. It is well known that the power spectra of natural photographic images tend to follow a reciprocal power law [26]:

$$S(f) \propto \frac{k}{f^{\alpha}}, \tag{2.1}$$

where $\alpha > 0$ determines the rate of spectral fall-off of the process.

Other phenomena that can be described by this law include the extreme case of white processes ($\alpha = 0$) which exhibit no correlation over time or space, and random walks (e.g., Brownian motion where $\alpha = 2$, which is the integral of white noise). Johnson [27] first observed a so called "*1/f*" phenomenon while studying shot noise in vacuum tubes. Processes that can be accurately described as "*1/f*" arise in such widely-varying disciplines as biological evolution [28], animal population studies [29], economics [30], personal growth and development [31], and musical loudness and pitch [32], among many others. The wide range of occurrences of the *1/f* phenomenon may be attributed to deep natural laws that reflect the self-similarities of certain signal measurements over scales and the behavior of equilibrium systems. Formal

mathematical frameworks such as fractional Brownian motion models [33,34], fractals [35], and iterated function systems [36] have been deeply developed, yet the physical origins of *1/f* phenomena are often poorly understood. For example, although images of natural scenes are enormously diverse, their power spectra can be reliably described as *1/f* [1,26,37], reflecting statistical regularities underlying their correlation structure, yet the origin of this behavior is not known.

Here I am primarily interested in the *1/f* image model in regards to its implications regarding the correlation structure of bandpass natural images. My interest in this topic is motivated by the successes that have been obtained on perception-driven image analysis problems using spatial NSS models, and which might be furthered by expanding these models. This may also lead to insights on how natural correlations may drive spatial interactions between visual cortical neurons [1,38–40]. Keshner [41] derived models of the stationary autocorrelation functions of one-dimensional *1/f* processes, arriving at a power law of reciprocal separation. In the following, I develop a similar expression for the peak correlation between bandpass image samples, using a stabilized reciprocal power law.

2.2 Background Related to VQA Databases

It is necessary that VQA algorithms be trained and/or tested on extensive subjective video quality data sets so that it maybe asserted that they reflect or are capable of closely replicating human judgments. As a result, over

the past decade numerous researchers have designed and built VQA databases.

2.2.1 Conventional Laboratory VQA Databases

The LIVE VQA Database [42] contains 10 pristine high-quality videos subjected to 4 distortion types: MPEG-2 compression, H.264 compression, H.264 bitstreams suffering from IP, and wireless packet losses. The resource in [43] offers 156 video streams suffering from H.264/AVC artifacts and wireless packet losses. The LIVE QoE Database for HTTP-based Video Streaming [44], studies the quality of experience of users under simulated varying channel induced distortions, and the LIVE Mobile Video Quality Database [45] includes channel induced distortions and dynamically varying distortions, such as varying compression rates. More recent databases include the TUM databases [46, 47], which target H.264/AVC distortions on a few contents (4 and 8); and the MCL-V [48] database consists of 12 video source clips and 96 distorted videos, targeting distortions related to streaming (compression, and compression followed by scaling). The MCL video quality database contains 200 raw sequences targeting compression artifacts [49]. Most available video quality databases were conducted under highly-controlled laboratory conditions by introducing sets of graded simulated impairments onto high-quality videos. Given questions that arise regarding the realism and accuracy of representation of synthetic distortions, researchers have also conducted studies on the quality perception of authentic, real-world distortions such as distortions that occur during video capture [50, 51].

2.2.2 Crowdsourced VQA Databases

Crowdsourcing is a portmanteau of the words crowd and outsourcing. The term was first used in 2006 to describe the transfer of certain kinds of tasks from professionals to the public via the Internet. Crowdsourcing has recently proved to be an efficient and successful method of obtaining annotations on images regarding content [52], image aesthetic [53] and picture quality [54]. An early effort to crowdsource video quality scores was reported in [55]. The authors proposed a crowdsourced framework, whereby pairwise subjective comparisons of the Quality of Experience of multimedia content (audio, images and videos) could be recorded.

The authors of [54] conducted a large-scale, comprehensive study of real-world picture quality and showed that their results were quite consistent with the results of subjective studies conducted in a laboratory.

This success of latter study [54] has inspired my work here, with a goal to build a large, diverse and representative video database on which I crowdsourced a large-scale subjective video quality study. I encountered many difficulties along the way, many of which were significantly more challenging than in the previous picture quality study [54]. The issues encountered from simple participant problems (distraction, reliability and a imperfect training [56]), to more serious issues such as variations in display quality, size and resolution, to very difficult problems involving display hardware speed and bandwidth conditions on the participant's side. I carefully designed a framework in Amazon Mechanical Turk (AMT) to crowdsource the quality scores while accounting

for these numerous factors, including low bandwidth issues which could result in video stalls, which are very annoying during viewing and can heavily impact the experienced video quality.

Previous crowdsourced video quality studies have not addressed the latter very important concern. For example, in the study in [57], the participants were allowed to use either Adobe Flash Player or HTML5 to display videos, depending on the compatibility of their browser. However, in their methodology, no assurance could be made that the videos would fully preload before viewing, hence there was no control over occurrences of frame freezes or stalls, or even to record such instances on the participants' end, as Flash Player does not have this option, and some browsers disable this option for HTML5 video element. When the videos are not preloaded and are streamed instead, interruptions and stalls are often introduced, whereas the study in [57] did not report any effort to record whether such events took place. The early QoE crowdsource framework [55] also did not report any accounting of this important factor.

A significant and recent crowdsourced VQA database was reported in [58], providing an important new resource to the video quality community. In this study, a subject was asked to rate any number of videos within the range 10 to 550 videos. I would like to note that viewing as many as videos as the upper end of this range is likely to produce fatigue, which affects performance. Generally, it is advisable to restrict the number of watched videos per session so that the session time does not exceed 30-40 mins including training, to reduce

fatigue or loss of focus [59]. Also, I observed that the study participants in [58] were allowed to zoom in or out while viewing, which can introduce artifacts on the videos, and can lead to interruptions and stalls. Scaling a video up or down is computationally expensive. Under these conditions, stalls could occur even if a video was fully preloaded into memory before being played. Downscaling and upscaling artifacts, and video stalls are factors that significantly impact the perceived video quality. Detecting whether stalls occurred is also critical. It appears that the authors of the study did not account in any way for stalls, which is highly questionable, since stalls can deeply impact reported subjective quality. These types of issues underline the difficulty of online video quality studies, and the need for careful design of the user interface, monitoring of the subjects, and the overall supporting pipeline used to execute a large-scale study.

A few other video crowdsourcing methods have been reported, at much smaller scales without addressing the difficult technical issues [60–64] described in the preceding.

2.3 Background Related to Video Quality Prediction

Once VQA databases are available, the next important problem is to design models and algorithms that can be used to automatically assess the quality of videos in high agreement with human opinions. I review here relevant background relared to video quality prediction.

2.3.1 Review of the VQA Algorithms

VQA algorithms are classified based on the amount of information that the algorithm has access to:

- Full reference (FR) VQA algorithms, which compare a distorted version of a source to its pristine reference. Notable commercially deployed FR VQA models include SSIM [65], MS-SSIM [66], VMAF [67], VIF [68], and MOVIE [69].

- Reduced reference (RR) VQA algorithms which do not have access to an entire reference video, such as ST-RRED [70] and VQM [71].

- Blind or no-reference (NR) VQA algorithms do not utilize any information to predict quality beyond the distorted videos. A notable NR VQA model that supplies good general performance is the Video-Blind Image Integrity Notator using DCT-Statistics (V-BLIINDS) [72]. The predictor that I describe in this dissertation also falls within this category of models, hence I discuss these in greater detail in the following.

The oldest VQA model is the FR PSNR/MSE, which is easy to implement and to compute, and is applied on a frame basis. Unfortunately, PSNR fails to account for perception, and cannot be used alone to analyze degradations that develop even over time intervals, or moving artifacts, or distortions that might be revealed by analyzing temporal change. As a result, they do not correlate well with human perceptions of video quality. There are a variety of

more successful FR models that embed models of how humans perceive visual distortions, and or the statistics of visible distortions in digital videos. Early models include the Visual Differences Predictor (VDP) [73], the classical Mannos and Sakrisons model [74], the Sarnoff JND Model [75], and the Moving Picture Quality Metric (MPQM) [76]. Subsequent models include the Structural Similarity (SSIM) [6], Multiscale- SSIM (MS-SSIM) [66], Visual Information Fidelity (VIF) [68], and the Visual Signal-to-Noise Ratio (VSNR) [77]. Another group of FR models makes use of more sophisticated temporal measurements such as TetraVQM [69], MOVIE [69], SpatioTemporal-Most Apparent Distortion (ST-MAD) [78], and Spatio-Temporal Reduced Reference Entropic Differences (STRRED) [70].

2.3.2 NR VQA Models

The "blind" or NR VQA problem is useful in any scenario where pristine no reference signal is unavailable , as when a consumer captures a video of a scene with a digital camera or smart phone. Several NR VQA models have been proposed. One of the most commonly used, and commercialized models is NIQE [8], which is a simple, quick and "completely" blind VQA model that relies only on assessing the naturalness of video frames using a classical natural scene statistic (NSS) model. It is notable for not requiring any training on any distorted signals, or on human opinions on them. When used to conduct VQA, NIQE is applied on a frame basis, on which the mean subtracted contrast normalized (MSCN) coefficients are computed. The MSCN

coefficients of high-quality of images or video frames strongly tend to follow a generalized Gaussian distribution (GGD). However, the presence of distortions usually causes this property to break down. However, the MSCN coefficients of distorted images and video frames are often effectively modeled as following a generalized Gaussian distribution whose variance and shape parameters vary systematically according to the nature of the distortions. When the MSCN features are computed on a set of distorted images annotated with human labels, and these are used to train a regressor, then a model called BRISQUE [7] is arrived at.

More sophisticated predictors that incorporate temporal information also exist. Notable examples include V-BLIINDS [72], which makes use of a measure of motion coherency, a simple estimate of egomotion, as well as NSS model features (spatial and temporal DCT features, sub-band features, DC coefficients and spatial naturalness) to predict quality. This blind algorithm also requires training. Another model called VIIDEO [79] measures losses of statistical regularity observed on natural videos when they are disturbed by distortions. VIIDEO models inter sub-band correlations over local and global time spans. The model described in [80] extends the ideas behind V-BLIINDS [72], by creating quality-aware 3D-DCT features which are used to predict video quality. The NR model FC [81] uses a variety of spatial and temporal information indices to predict quality. The NR model in [82] relies on statistical artifact measurement, while [83] describes a deep convolutional network trained to blindly predict video quality.

Chapter 3

Contributions

In this dissertation, I tackled two important main problems: exploring the bivariate NSS of images and tackling the blind VQA problem. The contributions in my dissertation are summarized as follows:

1. Modeling the bivariate NSS model of images in closed-form .
2. Using the bivariate NSS as tool to tackle a couple of quality assessment related problems (blindly prediction of image quality and 3D visual discomfort prediction of stereo images).
3. Scaling up subjective video quality studies by building a new robust framework for conducting crowdsourced VQA studies.
4. Building the largest VQA database across several key aspects such as the size of the database, the number of unique contents, distortions, combination of distortions, capture devices, resolutions, orientations, and subjects who provided quality labels.
5. Creating a "completely blind" video quality model that relies on a unique set of directional spatio-temporal NSS features, and which does not require any kind of training.

Chapter 4

Modeling the Bivariate NSS of Images in Closed-Form

In this chapter, I present the details of the model I built to capture the NSS of images in closed-form beginning with the preprocessing steps of bandpass decomposition and divisive normalization. Along the way, I demonstrate the various processing steps used in the model using high quality images from the pristine subset of LIVE Image Quality Assessment database [84]. I also study the model properties on white noise and the way the model fits are affected when the images are modified by common distortions. [1]

4.1 Processing Steps of my Model

A flow diagram of the involved processing in my model is shown in Fig. 4.1. Each step is presented in details next.

[1] This chapter appears in the following paper: Zeina Sinno, Constantine Caramanis, Alan C. Bovik: "Towards a Closed Form Second-Order Natural Scene Statistics Model" in the IEEE Transactions on Image Processing 27(7): 3194-3209 (2018). Zeina Sinno has studied the model and performed the full experimental analysis of the works described therein.

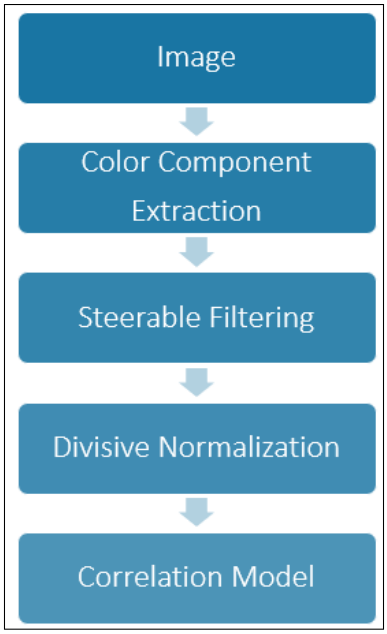


Figure 4.1: Image pre-processing used in the NSS correlation model.

4.1.1 Steerable Filters

The NSS model that I use and develop is based on luminance images that have been subjected to bandpass processing. While the model appears to hold over a wide range of bandpass operations (Gabor, wavelet, etc.), I use steerable filters [85] in my simulations, owing to their simple, easily manipulated form, their invariance to content translations, and their good fit as a frequently used model of bandpass simple cells in primary visual cortex. A steerable filter at a given frequency tuning orientation θ_1 is defined by:

$$F_{\theta_1}(\boldsymbol{x}) = cos(\theta_1)F_x(\boldsymbol{x}) + sin(\theta_1)F_y(\boldsymbol{x}), \tag{4.1}$$

where $\boldsymbol{x} = (x, y)$, and F_x and F_y are the gradient components of the two-dimensional unit-energy bivariate isotropic gaussian function:

$$G(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}}, \tag{4.2}$$

having scale parameter σ. Steerable filter based decompositions, such as steerable pyramids [86] yield substantially spatially decorrelated responses when applied to high-quality photographic images.

Modifying the scale parameters σ of the bivariate gaussian derivative functions (F_x and F_y) enables the construction of a multi-scale bandpass image decomposition broadly resembling the responses of populations of simple cells in cortical area V1. Other filter models could be used equally well to obtain bandpass orientation and radial frequency responses, such as Gabor filters, but the steerable filters present advantages of simple definition and efficient computation. The radial frequency bandwidth of the steerable filter (1) is fairly narrow (about 2.6 octaves). In the following development and testing of the bivariate correlation model, each analyzed image is passed through steerable filters of scales $\sigma \in \{1, 2, 3, \ ..., 15\}$ and over 15 frequency tuning orientations $\theta_1 \in [0, \pi/15, 2\pi/15, ..., \pi]$, yielding 225 bandpass responses. The bandpass images were computed on all 29 pristine images from the LIVE Image Quality Assessment database, yielding a total of 6525 bandpass filtered image responses.

4.1.2 Divisive Normalization

Divisive normalization was then applied on all of the steerable filter responses. When applied to naturalistic photographic images that have been bandpass filtered, normalization by the energy of the local signal has been observed to gaussianize and further decorrelate the image data [16, 87]. The divisive normalization model used here is:

$$u_j(\boldsymbol{x}) = \frac{w_j(\boldsymbol{x})}{\sqrt{t + \sum_{\boldsymbol{y}} g(j(\boldsymbol{y}), w_j(\boldsymbol{y}))^2}}, \tag{4.3}$$

where w_j are the steerable filter responses for filters indexed by j, u are the coefficients obtained after divisive normalization, and $t = 10^{-4}$ is a stabilizing saturation constant. The weighted sum in the denominator is computed over a spatial neighborhood of pixels from the same sub-band, where $g(x_i, y_i)$ is a circularly symmetric Gaussian function having unit volume. To match the increase in scale applied at the steerable filtering step (translated by increasing σ), the variance of $g(x_i, y_i)$ is also increased linearly as a function of σ. Furthermore, I note that this step is also a good functional model of the non-linear adaptive gain control of V1 neuronal responses in the visual cortex [38]. Divisive normalization causes the subband statistics of good quality natural images to become strongly Gaussianized. If the images are distorted, then the bandpass distribution tends away from Gaussian [9].

4.1.3 Bivariate Density Model

Following Su *et al.* [22], I used a multivariate generalized Gaussian distribution (MGGD) to model the joint histogram of a pair of divisively normalized bandpass image samples located at different spatial (pixel) locations. Methods for estimating the parameters of MGGD model fits to multidimensional image histograms are studied by Pascal *et al.* [88]. The probability density function of the MGGD is:

$$p(\boldsymbol{x}; \boldsymbol{M}, \eta, s) = \frac{1}{|\boldsymbol{M}|^{\frac{1}{2}}} g_{\eta,s}(\boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{M}^{-\boldsymbol{1}} \boldsymbol{x}), \tag{4.4}$$

where $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{M}$ is an $N \times N$ scatter matrix, η and s are scale and shape parameters respectively, and $g_{\eta,s}(.)$ is the density generator:

$$g_{\eta,s}(y) = \frac{s\Gamma(\frac{N}{2})}{(2^{\frac{1}{s}}\pi\eta)^{\frac{N}{2}}\Gamma(\frac{N}{2s})} e^{-\frac{1}{2}(\frac{y}{\eta})^s}, \tag{4.5}$$

where Γ is the digamma function and $y \in \mathbb{R}^+$. Note that when $s = 0.5$, (4.5) becomes a multivariate Laplacian density function, and when $s = 1$, it becomes multivariate Gaussian density. Here I fix $s = 1$, where η controls the spread of the density function.

While pairs of Gaussian random variables are not necessarily jointly Gaussian, pairs of image samples that have been subjected to bandpass processing followed by divisive normalization are observed to be reliably jointly Gaussian. The reason for the Gaussianity of images processed in this perceptually relevant manner remains elusive. It cannot be explained as a consequence

of the Central Limit Theorem (CLT), since the only additive process (linear filtering) is on strongly correlated, raw image samples rather-than on uncorrelated or weakly correlated variances, as required by the CLT. Moreover, the outcome of the linear filtering is decidedly non-Gaussian, and instead is distributed with much heavier tails, typically described as leptokurtic generalized Gaussian [87]. The shape of these empirical non-Gaussian "sparsity" densities is typically attributed to the imaging projection of a world that is smooth nearly everywhere (yielding heavily massed bandpass samples near or at zero), except where (blurred) singularities occur (resulting in large responses defining the heavy tails). Gaussianity finally arises as a consequence of a process of local divisive normalization by neighboring bandpass image energy [87, 89]. While this ultimate Gaussianity remains unexplained, there may be connections with theoretical processes defined as quotients of highly correlated quantities, such as the Fisz transform [90, 91].

The bivariate empirical histograms of the sub-band coefficients of natural images are thus modeled here as following a bivariate generalized Gaussian distribution (BGGD), by setting $N = 2$. This also presumes that the images have not been distorted, which may change their statistics. In all of the following, the parameters of the BGDD were estimated using the efficient maximum likelihood estimation method of [88]. I systematically applied this modeling process to all of the bandpass normalized images.

To remove any undesirable border filter effects, I cropped 10 pixels from each image's four borders, defined a window at a fixed position within the

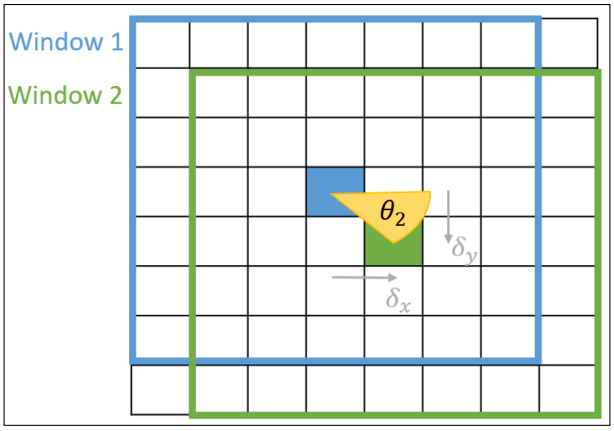
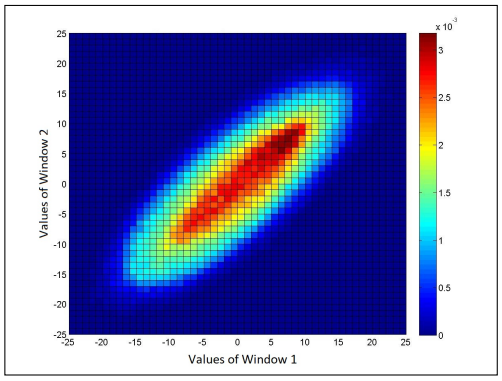


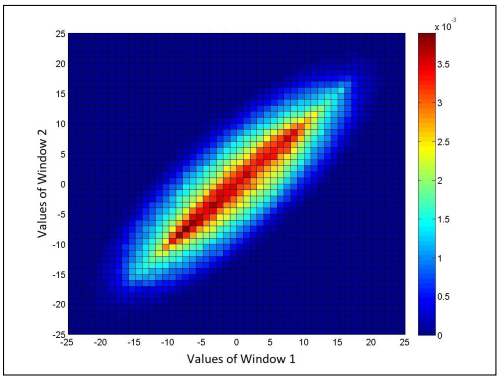
Figure 4.2: An illustration of an image after the divisive normalization and steerable filtering (of fixed σ and θ_1 values) are applied, with the two sliding windows, and how θ_2 is computed.

cropped image (Window 1) and another sliding window of the same dimensions (Window 2). Denote the distance between the center of the two windows of bandpass, normalized image samples of interest by d, and the angle between them by θ_2, as illustrated in Figure 4.2. Next, define the relative angle $\theta_2 - \theta_1$, where θ_1 is the sub-band tuning of the bandpass filter orientation relative to the horizontal axis. The bivariate histogram takes predictable shapes. For example, when the relative angle $\theta_2 - \theta_1 = 0$, the bivariate joint histogram takes a highly eccentric elliptical shape indicating a strong degree of a correlation, whereas when the relative angle is increased, the bivariate histogram becomes more circular. Figure 4.3 plots bivariate histograms as intensity for the case of $d = 1$ and $\theta_2 = \pi/2$. As I discuss further, I observe similar histogram shape trends for longer separations d and for all other spatial angles θ_2.

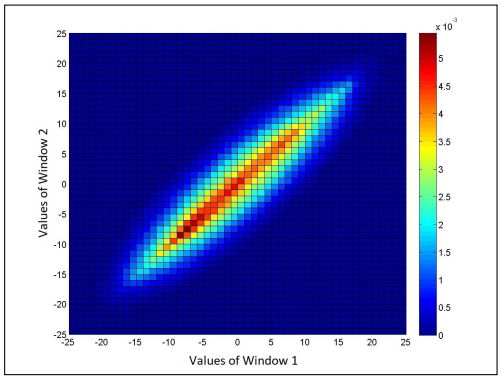
The bivariate model (4.4) is a closed form, except for the elements of



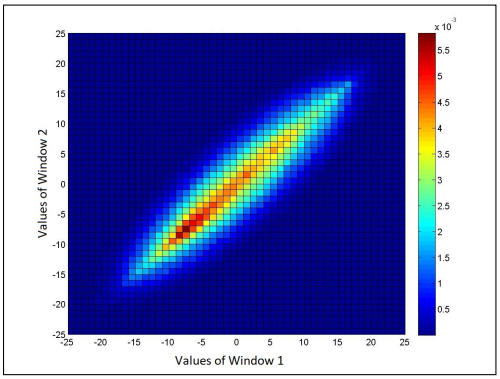
(a) $\theta_1 = 0$, and $\theta_2 - \theta_1 = -1.57$



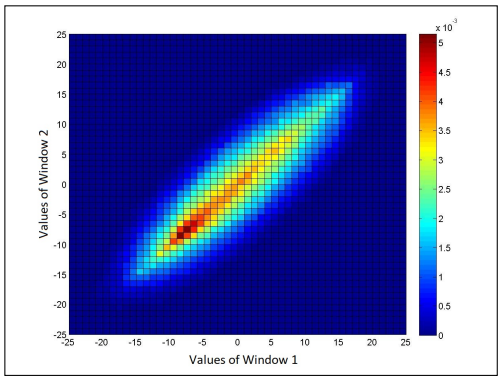
(b) $\theta_1 = \frac{3}{15\pi}$, and $\theta_2 - \theta_1 = -0.94$



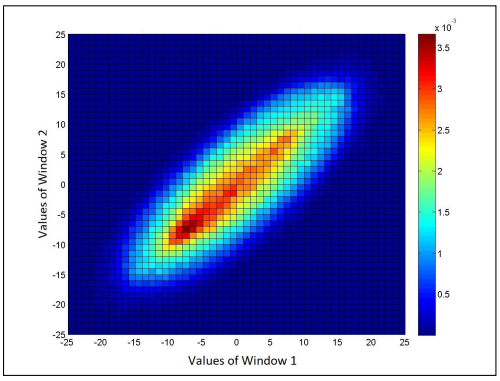
(c) $\theta_1 = \frac{6}{15\pi}$, and $\theta_2 - \theta_1 = -0.31$



(d) $\theta_1 = \frac{9}{15\pi}$, and $\theta_2 - \theta_1 = 0.31$



(e) $\theta_1 = \frac{11}{15\pi}$, and $\theta_2 - \theta_1 = 0.73$



(f) $\theta_1 = \frac{14}{15\pi}$, and $\theta_2 - \theta_1 = 1.36$

Figure 4.3: Bivariate joint histograms of a steerable filter response at distance $d = 1$, scale $\sigma = 2$, tuned to spatial orientation $\theta_2 = \pi/2$ for various spatial angular differences θ_1. Each plot presents the probability of the values that two pixels separated by d and θ_2 will take.

the scatter matrix $\boldsymbol{M}$. The scatter matrix defines the covariance matrix of the bivariate model. To complete the closed form model, I studied the Pearson correlation function between the two windows. I obtain the correlation as the gradient of the covariance of the two entities and the product of their standard deviations. The two windows were separated by horizontal and vertical separations δ_x and δ_y which I varied over the integer range from 1 and 25, i.e. distances of $\sqrt{\delta_x^2 + \delta_y^2}$ at spatial orientations $\theta_2 = \arctan(\frac{\delta_y}{\delta_x})$ (relative to the horizontal axis). I limited the range of θ_2 to $[0, \pi[$ since the quantities being measured are symmetrically defined and are π periodic.

The tuning orientation θ_1 is the frequency tuning orientation of the steerable filter. I used a discrete set of 15 sub-band orientations $\{0, \frac{\pi}{15}, \frac{2\pi}{15}, ..., , \frac{14\pi}{15}\}$ to build my model.

The correlation function model expresses a periodic behavior in the relative angle $\theta_2 - \theta_1$, which can be well modeled as:

$$\rho(d, \sigma, \theta_2) = A(d, \sigma, \theta_2)cos(2(\theta_2 - \theta_1)) + c(d, \sigma, \theta_2) \tag{4.6}$$

where $A(d, \sigma, \theta_2)$ is the amplitude, $c(d, \sigma, \theta_2)$ is an offset, d is the spatial separation between the target pixels, σ is the steerable filter spread parameter, and θ_2 is as before. Generally, the shapes of ρ, A, and c vary in a consistent way with d, σ and θ_2, as we shall see.

Figure 4.4 plots the average correlation function of several processed images from the set of LIVE reference images, as a function of $\theta_2 - \theta_1$, over 4 scales for $\theta_2 = \pi/2\ rad$ and $d = 1$. From this plot, it may be observed

that the maximum correlation value P =max(ρ) that is attained, occurs (as expected) when $\theta_2 - \theta_1 = 0$, falling monotonically from this maximum value as the absolute relative angle is increased to $\pi/2$. Figure 4.4 also shows that the correlation increases with the scale factor σ, which I have observed over all studied spatial orientations θ_2 and spatial separations d. This is to be expected, since as σ is increased, the filter bandwidths decrease, which tends to increase in-band correlations.

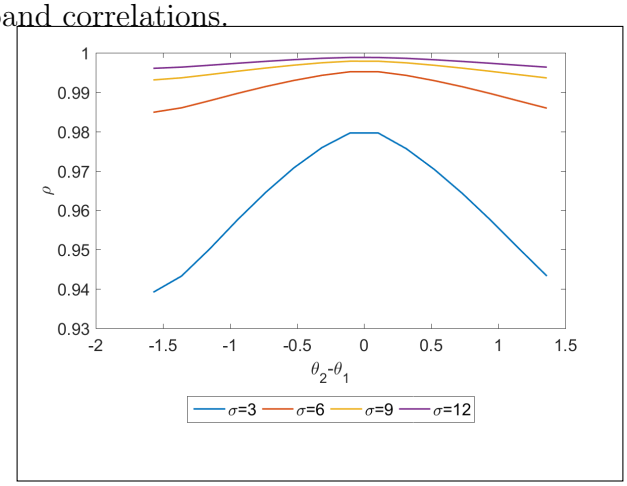


Figure 4.4: Average correlation function of the luminance components of natural images plotted against relative angle $\theta_2 - \theta_1$, for $\theta_2 = \pi/2\ rad$, $d = 1$, and $\sigma = 3,\ 6,\ 9$, and 12.

As the spatial separation d is increased, the correlation also drops, as shown in Fig. 4.5, where the empirical correlations are plotted for a fixed scale σ and spatial orientations θ_2, over several values of the spatial separation d .

As a further illustration of the correlation function's behavior, Figure 4.6 plots the correlations of adjacent samples measured at the same scale ($\sigma = 2$) and spatial separations but different spatial orientations $\theta_2 \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$.

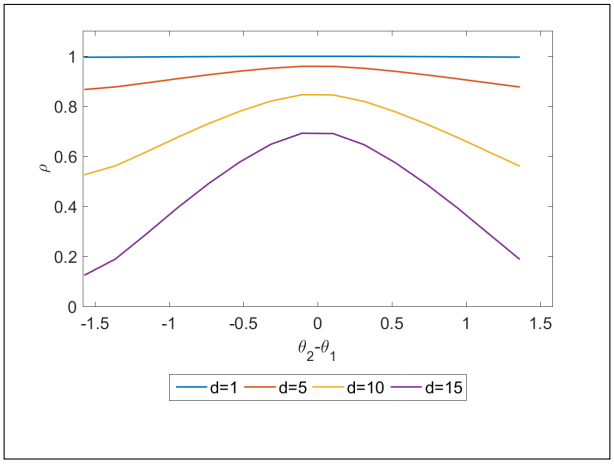


Figure 4.5: Average correlation function of the luminance components of natural images plotted against relative angle $\theta_2 - \theta_1$, for $\theta_2 = \pi/2\ rad$, $\sigma = 10$ and $d = 1,\ 5,\ 10$, and 15.

Note that the sample separation takes two values: $d = 1$ for $\theta_2 \in \{0, \frac{\pi}{2}\}$ and $d = \sqrt{2}$ for $\theta_2 \in \{\frac{\pi}{4}, \frac{3\pi}{4}\}$. From the plot, it may be seen that horizontally and vertically related pixels ($\theta_2 = \pi/2\ rad$ and $\theta_2 = 0\ rad$) are more correlated than diagonally related pixels ($\theta_2 = \pi/4\ rad$ and $\theta_2 = 3\pi/4\ rad$), which is also expected owing to the different spatial separations. However, the correlation also likely increases along the cardinal directions because of the preponderance of horizontal and vertical structures in real-world images [92].

In order to better understand and to complete my model of the correlation function ρ in (4.6), I also model the amplitude and offset functions A and c. To do so, I define the peak correlation function:

$$P = \max(\rho) = A + c. \tag{4.7}$$

wherein I may rewrite (4.6) as:

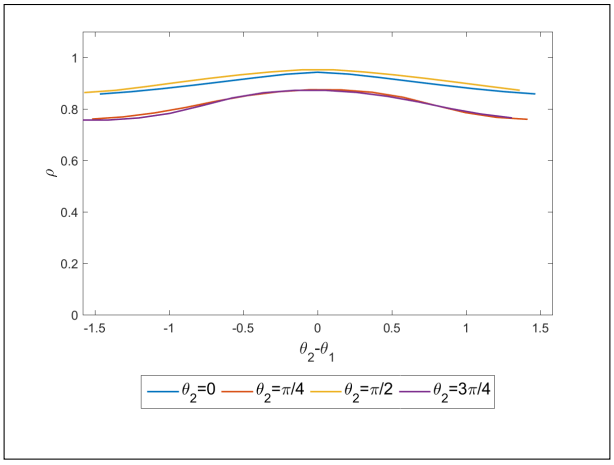


Figure 4.6: Average correlation function of the luminance components of bandpass, divisively normalized natural images for the case of adjacent pixels (horizontal, vertical, diagonal) plotted against relative angle, for $\sigma = 2$ for $\theta_2 - \theta_1$ for $\theta_2 = 0$, $\pi/4$, $\pi/2$, and $3\pi/4$.

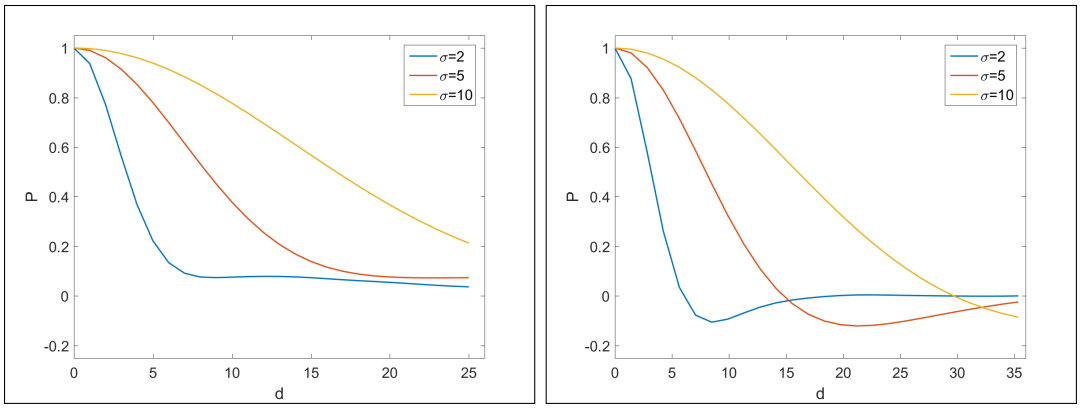
$$\rho(d, \sigma, \theta_2) \;=\; A(d, \sigma, \theta_2) cos(2(\theta_2 \,-\, \theta_1)) \,+\, [P(d, \sigma, \theta_2) \,-\, A(d, \sigma, \theta_2)] \quad (4.8)$$

I did not impose any constraints on the values of A and P when fitting ρ. I have observed the values of A to be positive except in a few instances where the correlation is very small (at large spatial separations) or large and flat (at small separations and large scales). In those cases, A took slightly negative values (10^{-3}).

As mentioned earlier in the background section, Lee, Mumford and Huang [21] systematically observed that the sample covariances of bandpass image pixels follow an approximate reciprocal power law, of the form $\frac{1}{|d|^b}$, which, like white processes, cannot be realized. Similarly, Keshner [41] remarks on the fact that the nonstationary autocorrelation function of $1/f$ processes take a reciprocal form, and that a practical stationary model might be obtained

by modifying the autocorrelation model near the origin. Here, I take a different approach, whereby I model the peak correlation function as having a general version of the form $\frac{1}{|d|^{\beta}+1}$.

Figure 4.7 plots the empirical peak correlation function P against the sample separation d for a few values of σ and θ_2. As expected, the measured correlations decrease rapidly from a peak value of 1 as the spatial separation d increases; which is natural since one should expect reduced correlations between pixels as the spatial separation increases. There is a slight observed undershoot, especially for small σ values, which is likely a consequence of unsmoothness of the applied filter, but this is small and difficult to model, hence I neglect this minor behavior.



(a) (b)

Figure 4.7: Peak function $P(d, \sigma, \theta_2)$ plotted against pixel separation d for $\sigma = 2$, 5, and 10 for (a) $\theta_2 = 0$ and (b) $\theta_2 = \frac{\pi}{4}$ (rad).

The general form of my stabilized peak correlation model is as follows:

given a fixed spatial orientation θ_2 and a scale σ, define

$$\hat{P}(d, \sigma, \theta_2) = \frac{1}{(\frac{d}{\alpha_0(\theta_2)*\sigma})^{\beta_0} + 1} \tag{4.9}$$

where $\{\alpha_0, \beta_0\}$ are parameters that control the shape and fall-off of the peak correlation function, and which depend on the spatial orientation θ_2.

I discuss the validation and application of my model (4.9) further along, but first I will look at the other function comprising the correlation model (4.8).

Figure 4.8 plots the amplitude function $A(d, \sigma, \theta_2)$ against d for few scales σ and spatial orientations θ_2. The graph of A rises from the value 0 at $d = 0$, then decreases with increasing separation. Given the similarity of the graph of A to the difference of two functions of the same form but different scales, and the close relationship between A and P, I model A as the difference of two functions of the form (4.9):

$$\hat{A}(d, \sigma, \theta_2) = \frac{1}{(\frac{d}{\alpha_1(\theta_2)*\sigma})^{\beta_1(\theta_2)} + 1} - \frac{1}{(\frac{d}{\alpha_2(\theta_2)*\sigma})^{\beta_2(\theta_2)} + 1} \tag{4.10}$$

where $\{\alpha_1, \beta_1, \alpha_2, \beta_2\}$ are parameters that are functions of θ_2 that control the shape of A.

My goal next is then to find, for a fixed spatial orientation θ_2, the values of the parameters $\{\alpha_0, \beta_0\}$ that produce the best fit to (4.9), and the parameters $\{\alpha_1, \beta_1, \alpha_2, \beta_2\}$, that yield the best fit to (4.10), in the least mean squared error sense. I form two optimization systems for P and A that account for scale to find those optimal values. The optimization systems minimize the

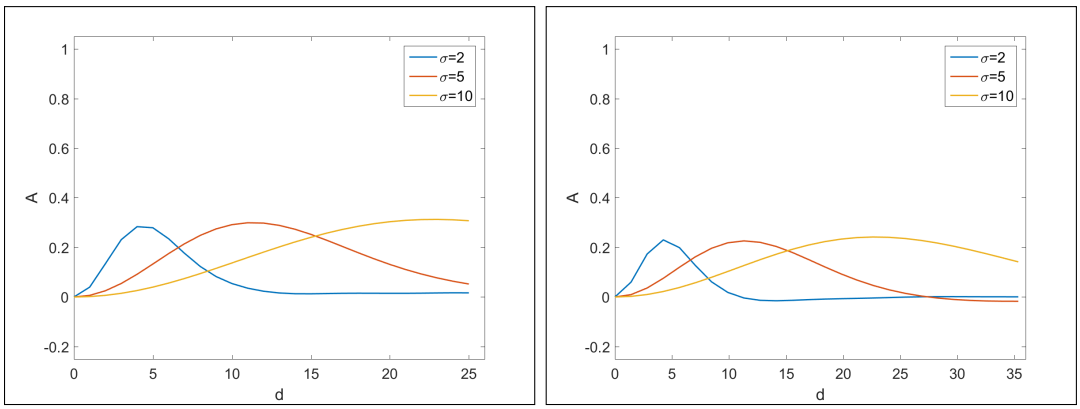


Figure 4.8: Amplitude function $A(d, \sigma, \theta_2)$ plotted against pixel separation d for $\sigma = 2$, 5, and 10 for (a) $\theta_2 = 0$ and (b) $\theta_2 = \frac{\pi}{4}$ (rad).

summed squared errors of the peak and amplitude. To accomplish this, I apply unconstrained nonlinear regression using the quasi newton method [93]. I restrict my modeling of the correlation to a span of dimensions 25×25 so that $d \in [0, \sqrt{1250}]$, since the peak correlation becomes negligible if d is increased further. The four functions $P(d, \sigma, \theta_2)$, $A(d, \sigma, \theta_2)$, $\hat{P}(d, \sigma, \theta_2)$, and $\hat{A}(d, \sigma, \theta_2)$ form vectors of size $m \times 1$, where m is the number of occurrences of θ_2 inside the span of interest. Denote by D the set of distances for a given spatial orientation θ_2. For the case $\theta_2 = 0$ or $\pi/2$, $D = \{0, 1, 2, 3, ..., 24, 25\}$. For the case $\theta_2 = \pi/4$ or $3\pi/4$, $D = \{0, \sqrt{2}, \sqrt{8}, \sqrt{18}, ..., \sqrt{1152}, \sqrt{1250}\}$. My optimization systems are then expressed as:

$$\min_{\alpha_0, b_0} \quad \sum_{d \in D} \sum_{\sigma=2}^{15} (P(d, \sigma, \theta_2) - \hat{P}(d, \sigma, \theta_2))^2 \tag{4.11}$$

and

$$\min_{\alpha_1, b_1, \alpha_2, \beta_2, b_2} \quad \sum_{d \in D} \sum_{\sigma=2}^{15} (A(d, \sigma, \theta_2) - \hat{A}(d, \sigma, \theta_2))^2 \tag{4.12}$$

Table 4.1 gives the optimal parameters yielding the best average correlation fit to (4.11) and (4.12) over all of the (luminance) images in the LIVE reference image set over the 8 most frequently occurring spatial orientations θ_2. It may be observed that the fitting parameters fall within narrow ranges, the exceptions being the peak correlation parameters (α_0, β_0) which deviate a little more along the cardinal orientations, and to a lesser degree, along the diagonal orientations. This is not unexpected given the well-known prevalence of horizontal, diagonal, and vertical oriented structures in the visual environment [92]. What is perhaps surprising is the high degree of uniformity of the other parameters against orientation, particularly those of the amplitude function (4.10). I also computed these parameters over the larger set of values $\theta_2 = \{0.000,\ 0.785,\ 1.571,\ 2.356,\ 0.464,\ 1.107,\ 2.034,\ 2.678,\ 0.322,\ 0.588,\ 0.983,\ 1.249,\ 1.893,\ 2.159,\ 2.554,\ 2.820,\ 0.245,\ 0.644,\ 0.927,\ 1.326,\ 1.816,\ 2.214,\ 2.498,\ 2.897\}$. These values occur at least 5 times in the area of interest. Values of θ_2 where there was insufficient data (viz., pairs of pixels at those

orientations) are left out to conduct the optimization. I computed the optimal parameters $\alpha_0, \beta_0, \alpha_1, \beta_1, \alpha_2$, and β_2 for this set of θ_2 values for each $\sigma \in \{1, 2, 3, \ ..., 15\}$. Since this is a sizeable amount of tabulated data, I make it available at the following link:

http://live.ece.utexas.edu/research/bivariateNSS/index.html.

Table 4.1: Optimal values of $\alpha_0, \beta_0, b_0, \alpha_1, \beta_1, b_1, \alpha_2, \beta_2$, and b_2 for the 8 most frequently occurring values of θ_2 on the LIVE IQA reference luminance images.

θ_2	α_0	β_0	α_1	β_1	α_2	β_2
0.000	1.623	2.628	3.197	3.336	2.092	2.274
0.464	1.519	3.325	3.307	3.501	2.383	2.204
0.785	1.528	3.771	3.302	3.450	2.456	2.252
1.107	1.724	3.175	3.357	3.542	2.370	2.173
1.571	2.210	2.181	3.267	3.171	1.978	2.313
2.034	1.718	3.207	3.358	3.545	2.371	2.161
2.356	1.522	3.767	3.293	3.456	2.448	2.253
2.678	1.506	3.351	3.288	3.471	2.382	2.203

4.1.4 Model Validation

Next, I validate my model by examining the closeness of fit of the models $\hat{P}$, $\hat{A}$ and $\hat{\rho}$ to the empirical functions P, A and ρ.

4.1.4.1 Validation of A and P

I computed the mean squared error (MSE) between the reconstructed peak and amplitude correlation functions $\hat{P}$ and $\hat{A}$, relative to the empirical average functions P and A that were computed and measured, respectively, on the LIVE Image Quality Assessment Database [84] luminance images across

integer scales $\sigma \in \{2, 3, \ ..., 15\}$. The MSE between P and $\hat{P}$ for a fixed scale σ and orientation θ_2 is defined as:

$$MSE_P = \sum_{d \in D} \frac{(P(d, \sigma, \theta_2) - \hat{P}(d, \sigma, \theta_2))^2}{|D|}, \tag{4.13}$$

where $|D|$ is the cardinality of D. Similarly, for a fixed scale σ and orientation θ_2, the MSE of between A and $\hat{A}$ is defined as:

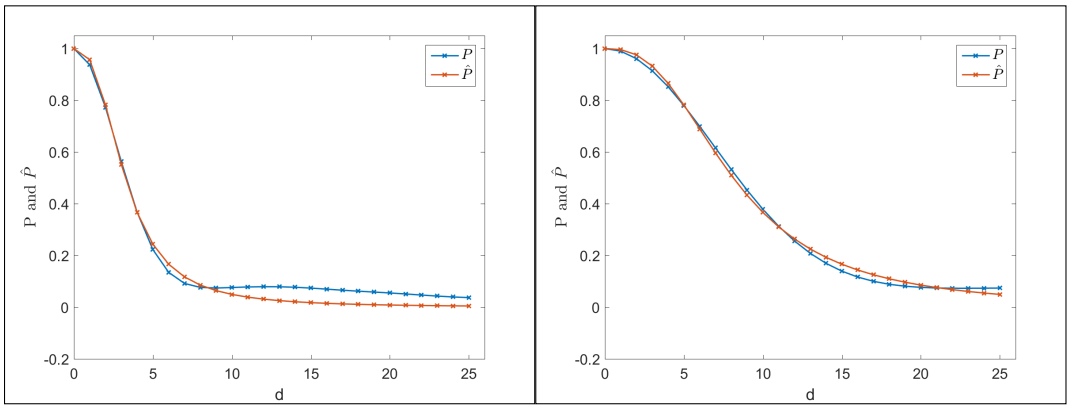
$$MSE_A = \sum_{d \in D} \frac{(A(d, \sigma, \theta_2) - \hat{A}(d, \sigma, \theta_2))^2}{|D|}. \tag{4.14}$$

The largest errors between P and $\hat{P}$ and A and $\hat{A}$ over all pairs (σ, θ_2) were on the order of 10^{-3}. The results for the considered (σ, θ_2) pairs can be found at the same link as above. Examples of the empirical functions P and A are shown in Fig. 4.9 and Fig. 4.10, which visually illustrate the goodness of my model in capturing P and A. It is worth remarking that the results obtained by finding the best-fitting $\hat{P}$ and $\hat{A}$ on the average empirical correlation data, were as good as those obtained by finding the best fits on the empirical correlations from each of the naturalistic images in the LIVE Image Quality Assessment [84] database and their corresponding best fits $\hat{P}$ and $\hat{A}$.

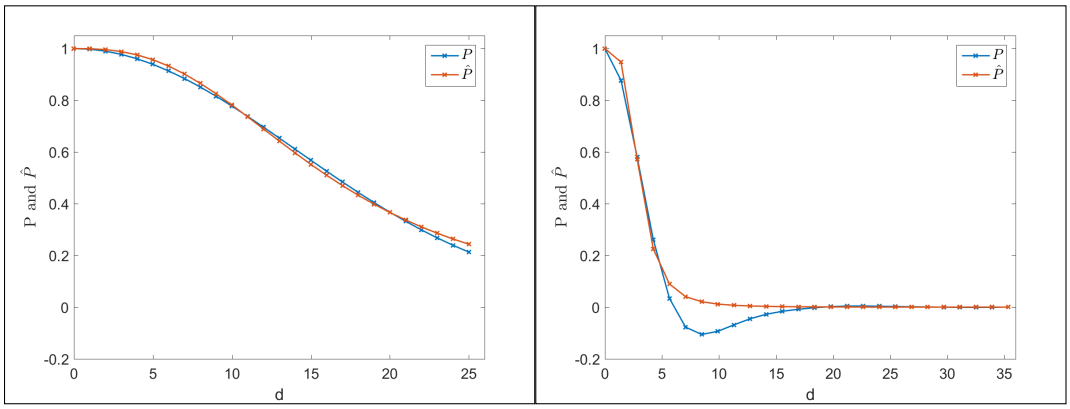
Furthermore, for each fixed scale σ and spatial orientation θ_2, I computed the χ^2 test statistic:

$$\chi^2_A = \sum_{d \in D} \sum_{i=1}^{29} \frac{(A_i(d, \sigma, \theta_2) - \hat{A}(d, \sigma, \theta_2))^2}{\hat{A}(d, \sigma, \theta_2)} \tag{4.15}$$

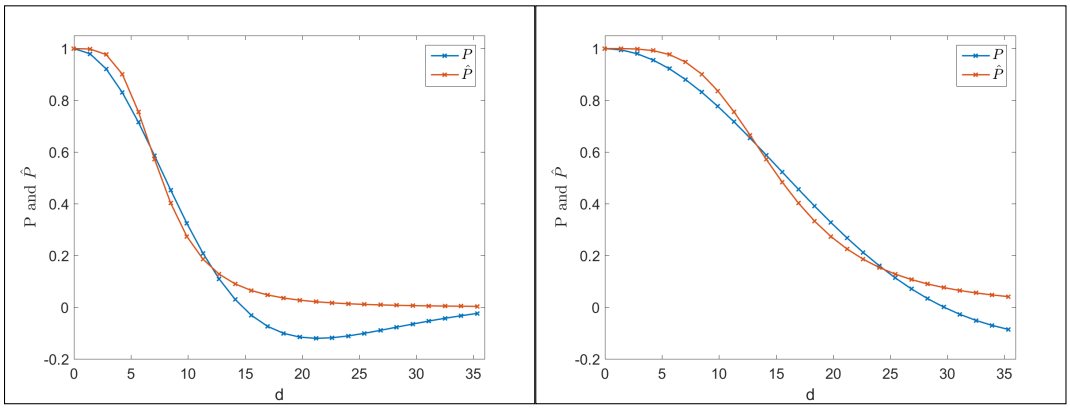
where $A_i(d, \sigma, \theta_2)$ is the empirical amplitude function of the i^{th} naturalistic image from the LIVE Image Quality Assessment [84] for fixed σ and θ_2 values,



(a) $\sigma = 2, \theta_2 = 0$ (b) $\sigma = 5, \theta_2 = 0$

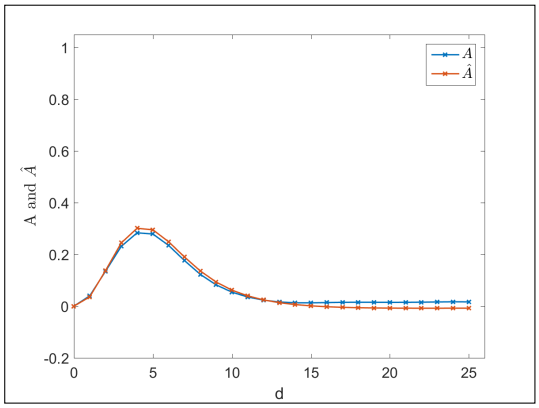


(c) $\sigma = 10, \theta_2 = 0$ (d) $\sigma = 2, \theta_2 = \pi/4$

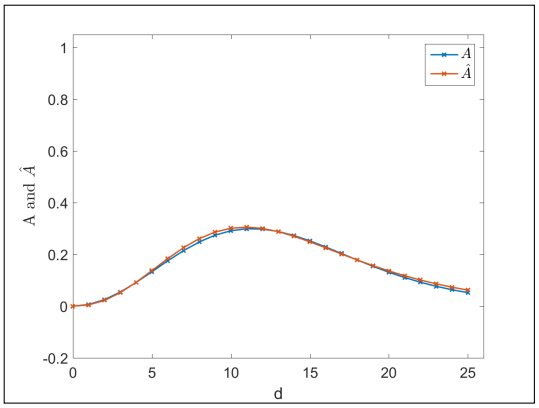


(e) $\sigma = 5, \theta_2 = \pi/4$ (f) $\sigma = 10, \theta_2 = \pi/4$

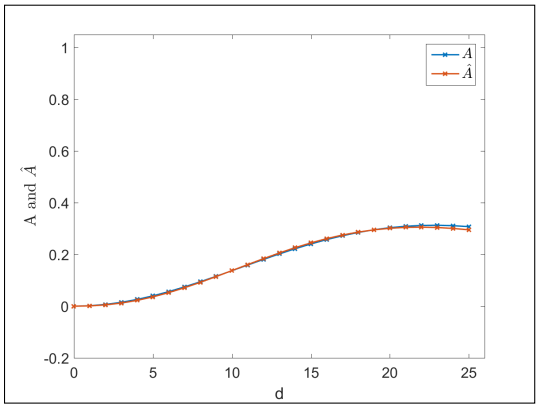
Figure 4.9: Examples of best-fitting peak correlation model $\hat{P}$ to the peak correlation P of the average empirical correlation P.



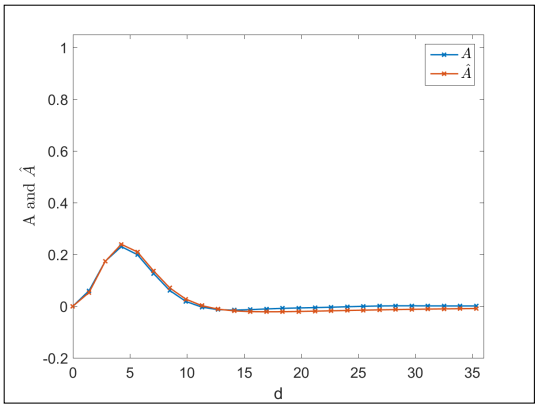
(a) $\sigma = 2, \theta_2 = 0$



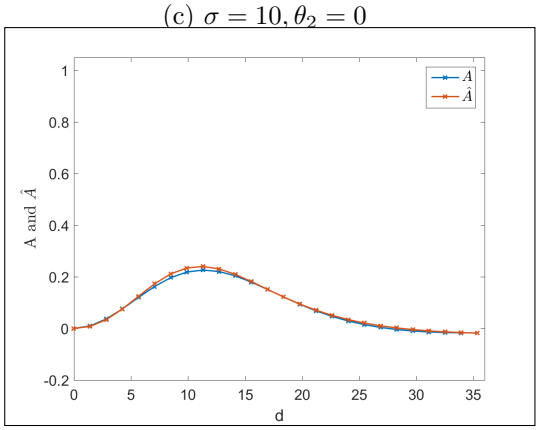
(b) $\sigma = 5, \theta_2 = 0$



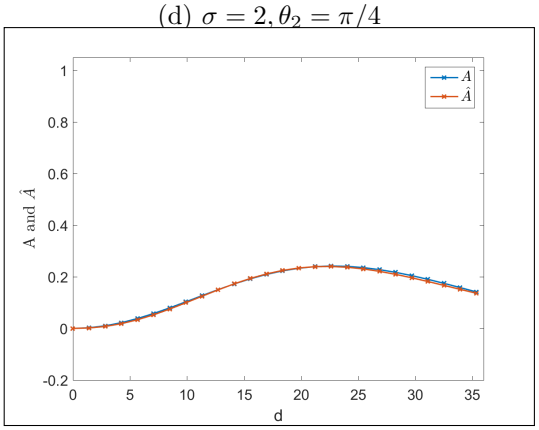
(c) $\sigma = 10, \theta_2 = 0$



(d) $\sigma = 2, \theta_2 = \pi/4$



(e) $\sigma = 5, \theta_2 = \pi/4$



(f) $\sigma = 10, \theta_2 = \pi/4$

Figure 4.10: Examples of best-fitting amplitude correlation model $\hat{A}$ to the peak correlation A of the average empirical correlation A.

and $\hat{A}(d, \sigma, \theta_2)$ was obtained by finding the best fit to the amplitude function of the average empirical correlation. Likewise the χ^2 statistic for the peak correlation was also computed:

$$\chi^2_P = \sum_{d \in D} \sum_{i=1}^{29} \frac{(P_i(d, \sigma, \theta_2) - \hat{P}(d, \sigma, \theta_2))^2}{\hat{P}(d, \sigma, \theta_2)}, \tag{4.16}$$

where $P_i(d, \sigma, \theta_2)$ is the empirical peak correlation function of the i^{th} naturalistic image from the LIVE Image Quality Assessment [84] for σ and θ_2 fixed, and $\hat{P}(d, \sigma, \theta_2)$ was obtained by finding the best fit to the peak of the average correlation.

The results of the χ^2_P and χ^2_A tests for the 8 most frequently occurring values of θ_2 are presented in Tables 4.2 and 4.3 respectively. The values are in general small (on the order of 10) except at the smallest scales of σ. This is not unexpected, as highly localized (less smoothed) measurements of the correlation will be less certain. However, I have observed the functional fits to be reasonably good, even when $\sigma = 1$. The somewhat less consistent behavior of the results when $\sigma = 1$ is likely due to two reasons: first, the spatial implementation of the steerable filters begins to become degenerate at that scale, leading to poorer localization properties than afforded by larger Gaussian envelopes, and second, the presence of high frequency noise, including quantization, present even in high-quality pictures, may affect the steerable responses as well.

Table 4.2: χ^2_P results for the 8 most frequently occurring θ_2 on the LIVE Image Quality Assessment reference luminance images.

	$\theta_2 = 0$	$\theta_2 = 0.464$	$\theta_2 = 0.785$	$\theta_2 = 1.107$	$\theta_2 = 1.571$	$\theta_2 = 2.034$	$\theta_2 = 2.356$	$\theta_2 = 2.678$
σ =1	667.17	177.88	2063.83	166.83	321.36	141.25	2051.45	157.31
σ =2	309.70	126.07	670.77	105.02	162.62	112.74	657.30	115.90
σ =3	170.18	125.16	455.59	124.99	101.75	123.38	382.69	116.03
σ =4	106.40	108.16	476.94	93.09	73.82	92.76	465.55	116.56
σ =5	79.20	86.79	521.75	57.49	60.19	59.49	549.10	102.24
σ =6	65.15	64.17	490.16	33.70	50.72	35.21	503.69	78.63
σ =7	54.00	45.57	394.72	20.73	41.62	21.76	378.95	54.75
σ =8	43.89	31.92	277.88	13.49	32.88	14.37	247.53	35.99
σ =9	35.02	23.14	180.06	9.62	25.59	10.24	149.55	23.99
σ =10	27.54	17.45	110.65	7.09	19.50	7.69	85.80	16.64
σ =11	21.50	13.56	69.57	5.30	14.71	5.92	51.05	12.16
σ =12	16.71	10.59	45.95	4.06	11.13	4.63	32.55	9.22
σ =13	12.91	8.23	32.28	3.18	8.45	3.69	22.91	7.18
σ =14	10.05	6.31	24.16	2.50	6.54	2.93	17.54	5.61
σ =15	7.83	4.79	18.67	2.00	5.15	2.36	13.98	4.39

Table 4.3: χ^2_A results for the 8 most frequently occurring θ_2 on the LIVE Image Quality Assessment reference luminance images.

	$\theta_2 = 0$	$\theta_2 = 0.464$	$\theta_2 = 0.785$	$\theta_2 = 1.107$	$\theta_2 = 1.571$	$\theta_2 = 2.034$	$\theta_2 = 2.356$	$\theta_2 = 2.678$
σ =1	-81.17	-2.93	0.05	-3.52	-206.53	-2.73	0.42	-2.45
σ =2	-89.16	1.90	1.28	0.64	522.65	1.74	2.28	0.93
σ =3	-51.07	-3.27	3.41	-6.45	20.09	-2.54	4.56	-2.85
σ =4	12.66	5.20	7.90	2.21	8.68	3.16	6.69	4.58
σ =5	8.24	5.05	9.25	5.57	6.34	4.02	8.42	5.95
σ =6	6.40	2.72	13.13	2.68	5.21	2.33	13.18	2.97
σ =7	5.48	2.33	9.17	1.99	4.60	2.17	8.20	2.38
σ =8	4.91	2.23	7.33	1.71	4.16	2.23	7.03	2.22
σ =9	4.44	2.20	6.59	1.54	3.80	2.29	6.76	2.15
σ =10	3.81	2.17	6.22	1.40	3.42	2.31	6.72	2.07
σ =11	3.28	2.11	5.85	1.28	3.13	2.26	6.58	1.97
σ =12	2.99	2.06	5.48	1.18	2.90	2.15	6.40	1.91
σ =13	2.83	1.98	5.11	1.10	2.64	2.00	6.21	1.88
σ =14	2.93	1.90	4.77	1.05	2.55	1.88	5.95	1.89
σ =15	3.17	1.84	4.48	1.06	2.56	1.79	5.69	1.91

4.1.4.2 Validation of ρ

To validate the correlation model, ρ, I followed a similar approach. The MSE for a fixed scale σ, distance d, and orientation θ_2, is defined as:

$$MSE_\rho = \sum_{\theta_1=0,\frac{\pi}{15}}^{\frac{14\pi}{15}} \frac{(\rho(\theta_2-\theta_1)-\hat{\rho}(\theta_2-\theta_1))^2}{15}. \tag{4.17}$$

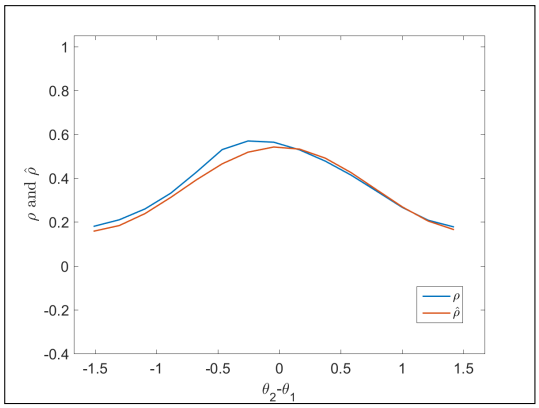
I computed the MSE values between the model correlation function $\hat{\rho}$ and the average empirical correlation function ρ on the luminance components of the LIVE IQA dataset. Again, the largest error was on the order of 10^{-3} for $\theta_2 = 0$. I observed similar results across other θ_2 values, which are not included here for lack of space, but could be found on http://live.ece.utexas.edu/research/bivariateNSS/index.html.

Figure 4.11 plots the best-fitting model correlation $\hat{\rho}$ along with the empirical correlation function for a variety of randomly selected values of d, θ_2 and σ.

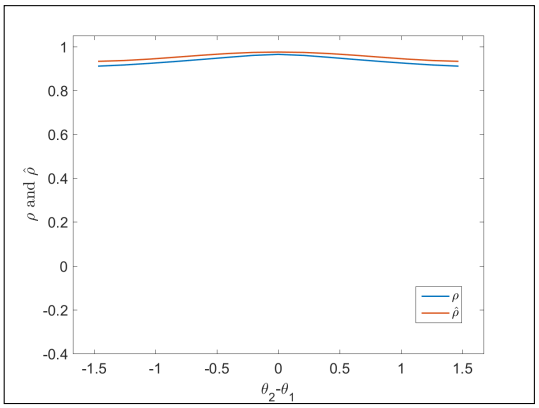
I also performed the χ^2 test for θ_2 and d fixed. The statistic is computed as:

$$\chi^2_\rho = \sum_{\theta_1=0,\frac{\pi}{15}}^{\frac{14\pi}{15}} \sum_{i=1}^{29} \frac{(\rho_i(d,\sigma,\theta_2)-\hat{\rho}(d,\sigma,\theta_2))^2}{\hat{\rho}(d,\sigma,\theta_2)}, \tag{4.18}$$

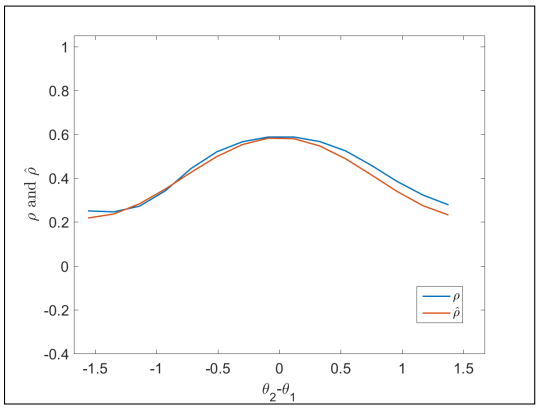
where $\rho_i(d,\sigma,\theta_2)$ is the correlation of the i^{th} naturalistic image from the LIVE Image Quality Assessment reference set for given values of d, σ and θ_2, and $\hat{\rho}(d,\sigma,\theta_2)$ is the best fit of the average correlation. Due to the lack of space, I only present the median results for $\rho_i(d,\sigma,\theta_2)$ as a function of σ and d for



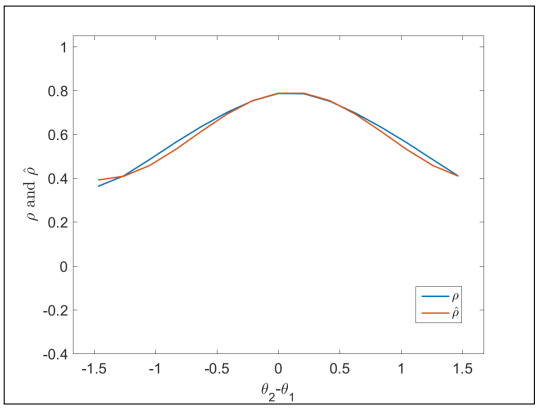
(a) $d = 20.13, \sigma = 14, \theta_2 = 2.678$



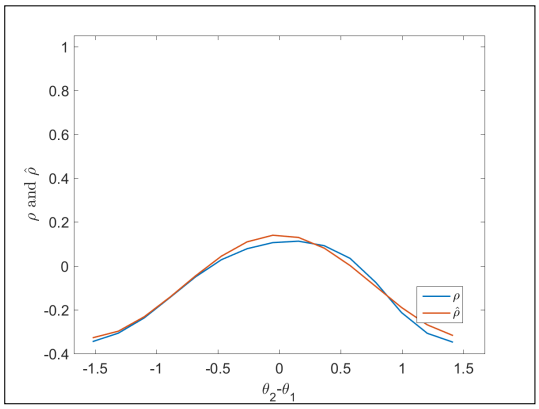
(b) $d = 2, \sigma = 5, \theta_2 = 0$



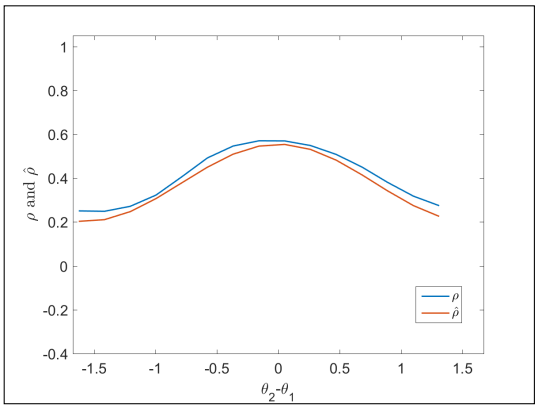
(c) $d = 10, \sigma = 7, \theta_2 = 2.214$



(d) $d = 12, \sigma = 10, \theta_2 = 1.571$



(e) $d = 29.70, \sigma = 12, \theta_2 = 0.7854$



(f) $d = 4.24, \sigma = 3, \theta_2 = 2.356$

Figure 4.11: Graphs of the model and empirical correlation functions ρ and $\hat{\rho}$ plotted against $\theta_2 - \theta_1$ for various values d, θ_2 and σ values.

the case of $\theta_2 = 0$, in Tables 4.4 and 4.5 respectively. Results for other angles (occurring less frequently) can be found at the same web link given earlier.

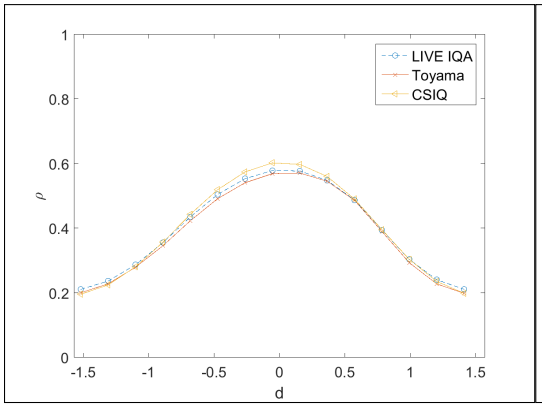
Table 4.4: Median χ^2_ρ with respect to the average luminance correlation for $\theta_2 = 0$ on the LIVE IQA reference images as a function of the scale parameter σ.

$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$	$\sigma = 6$	$\sigma = 7$	$\sigma = 8$	$\sigma = 9$	$\sigma = 10$	$\sigma = 11$	$\sigma = 12$	$\sigma = 13$	$\sigma = 14$	$\sigma = 15$
159.19	116.48	10.69	0.39	0.62	0.15	0.96	2.90	2.26	2.90	2.41	2.74	2.19	3.20	1.54

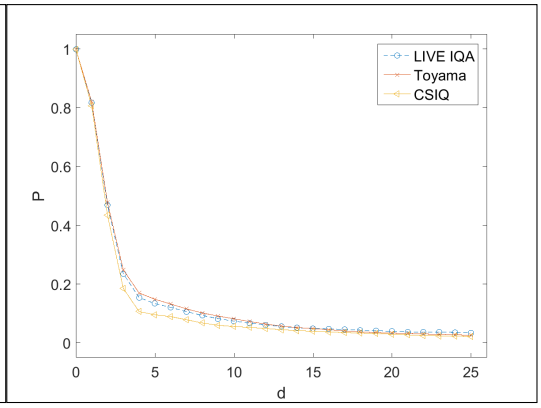
Table 4.5: Median χ^2_ρ with respect to the average luminance correlation for $\theta_2 = 0$ on the LIVE IQA reference images as a function of the spatial separation d.

$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$	$d = 9$	$d = 10$	$d = 11$	$d = 12$	$d = 13$	$d = 14$	$d = 15$	$d = 16$	$d = 17$	$d = 18$	$d = 19$	$d = 20$	$d = 21$	$d = 22$	$d = 23$	$d = 24$	$d = 25$
0.01	0.06	0.15	0.31	0.53	0.81	0.97	1.41	1.68	3.18	2.74	7.39	5.35	17.61	7.52	10.69	9.91	34.62	17.71	37.60	27.36	27.42	86.23	18.57	44.70

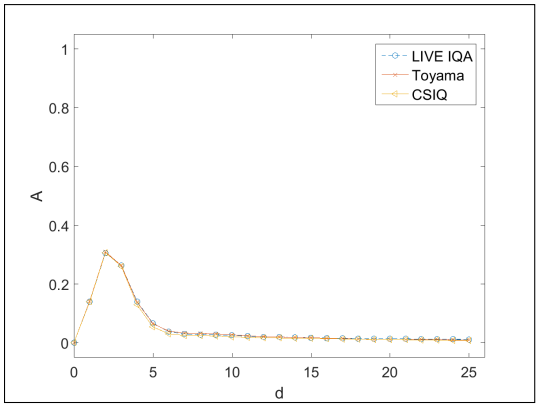
The very low MSE values, the low values of χ^2_ρ, and the apparent good functional fits shown in the plots validates the accuracy of my model. In a few instances, the values of χ^2_ρ took larger values, as a byproduct of numerical instability when computing (4.18): the appearance of small values in the denominator of (4.18) resulted in larger values of χ^2_ρ. However, even in those cases, I still observed excellent alignment between the empirical data and the functional fits. I also found that the model correlations computed on the individual pristine LIVE reference luminance images yielded similar measurements of goodness of fit.



(a) $\rho(d = 1, \sigma = 1, \theta_2 = \pi/2)$ vs $\theta_2 - \theta_1$



(b) $P(d, \sigma = 1, \theta_2 = \pi/2)$ vs d



(c) $A(d, \sigma = 1, \theta_2 = \pi/2)$ vs d

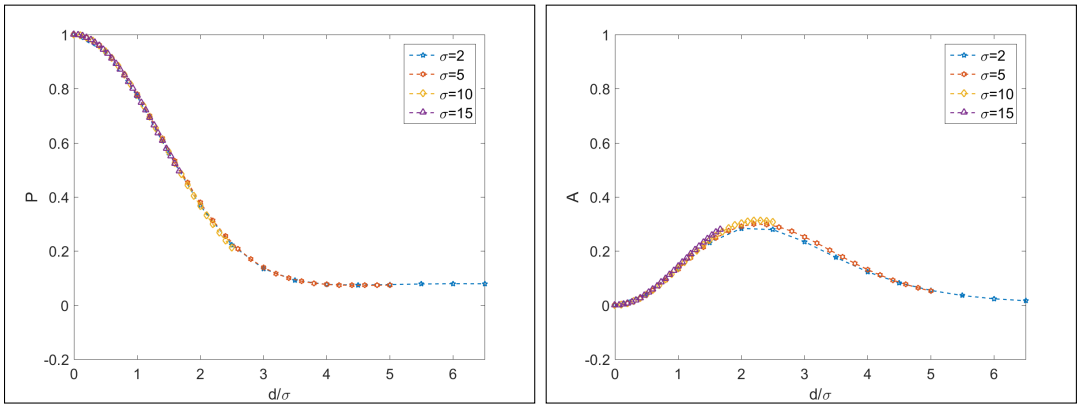
Figure 4.12: Comparison of the behavior of the model over LIVE IQA, Toyama and CSIQ.

4.1.5 Validation on other databases

As an additional way to validate my model, I studied its behavior of on other databases; first on the CSIQ database [94] which contains 30 pristine images and second on the Toyama Database [95] which contains 14 pristine images. As depicted by the example in Fig. 4.12 I obtained a great overlap between the average correlation, amplitude in peak of the different databases. Also I observed small χ^2_ρ, χ^2_P and χ^2_A values between the mean case from the LIVE IQA database and the images from the other databases.

4.1.6 Scale Invariance

Several aspects in the environment are statistically self-similar, meaning that their structure is invariant over multiple scales. An observed property of natural images is the invariance of their statistics with respect to the scale at which the image is observed. For example, the power spectrum of images is invariant to scaling [96], which implies a similar correlation scale-invariance property. Also, many natural structures are scale-invariant [35]. Figure 4.13 plots P and A against the scaled spatial separation for several values of the scale, for the case $\theta_2 = 0$. Excellent alignment of the plots across scales is observed, in agreement with the scale-invariance property, over all θ_2 values. To conserve space, I present the results for only a few scales in Fig. 4.13, however I have observed invariance to also hold over all the other considered scales [2, 15].



(a) (b)

Figure 4.13: Plots of (a) peak function and (b) amplitude correlation function for $\theta_2 = 0$ rad, for several values of scale σ, illustrating the scale invariance of these functions.

4.1.7 White Noise Images

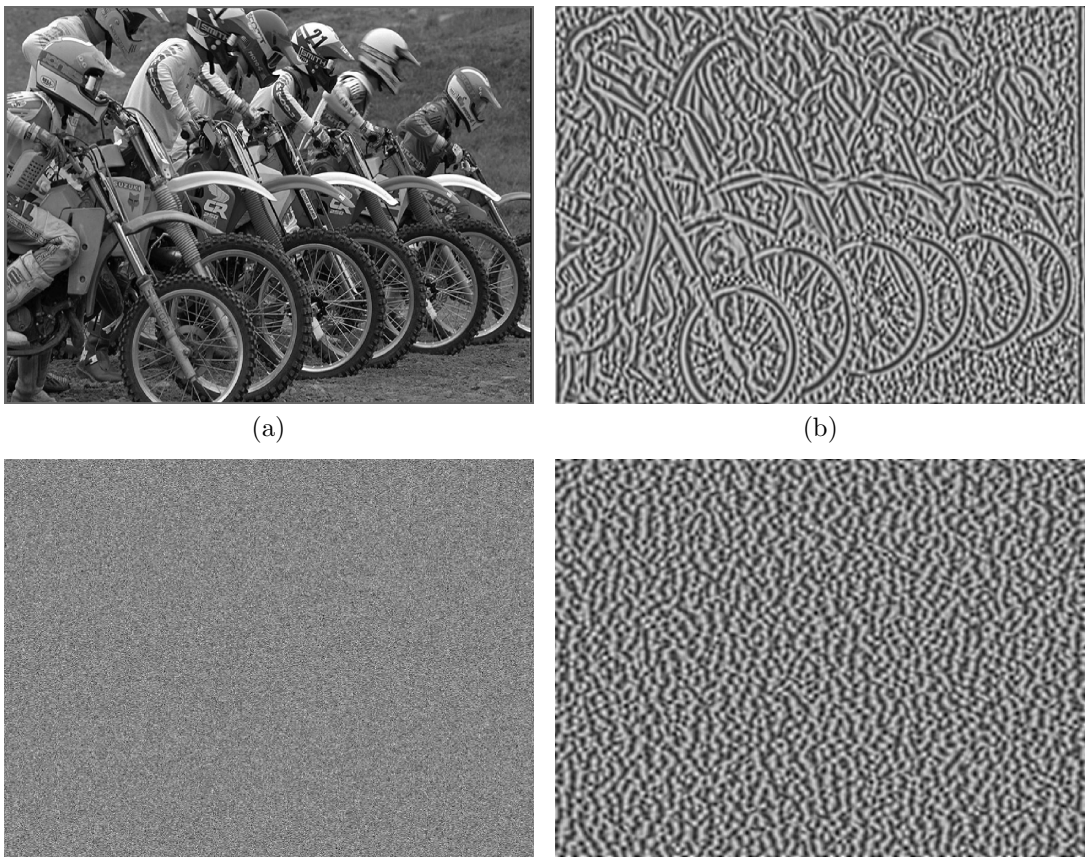
Next I study how my correlation model applies to composed images of simulated white noise (i.e, random matrices). I conduct this analysis both as an experimental control and as a way to better understand the properties of my model. The relative correlation structure of bandpass filtered and normalized white noise against that of natural images is of interest. For example, while the perceptually relevant processing used in my model is known to decorrelate natural images, which are otherwise strongly correlated, instead it introduces correlation on white noise images.

However, the processed white noise images still exhibit less correlation than processed natural images, as may be observed in Fig. 4.15 (a). Note that both the correlation and peak correlation functions of the (bandpass, normal-

ized) processed natural images are everywhere higher than for the processed white noise. Overall, I have found the parametric fits (and associated parameters) to natural images and white noise to be quite different and to obey the ordering observed in Fig. 4.15 (b). This serves not only to validate the unique characteristics of high-quality natural images processed in this manner (like those in the LIVE reference dataset), but also raises the question of how the model applies to distorted images, and how it might be exploited to analyze them. For example, they might be exploited to augment or improve upon existing image quality prediction models and algorithms [5].

4.2 Behavior of the model in the presence of distortions

Distortions lead to consistent changes in the behavior of bandpass image fits to univariate NSS models [9]. Next, I examine how my correlative model behaves in the presence of image distortion. To study this, I applied the model to both reference and distorted images taken from the LIVE IQA database [84]. The database contains images impaired by gaussian blur, JPEG compression, JPEG 2000 compression, fast fading channel noise and additive white noise. I begin by using a simple example image from the LIVE IQA database to demonstrate my observations along the way; image "Woman Hat" is shown in Fig. 4.16(a). While I restrict ourselves to commenting on "Woman Hat," I have observed very similar results on distorted versions of all the other LIVE IQA images.



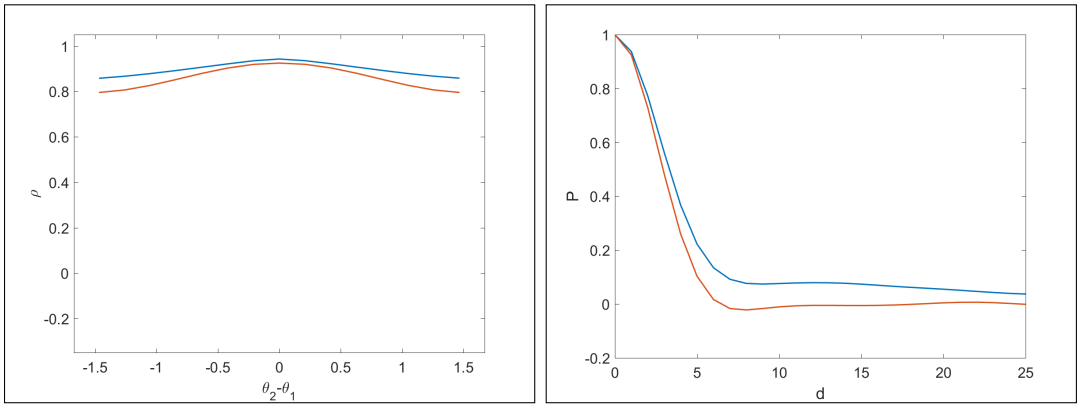
(a) (b)

(c) (d)

Figure 4.14: A sample natural image (a) before and (b) after bandpass filtering and normalization. Similar for white noise image (c) and processed version of it (d).

4.2.1 Impact of Distortions on Correlation, Amplitude and Peak

I begin by visualizing the correlation, peak and amplitude functions as they are modified by distortion. Figures 4.17-4.19 show plots of ρ, P, and A, respectively, on the images modified by the distortions. The observations that will be drawn from the presented examples are generalized across other scales,



(a) (b)

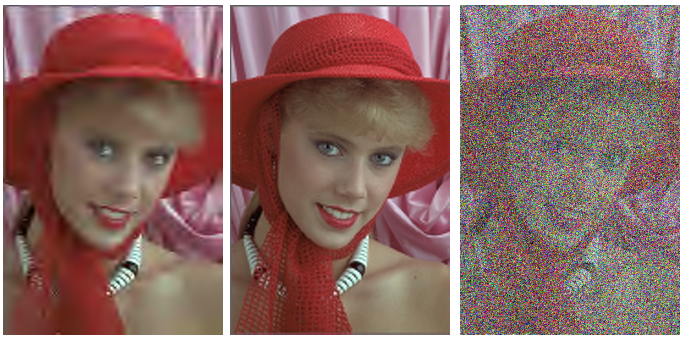
Figure 4.15: Graphs of (a) correlation function ρ plotted against $\theta_2 - \theta_1$ for $\sigma = 2$, $d = 1$, and $\theta_2 = 0$ (b) peak correlation function P plotted against d for $\sigma = 2$ and $\theta_2 = 0$. Each plot shows the result of processing natural images (in blue) and white noise (in red).

σ, and spatial orientation θ_2.

4.2.1.1 Blur

Increases in blur were produced by increasing the space constant of the applied gaussian filter σ, which generally leads to worsening degradation of the perceptual image quality, which are reflected in drops in the Structural Similarity Index (SSIM) [65] between the blurred images and the undistorted original values. As expected, blur leads to an increase in the correlation functions of the bandpass normalized images, as can be seen in Fig. 4.17(a). The reductions of detail and diversity as a consequence of low-pass smoothing (Fig. 4.16(b)) progressively increases the correlation as the filter bandwidth

(a) Distortion Free (b) Blur (c) JPEG



(d) JPEG 2000 (e) Fast Fading (f) White Noise

Figure 4.16: Image "Woman Hat" and several distorted versions of it.

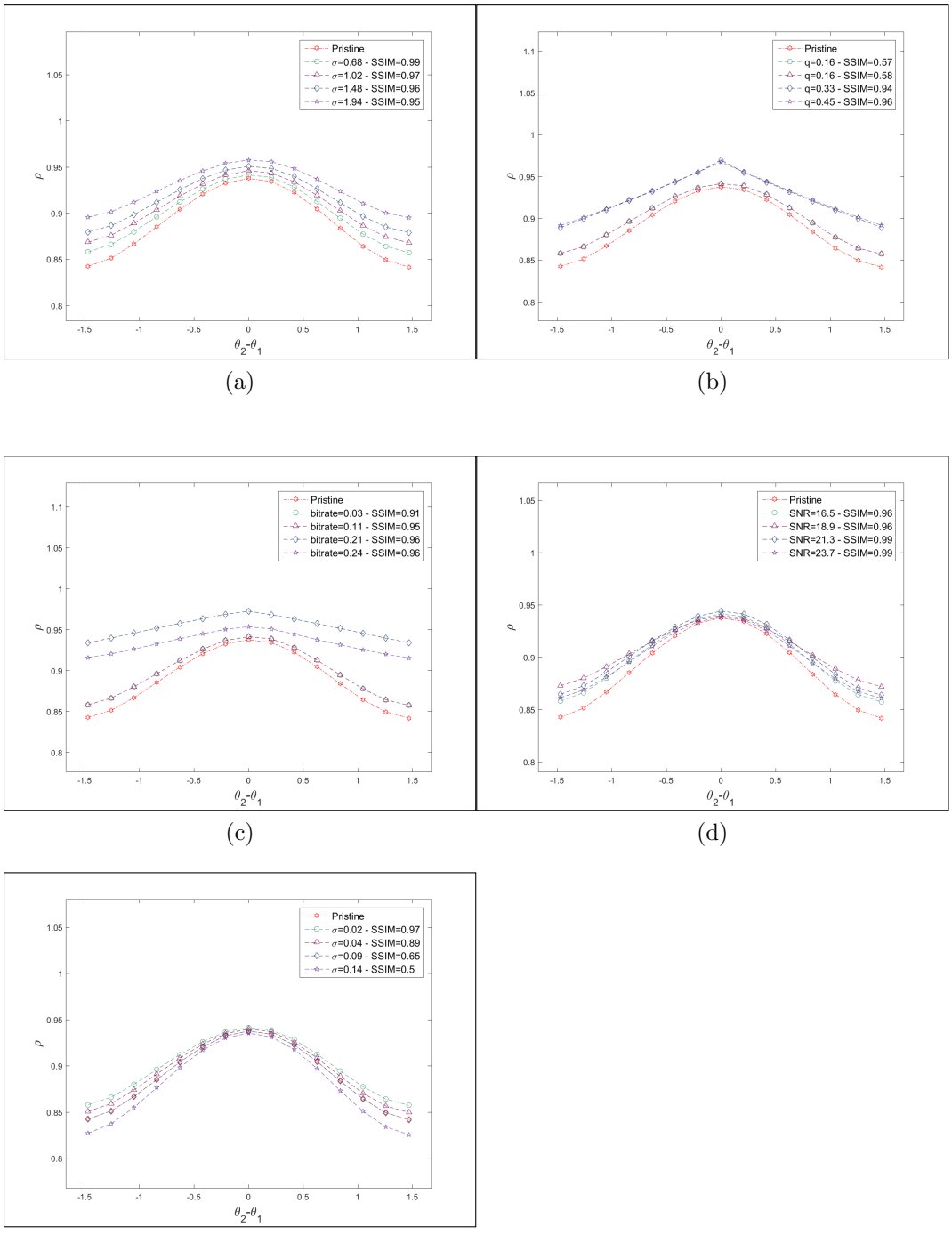


Figure 4.17: Plots of the correlation function of image "Woman Hat" subject to (a) blur; (b) JPEG; (c) JPEG 2000; (d) Fast fading; (e) White noise.

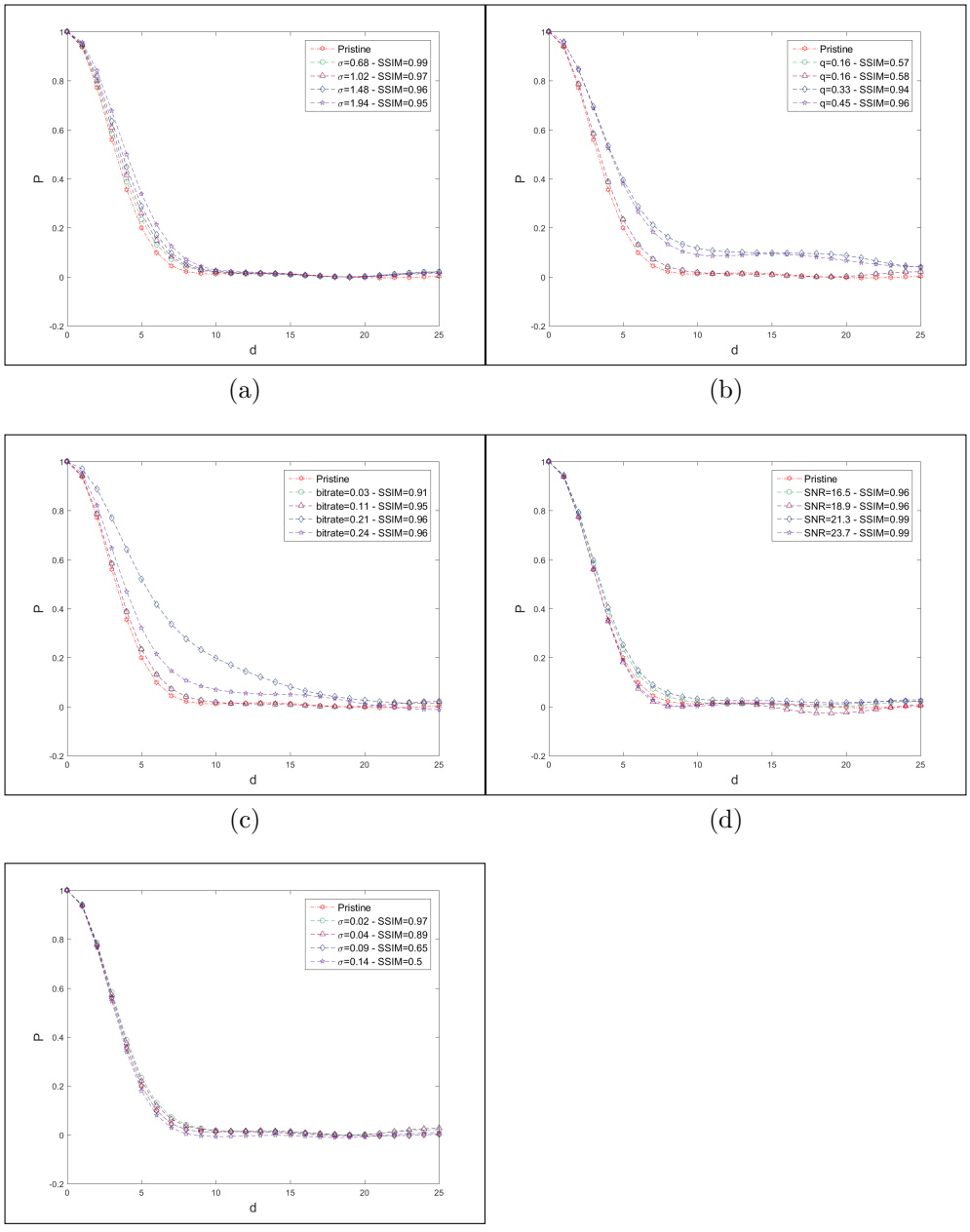
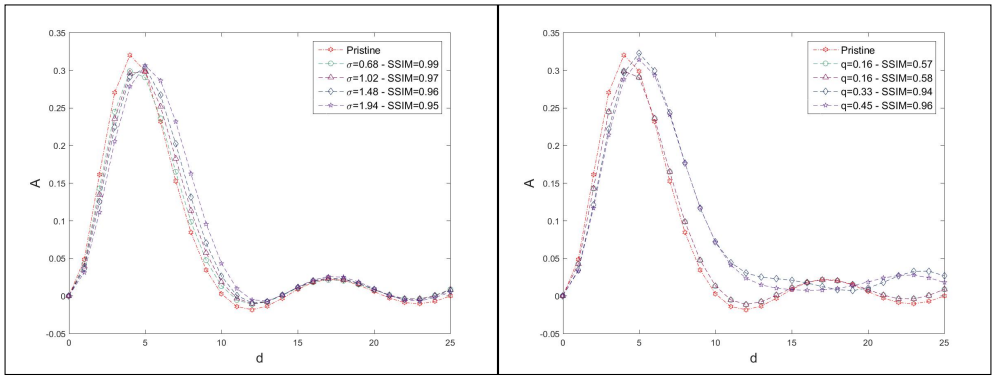
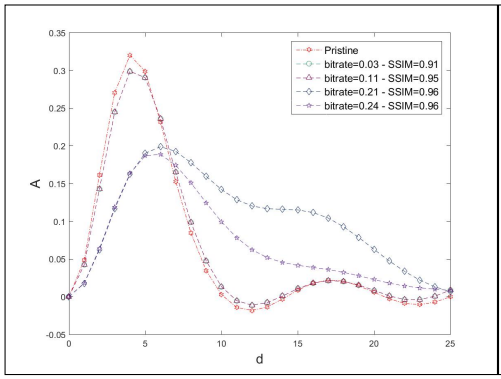


Figure 4.18: Plots of the peak function of image "Woman Hat" subject to (a) blur; (b) JPEG; (c) JPEG 2000; (d) Fast fading; (e) White noise.

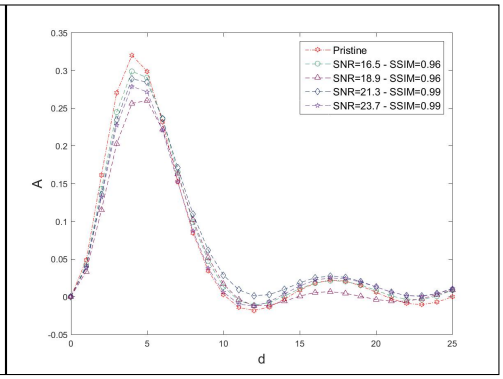


(a)

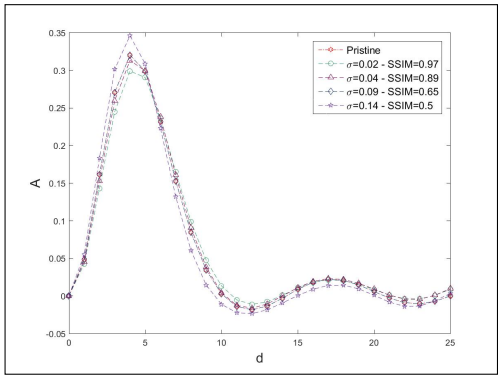
(b)



(c)



(d)



(e)

Figure 4.19: Plots of the amplitude function of image "Woman Hat" subject to (a) blur; (b) JPEG; (c) JPEG 2000; (d) Fast fading; (e) White noise.

is decreased. The increase in the correlation is monotonic with the level of blur. Furthermore, at small spatial separations, the values of the peak and amplitude functions increase, as can be seen in Fig. 4.18(a) and Fig. 4.19(a), respectively.

4.2.1.2 JPEG

Increases in JPEG compression is controlled by decreases in the JPEG coefficient quantization q, which in turn leads to reduction of the SSIM values between compressed and original images. JPEG distortion also leads to an increase in the computed correlation function since it causes both over-smoothing and blocking artifacts, and hence greater degrees of local homogeneity, as may be observed in Fig. 4.16(c). The increase in the correlation is monotonic, as may be seen by comparing the plots in Fig. 4.17(b). As a result, the peak and amplitude functions also increase in value, as in Fig. 4.18(b) and in Fig. 4.19(b) respectively. This increase is not limited to small spatial separations, unlike the case with blur. The values of A and P remain high even for larger separations when measured on heavily compressed images.

4.2.1.3 JPEG 2000

Ringing and blur are two common artifacts that afflict JPEG 2000 compressed images as may be observed in Fig. 4.16(c). Generally, the correlation is increased as shown in Fig. 4.17(c). However this increase is not monotonic with increased compression. This may be partially explained by the fact that

multiple parameters control JPEG 2000, such as the assigned weighting. The LIVE IQA database [84] does not indicate the weighting used on each image. P and A are impacted in ways similar to JPEG, as may seen in Fig. 4.18(c) and Fig. 4.19(c), respectively. The peak and amplitude functions in these cases exhibit bumps that are possibly caused by the ringing distortions.

4.2.1.4 Fast Fading

The fast fading category in LIVE IQA [84] is a complex, difficult distortion that is modeled as JPEG-2000 compression followed by fast fading bit errors. It also leads to increases in the correlation functions shown in Fig. 4.17(d). The behavior is not entirely monotonic owing to the complexity of the distortion. Generally, however, there is a resemblance in the correlation plots of JPEG 2000 and fast fading channel noise, since both contain JPEG 2000 compression artifacts (Fig. 4.16(d)). However, the increases in the peak and amplitude values are less subtle as compared to JPEG 2000, as depicted in Fig. 4.18(d) and in Fig. 4.19(d). This is because low compression (2.5 bits per pixel) was used to generate the JPEG 2000 distortion on all of the fast fading data, leading to less harsh ringing or blur artifacts as compared to the pure JPEG 2000 distortions.

4.2.1.5 White Noise

White noise of standard deviation σ was added to the R, G and B components. This leads to a decrease in the SSIM values. As a general trend,

white noise leads to a decrease in the correlation functions, as expected. The peak correlation function is not impacted, as shown in Fig. 4.18(e). The amplitude functions appears to absorb most of the variation, as seen in Fig. 4.19(e), where the amplitude at small distances is higher. The exception to this general observation occurs at small standard deviations. In this case, the correlation slightly increased.

4.2.2 Impact of Distortions on Model Parameters α_0, β_0, α_1, β_1, α_2, and β_2

Understanding how the values of α_0, β_0, α_1, β_1, α_2, and β_2 are impacted as function of the distortions is less straightforward. It is not clear yet how changing trends in the different parameters impact A, P, and ρ. Figure 4.20 shows these parameters against the various considered distortions. In the presence of distortions, the distributions of the values of these different parameters are modified. Some parameters seem to respond to distortions better than others; by comparing the boxplots of α_2 and β_1, for example, it may be observed that the distribution of values of α_2 changes more drastically than does the distribution of values of β_1. In the near future, I will describe ways to use the parameters α_0, β_0, α_1, β_1, α_2, and β_2 as features to build correlation-based models that are able to automatically assess the perceptual quality of images.

Towards exploring the utility of correlation features for a wide array of possible distortion-sensitive applications (quality assessment, denoising, de-

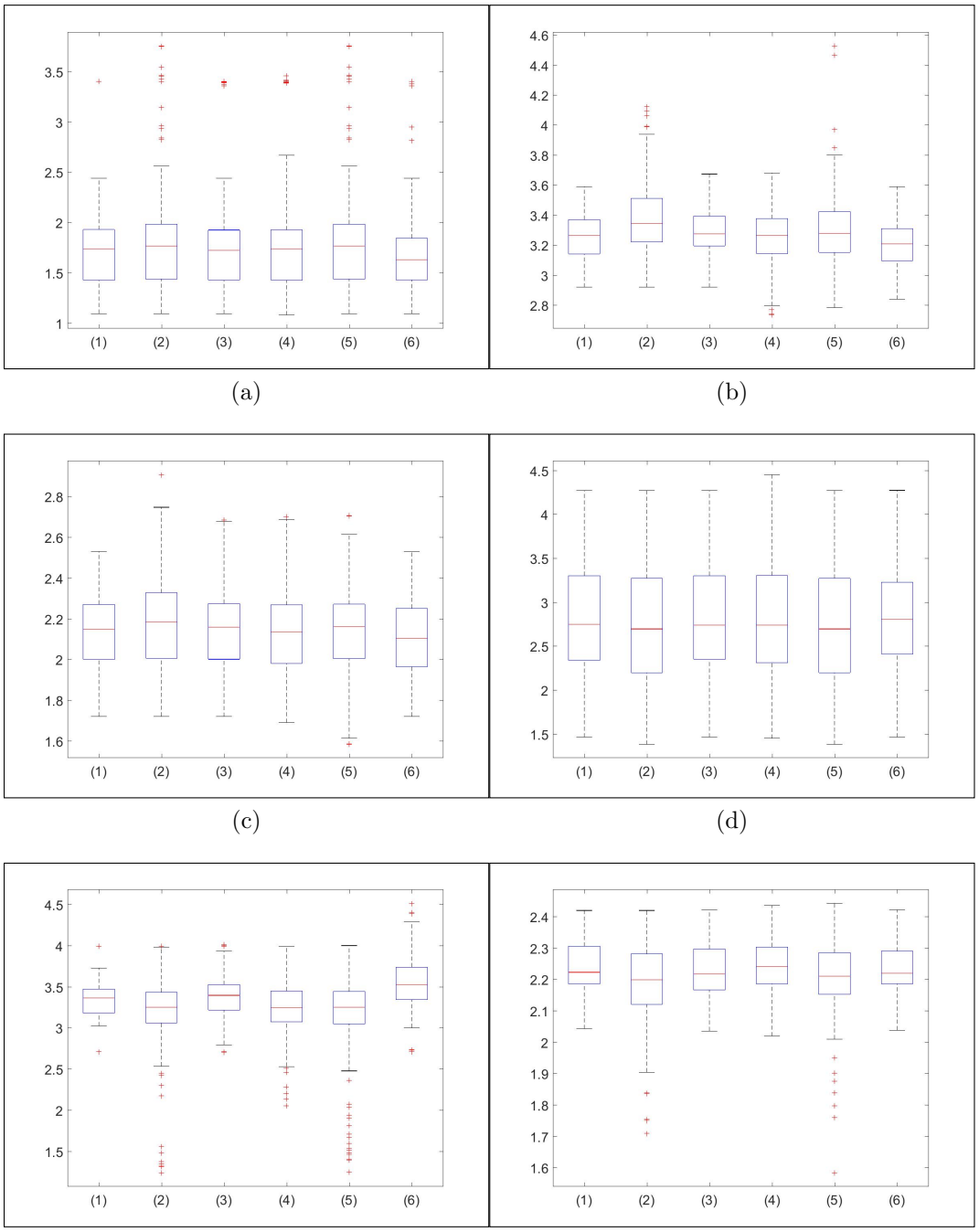


Figure 4.20: Boxplots of the different parameters (a) α_0; (b) β_0; (c) α_1; (d) β_1; (e) α_2; and (f) β_2 for the various applied controlled distortions; (1) Undistorted; (2) Blur; (3) JPEG; (4) JPEG 2000; (5) Fast Fading; (6) White Noise.

blurring, deblocking, etc), I built a system to classify images by distortions. To do this, I focused on the set of distortions that are common to the LIVE IQA database [84] and the TID database [97]: JPEG 2000, JPEG, White Noise and Gaussian Blur. I partitioned the data into random 80%-20% training-testing splits, on which I trained a Support Vector Machine (SVM) [98] to classify images by distortions. I used $\{\alpha_0, \beta_0, \alpha_1, \beta_1\}$ at $\theta_2 = 0$ and $\theta_2 = \frac{\pi}{2}$ as training features (8 features). The experiment was repeated over 100 iterations, yielding a median correct classification rate of 70% on the LIVE IQA database [84] and 71% on the TID database [97]. These are good results using only a sparse set of second-order features. Including simple first-order (univariate) NSS features such as the shape and the variance parameters of the Mean Subtracted Contrast Normalized (MSCN) coefficients [7] improves the classification accuracy. Using the shape and variance parameters at scale 1 and the shape parameter at scales 2 (as defined in [7]) increased the correct classification rate to 85% on the LIVE IQA database [84] and to 86% on the TID database [97], using a total of 11 features. Details regarding the individual distortion classification performance and the inter-class accuracies are given in Table 4.6.

4.3 Concluding Remarks

In this chapter I built a simply, parametric bivariate natural scene statistic model of images and demonstrated its validity on a well-known set of high quality images. My new model is global and is able to accurately capture

Table 4.6: Distortion Classification Performance

	JPEG2000	JPEG	White Noise	Gaussian Blur	Overall
Bivariate Model (LIVE IQA)	69%	67%	71%	75%	70%
Univariate + Bivariate Model (LIVE IQA)	85%	78%	94%	88%	85%
Bivariate Model (TID)	75%	50%	90%	74%	71%
Univariate + Bivariate Model (TID)	77%	80%	100%	90%	86%

the correlation behavior of natural images as well distorted images. Moving forward, I will be presenting in the next chapter a new blind image quality prediction model and 3D visual discomfort predictor that makes use of the bivariate NSS features presented above.

Chapter 5

The Bivariate NSS Model as a Tool to Tackle Image Quality Problems

In this chapter, I use the bivariate NSS model I built to tackle the blind image quality (IQA) and 3D visual discomfort for stereo images. [1]

5.1 Blind IQA Predictor

Digital images have witnessed tremendous growth as a medium for representation and communication. Since human observers are the ultimate receivers of the visual information in images, subjective experiments using human observers remains the most reliable way to assess the quality of an image. Given that 1.3 trillion images were captured in 2017 [99], relying on human observers to assess picture quality is unrealistic. Building models that predict the quality of images in accordance with human observers is a more

[1] This chapter appears in the following papers:
1) Zeina Sinno, Constantine Caramanis, Alan C. Bovik: "Second Order Natural Scene Statistics Model of Blind Image Quality Assessment" in the IEEE International Conference on Acoustics, Speech, and Signal Processing: 1238-1242 (2018).
2) Zeina Sinno, Alan C. Bovik: "Predicting 3D visual discomfort using natural scene statistics and a binocular model" in the International Society for Optics and Photonics - Applications of Digital Image Processing XLI Vol. 10752, p. 107520G, (2018).
Zeina Sinno has designed the two described models and performed their full experimental analysis.

feasible solution to this problem. The study of blind (no-reference) IQA models involves building learned predictors that deploy low-level image descriptors as inputs. Many models have been developed that extract distortion specific features [100, 101], and [102], while others train learning machines on NSS features computed from distorted images. Examples of this approach include [7], and [103]. Other notable models, such as Ye *et al.* [104] learn visual code words predictive of image quality, and the completely blind model [8], which measures a distance between distorted and pristine NSS features, without requiring any training on either distorted images or on human opinion scores. Saha *et al.* [105] used visibility measured over multiple scales to predict picture quality.

The univariate statistics of bandpass-filtered images provide powerful features that drive many successful IQA algorithms [7,8,106]. Motivated by the observation that distortions lead to systematic and predictable perturbations of my correlation models' features, it is natural to consider whether they can be used to predict the perceptual quality of images.

I combined univariate and bivariate NSS features to build a no-reference IQA model that strongly competes with existing models.

5.1.1 Model Features

I begin this section by presenting the considered bivariate features.

5.1.1.1 Bivariate Features

I studied the quality-predictive efficacies of the parameters $\{\alpha_0, \beta_0, \alpha_1, \beta_1, \alpha_2, \beta_2\}$ as defined in (4.9) and (4.10) over multiple spatial angles θ_2. I found α_0, β_0 α_1, and α_2 to be the most responsive to distortion, and hence the most useful for quality prediction. I also found that only using parameters computed along the horizontal, vertical, and diagonal was sufficient; adding more angles did not further boost performance.

I noticed that applying non-linear operations to these raw features boosts performance. Denote by $\{F_1,F_2,...,F_8\}$ the quality predictive features derived from my correlation model. Noting that $\alpha_i(\theta_2)$, $\beta_i(\theta_2)$; $i = 1, 2, 3$ are all functions of θ_2, then define $F_1 = {\beta_0}^{-1}(0)$; $F_2 = {\beta_0}^{-1}(\frac{\pi}{4})$; $F_3 = {\beta_0}^{-1}(\frac{\pi}{2})$; $F_4 = {\beta_0}^{-1}(\frac{3\pi}{4})$ where $\beta^{-1} = \frac{1}{\beta}$; and also $F_5 = {\alpha_0}^{-1}(0){\alpha_0}^{-1}(\frac{\pi}{2})$; $F_6 = {\beta_0}^{-1}(0){\beta_0}^{-1}(\frac{\pi}{2})$; $F_7 = {\alpha_1}^{-1}(0){\alpha_1}^{-1}(\frac{\pi}{2})$; and $F_8 = {\alpha_2}^{-1}(0){\alpha_2}^{-1}(\frac{\pi}{2})$.

As a first test, I trained a Support Vector Regression (SVR) model using a radial basis function and 80-20 split on the LIVE Challenge images using 8 features as input, obtaining Pearsons linear correlation coefficient (PLCC) and Spearmans rank ordered correlation coefficient (SROCC) both in the range of 0.3. Clearly, taken alone the bivariate features are insufficient predictors of picture quality. However, I find that they are usefully complementary to existing univariate features for the IQA task.

5.1.1.2 Mean Subtracted Contrast Normalized Features

Rather than using my model features in isolation, I combine them with univariate NSS that have been used successfully for blind picture quality prediction. The motivation behind this reasoning is that the bivariate correlation model is not standalone, rather it extends existing univariate NSS models and completes a bivariate density model. I computed the mean subtracted, contrast normalized (MSCN) coefficients as used in the BRISQUE model [7]. Given an image I, process luminances via local mean subtraction and divisive normalization, similarly to (4.3). Mittal *et al.* [7] showed that those MSCN coefficients are disturbed by the presence of distortion. I extracted the arithmetic mean $\dot{\mu}$, sample kurtosis κ, and skewness γ, from the luminance image to obtain ($\dot{\mu}_L$, κ_L, γ_L), and form the chrominance component a* from the CIELAB space ($\dot{\mu}_a$, κ_a, γ_a), and used them as additional features in my predictor.

Furthermore, [7] showed that the histogram of the MSCN coefficients of both pristine and distorted images are modeled as fitting a zero-mean generalized gaussian density (GGD):

$$f(x;\phi;\gamma^2) = \frac{\phi}{2\eta\Gamma(1/\phi)} \exp[-(\frac{|x|}{\eta})^{\phi}] \tag{5.1}$$

where

$$\eta = \gamma\sqrt{\frac{\Gamma(\frac{1}{\phi})}{\Gamma(\frac{3}{\phi})}} \tag{5.2}$$

and $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(a) = \int_0^{\infty} t^{a-1}e^{-t}dt \quad a > 0. \tag{5.3}$$

The shape parameter ϕ controls the shape of the distribution, while η controls its variance. I used the moment matching approach to estimate these two parameters from the histograms of each considered image's MSCN coefficients [107].

Denote by ϕ_L and ι_L the shape and the variance features at scale 1 of the luminance component, and by ϕ_Y and ι_Y the scale and shape parameters of scale 1 from the yellow color channel component. The yellow color channel component is computed on an RGB image I as:

$$Y = \frac{R+G}{2} - \frac{|R-G|}{2} - B \tag{5.4}$$

Sinno [108] *et al.* observed that the height of the peak at zero of the histogram of the MSCN coefficients is highly correlated with how well exposed the image is. A small peak indicates that the image is well-exposed, whereas a high peak means that the image is poorly-exposed (under exposed or over exposed). Furthermore, they used this information to correct for underexposed and overexposed regions in an image using Laplacian pyramid fusion of multiple shots of the same scene, but varying in exposure. I used the peak at zero of the histogram of MSCN coefficients as a feature in my model, and denote it by δ.

I also considered the pairwise products of neighboring MSCN coefficients along four orientations (H), vertical (V), main-diagonal ($D1$) and secondary-diagonal ($D2$), similarly to [7]. As shown in [7], the histograms

of the pairwise MSCN coefficients are well modeled as asymmetrical generalized gaussian distributed (AGGD):

$$f(x;\nu,\eta_l^2,\eta_r^2) = \begin{cases} \frac{\nu}{(\eta_l+\eta_r)}\exp[-(\frac{-x}{\eta_l})] & \text{x} < 0 \\ \frac{\nu}{(\eta_l+\eta_r)}\exp[-(\frac{-x}{\eta_r})] & \text{x} \geq 0 \end{cases} \tag{5.5}$$

where

$$\eta_l = \iota_l\sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}} \tag{5.6}$$

$$\eta_r = \iota_r\sqrt{\frac{\Gamma(\frac{1}{\nu})}{\Gamma(\frac{3}{\nu})}}. \tag{5.7}$$

The shape parameter ν controls the 'shape' of the distribution while η_l^2 and η_r^2 are scale parameters that control the spread on each side of the mode, respectively. The parameters (ν,η_l^2,η_r^2) are also estimated using moment-matching [109]. Next, I also created a reduced resolution version of the luminance image by low pass filtering followed by downsampling by a factor of two, then followed the same procedure as above to obtain (ν,η_l^2,η_r^2) at the new scale. In my predictor, I used (ν,η_l^2) as features over the four orientations H, V, $D1$, and $D2$, which I denote by $(\nu_H,{\eta_l}^2{}_H)$, $(\nu_V,{\eta_l}^2{}_V)$, $(\nu_{D1},{\eta_l}^2{}_{D1})$ and $(\nu_{D2},{\eta_l}^2{}_{D2})$. This yields 8 additional features.

Combining the correlation features $F_1 - F_2$ with the MSCN features yields a total of 27 features, as summarized in Table 5.1.

5.1.2 Quality Evaluation

As a resource to learn a blind IQA model using my model, I used the recent LIVE in the Wild Image Quality Challenge Database ("LIVE Chal-

Table 5.1: Features used in the bivariate image quality prediction model.

Feature ID	Feature Description
F_1 - F_4	${\beta_0}^{-1}(0),\ {\beta_0}^{-1}(\frac{\pi}{4}),\ {\beta_0}^{-1}(\frac{\pi}{2}),\ {\beta_0}^{-1}(\frac{3\pi}{4})$
F_5	${\alpha_0}^{-1}(0){\alpha_0}^{-1}(\frac{\pi}{2})$
F_6	${\beta_0}^{-1}(0){\beta_0}^{-1}(\frac{\pi}{2})$
F_7	${\alpha_1}^{-1}(0){\alpha_1}^{-1}(\frac{\pi}{2})$
F_8	${\alpha_2}^{-1}(0){\alpha_2}^{-1}(\frac{\pi}{2})$
F_9 - F_{11}	$\dot{\mu}_L,\ \kappa_L,\ \gamma_L$
F_{12} - F_{14}	$\dot{\mu}_a,\ \kappa_a,\ \gamma_a$
F_{15}	δ
F_{16} - F_{17}	$\phi_L,\ \iota_L$
F_{18} - F_{19}	$\phi_Y,\ \iota_Y$
F_{20} - F_{21}	$\nu_H,\ \eta^2_{l\,H}$
F_{22} - F_{23}	$\nu_V,\ \eta^2_{l\,V}$
F_{24} - F_{25}	$\nu_{D1},\ \eta^2_{l\,D1}$
F_{26} - F_{27}	$\nu_{D2},\ \eta^2_{l\,D2}$

lenge") [54]. This database contains 1162 images captured using mobile devices. This database is a unique and difficult test of blind IQA predictors. Using a regression module, I constructed a mapping from the feature space (Table 5.1) to human ratings, resulting in a measure of image quality. I used a support vector regressor (SVR) [98] that has been successfully deployed in many prior image quality models [7, 9]. I used the LIBSVM package [110] to implement the SVR with a radial basis function (RBF) kernel and to predict the MOS scores. I split the images randomly and used 80% of it for training and the rest for testing, then I normalized my features, and fed them into the SVR module to predict the MOS score. I repeated the process 50 times. I obtained a median Pearsons linear correlation coefficient (PLCC) of 0.73 and a Spearmans rank ordered correlation coefficient (SROCC) of 0.69 against MOS.

Table 5.2 compares the performances of various reported algorithms. With 27 features only, my correlation-enhanced model was able to outperform the other leading models, demonstrating the power of the bivariate features. The performance of my model was only approached by FRIQUEE, which uses a large number of features (more than 20× as many). Notably, the correlation features substantially boosted the performance of simple BRISQUE [7]. An interesting extension will be to apply the model to temporal pictures, such as video frame differences which present highly regular statistical structures [68].

Table 5.2: Comparison of Image Quality Models on the LIVE Challenge Database.

	Number of Features	PLCC	SROCC
Bivariate Model	27	**0.73**	0.69
FRIQUEE [111]	584	0.72	**0.72**
BRISQUE [7]	36	0.61	0.60
C-DIVIINE [112]	82	0.66	0.63
DIVIINE [9]	88	0.56	0.51
BLIINDS-II [106]	**24**	0.45	0.48
NIQE [8]	36	0.48	0.42

I also tested the performance of my model on the LIVE IQA database [6]. The results are summarized in Table 5.3. My predictor also outperformed on this database too.

I also performed the p statistical significance test on the different groups of features used by my predictor and I was able to verify that the features deliver statistically significant superior performance. As an additional test, I

Table 5.3: Comparison of Image Quality Models on the LIVE IQA Database.

	Number of Features	PLCC	SROCC
Bivariate Model	27	**0.96**	**0.96**
FRIQUEE [111]	584	0.93	0.95
BRISQUE [7]	36	0.94	0.94
C-DIVIINE [112]	82	0.94	0.95
DIVIINE [9]	88	0.93	0.92
BLIINDS-II [106]	**24**	0.92	0.91
NIQE [8]	36	0.92	0.91

removed each group of features in my predictor to understand their individual contributions. Removing the BRISQUE derived luminance based features had the greatest impact on performance, followed by my bivariate NSS features. This is not unexpected, because the univariate NSS model in BRISQUE is complemented by my bivariate NSS correlation model.

5.2 3D Visual Discomfort Predictor

For the case of 2D images, NSS models allowed us to understand how the HVS can efficiently process gigantic amounts of visual data [1]. Recent efforts has been directed towards understanding the joint statistics of multiple pixels [24, 25]. It has been established previously in the previous chapter and in [113] that such statistics can be captured in closed-form for luminance 2D images, as well as for the case of chromatic images [114]. I also just demonstrated that such features can be used to derive new methods for predicting the quality of images [115].

Here I describe how the bivariate NSS model [113] can be applied to

3D images. But before doing so, I will start with an overview of the 3D visual discomfort problem to provide the necessary background.

5.2.1 Overview of the 3D Visual Discomfort Problem

The positioning of the two eyes in the front of the head, horizontally aligned but separated, allows them to obtain slightly different retinal images. The brain combines the left and the right images to obtain a fused, 'cyclopean' image. Based on the distance between corresponding points in the left and right images, disparity information is extracted and depths are computed. This ability facilitates a variety of exteroceptive and visuomotor tasks [116], especially in the comprehension of complex visual presentations and those requiring hand-eye coordination [117].

The binocular processing centers of the brain capture differences between the left and the right images obtained from the eyes. For humans to perceive depth correctly, the two images need to align closely. If for some reason they do not, then visual discomfort may be experienced [118]. 3D visual discomfort can take several different symptoms, including eye strain, nausea, fatigue and headaches [119]. There are several explanations of experienced visual discomfort when viewing stereo displays, including the eyewear required to present images to the two eyes, ghosting or cross-talk between the images, misalignment of the images, inappropriate head orientation, vergence-accommodation conflicts, visibility of flicker or motion artifacts, and visual-vestibular conflicts [118]. The vergence-accommodation conflict has often been

identified as the primary culprit causing visual fatigue [120, 121].

Vergence describes the mechanism of binocular eye-movement that directs the eyes towards an object. When fixation on an object moves closer or farther, the eyes converge or diverge, respectively. Accommodation describes the mechanism of adjusting the focal power of the crystalline eye lens to acquire a clear and sharp retinal image of an object. As the object moves closer or farther, the focal power increases or decreases respectively [122]. In a natural environment, both mechanisms are coupled and occur in parallel. More importantly, the amount of accommodation required to put an object into focus is proportional to the amount of of vergence needed to fixate on the object [123]. The human visual system (HVS) has evolved towards associating these processes neurologically; triggering vergence stimulates accommodation, and vice versa. Stereoscopic displays stimulate accommodation and vergence in an unnatural way, resulting in vergence-accommodation conflicts.

Over the past decade, safety and health issues related to stereo images and videos have been well studied. A significant aim for 3D camera acquisition makers and display manufacturers is to characterize the visual discomfort of stereo images accurately in an attempt to reduce it or eliminate it. Several efforts have been made in the literature to create such models. Early on, Nojiri *et al.* [124, 125] found a close correlation between the range of parallax distribution and the degree of visual discomfort. In particular, they found that the reconstructed scene should be positioned behind the screen to deliver a more comfortable viewing experience. Yano *et al.* [126] measured the degree

of visual fatigue from the change of accommodation response before and after viewing stereoscopic images. Choi *et al.* [127] used a Principal Component Analysis (PCA) approach to understand factors contributing to visual fatigue in stereoscopic videos including spatial complexity, depth position, temporal complexity, scene movement, depth gradient, crosstalk, and brightness. Kim *et al.*'s model [128] characterized horizontal and vertical disparities. In [129], Park *et al.* described a predictor which combines features from a neural population coding model with the statistics of horizontal disparity maps. Kim *et al.* in [130] described a more advanced second-order system model that forms a transfer functions integrating information about the optical nerve, the accommodation and vergence neural pathways, the oculomotor plant, and visual area MT. In [131], Oh *et al.* constructed different maps and extracted their features to create their model. The considered maps are the degree of out-of-focus map starting from the focal distance, the Panum's fusional area map representing how well the 3D object is fused, the stereoscopic map representing the output responses induced by processes of accommodation and vergence, and the conflict response map which accounts for the disagreement between the response when viewing stereo images on a flat screen vs in a natural environment. In [132], Park *et al.* proposed a model which accounts for both accommodation and vergence to predict quality, by making use of the physiological optics of binocular vision and foveation.

The models described above are all perception based. Deep convolutional neural networks (CNN) have also recently been applied to the problem.

Very recently the authors in [133] used a CNN which is fed a disparity map to predict visual discomfort. Using NSS as a tool to assess visual discomfort has not been exploited yet. NSS has proved to be a powerful strategy for gaining insight into the HVS by measuring and analyzing the physical regularities of the natural environment, as the HVS has evolved based on the natural environment that we perceive. The power of this approach is that by characterizing the physical regularities in the visual environment, then we can gain insight into how those regularities could be exploited to perform visual tasks.

To be able to exploit the bivariate NSS model to tackle the 3D visual discomfort problem, I first look at the bivariate NSS of disparity maps, in attempt to first observe whether the model still holds in that case and to understand how those statistics correlate with visual discomfort. My hypothesis is that the statistics of natural disparity maps would likely not cause visual discomfort. I extract NSS features of disparity maps, and combine them with a subset of the features of the binocular model in [132] to create a new 3D visual discomfort predictor.

5.2.2 Bivariate NSS Modeling of Disparity Maps

The IEEE-SA database [134] was used as a basis for developing my model. The database contains 160 scenes, captured with different disparity ranges, resulting in 800 S3D images pairs, each associated with a Mean Opinion Score (MOS). The resolution of all the images is 1920×1080. The content of this database is diverse, containing a wide variety of objects. The scenes were

captured indoors and outdoors. A flow chart of the model is presented in Fig. 5.1.

5.2.2.1 Applying the Bivariate NSS Model on Stereo Images

The input to my model are left and right stereo pairs. The HVS infers depth information from the observed left and right images. To extract this information, I computed disparity maps. I used a classical technique described in [135, 136] to obtain a single disparity map for every image pair. Moving forward, the processing steps were similar to the case of the visible light images but with slight modifications in the parameters.

I applied steerable filtering on the disparity maps while varying $\sigma \in \{2, 3, \ ..., 7\}$ instead of $\sigma \in \{1, 2, 3, \ ..., 15\}$, as it was the case for visible light images. This is because I observed that adding more scales did not improve the results, hence this allowed me to save on the computation time. The frequency tuning orientations of the steerable filters $\theta_1 \in [0, \pi/15, 2\pi/15, ..., \pi]$, remained the same as previously. At the end of this step, I obtained 90 bandpass responses for every disparity map. Next, divisive normalization was applied similarly to the case of visible light images. I performed the correlation modeling similarly to the case of visible light images but I limited the values of δ_x and δ_y over the integer range from 0 and 19 instead of 0 and 25. This is justified by the fact that disparity maps contain less structure as compared to visible light images, so limiting the value of d , which represents the spatial separation between the two windows getting correlated to a smaller value is more ade-

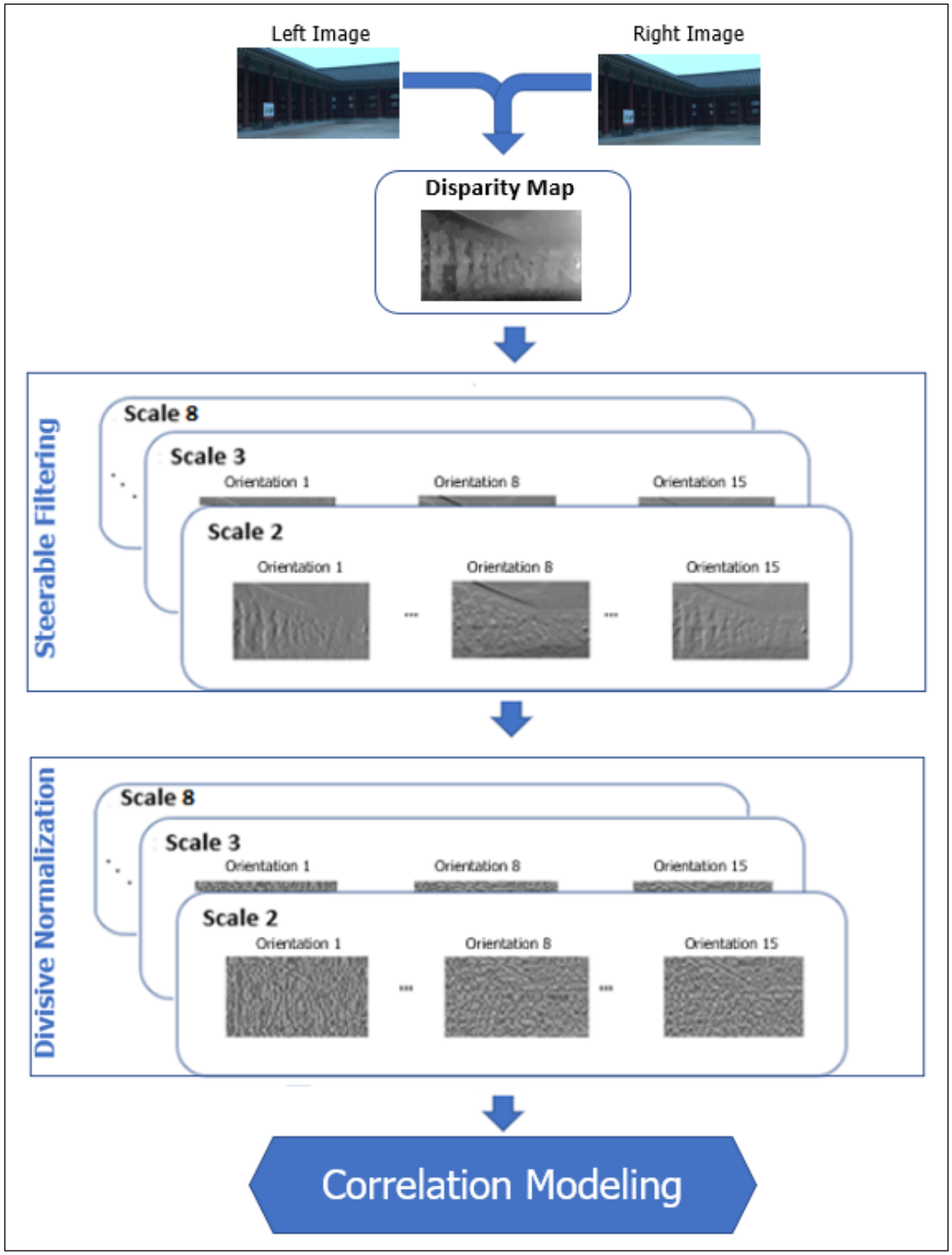


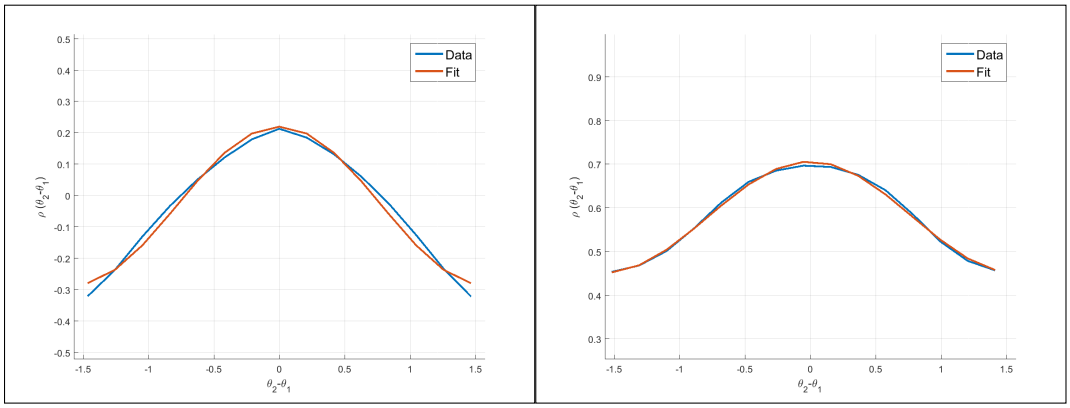
Figure 5.1: Flow chart of the 3D Visual Discomfort Predictor.

quate. As a result, d took values between 0 and $\sqrt{19^2 + 19^2} = \sqrt{722} = 26.87$. I considered the 8 most frequent values of θ_2 : $[0, 0.785, 1.570, 2.356]$ occurring 19 times in the considered window and $[0.436, 1.107, 2.034, 2.677]$ occurring 9 times there. I excluded the less frequent values of θ_2, because including those does not add too much value. The computations of the correlation ρ, peak P and amplitude A were performed as before, as well as the optimization operations to obtain the values of $\{\alpha_0, \beta_0\}$ to reconstruct P and the values of $\{\alpha_1, \beta_1, \alpha_2, \beta_2\}$ to reconstruct A. Fig. 5.2, Fig. 5.3 and Fig. 5.4 present sample correlation, peak and amplitude plots respectively obtained from the data and their corresponding fits, from the stereo pair ISS1-0-L.png and ISS1-0-R.png from the IEEE-SA database.

The great overlap between the empirical correlation data, its fit and its reconstruction demonstrates the validity of my model for the case of disparity maps as well. Similar observations can be made across all the images of the IEEE-SA database [134], and across different d, θ_2 and σ values. This claim is also validated by computing the Mean Squared Error between the empirical data and its reconstruction.

5.2.3 Discomfort Prediction

Motivated by the observation that stereoscopic displays stimulate accommodation and vergence in an unnatural way resulting in the vergence-accommodation conflict, I combined binocular vergence and accommodation features computed based on the disparity maps along with the NSS features



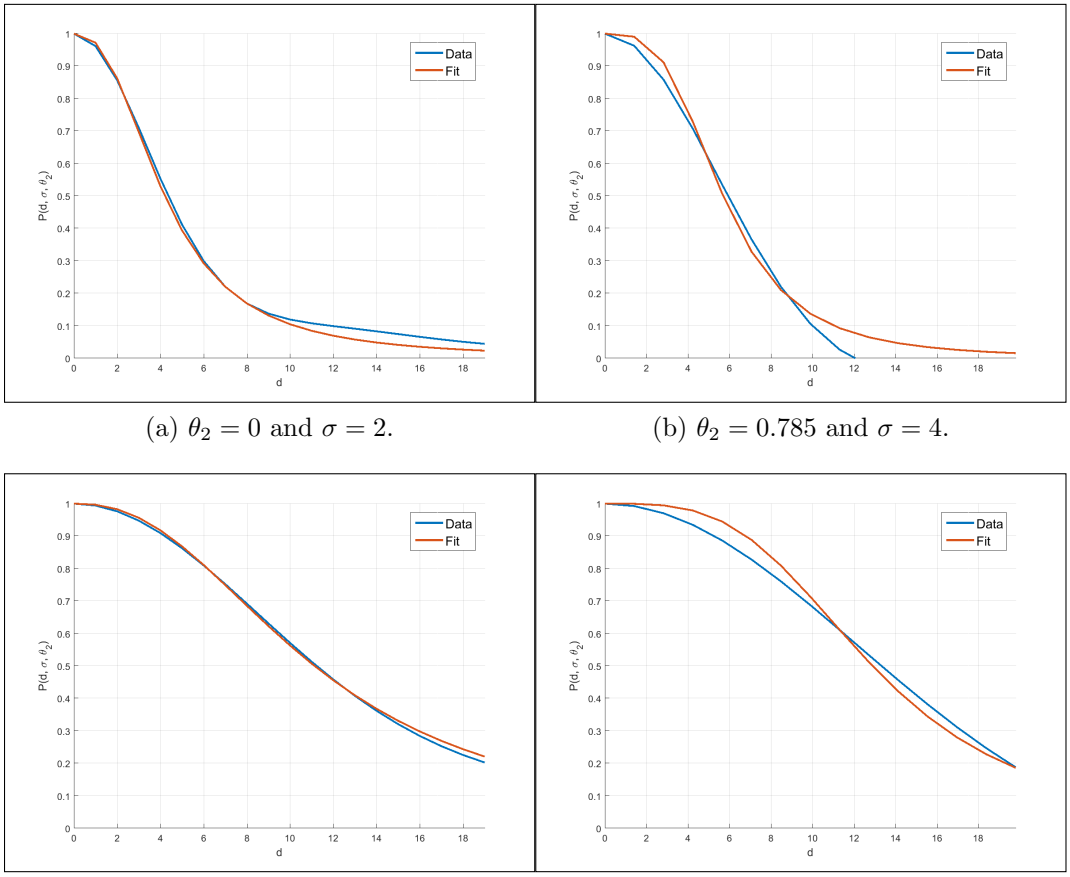
(a) $d = 7$, $\theta_2 = 0$ and $\sigma = 2$.

(b) $d = 4.24$, $\theta_2 = 0.785$ and $\sigma = 3$.

(c) $d = 12$, $\theta_2 = 1.507$ and $\sigma = 5$.

(d) $d = 12.72$, $\theta_2 = 2.352$ and $\sigma = 7$.

Figure 5.2: Sample correlation plots obtained from the data and their corresponding fits, obtained from the stereo pair ISS1-0-L.png and ISS1-0-R.png from the IEEE-SA database. The sample correlation plots span various d, θ_2 and σ values.



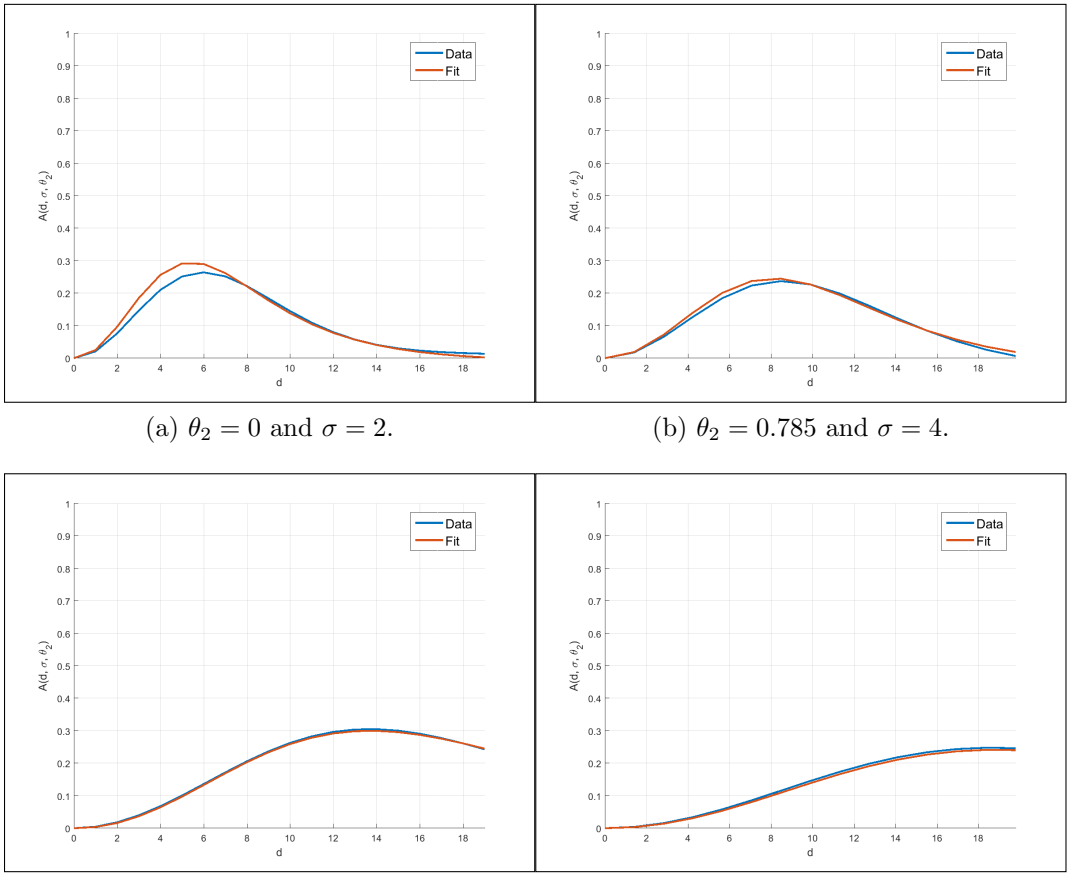
(a) $\theta_2 = 0$ and $\sigma = 2$.

(b) $\theta_2 = 0.785$ and $\sigma = 4$.

(c) $\theta_2 = 1.507$ and $\sigma = 3$.

(d) $\theta_2 = 2.352$ and $\sigma = 7$.

Figure 5.3: Sample empirical P and their fits, obtained from the stereo pair ISS1-0-L.png and ISS1-0-R.png from the IEEE-SA database. The sample P plots span various θ_2 and σ values.



(a) $\theta_2 = 0$ and $\sigma = 2$.

(b) $\theta_2 = 0.785$ and $\sigma = 4$.

(c) $\theta_2 = 1.507$ and $\sigma = 3$.

(d) $\theta_2 = 2.352$ and $\sigma = 7$.

Figure 5.4: Sample empirical A and their fits, obtained from the stereo pair ISS1-0-L.png and ISS1-0-R.png from the IEEE-SA database. The sample A plots span various θ_2 and σ values.

(α_0, β_0, α_1, β_1, α_2, β_2) along different spatial orientations θ_2 values to create a regression module that would map all those features to human ratings in the IEEE-SA [134] database.

5.2.3.1 Bivariate Natural Scene Statistics Depth Features

I studied the correlation between the bivariate NSS depth features $\{\alpha_0$, β_0, α_1, β_1, α_2, $\beta_2\}$ at the 8 most frequently occurring spatial orientations $\theta_2 \in [0, 0.436, 0.785, 1.107, 1.570, 2.034, 2.356, 2.677]$. As a first test, I trained a Support Vector Regression (SVR) model [98] using a radial basis function and 80-20 split on the IEEE-SA database [134] using 46 features as input, obtaining Pearson's linear correlation coefficient (PLCC) and Spearmans rank ordered correlation coefficient (SROCC) both in the range of 0.71 and 0.63 respectively, when the experiment was repeated over multiple iterations. I found that taking a subset of those features and complementing them with other binocular model features helped improve performance. In particular, I observed that a combination of the bivariate features at $\theta_2 = 0$ and $\theta_2 = 2.677$ correlated the most with the MOS. I denote these features by F_1 - F_{12}

5.2.3.2 Binocular Model Features

I complemented my model with four other statistical based binocular model features. The authors in [132] proposed perceptual features to evaluate visual discomfort. I considered a subset of their proposed features, in particular the ones related to disparity maps which were reported to have a PLCC of

0.83 and an SROCC of 0.76 over multiple iterations [132]. I used the method described in [135,136] to extract the disparity maps. The main motivation behind the use of the considered features relates to the close correlation between visual discomfort and the parallax distribution [124,125]. If the reconstructed scene is positioned behind the screen then the viewing experience is comfortable. In that case, the eyes diverge, and the disparity is positive. Otherwise, the eyes converge, as the disparity is negative, leading to visual discomfort. Thus the correlation between the sign of disparity and visual discomfort. The features that I included capture the sign of the disparity. The first feature represents the mean of positive disparities defined by:

$$F_{13} = \begin{cases} \frac{1}{N_{Pos}} \sum_{D(n)>0} D(n), & \text{if } N_{Pos} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.8}$$

where $D(n)$ is the n^{th} smallest value in the disparity map and N_{Pos} is the number of positive elements in the map.

In a similar fashion, I let the mean of negative disparities be another feature defined by:

$$F_{14} = \begin{cases} \frac{1}{N_{Neg}} \sum_{D(n)\leq 0} D(n), & \text{if } N_{Neg} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.9}$$

where N_{Neg} is the number of negative elements in the map.

I also include the mean of the lowest 5^{th} and highest 95^{th} percentiles as additional features, defined as:

$$F_{15} = \frac{1}{N_{P_{5^{th}}}} \sum_{D(n) \leq D(P_{5^{th}})} D(n) \tag{5.10}$$

where $N_{P_{5^{th}}}$ is the number of elements smaller than or equal to the 5^{th} percentiles in $D(n)$. And:

$$F_{16} = \frac{1}{N_{P_{95^{th}}}} \sum_{D(n) \geq D(P_{95^{th}})} D(n) \tag{5.11}$$

where $N_{P_{5^{th}}}$ is the number of elements greater than or equal to the 95^{th} percentiles in $D(n)$.

Taking the combination of all the 16 features results in a predictor that is summarized in Table 5.4.

5.2.4 Results

Using a regression module, I constructed a mapping from the feature space, (Table 5.4) to the MOS, resulting in a measure of 3D visual discomfort. I used a support vector regressor (SVR) [98], in particular the LIBSVM package [110] to implement the SVR with a radial basis function (RBF) kernel and to predict the MOS scores. I split the images randomly and used 80% of it for training and the rest for testing, then I normalized my features, and fed them into the SVR module to predict the MOS score. I repeated the process 50 times. I obtained a median Pearsons linear correlation coefficient (PLCC) of about 0.89 and a Spearmans rank ordered correlation coefficient (SROCC) of

Table 5.4: Summary of the features used in the predictor.

Feature	Description
F_1	α_0 at $\theta_2 = 0$ rad
F_2	β_0 at $\theta_2 = 0$ rad
F_3	α_1 at $\theta_2 = 0$ rad
F_4	β_1 at $\theta_2 = 0$ rad
F_5	α_2 at $\theta_2 = 0$ rad
F_6	β_2 at $\theta_2 = 0$ rad
F_7	α_0 at $\theta_2 = 2.677$ rad
F_8	β_0 at $\theta_2 = 2.677$ rad
F_9	α_1 at $\theta_2 = 2.677$ rad
F_{10}	β_1 at $\theta_2 = 2.677$ rad
F_{11}	α_2 at $\theta_2 = 2.677$ rad
F_{12}	β_2 at $\theta_2 = 2.677$ rad
F_{13}	mean of positive disparities
F_{14}	mean of negative disparities
F_{15}	mean of the smallest 5% of the values in the disparity map
F_{16}	mean of the largest 95% of the values in the disparity map

about 0.83 against MOS. Table 5.5 compares the performances of other various reported algorithms. The performance of my model was only approached by DeepVDP [133], which uses a complex convolutional neural network.

5.3 Concluding Remarks

In this chapter, I first built a new predictor for the IQA problem by combining quality-predictive features from a new bivariate NSS correlation model

Table 5.5: Mean PLCC and SROCC and their standard deviations over the IEEE-SA database, with 80-20% splits, over 50 iterations.

Model	PLCC	SROCC
Nojiri *et al.* [124]	0.6854 ± 0.0788	0.6108 ± 0.0732
Yano *et al.* [126]	0.3988 ± 0.0748	0.3363 ± 0.0798
Choi *et al.* [127]	0.6509 ± 0.0783	0.5851 ± 0.0798
Park *et al.* [129]	0.8310 ± 0.0526	0.7534 ± 0.0498
Kim *et al.* [130]	0.7018 ± 0.0771	0.6151 ± 0.0700
Oh *et al.* [131]	0.8590 ± 0.0452	0.7887 ± 0.0405
Park *et al.* [132]	0.8524 ± 0.0482	0.7785 ± 0.0451
Oh *et al.* [133]	0.8849 ± 0.0283	0.8164 ± 0.0254
Proposed Model	**0.8884** ± 0.0197	**0.8264** ± 0.0292

with known BRISQUE univariate NSS features. The resulting new IQA model was shown to outperform top performing blind image quality assessment models. I also studied the bivariate NSS of disparity maps, and modeled them in closed form. I showed that using 6 features per spatial orientation allows us to capture those statistics with very small error. I demonstrated a close relationship between those statistics and 3D visual discomfort. I combined those features along with other simple statistics of disparity maps related to positive and negative disparities to create a powerful 3D visual discomfort predictor that outperforms state of the predictors, which are perceptually based and/or use deep convolutional networks.

Provided the success of bivariate NSS in image quality related tasks, I decided to expand this for video. Moving forward, I will be tackling the blind

VQA problem. I will be demonstrating in a few chapters that studying the NSS of the shifted frame difference of the frames is indeed a powerful tool for building models that can blindly predict the quality of videos. The next chapter discusses how I built a representative benchmark that would allow me to assess the performance of such models.

Chapter 6

Scaling Up Subjective Studies: The LIVE Video Quality Challenge Database

In this chapter, I present the construction of the LIVE-VQC database [137,138] and the design of an adequate framework for scaling up the collection of the subjective scores [139]. I also discuss the results of the large-scale crowdsourced video study that I conducted, resulting in more than 205000 opinion scores on 585 diverse videos containing complex authentic distortions. I also evaluate the performances of prominent blind VQA algorithms on the new database. [1]

6.1 Construction of the LIVE-VQC Database

While considerable effort has been applied to the VQA problem for high-end streaming video (e.g., Internet television), much less work has been done on videos captured by mobile and digital cameras by casual users. My

[1] This chapter appears in the following papers:
1) Zeina Sinno, Alan C. Bovik: "Large-Scale Study of Perceptual Video Quality" in the IEEE Transactions on Image Processing 28(2): 612-627 (2019).
2) Zeina Sinno, Alan C. Bovik: "Large Scale Subjective Video Quality Study" in the IEEE International Conference on Image Processing: 276-280 (2018).
Zeina Sinno has designed and constructed the framework, collected the data and performed full experimental analysis of the works described therein.

objective was to create a resource to support research on this very large-scale, consequential topic. My specific aim is to offer a large, high-quality dataset of authentically captured and distorted videos, and a large corpus of science-quality psychometric video quality scores.

6.1.1 Content Collection

My data was collected with the assistance of 80 largely naïve mobile camera users from highly diverse age groups, gender, and social, cultural and geographic diversity. I requested the collaborators to upload their videos just as captured, without any processing (for example by video processing 'apps' like Instagram or Snapchat). Only videos having durations of at least 10 seconds were accepted. No instructions regarding the content or capture style was provided, other than it reflect their normal use.

Most of the video contributors were volunteers, including acquaintances of LIVE members; i.e, from family, friends, friends of friends, and so on, from around the world, while the rest ($\sim$18%) were students solicited from the undergraduate and graduate population at The University of Texas at Austin. The number of videos provided by each contributing videographer varied but none contributed by more than 9% of the video set, to ensure diversity of method, content and style. The contributors spanned a wide age range (11 to 65 years old), and were divided about evenly by gender. The content was shot on all the populated continents and in many countries, including Australia, U.S.A., Mexico, Peru, Panama, Colombia, Bolivia, India, Malaysia, Vietnam,

China, South Korea, Germany, Norway, Switzerland, Poland, Sweden, U.K., Portugal, Turkey, Lebanon, the United Arab Emirates, Oman, Tunisia, Egypt and more. More than 1000 videos were gathered, cropped to 10 seconds while seeking to preserve 'story' continuity, culled to remove redundant content captured by a same user and videos with disturbing content (e.g. a scene of surgery). After this cleaning, I was left with 585 videos.

As exemplified in Fig. 6.1, the obtained video content is quite diverse, and includes scenes of sports games, music concerts, nature, various human activities (parades, dancers, street artists, cowboys etc.), and much more. The scenes were captured under different lighting conditions (different times of day and night), and include both indoor and outdoor scenes. Widely diverse levels of motion (camera motion and in-frame motion) are present and often contibute to complex, space variant distortions. While it is very difficult to categorize real-world picture and video distortions with any precision, owing to their intrinsic mutability, their tendency to commingle, many distortions have been observed including, for example, poor exposures, and a variety of motion blurs, haziness, various imperfect color representations, low-light effects including blur and graininess, resolution and compression artifacts, diverse defocus blurs, complicated combinations of all of these, and much more. The interactions of multiple artifacts also give rise to very complex, difficult to describe composite impairments. Often visible distortions appear, disappear, or otherwise morph during a video, as for example, temporary autofocus blurs, exposure adjustments, and changes in lighting. As such, I made no attempt

to supply distortion labels to the videos.

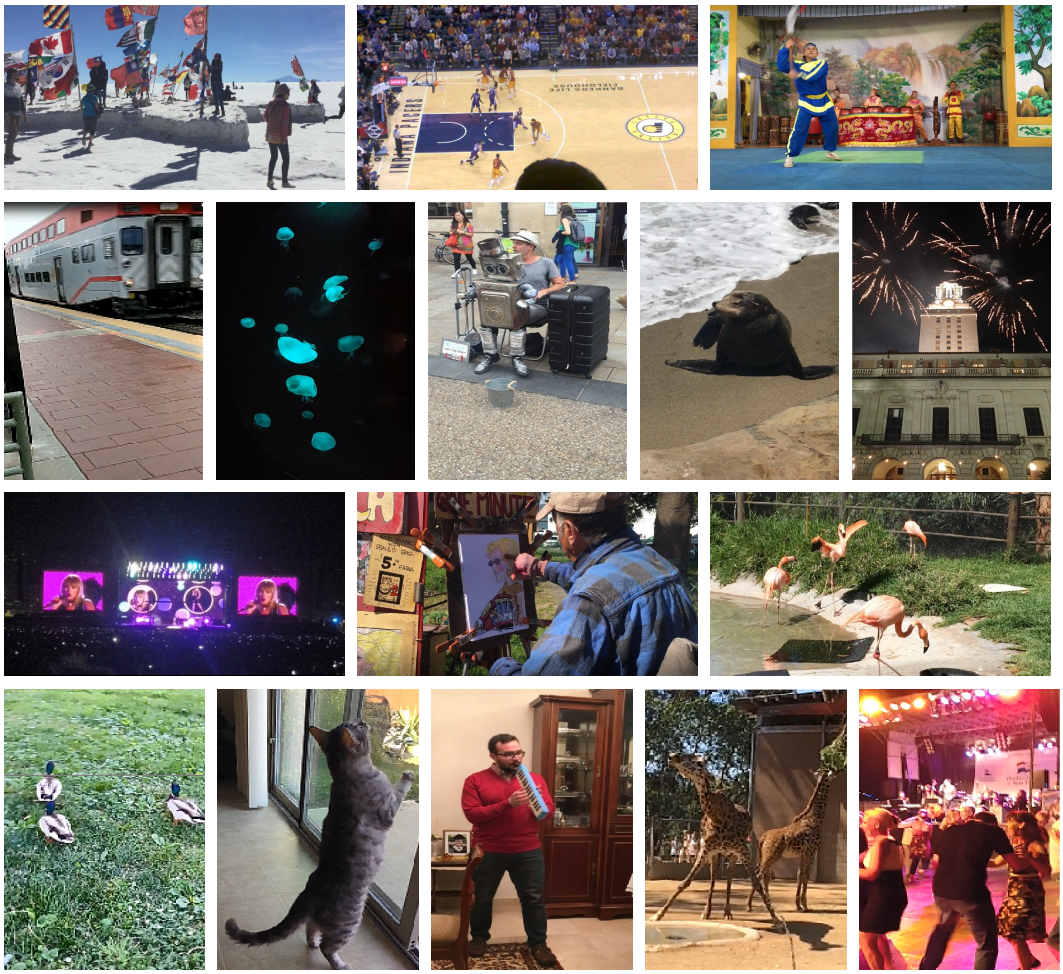


Figure 6.1: Screenshots of frames from some of those presented during the study.

6.1.2 Capture Devices

A taxonomy of the mobile devices used to capture the videos is given in Table 6.1. Unsurprisingly, the majority of these were smartphones. A total

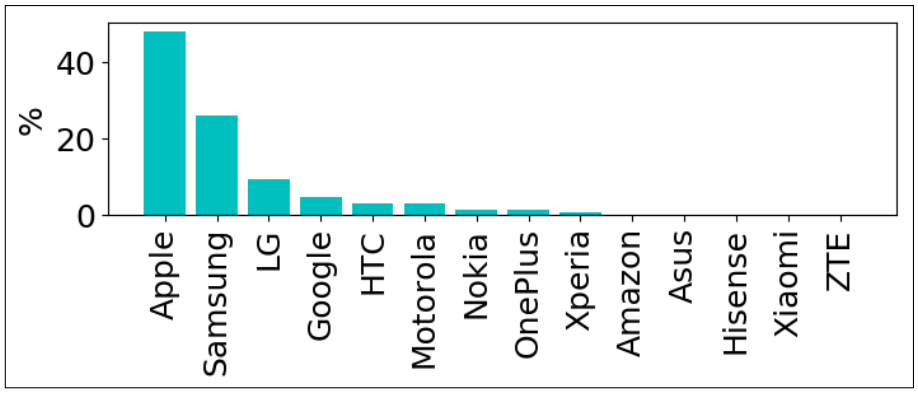


Figure 6.2: Distribution of viewed videos grouped by device brand.

of 101 different devices were deployed (some users provided videos captured by multiple devices), including 43 different models. The commercial releases of the devices varied between 2009 and 2017, although most of the videos were captured using devices that were released after 2015 and beyond.

Figure 6.2 depicts the distribution of the viewed videos grouped by brand. As expected [140] the majority of these videos ($\sim$ 74%) were captured using commercially dominant Apple and Samsung devices.

6.1.3 Video Orientations and Resolutions

I imposed no restrictions on the orientation of the camera device during capture (or after), and 23.2% of the videos in the database were taken in portrait mode and the other 76.2% in landscape mode. The majority of the videos shot in portrait and are of high resolutions (1080×1920 and 3840×2160) which cannot be fully displayed by most available displays without down-

Table 6.1: Number of videos captured by each type of camera devices.

Make	Model	Number of Videos
Amazon	Fire HDX	1
Apple	Ipad Pro	2
Apple	Iphone 3GS	1
Apple	Iphone 4	14
Apple	Iphone 4S	2
Apple	Iphone 5	25
Apple	Iphone 5s	49
Apple	Iphone 6	48
Apple	Iphone 6s	107
Apple	Iphone 6s plus	5
Apple	Iphone 7	17
Apple	Iphone 7 plus	3
Apple	Ipod touch	8
Asus	Zenfone Max	1
Google	Pixel	7
Google	Pixel XL	20
Hisense	S1	1
HTC	10	13
HTC	M8	5
Huawei	Nexus 6P	10
LG	G3	3
LG	G4	2
LG	Nexus 5	50
Motorola	E4	1
Motorola	Moto G 4G	3
Motorola	Moto G4+	1
Motorola	Moto Z Force	12
Nokia	Lumia 635	5
Nokia	Lumia 720	3
OnePlus	2	4
OnePlus	3	4
Samsung	Core Prime	5
Samsung	Galaxy Mega	1
Samsung	Galaxy Note 2	21
Samsung	Galaxy Note 3	5
Samsung	Galaxy Note 5	72
Samsung	Galaxy S3	4
Samsung	Galaxy S5	25
Samsung	Galaxy S6	14
Samsung	Galaxy S8	6
Xiaomi	MI3	1
Xperia	3 Compact	3
ZTE	Axon 7	1

scaling them. The median display configuration in use today appears to be 1280×720 [141]. To ensure compatibility, all portrait videos of resolutions 1080×1920, 2160×3840, and 720×1080 were downscaled using bicubic interpolation to 404×720, so that they could be displayed at the native display resolutions of all subjects accepted to participate in the study.

Among the videos in landscape mode, many were of resolutions that cannot be displayed by viewers (those that were 1920×1080 and 3840×2160). At the time the study was conducted, it was estimated that only between 10-20% of global web viewers possessed high resolution displays equal to or exceeding 1920×1080 [141]. As a way of accessing both high and low resolution display groups, I decided to downscale a portion of the large resolution videos to 1280×720 to better distribute the scoring tasks, since I expected relatively few participants to be capable of viewing high resolutions. Thus, 110 videos randomly selected videos were maintained at resolution of 1920×1080, while the remaining 1920×1080 and higher resolution videos were downsampled to 1280×720 using bicubic interpolation. I ended up with 18 different resolutions in my database, as summarized in Table 6.2.

Table 6.2: Video resolutions in the database

1920×1080	1280×720	960×540	800×450	480×640	640×480
404×720	360×640	640×360	352×640	640×352	320×568
568×320	360×480	480×360	272×480	240×320	320×240

The predominant resolutions were 1920×1080, 1280×720 and 404×720, which together accounted for 93.2% of the total, as shown in Fig. 6.3. The

other resolutions combined accounted for 6.8% of the database.

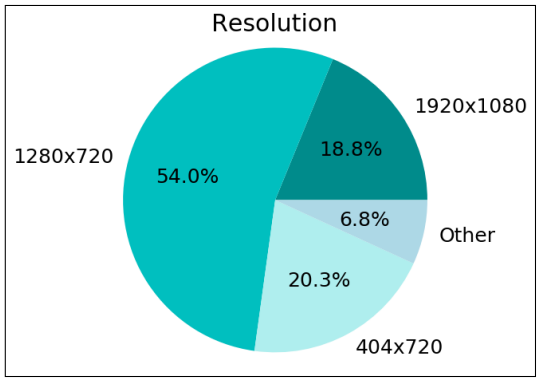


Figure 6.3: Distribution of the resolutions viewed by the AMT workers.

I managed the videos shown to each worker by first detecting their display resolution by executing code in the background. If their display resolution was at least 1920×1080, then half of the videos they evaluated would have this resolution, whilst the rest were randomly selected from the rest of the database. All of the other participants viewed randomly selected videos having resolutions less than 1920×1080.

6.2 Crowdsourcing the Subjective Scores

Given that real users nearly always view single videos, rather than side-by-side pairs, and since, in every case I have only a single, authentically distorted version of each content, I deployed a single stimulus presentation protocol. Since the study is crowdsourced and conducted in the wild, I could

not apply many ITU standard recommendations (e.g. [59]) when conducting the subjective studies [139]. I did, however abide by agreed-on principles regarding timing, stimulus presentation, and subject training, as detailed in the following.

6.2.1 Participation Requirements

I first list the participation requirements, then explain each in detail. To be eligible to participate, the worker should:

1) Have an AMT reliability score above 90% (reliability constraint).

2) Have not participated previously in the study (unique worker constraint).

3) Use a non-mobile display device, *viz.* desktop and laptops are allowed, while mobile phones and tablets are not (display constraint).

4) Use a display having a minimum resolution of 1280×720 (resolution constraint).

5) Use a recently updated supported browser. The study supports Google Chrome, Safari, Mozilla Firefox, and Opera. Internet Explorer, Microsoft Edge and other browsers were not allowed (browser constraint).

6) Have a good Internet capacity (connectivity constraint).

7) Use a device with adequate computational power (hardware constraint).

Explanations supporting these choices are as follows:

6.2.1.1 Reliability constraint

AMT records how many jobs each worker has completed, and how many jobs were accepted, to determine the acceptance ratio, known as the reliability score. Because of the subtlety of many distortions and to better ensure subject assiduity, I only allowed workers having an acceptance rates exceeding 90% to participate.

6.2.1.2 Unique worker constraint

I imposed this condition to avoid any judgment biases that might arise if workers rated videos more than once.

6.2.1.3 Device constraint

I enforced this condition for two reasons. First, mobile browsers do not support preloading videos, which is a major concern. Second, it is not possible to control the resolutions of videos displayed on mobile browsers, since they must be played using a native player on the device where are upscaled or downscaled, then played in full screen mode, whereby additional, unknown artifacts are introduced.

6.2.1.4 Resolution constraint

I required the worker display resolutions to be at least 1280×720 (720p) as discussed earlier.

6.2.1.5 Browser constraint

As of early 2018, Internet Explorer and Microsoft Edge do not support video preloading in HTML5. For this reason, I did not allow users of those browsers to take part of the study. Google Chrome, Safari, Mozilla Firefox, and Opera support this option starting at a certain version. I verified that the browser (and the version) used by each worker was compatible with the HTML5 video preloading attribute. I verified that each session could proceed with smooth preloading, thereby eliminating the possibility of bandwidth-induced stalling (rebuffering) events.

6.2.1.6 Connectivity constraint

Poor Internet connectivity or slow bandwidths can cause annoying delays as the videos are loading leading to possible frustration and loss of focus on the part of the subjects. Under extremely poor bandwidth conditions, it can also lead to timeouts in the connection established between the server where the videos are stored and the worker's side. Internet bandwidth is stochastic and unpredictable as it changes over time. In rare cases, a timeout can be emerge at the users' side, with good bandwidth conditions. For example, a new device may join the network and initiate a large download. In such a case, a sudden drop in the bandwidth could be experienced. To minimize these problems, I tracked the loading progress of each video and acted accordingly. Each video was requested at least 30 seconds before it was needed by the subject. If the connection was not successfully established the first

time, a second attempt was made 10 seconds later, and a third 10 seconds after that. If the connection again failed, the session was terminated and the worker was informed. Once a connection was successfully established and the loading commenced, if it was detected that the loading progress halted for a certain interval of time, the connection with the server was terminated and a new one established. This was allowed to occur only once. As a global constraint, the duration of each study session was not allowed to exceed 30 minutes. This helped to filter out corner cases where connections were successfully established but the loading progress was very slow. I also implemented two temporal checkpoints to track the progress of each worker. After a third of the session videos were viewed, if it was detected that more than 10 minutes (one third of the allowed time) had elapsed, then a warning message was displayed informing the worker that they might not be able to complete the test on time. The second checkpoint occurred after two thirds of the content had been viewed, warning them if 20 minutes had passed. I encouraged the workers (as part of the training process) to close any other windows or tabs open in the background before commencing the study to avoid draining their bandwidth. They were reminded of this again if their progress was slow or if they were experiencing large delays. Before launching the study, I extensively tested the framework under highly diverse bandwidth conditions and scenarios to ensure its efficacy.

6.2.1.7 Hardware constraint

Slow hardware or poor computational power can lead to "frame freeze" stalls while a video is played. To minimize the frequency of these occurrences, I encouraged each worker (via the training instructions) to close any programs running in the background, and if they were using a laptop, to ensure that it was plugged into an outlet to further promote better performance.

6.2.2 Viewed Content

During a single session, each subject viewed a total of 50 videos: 7 training and the remaining 43 during the rating process (Fig. 6.4).

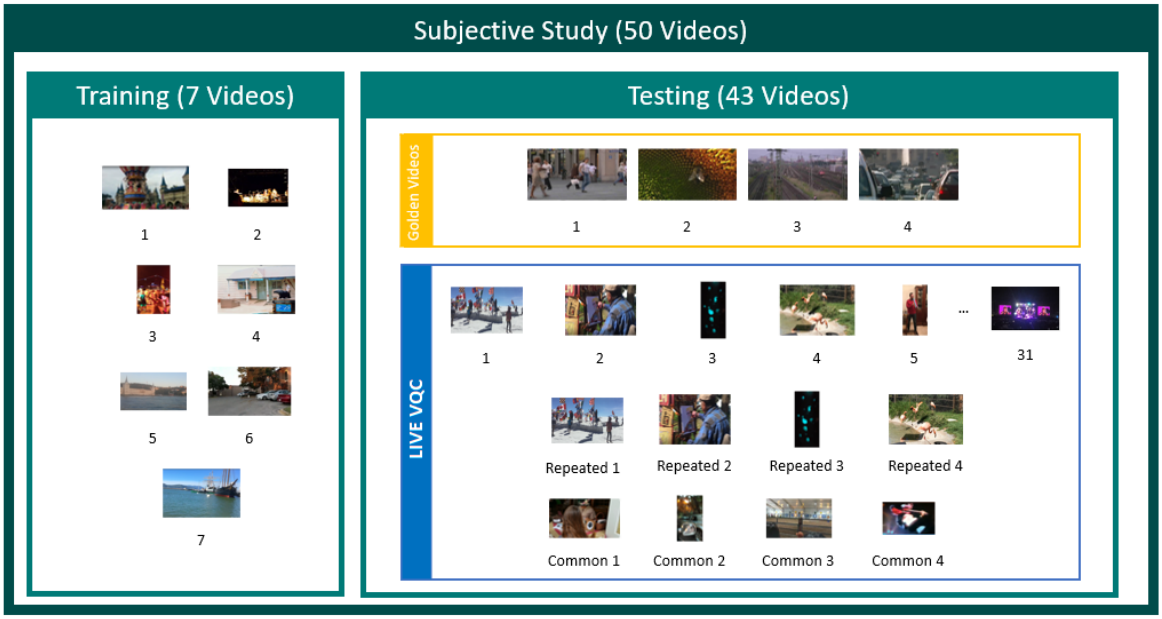


Figure 6.4: Chart showing the categories of videos seen by a subject viewing session.

The videos displayed during the training process were selected to broadly span the ranges of video quality containing within the database, and were of mixed resolutions, to prepare the subjects for their later viewing. The training varied slightly with the detected sizes of the workers' displays. Those viewers using displays of resolution of 1920×1080 or higher were presented with two videos matching their display, along with a mixture of videos of smaller resolutions (1280×720 and less). Those subjects having display resolutions lower than 1920×1080 were presented with videos of mixed resolutions no higher than 1280×720.

The 43 videos viewed during the judgment (test) phase included:

- 4 distorted videos drawn from the LIVE Video Quality Assessment Database [42], which I will refer to as the "golden videos." These videos were previously rated by human viewers in the tightly controlled study [42], and are used, along with the prior subjective scores from [42], as a control to validate the subjects' ratings.

- 31 videos randomly selected from the new distorted video database. If the worker had a display resolution no less than 1920×1080, then 18 videos were drawn from the pool of videos having a resolution of 1920×1080, and the remaining 13 videos selected from the other, lower resolution videos.

- 4 videos randomly selected from the same pool of 31 videos as above, but repeated at relatively displayed moments as a control.

- 4 videos selected from the database were viewed and rated by all of the workers.

The 43 videos were placed in re-randomized order for each subject.

6.2.3 Experimental Flow

Each subjective study session followed the workflow depicted in Fig. 6.5. I now describe each step in detail.

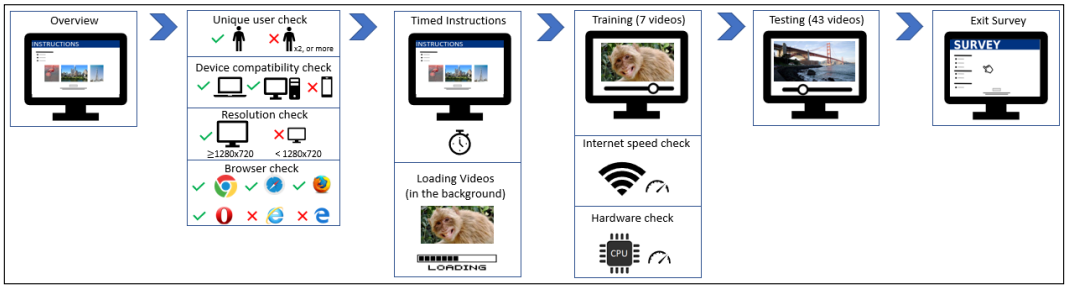


Figure 6.5: Subjective study workflow.

Step 1: Overview

Once a worker with a reliability score exceeding 90% selected my study to preview it, s/he was prompted to an overview page describing the task, the requirements to participate (conditions 2-7), the instructions on how to rate a video, and a few example videos to give them a clearer sense of the task. The worker was instructed to rate the videos based on how well s/he believes the presented video quality compares to an ideal, or best possible video of the same content. Several example videos were then played to demonstrate exemplars

of some of the video distortions such as under exposure, stalls, shakes, blur and poor color representation. The worker was informed that other types of distortions exist and would be seen, so the worker would not supply ratings based only on the exemplar types of distortions, but would instead rate all distortions.

Step 2: Eligibility check

If the user accepted to work on the 'hit', and it was determined whether s/he did not previously participate in the study, and that s/he met conditions 2)-5) above. If the worker did not meet any of those conditions, a message was displayed indicating which condition was unmet, and that s/he was kindly requested to return the hit. If it was the case that any of conditions 3)-5) was unmet, then the displayed message invited the worker to try working on the hit again, but using another device/ browser etc., depending on which condition was not met. During this step, the browser zoom level was adjusted to 100% to prevent any downscaling or upscaling artifacts from occurring when the videos are played.

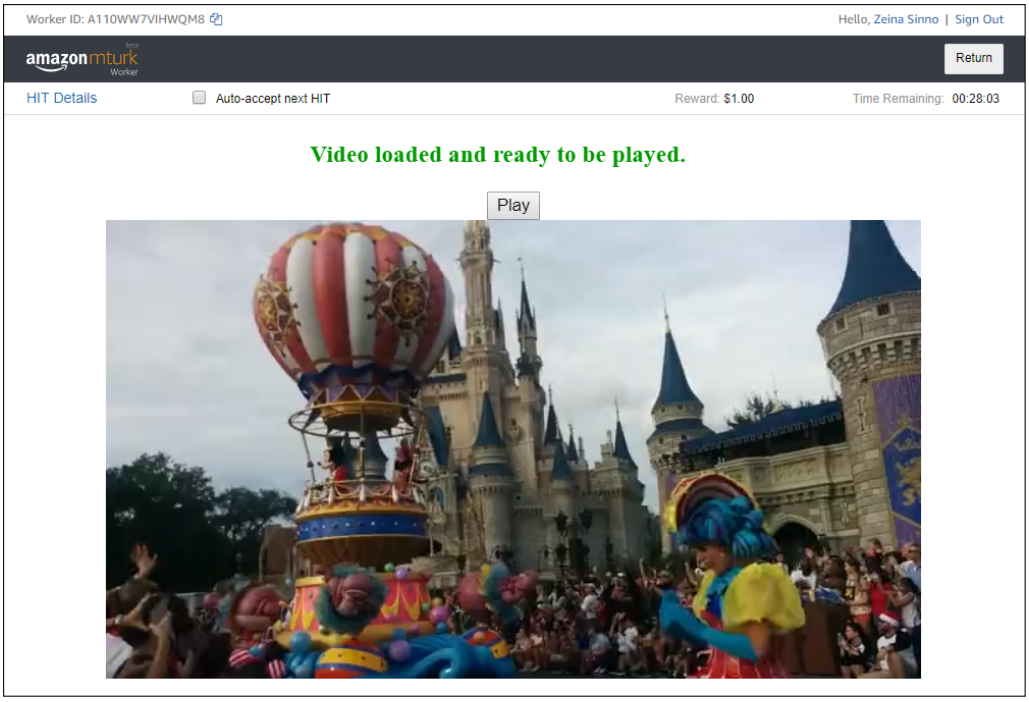
Step 3: Timed Instructions

If the worker was able to proceed, the instruction page was displayed again, with a countdown timer of one minute. Once the countdown timer reached zero, a proceed button would appear at the bottom of the page, thereby allowing the worker to move forward. The instructions were repeated

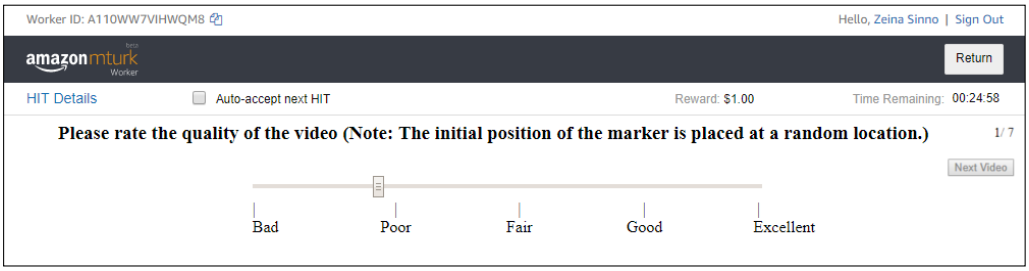
because while the study was in progress (at the end of 2017), AMT was in the process of migrating towards a new user interface that allowed the users to accept a 'hit' without first previewing it. Additionally, some workers used scripts while working on AMT, such as Microworkers and Tampermonkey, which would auto-accept hits on behalf of the workers when posted. Hence, some of the workers would not have had the opportunity to read the instructions if they were not repeated. While the instructions were being repeated, the first three videos began loading in the background, and the videos that were to be displayed during the testing phase were determined.

Step 4: Training

Once a subject clicked on the *Proceed* button, a message was displayed indicating that the training phase was about to start. This phase consisted of 7 videos. A screenshot of the interface featuring each video to be rated (during both the training and test phase) is shown in Fig. 6.6. As shown, the video controls were hidden and inaccessible, to prevent less dedicated workers from pausing, replaying or skipping the videos. Before a video was fully loaded, a message was displayed showing the loading progress. Once the video was fully loaded, a message informed the user that "Video loaded and ready to be played." At this point, the zoom level was checked to determine whether it was at 100%, and was then adjusted if need to be. The video was then played in entirety (while being muted) if the page was displayed in full screen mode. Otherwise, a message was displayed directing the worker to adjust it and to



a) Step 1: viewing a video.



b) Step 2: rating the video

Figure 6.6: Screenshot of the interface used to rate the videos.

again press the play button afterwards. This had the additional benefit of reducing worker multitasking, which could distract attention from the video task.

Once each video finished playing, it disappeared, revealing the rating interface shown in Fig. 6.6(b). A continuous bar allowed the workers to rate the quality of the videos, where a Likert-like scale with 5 marks; Bad, Poor, Fair, Good, and Excellent was provided to generally guide their ratings. The initial position of the cursor was randomized. This was mentioned in the instructions and was also indicated in a note above the rating bar. Once each video finished playing, the user moved the cursor to the appropriate position. The user was not allowed to proceed to the next video unless s/he moved the position of the cursor. Once a change in the cursor location was detected, the *Next Video* button became clickeable. Once clicked, the worker moved to a new page, with a new video to be rated and the process continued until the last video had been rated.

A number of processes were ongoing as the user was viewing/rating rating each video. For example, the following videos to be rated next start would begin loading in the background, as described in the previous section. During the training process, the play duration of each video was measured to assess the workers' play capability. There are many ways that stalls could occur while a video is playing. If a worker's hardware CPU was slow, if other programs were running in the background (CPU is busy) or if the Internet connection was poor, then stalls or frame freezes could (and did) occur. Required background tasks (such as loading the videos to be played next) added processing overhead, while slower Internet bandwidths required increased processing overhead, further impacting foreground performance. During the training process,

7 videos of 10 seconds duration each were played. Importantly, the workers were not able to proceed further if it took more than 15 seconds to play any of the 7 videos or if any 3 of the 7 videos each required more than 12 seconds to play. Adopting this strategy guaranteed that most of the training videos were played smoothly, and also allowed us to eliminate workers who were unlikely be able to successfully complete the 'hit.'

Step 5: Testing

After the training phase was completed, a message was displayed indicating that the video rating phase was about to begin. The testing phase was very similar to the training phase; the videos were displayed, controlled and rated in the same way. However, the testing phase required 43 videos to be rated, instead of 7.

Once a third of the study was completed, (10 testing videos rated), if the progress of the worker was sufficient, then the following message was displayed: "You have completed one third of the overall study! Good Job :-) Keep up the Good Work!" As shown in [142], providing workers with motivational feedback can encourage them to provide work of better quality. If the progress of the worker was slow (>10 minutes had passed), the following message was displayed "You have completed one third of the overall study but your progress is slow. Are the videos taking too long to load? If so, make sure to close any programs running in the background." A similar message was displayed after two thirds of the study was completed.

Step 6: Exit Survey

Once the worker finished rating all of the videos, s/he is directed to the exit survey so that information regarding the following factors could be collected:

- the display,
- viewing distance,
- gender and age of the worker,
- country where the task study was undertaken,
- whether the worker needs corrective lenses, and if so, if s/he wore them.

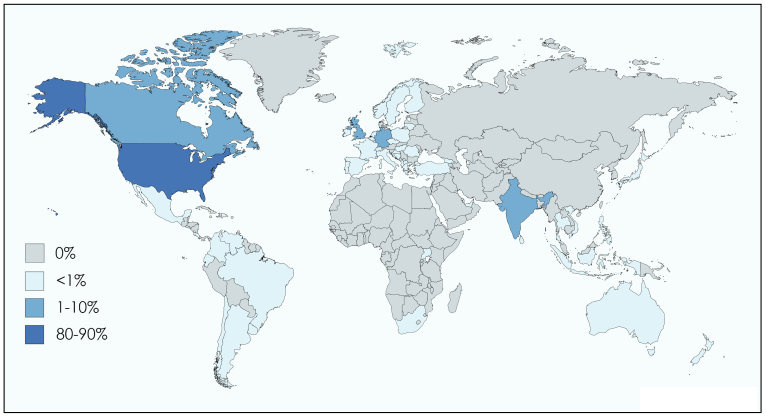
The subjects were also asked whether they had any additional comments or questions. At the same time, information was automatically collected regarding the display resolution.

6.2.4 Human Subjects

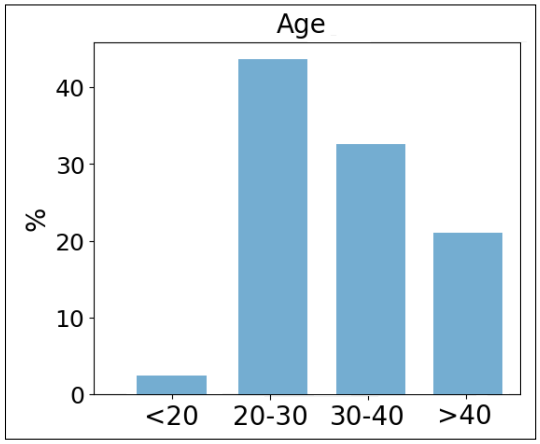
6.2.4.1 Demographic Information

The study participants were workers from AMT having approval rates exceeding 90% on previous studies. A total of 4776 subjects took part in the experiment. The participants were from 56 different countries as highlighted in Fig. 6.7(a) with the majority being located in the United States and India (together accounting for 91% of the participants). Figure 6.7(b) shows the age

distribution of the participants. About half of the participants were of each gender (46.4% male versus 53.6% female).



(a)



(b)

Figure 6.7: Participant demographics (a) Countries where participants were located; (b) Age distribution of the participants.

6.2.4.2 Viewing Conditions

As with any crowdsourced study, the participants operated under diverse viewing conditions; including locations, visual correction, viewing dis-

tances, browser, display device, resolution, ambient lighting, and so on. Figure 6.8 presents statistics that I collected regarding some of the aspects of subject viewing. As shown in Fig. 6.8(a), the majority of the participants had normal or corrected-to-normal vision (e.g., glasses or contacts). A tiny percentage (2.5%) had abnormal, uncorrected vision, and I excluded their results. The participants used mostly laptop and desktop monitors to view the videos (Fig. 6.8(b)), and were mostly positioned between 15 and 30 inches from the display (Fig. 6.8(c)). The subjects used 83 different display resolutions which ranged between 1280×720 and 3840×2160 pixels, as plotted in Fig. 6.8(d). Of these, 31.15% had display resolutions of at least 1920×1080, while the rest had lower resolution displays.

6.2.4.3 Compensation

On average the subjects require 16.5 minutes each to complete the study. The financial compensation given for completing the hit was one US dollar. I wanted to attract high-quality workers, hence maintaining a good AMT reputation was important. There exists a variety of forums, like TurkerNation.com, where AMT workers can share their experiences of AMT hits. These forums build a valuable sense of community among Turk workers, and helps to protect them from unfair or bad practices. I noticed that a small number of workers were uneasy about being unable to complete the study because of some of the eligibility requirements that I imposed. This was especially true when, because of hardware inadequacy, subjects were asked to

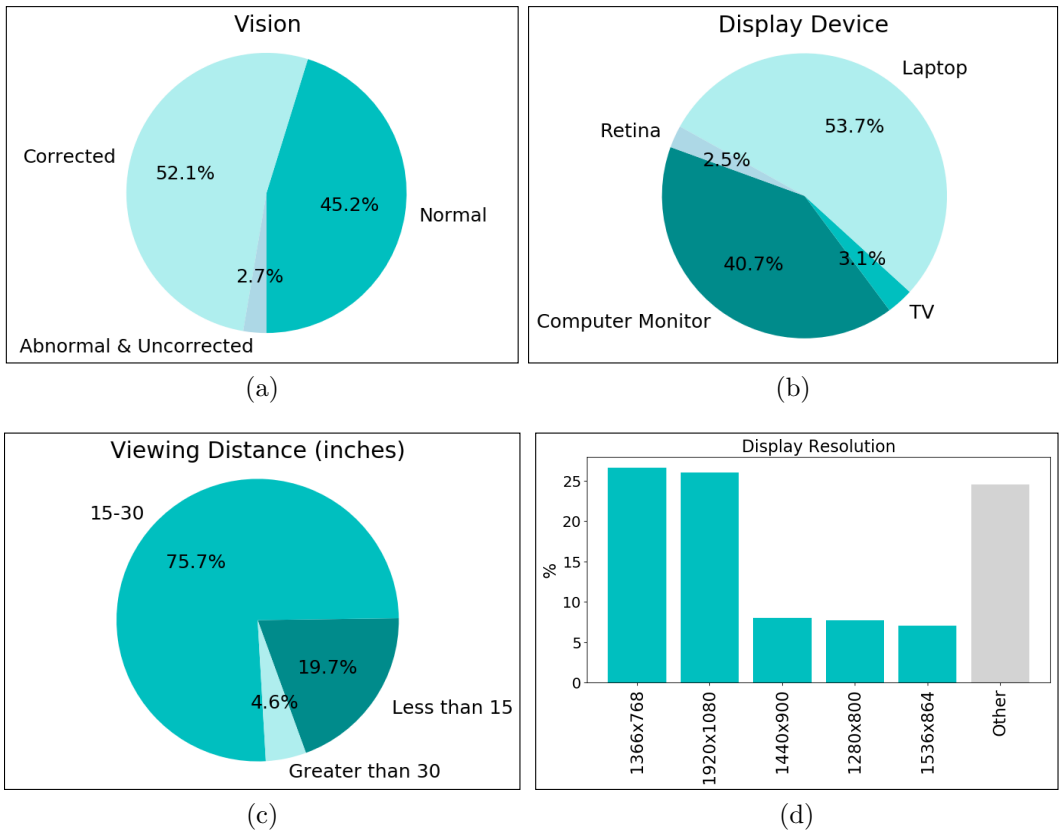


Figure 6.8: Participant statistics: (a) Visual correction; (b) type of display device; (c) approximate viewing distance; (d) display resolution.

return the hit during training. I notified a worker that s/he would be ineligible to continue working on the hit as soon as a problem was detected (*viz.* when a video stalled >5 seconds or when 3 videos stalled >2 seconds during training). The training instructions did inform the subjects that they could be asked to return the hit if any eligibility requirement was not met. I did not compensate any "less dedicated" worker who skipped any video by re-enabling the

controls of the video; either by using Javascript commands or by modifying browser settings, since I wanted to discourage workers from attempting such practices. Interestingly, about 2% of the workers were "skippers" and were not compensated.

Adopting a strategy to reject subjects on the fly that were not providing consistent results was a more challenging issue. The previous large image quality crowdsourced study in [54] repeated some of the images to determine how consistent a subject would rate the same distorted content, and rejected inconsistent subjects. I adopted a similar strategy, with important modifications, by repeating 4 of the videos (at random relative spacings) to measure intra-subject consistency. However this measurement problem was more complex in my study than in [54] since hardware-related stalls; although greatly reduced in frequency, could still occur. Thus, a video and its time-displaced repeat could each present with or without stalls (of different locations and durations), thereby greatly affecting their perceived quality. I noticed that the workers were generally providing very similar ratings on videos viewed twice, when no stalls occurred, which I attribute at least in part to only including participants having high reliability scores. When stalls occurred, the results were harder to interpret. I did not want to reject the workers unfairly, hence I decided to adopt a strategy similar to that used in the crowdsourced video annotation study [142], where the authors compensated the workers regardless of the consistency of their results, arguing that keeping a good reputation among workers was worth the overhead cost, and helped motivate the workers

to produce high quality work. While I informed the workers that I reserved the right to reject poor work, I also adopted this method.

A summary of those participants that were not compensated is given in Table 6.3.

Table 6.3: Summary of the participants that were not compensated.

Participants Group	Filtering	Action(s)
Ineligible Participants	Device, display, resolution and browser information captured after the study overview. Bandwidth and hardware tests during training.	Asked to return the hit. Not compensated. Video scores not collected.
Video Skippers	Measure of the viewing duration.	Not compensated. Video scores excluded.

6.2.4.4 Subject Feedback

I provided the workers with space to give comments in the exit survey. The feedback that I received was generally very positive, which suggests that the workers successfully engaged in the visual task.

Among the 4776 workers who completed the study, 32% completed the additional comments' box. Among those, 55% wrote that they did not have any additional comments (e.g. no comment, none, not applicable), 31% described the test as good, nice, awesome, great, cool, enjoyable or fun. Some (13%) of the workers provided miscellaneous comments, e.g, that they noticed that some videos repeated, or provided additional information about their display, or wondered how the results would be used, or just thanked us for the opportunity to participate.

6.3 Subjective Data Processing and Results

Here I discuss handling of stalled videos, subject rejection, and the effects of the various experimental parameters on the study outcomes.

6.3.1 Video Stalls

As mentioned earlier, I adopted a strategy to identify, during the training phase, those subjects that were the most susceptible to experiencing video stalls. While I was able to substantially mitigate the video stall problem, I was not able to eliminate it entirely. Although I was able to eliminate network-induced video stalls by requiring that all videos pre-loaded before display, the computational power that was available to each participants' device to play and display the videos was a stochastic resource that was a function of other processes executing in the background. While I asked the workers to close any other windows, tabs or programs, there was no way to verify whether these instructions were followed. Moreover, other necessary foreground, and background processes related to high-priority operating system tasks could affect performance. Since network connectivity can be time-varying and unpredictable, further overhead may also have weighed on processor performance during poor connections.

Figure 6.9 plots the distribution of the video stall durations. It can be observed 92% of the videos had no stalls at all or had stalls that lasted for less than 1 sec. In fact, 77% of the videos played with no stalls at all.

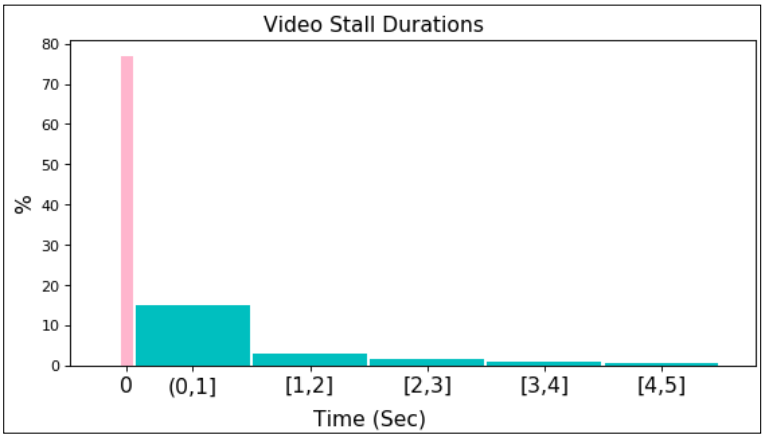


Figure 6.9: Distribution of the video stall durations.

6.3.2 Outlier Rejection

I first rejected the ratings of those users who indicated that they wore corrective lenses, but did not wear them during the course of the experiment. This accounted for 2.5% of all subjects. As mentioned earlier, 2% of subjects attempted to circumvent the experiment and did not watch the videos fully; their results were also excluded.

I also excluded the ratings of users whose computation/display problems were so severe that at least 75% of their viewed videos stalled, which eliminated ratings of 11.5% of the subject population. The remaining subjects viewed at least 11 out of the 43 test videos (usually many more) without experiencing any stalls. For the remaining video ratings, I applied the standard BT-500-13 (Annex 2, section 2.3) [59] rejection portion on the ratings of videos, played without any stalls. I found that only 23 subjects were outliers (0.5%) from among the entire population. This number seemed low that

I also studied the intra-subject consistency. By design, each subject viewed 4 repeated videos during the test phase; I examined the differences in these pairs of scores, as follows. The average standard deviation of all non-stalled videos was about 18. I used this value as a threshold for consistency: given a non-stalled video that was viewed twice, the absolute difference in MOS of the two videos was computed. If it was below the threshold, then the rating for the video was regarded as consistent. Otherwise, it was not. I repeated this analysis across all the 4 videos across all subjects, and found that the majority ($\sim$99%) of the subjects were self-consistent at least half of the time. It is important to emphasize that I excluded the stalled videos from the consistency analysis and when applying the subject rejection [59], because the presence of any stalls rendered the corresponding subject ratings non-comparable.

After rejecting the outliers, I was left with about 205 opinion scores for each video, without stalls, which is a substantial number. I computed the MOS of each video; Fig. 6.10 plots the histogram of these obtained MOS following all of the above-described data cleansing. It may be observed that the MOS distribution substantially spans the quality spectrum with a greater density of videos in the range 60-80.

6.3.3 Validation of Results

6.3.3.1 Golden Videos

During the testing phase of each subject's session, 4 distorted videos taken from the LIVE VQA Database [42] - the aforementioned "Golden Videos"

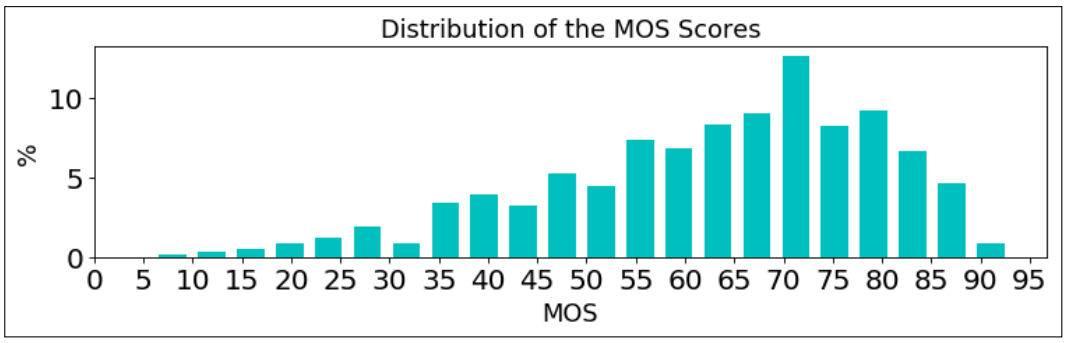


Figure 6.10: Distribution of MOS of the final set of video ratings.

- were displayed at random placements to each worker to serve as a control. The mean Spearman rank ordered correlation (SROCC) values computed between the workers' MOS on the gold standard images and the corresponding ground truth MOS values from the LIVE VQA was found to be 0.99. The mean absolute difference between the MOS values obtained from my study and the ground truth MOS values of the "Golden Videos" was 8.5. I also conducted a paired-sampled Wilcoxon t-test, and found that the differences between these to be insignificant at $p < 0.05$. A recent experiment [143] showed that the MOS collected in subjective studies tends to vary with the overall quality of the presented videos. The videos in LIVE-VQC database span a wider range of quality than the LIVE VQA Database [42], which only contains videos contaminated by only a few synthetic distortion types each at a few levels of severity. I believe that this explains the consistent shift in MOS across the 4 golden videos, when the outcomes from both experiments are compared.

The excellent agreement between the crowdsourced scores and the laboratory MOS significantly validates my experimental protocol.

6.3.3.2 Overall inter-subject consistency

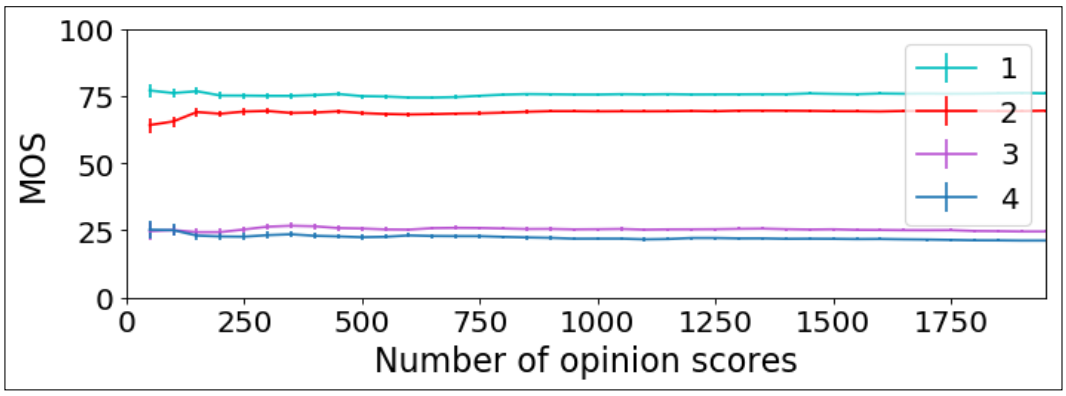
To study overall subject consistency, I divided the opinion scores obtained on each video into two disjoint equal sets, then I computed MOS values on each set. I conducted on all the videos, then computed the SROCC between the two sets of MOS. This experiment was repeated 100 times, and the average SROCC between the halves was found to be 0.984.

6.3.4 Impact of Experimental Parameters

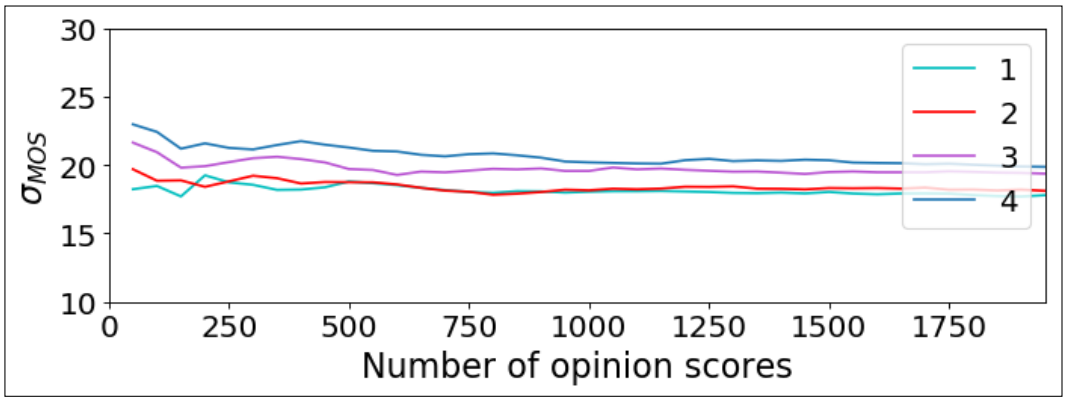
6.3.4.1 Number of subjects.

To understand the impact of the number of subjects on the obtained MOS, I considered the set of videos that were viewed by all subjects, and plotted the error bar plots of the associated MOS along with the standard deviation as a function of the number of ratings (up to 2000), as shown in Fig. 6.11.

I found that increasing the sample size behind beyond 200 did not improve or otherwise affect the figures. I collected slightly more than 200 opinion scores for each video (without stalls). I observed similar behaviors across the rest of the videos.



(a)



(b)

Figure 6.11: Plots of (a) error bars of MOS; (b) standard deviation of MOS, for the set of videos viewed by all of the subjects.

6.3.4.2 Stalls.

I computed the differential mean opinion scores (DMOS) between the non-stalled videos and the stalled videos:

$$DMOS = MOS_{\,without\ stalls} - MOS_{\,with\ stalls} \tag{6.1}$$

The DMOS is plotted against the video index in Fig. 6.12.

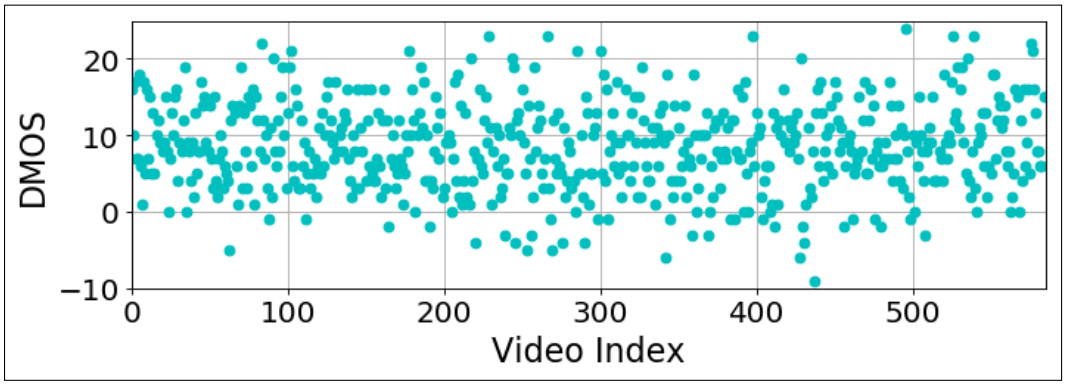


Figure 6.12: DMOS between the videos that played normally and the stalled videos.

Stalls nearly always resulted in a drop of MOS (for $> 95\%$ of the videos). Since it is difficult to assert a reason for a rare small increase in MOS on a given content when stalled, I simply regard those events as noise in the data.

While I have not included an analysis of the stalled video ratings here, I still regard the data as valuable, even though I was only able to collect the per-video total stall durations (but not the number or locations of the stalls). In future work, I plan to analyze stalled video ratings as well, with an eye towards helping guide the development of models that can account for stall occurrences when predicting the video quality.

6.3.5 Worker Parameters

6.3.5.1 High vs low resolution pools

The low resolution group (subjects with display resolutions less than 1920×1080) rated 475 of the 585 videos in my database. Whereas, the high

resolution pool (resolutions of at least 1920×1080) rated all 585 videos. I studied the inter-subject agreement between the two groups over the common set of videos. I computed the SROCC over the MOS obtained from both groups and obtained a value of 0.97. The mean difference in MOS between the two sets was close to 1, which might be attributed to statistical noise. The high inter-subject agreement between the two groups is important, as it shows that the subjects from the high resolution pool (accounting for 31.15% of the total population), who had seen videos of higher resolutions, did not rate the low resolution videos differently than did the low resolution pool of participants.

6.3.5.2 Participants' Resolution

As can be observed in Fig. 6.8(d), the two dominant resolution groups were 1366x768 and 1920×1080. The other resolutions occurred less frequently (less than 10% of the time). I studied the influence of resolution on the distribution of MOS for the two most dominant resolutions. The SROCC between the two classes was 0.95, while the mean difference in MOS between the two sets was close to zero. This result further supports the belief that the participants' display resolutions did not significantly impact the video quality ratings they supplied.

6.3.5.3 Participants' Display Devices

Laptops and computer monitors most often used as display devices (Fig. 6.8(b)). I also studied the influence of the display device, and found that it did not noticeably impact the MOS either (the SROCC between the two groups was 0.97 and the mean difference in the MOS was close to 1).

6.3.5.4 Viewing Distances

Another parameter that I studied was the reported viewing distance. There were three categories: small (<15 inches), medium (15-30 inches) and large (>30 inches) viewing distances as shown in Fig. 6.8(c). I found that the viewing distance had only a small effect on the distribution of MOS. The SROCC between the three categories ranged between 0.91 and 0.97, while the average difference in the MOS was less than 1.

6.3.5.5 Other demographic information

I also analyzed the impact of subjects' demographics on the distribution of MOS. First, I did not find noticeable differences between the MOS distributions across the male and female populations. The SROCC between the two gender classes was 0.97, and the average difference between the MOS was about 2; female participants tended to give slightly lower scores as compared to male participants. It is possible that this might be attributed to biological differences in the perceptual systems between of the two genders; for example, it has been reported that females are more adept at distinguishing shades of

color [144].

Age did impact the MOS distribution. I compared the distributions of MOS for the four age ranges, and found that younger participants as a group delivered lower opinion scores than did older participants. These differences might be attributed to young participants having better vision [145], or it might related to differing expectations of younger and older viewers. As it can be observed in Table 6.4, the larger the difference between the age groups, the lower the SROCC. The difference between the MOS distributions becomes more subtle as the difference in the age gap increases; participants younger than 20 tended to assign lower quality scores than did participants older than 40.

Table 6.4: Spearman Correlation of the MOS distributions obtained between the different age groups.

	<20	20-30	30-40	>40
<20	1	0.84	0.82	0.79
20-30	0.84	1	0.97	0.94
30-40	0.82	0.97	1	0.96
>40	0.79	0.94	0.96	1

6.4 Performance of Video Quality Predictors

As mentioned earlier in the background section, I conducted this study with the aim to advance VQA research efforts, by providing a database that closely represents distorted videos encountered in the real world, along with a large number of accurate human opinions of them. In recent years, there

has been numerous efforts to develop blind VQA models. Noteworthy examples include simple frame-based Natural Scene Statistics (NSS) based models, NIQE [8] and BRISQUE [7], as well as more sophisticated predictors that incorporate more complex information such as motion. These include V-BLIINDS [72], VIIDEO [79], the 3D-DCT based NR-VQA predictor described in [80], the FC model [81], the statistical analysis model in [82], and the convolutional neural network model in [83].

To demonstrate the usefulness of my database, I evaluated the quality prediction performance of a number of leading blind VQA algorithms (whose code was publicly available). NIQE [8] and VIIDEO [79] are training-free models capable of outputting quality scores on video. V-BLIINDS [72] and BRISQUE [7], require training hence I learned mappings from their feature spaces to the ground truth MOS, using a support vector regressor (SVR) [98] that has been successfully deployed in many prior image and video quality models. I used the LIBSVM package [110] to implement the SVR with a radial basis function (RBF) kernel and to predict the MOS. I applied a 5-fold cross validation technique as described in [146]. To predict quality scores over the entire database, I aggregated the predicted values obtained from each fold. The NIQE [8] features were computed on non-overlapping blocks of size 96×96, then the computed NIQE distance is computed over frames and averaged over time, similar to how it was originally implemented in V-BLIINDS [72]. BRISQUE [7] was calculated over frames and averaged in time.

Figure 6.13 presents scatter plots, of NIQE [8] and VIIDEO [79] qual-

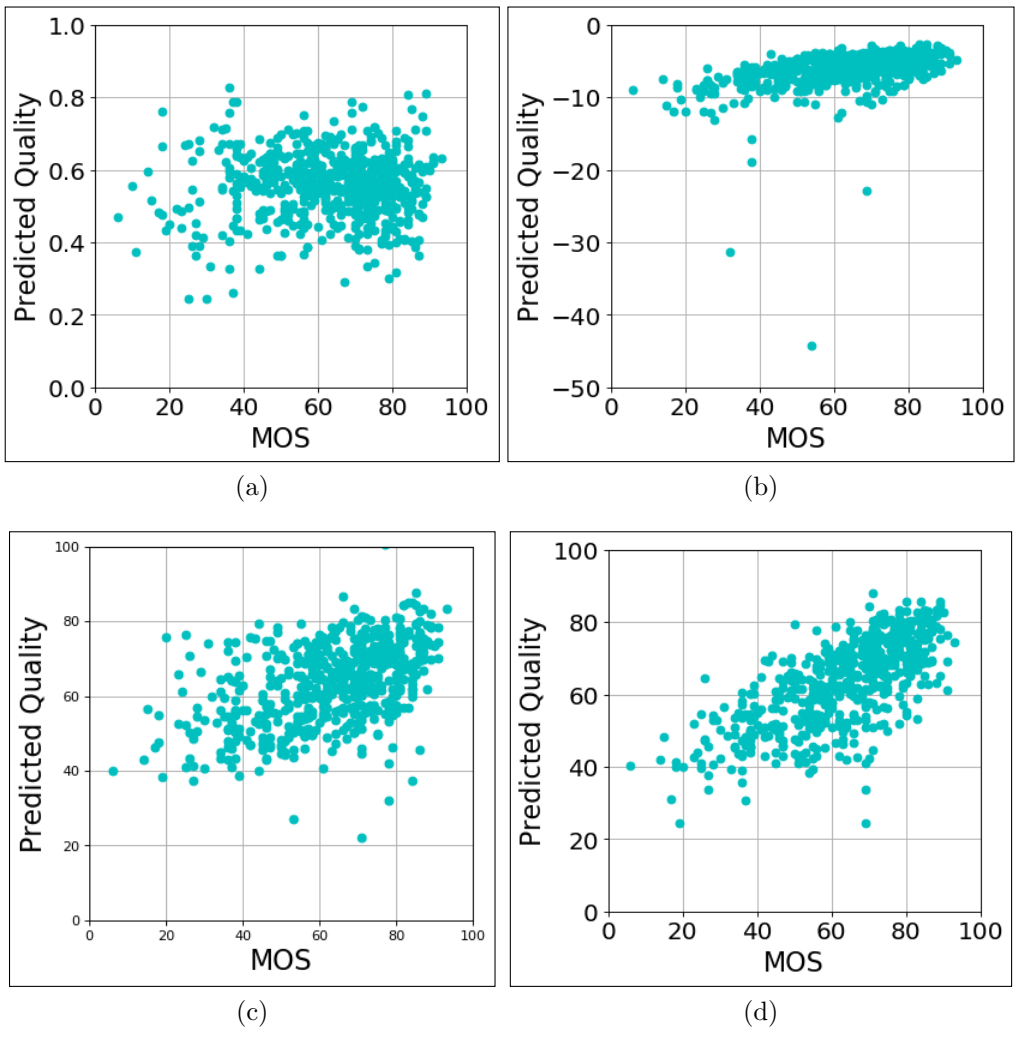


Figure 6.13: Scatter plots of the predicted quality scores versus MOS for four NR VQA models; (a) VIIDEO; (b) NIQE; (c) BRISQUE; (d) V-BLIINDS.

ity predictions, and V-BLIINDS [72] and BRISQUE [7] predictions obtained after the 5-fold cross validation. Since NIQE provides a distance measure that increases as the video becomes more distorted, I will instead analyze the quantity -NIQE for simpler visual comparison with other models. As may

be observed in Fig. 6.13(a), the predicted VIIIDEO scores correlated poorly with the ground truth MOS, while for the other models followed regular trends against the MOS, as shown in Fig. 6.13(b), 6.13(c) and 6.13(d).

I used three performance metrics to quantify the performance of the different VQA models. First, I computed the Pearson Linear Correlation Coefficient (PLCC) between the predicted quality values and MOS distributions, after applying a non-linear mapping as prescribed in [147] to the predicted quality values. Second, I computed the Root Mean Squared Error (RMSE) between the two distributions. Finally, I computed the SROCC values between the predicted quality values and MOS values. When evaluating V-BLIINDS [72] and BRISQUE [7], I randomly divided the videos into two disjoint sets (80%-20%). I used the larger set for training and the other one for testing, then I normalized my features, and fed them into the SVR module [98] to predict the MOS. I repeated this process 100 times, and computed the median PLCC, SROCC and RMSE values. A summary of the results obtained over all the models is given in Table 7.4. I could not run V-BLIINDS [72] successfully on all the videos, especially at lower resolutions. Unable to trace the source of this problem in the span of the current report, or resolve it, the results are reported on a subset of 553 out of the 585 videos. Note that computing the results on this subset led to a slim increase in the performance of NIQE [8] and BRISQUE [7]. VIIDEO's code [79]'s would not run successfully on 3 videos so I report the results for it on the remaining 582 videos. The results reported for NIQE [8] and BRISQUE [7] are on the full database,

since these algorithms could run successfully on all the videos. As may be observed, V-BLIINDS supplied the best performance in terms of the three performance metrics. However, there remain ample room for improvement, suggesting the need for developing better NR VQA models, capable of better assessing authentic, real-world video distortions.

Table 6.5: Performance Metrics Measured on the Compared VQA Models.

	PLCC	SROCC	RMSE
VIIDEO [79]	0.1366	-0.0293	16.851
NIQE [8]	0.5832	0.5635	13.857
BRISQUE [7]	0.6456	0.6072	12.908
V-BLIINDS [72]	0.7196	0.7083	11.478

6.5 Concluding Remarks

I have described the construction of a new "in the wild" video quality database, LIVE-VQC, containing 585 videos of unique contents, and impaired by authentic distortion combinations, captured by 80 users around the globe using 43 different device models. I also designed and built a crowdsourced framework to collect more than 205000 online opinion scores of the quality of these videos. The AMT subjects who participated were located in 56 different countries, represented genders about equally, and spanned a wide range of ages. The significant diversity of the subject pool raised many technical challenges owing to widely differing viewing conditions and resources. However, the framework I built proved to be robust against the many variables affecting the video rating process. While the VQA models that I tested did

not perform particularly well on the new database, this was not unexpected as existing NR VQA models were not been adequately engineered to deal with so many real-world distortions. To address this problem, in the next chapter, I will be presenting a completely blind NR VQA model.

Chapter 7

A Completely Blind Video Quality Predictor

In this chapter, I present a highly efficient, "completely blind" video quality model that relies on a unique set of directional spatio-temporal NSS features, and which does not require any kind of training. I call this model VIdeo Naturalness Assessor, or VINA. I begin this chapter by describing the features that define this model. [1]

7.1 VINA's features

Evolution has left a significant trace on the neurological resources of visual perception in response to the statistical properties of the physical natural environment [148], and images of it. Hence, the study of natural image and video statistics is highly relevant to understanding visual perception, including the perception of visual distortions [7, 8, 68, 72, 149, 150]. Following this philosophy, the VQA model I develop here utilizes measurements of the physical statistics of spatial frames as well as directional bandpass statistical

[1] A part of this chapter has been submitted in the following paper:
Zeina Sinno, Alan C. Bovik: "Spatio-temporal Measures of Naturalness" in the IEEE International Conference on Image Processing, 2019.
Zeina Sinno has constructed the model and collected the data and performed full experimental analysis of the works described therein.

space-time features to predict quality. An overview of the VINA model is shown in Fig. 7.1. I begin by describing the spatial features.

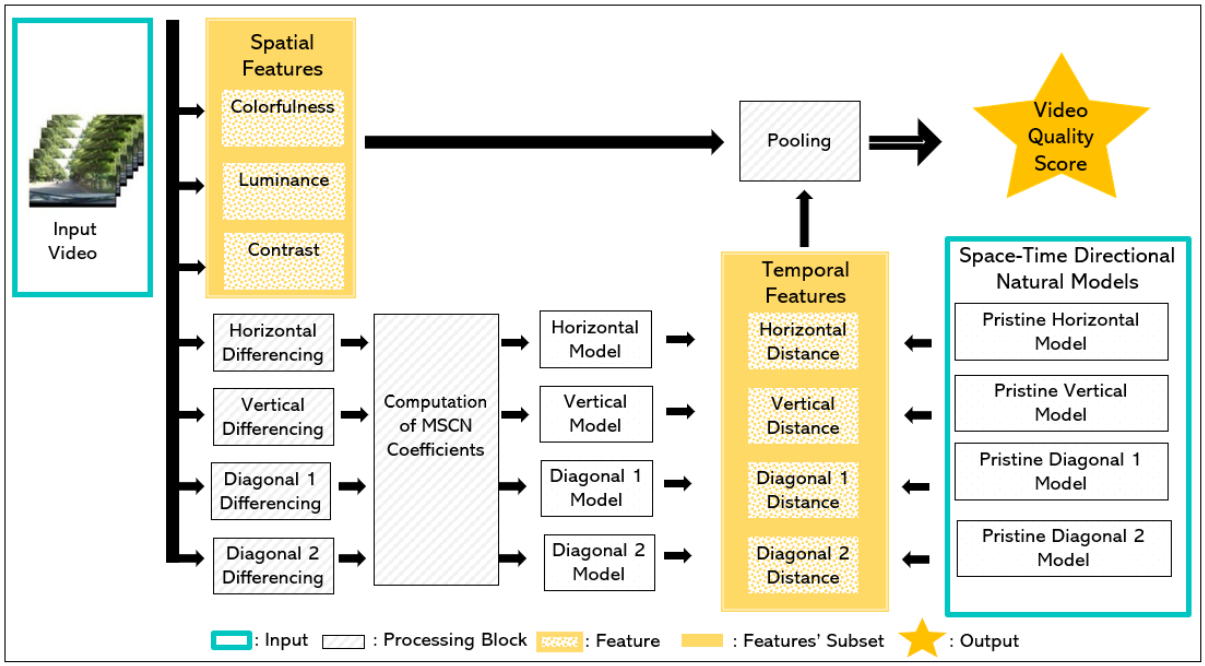


Figure 7.1: VINA's overview.

7.1.1 Spatial Features

It is difficult to assert whether certain attributes of image and videos relate more closely to *aesthetics* or to *quality*. For example, if an image is too dim or has limited color content, than it might be deemed as both lower in aesthetic quality, as well as of poor quality, even if it is not distorted in the usual sense. Towards capturing some of the attributes in this gray area that are not traditional distortions but still contribute to the percept of quality, I attempt to account for some of these properties. Subjects' ratings of image appeal

can be influenced by low-level image attributes [151]. For example, Savakis *et al.* [152] reported that human viewers found images that are colorful, well-lit, and of high contrast are considered appealing, while darker low-contrast images were considered less appealing. Studies on the perception of image naturalness [153, 154] have also revealed the importance of such attributes.

VINA incorporates kinds of low-level images non-traditional distortion properties that may also be viewed as aesthetic attributes: colorfulness, luminance and contrast. This choice of attributes agrees with the observations made in the subjective study [152].

7.1.1.1 Colorfulness

I used the popular index introduced in [155] to capture colorfulness information. This computationally approach is defined as the weighted sum of the mean and standard deviation of the cloud of pixel values along direction $rgyb$. Specifically, the colorfulness c of an image, or video frame by c is measured as follow.

First, define the rg and yb plane directions $rg = R - G$ and $yb = \frac{1}{2}(R + G) - B$, then compute the means μ_{rg} and μ_{yb} and standard deviations as and σ_{rg} and σ_{yb} along the rg and yb directions, respectively. Then, define the mean and standard deviation along the $rgyb$ direction as $\mu_{rgyb} = \sqrt{{\mu_{rg}}^2 + {\mu_{yb}}^2}$ and $\sigma_{rgyb} = \sqrt{{\sigma_{rg}}^2 + {\sigma_{yb}}^2}$, respectively.

Then finally

$$c = \sigma_{rgyb} + 0.3\mu_{rgyb}. \tag{7.1}$$

When processing videos, c is computed on a frame per frame basis so, hence denote the colorfulness of the n^{th} frame by $c(n)$, and the obtained sequence of colorfulness values for a video by $\{c\}$. Following the well-established strategy that the worst distortions, when localized in space, or in the case time, may contribute the most to perceived quality degradation, I deploy ranking methods on each of the basic low-level "aesthetic quality" features. Thus, let C_p be the mean of the lowest p^{th} percentile of colorfulness across all the frames of a video:

$$C_p = \frac{1}{N_p} \sum_{c(n) \leq c(p)} c(n) \tag{7.2}$$

where N_p is the cardinality of values of $c(n)$ in the lowest p^{th} percentiles in $\{c\}$.

I fixed the lower percentile to be the nominal value $p = 5$ for each low-level aesthetic feature to avoid any tuning bias. Hence, C_5 is the first feature in my model.

7.1.1.2 Luminance

The study in [152] also suggested that brighter images, e.g. of well lit scenes,are also more appealing to human viewers than dim images of poorly lit scenes. Hence I also compute average luminance $l(n)$ the average luminance on each frame indexed by n and use $\{l\}$ to denote the sequence of frame luminance values of a video.

Similarly to colorfulness, compute L_p, the mean of the lowest p^{th} per-

centile of the average frame luminances:

$$L_p = \frac{1}{N_p} \sum_{l(n) \leq l(p)} l(n). \tag{7.3}$$

This feature is motivated by the observation that the lower the average luminance of an image, the less desirable it often is. As mentioned above, I again fix $p = 5$, L_5 is the second feature of my model.

7.1.1.3 Contrast

Contrast is also a generic predictive feature of quality, as demonstrated in [152]. While there are several definitions of contrast available e.g., Michelson contrast [156], Weber-Fechner contrast [157] and the RMS contrast σ_{rms} [158], I utilize the more widely used σ_{rms} as defined in [158] as it is a basic statistics that is not captured by the natural scene models I will also be using. It also yields better quality prediction results.

The RMS contrast measure that I use is computed as:

$$\sigma_{rms} = \sqrt{\frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} (I - \overline{I})^2}, \tag{7.4}$$

where I is the luminance image, $i \in \{1, 2, ...w\}$ and $j \in \{1, 2, ...h\}$ are spatial indices, w and h are the frame height and width respectively and $\overline{I}$ is a local weighted average of the luminance:

$$\overline{I}(i, j) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} g_{m,n} I_{m,n}(i, j) \tag{7.5}$$

where $g_{m,n}$, $m = -M, ..., M, n = -N, ..., N$ is a 2D circularly-symmetric Gaussian weighting function sampled out to 3 standard deviations and rescaled to unit volume. I set M and N to 3. Note that (7.4) ha the interpretation of the average local contrast of I, which could be much less sensitive to contrast flattening local distortions. Let $\sigma_{rms}(n)$ denote (7.4) computed on frame n, and $\{\sigma_{rms}\}$ be the sequence of frame contrast values of a video.

Further, define the mean of the lowest p^{th} percentile of the $\{\sigma_{rms}\}$ values over all the frames of a video to be:

$$S_p = \frac{1}{N_p} \sum_{\sigma_{rms}(n) \leq \sigma_{rms}(p)} \sigma_{rms}(n) \tag{7.6}$$

where N_p is as defined before. As mentioned earlier, I take $p = 5$, S_5 becomes the third feature in my model.

7.1.2 Temporal Features

The main design of VINA is based on the premise that pristine, undistorted videos reliably present statistical regularities that are systematically and predictably degraded by distortions. Indeed, this has been demonstrated in highly successful FR [68, 159], RR [70, 160] and blind (NR) [8, 72, 106]. Recognizing that distortions of videos are intrinsically spatio-temporal phenomena, here I define first-of-a-kind space-time distortion-aware video NSS features.

Of course, videos are deeply affected by object motion, and various distortions are associated with motion, such as jitter, ghosting, motion compensation mismatches, all of which can render the perception of moving pixels un-

natural. Unfortunately, as has been previously observed [161] and commented on [72], the statistics of motion (optical flow) does not generally follow observed regularities on videos containing moving objects, whereas the statistics of *changes* over time, e.g., as captured by frame differences, are nicely regular [70]. My unique concept is to capture and analyze the natural statistics of frame differences that are oriented in space-time. It turns out that these are more reliable video predictions than temporal-only differences. Specifically, I devise a space-time directional natural video statistics (NVS) model based on the effects of distortions on the statistics of displaced frame differences along the four cardinal directions horizontal, vertical, and both diagonals.

7.1.2.1 Space-Time Directional Models

For a given video containing T luminance frames $\{I_1,\ I_2,\ I_3..., I_T\}$ of width and height dimensions w and h, define four directional temporal differences between each pair of adjacent frames as depicted in Fig. 7.2. At each frame index t and spatial coordinate (i, j), define the set of oriented spatial differences:

$$D_H(i,j)_t = I_t(i,j) - I_{t+1}(i,j-1) \tag{7.7}$$

$$D_V(i,j)_t = I_t(i,j) - I_{t+1}(i-1,j) \tag{7.8}$$

$$D_{D_1}(i,j)_t = I_t(i,j) - I_{t+1}(i-1,j-1) \tag{7.9}$$

$$D_{D_2}(i,j)_t = I_t(i,j) - I_{t+1}(i-1,j+1). \tag{7.10}$$

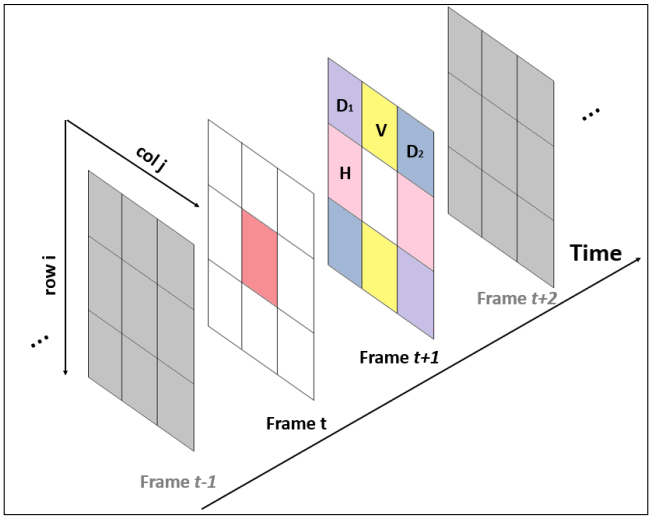


Figure 7.2: Depiction of pixels in frame t and the four pixels in frame $t+1$ it is differenced with.

Then define the local 4 directional average values $\overline{D_H}(i,j)_t$, $\overline{D_V}(i,j)_t$, $\overline{D_{D_1}}(i,j)_t$ and $\overline{D_{D_2}}(i,j)_t$ in the same way as (7.5), using the same weighting function g, as well as four variance fields:

$$\sigma^2_{(\cdot)}(i,j) = \sum_{i=1}^{w}\sum_{j=1}^{h}\left(D_{(\cdot)} - \overline{D_{(\cdot)}}\right)^2 \tag{7.11}$$

where $(\cdot)$ denotes any of the four cardinal directions (horizontal, vertical or one of the diagonals).

These are used to obtain the mean subtracted contrast normalized coefficients (MSCN):

$$\hat{D}_{(\cdot)}(i,j) = \frac{D_{(\cdot)}(i,j) - \overline{D_{(\cdot)}}(i,j)}{\sigma_{(\cdot)}(i,j) + 1}. \tag{7.12}$$

MSCN coefficients of images have been used to devise highly efficient NR IQA models [7, 8] and the MSCN of simple frame differences have been

used to conduct NR VQA of H264 compressed and rescaled videos [162]. In these algorithms, parametric fits of the feature distributions to generalized gaussian density (GGD) model are used to construct the NR IQA engines.

Following [8], each frame is partitioned into patches of size 96×96, within the MSCN coefficients and the distribution of those coefficients in each patch are then fitted using a generalized Gaussian distribution (GGD) function of zero mean using the moment matching function described in (7.12) are computed along each direction. The MSCN histograms are then each fit with a GGD of the form:

$$f(x;\alpha;\beta) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} \exp[-(\frac{|x|}{\beta})^{\alpha}] \tag{7.13}$$

where $\Gamma(\cdot)$ is the gamma function:

$$\Gamma(a) = \int_0^{\infty} t^{a-1}e^{-t}dt \quad a > 0, \tag{7.14}$$

using the moment matching function described in [107]. The shape parameter α controls the shape of the distribution, while β controls its variance. For each directional MSCN patch histogram, the best-fitting α and β parameters are thus computed, and then aggregated across all patches and all frames, by aggregating the pair of features (α,β) from each patch, in each frame of a video, yielding four directional NVS models of any given input video, which I denote by M_H, M_V, M_{D_1}, and M_{D_2}. Each vector NVS model is of size $p \times 2$, where p is the total number of 96×96 patches in a certain video: $p = \lfloor \frac{w}{96} \rfloor \times \lfloor \frac{h}{96} \rfloor \times T$, provided that the patches partition the frames without any overlap. Each directional model is thus obtained on all of the luminance frames of a video.

7.1.2.2 Construction of the Pristine Directional Models

As I will describe shortly, VINA predicts the quality of a video by making a statistical comparison to a pristine video model. To construct the pristine model, I collected about 600 high quality videos from sets of videos in established VQA databases, such as the LIVE Mobile Video Quality Database [45] and the MCL-V [48] database, along with other high resolution, high quality content obtained from across the web from sources such as Shutterstock [163], Videezy [164], Pixabay [165], and Videvo [166]. The pristine videos that I obtained from open source websites were mostly of resolution of 1920×1080 and above. Provided that consumer devices provide a wide range of resolutions as reflected by the LIVE-VQC database [137] in Table 6.2. As a check, I conducted a small scale subjective study involving 3 human subjects who each gave a binary evaluation of each of the set of collected videos as either visually distorted or not. I only retained those videos that were agreed to be non-distorted by all the subjects. Random sets of the resulting pristine videos were then downscaled using bicubic interpolation [167], yielding four collections of pristine videos of resolutions: 1280×720, 960×640, and 640×360, each containing about 50 videos. This was done to allow for resolution-specific prediction models.

Similar MSCN processing steps were applied to each of the pristine videos. In fact, the processing was identical to the one above, with one difference: following the approach in [8], the MSCN coefficients were only computed on the sharpest patches. Specifically, the sharpest 5% of the patches in each

frame were processed on videos of resolutions 1280×720, while the sharpest 25% of frame patches were processed on the videos of resolutions 960×640 and 640×360. A larger percentage was used on the lower resolution videos, in order to ensure that an adequate, representative quantity of lower-resolution patches were included. The best-fitting parameters (α,β) were collected and horizontally aggregated over space and time across all videos for each of the four orientations, yielding 4 pristine models P_H, P_V, P_{D_1}, and P_{D_2}. Each model was contained in a matrix of size 400000×2, since the total number of sharp patches across all pristine videos was 400000.

7.1.2.3 Measure of the Directional Naturalness

Similar to the approach described in NIQE [8], I computed a modified Mahalanobis distance [168] $\Delta_{(\cdot)}$ between each directional pristine model and the corresponding directional input video models:

$$\Delta_{(\cdot)}(\gamma_{M_{(\cdot)}}, \gamma_{P_{(\cdot)}}, \Sigma_{M_{(\cdot)}}, \Sigma_{P_{(\cdot)}}) = \sqrt{(\gamma_{M_{(\cdot)}} - \gamma_{P_{(\cdot)}})^T (\frac{\Sigma_{M_{(\cdot)}} + \Sigma_{P_{(\cdot)}}}{2})^{-1} (\gamma_{M_{(\cdot)}} - \gamma_{P_{(\cdot)}})}, \tag{7.15}$$

where $\gamma_{M_{(\cdot)}}$ and $\gamma_{P_{(\cdot)}}$ are the means of the models $M_{(\cdot)}$ and $P_{(\cdot)}$ respectively, and $\Sigma_{M_{(\cdot)}}$ and $\Sigma_{P_{(\cdot)}}$ are their respective covariance matrices. The resulting distances Δ_H, Δ_V, Δ_{D_1}, and Δ_{D_2} each capture the degree of directional naturalness (or lack thereof) relative to the pristine video behavior. A simple way to understand these quantities is that the higher their values, then the less natural the directional space-time changes are, which often arise from distorted object and/or distortion motion.

7.2 Pooling Mechanism

The method of feature pooling I use here of the temporal features was inspired by the completely blind NR IQA model NIQE [8], which considers a natural model of images in the form of an NSS feature dictionary, and computes the distance between an image to this model to obtain a measure of how natural or undistorted the image is. A lower distance is desirable, indicating that the image is less distorted, while a higher distance indicates that the image is more distorted. Unlike NIQE, my features are more diverse, hence I pool them into groups. My method of pooling is designed in a product form, so that a severe loss of quality affecting any single, or group of features can adequately reduce the overall quality prediction. As a way of normalizing each of the feature contributions to the product, I exponentiate each feature after scaling it by a normalization factor representative of the typical order of magnitude that feature value takes. Thus, C_5 and L_5 are each scaled by 100, while every other feature is scaled by 10. This avoids any single feature dominating the others. The signs of the temporal features were also flipped, since they each measure a naturalness distance, hence are negatively correlated. Finally, the overall VINA score is computed as:

$$VINA = e^{0.1C_5} \times e^{0.1L_5} \times e^{S_5} \times e^{-\Delta_H} \times e^{-\Delta_V} \times e^{-\Delta_{D_1}} \times e^{-\Delta_{D_1}} \tag{7.16}$$

While VINA in this form performs well, visualization of VINA scores when plotted is enhanced by the monotonic assignment $VINA \leftarrow (VINA)^{-0.1}$, which I use in all the following comparisons.

7.3 Results

To demonstrate the efficacy of my predictor, I first evaluated the quality prediction performance of each feature used and then compared VINA's overall performance against other leading VQA models.

7.3.0.1 Performance of the Individual Features

I studied the performance of my model by first computing the Pearson Linear Correlation Coefficient (PLCC) between the values of the features and MOS, after applying a non-linear mapping as prescribed in [147] to the predicted quality values. Second, I computed the Root Mean Squared Error (RMSE) between the two distributions. Finally, I computed the SROCC values between the features and MOS values. As mentioned previously, I deployed a ranking method on each of the basic low-level spatial features. I found that the lower the percentile p is, the better the performance, so I fixed $p = 5\%$. I also studied the performance of VINA as a function of the percentile p. Table 7.1 tabulates the prediction performances of the colorfulness, luminance and contrast features when applied in isolation, clearly promoting the effectiveness of lower values of p. I also report the performance of the overall VINA predictor as a function of p in Table 7.2. My choice of $p = 5$ is slightly sub-optimal but is likely more robust in practice given that it averages several feature values, rather than relying on an extreme value.

I also plotted the scatter of C_5, L_5 and S_5 against the ground truth MOS in Fig. 7.3.

Table 7.1: Performance of C_p, L_p, and S_p as a function of p.

	C_1	C_5	C_{50}	C_{100}	L_1	L_5	L_{50}	L_{100}	S_1	S_5	S_{50}	S_{100}
PLCC	0.2607	**0.1605**	0.1166	0.1138	0.5311	**0.5267**	0.5012	0.4771	0.5913	**0.5804**	0.5360	0.5066
SROCC	0.2014	**0.1830**	0.1146	0.0651	0.4199	**0.4119**	0.3560	0.3121	0.5782	**0.5621**	0.5060	0.4608
RMSE	16.46	**16.84**	16.94	16.95	14.45	**14.50**	14.76	14.50	13.76	**13.89**	14.40	14.71

Table 7.2: VINA's performance as a function of p.

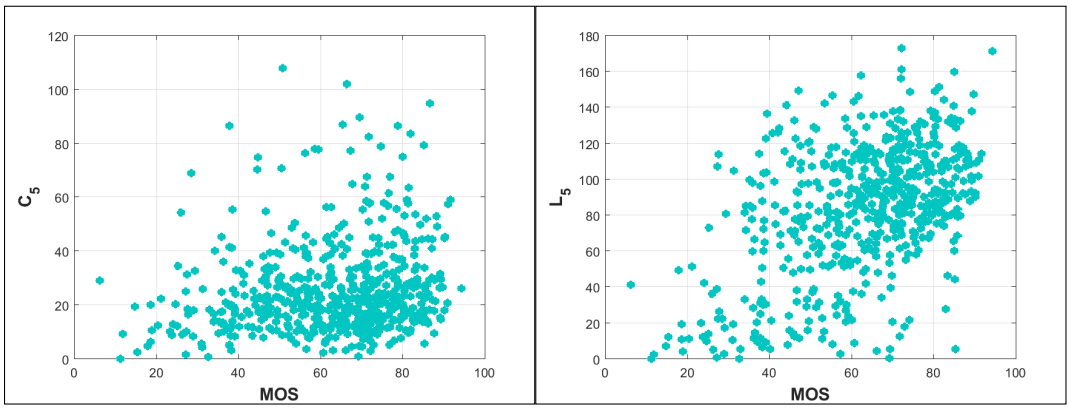
	$p = 1$	$p = 5$	$p = 50$	$p = 100$
PLCC	0.6863	**0.6808**	0.6585	0.6424
SROCC	0.6689	**0.6619**	0.6302	0.6050
RMSE	12.41	**12.50**	12.84	13.07

A degree of linearity may be observed in the trends of the features, although there is a fair degree of spread. This is not unexpected, since these "aesthetic quality" features do not capture the local distortion artifacts that dominate quality perception. Instead, they contribute somewhat weaker, albeit complementary (and hence valuable) quality-aware information. Among these, C_5 the weakest quality predictor, while S_5 is the strongest, which not surprising given that it relates to image contrast.

I repeated the same analysis on the spatio-temporal features, using their negative values $-\Delta_H$, $-\Delta_V$, $-\Delta_{D_1}$ and $-\Delta_{D_2}$. Table 7.3 lists the SROCC, PLCC and RMSE values computed between these features and MOS. I also plotted those features as a function of MOS in Fig. 7.4. The spatio-temporal features performed better than the spatial aesthetic ones that is more richly sensitive to distortion along multiple spatial-temporal orientation, which could arise from motion perturbations, moving artifacts, and local flickers. It is interesting that each of the spatio-temporal features performed better than NIQE [8] (see Table 7.4).

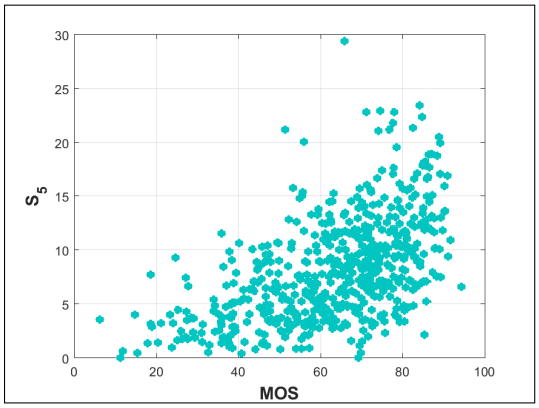
Table 7.3: Performance of Δ_H, Δ_V, Δ_{D_1} and Δ_{D_2} vs MOS.

	$-\Delta_H$	$-\Delta_{D_1}$	$-\Delta_V$	$-\Delta_{D_2}$
PLCC	0.6054	0.6062	0.6069	0.6048
SROCC	0.5968	0.5986	0.5994	0.5963
RMSE	13.58	13.57	13.56	13.58



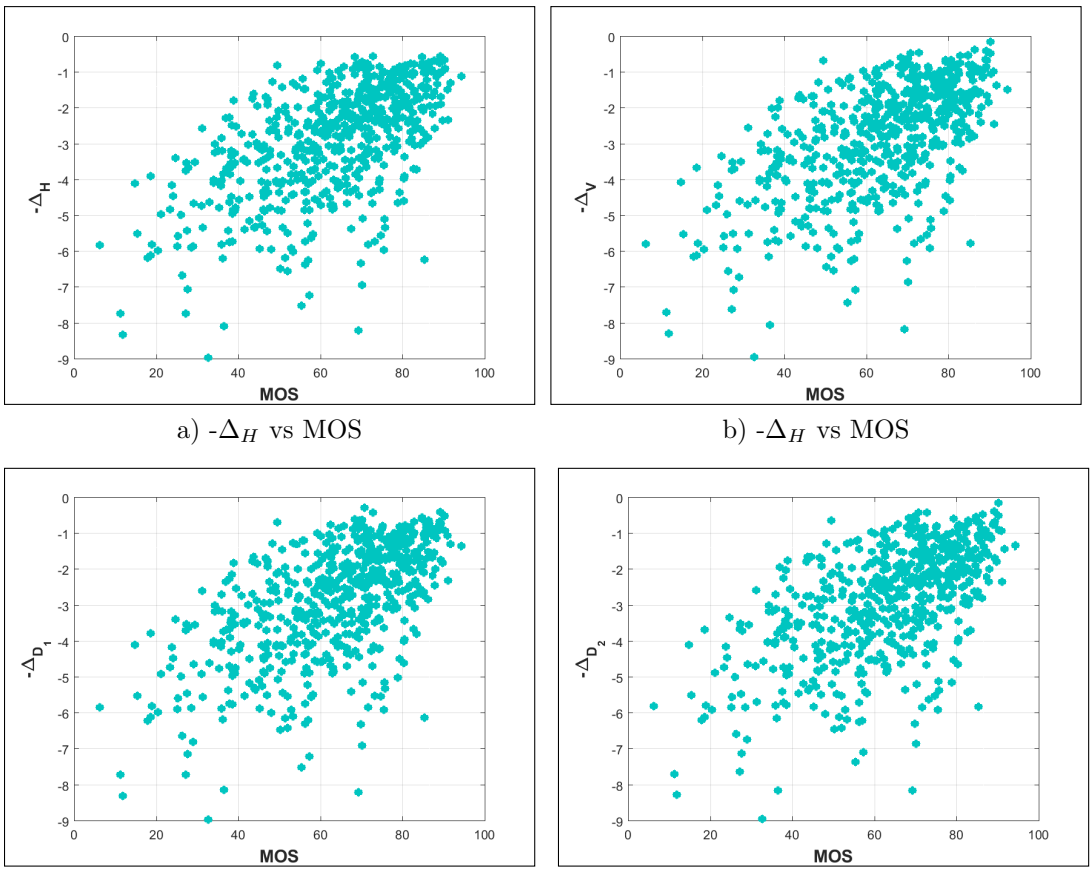
a) C_5 vs MOS

b) L_5 vs MOS



c) S_5 vs MOS

Figure 7.3: Scatter plots of C_5, L_5, and S_5 vs MOS.



a) -Δ_H vs MOS

b) -Δ_H vs MOS

c) -Δ_{D_1} vs MOS

d) -Δ_{D_2} vs MOS

Figure 7.4: Scatter plots of the temporal distances Δ_H, Δ_V, Δ_{D_1} and Δ_{D_2} vs MOS.

7.3.0.2 Performance of VINA

I plotted the distribution of the numerical quality predictions produced by VINA against MOS in Fig. 7.5 and observed a general linear trend. Furthermore, I evaluated the quality prediction performance of VINA against a number of leading blind and completely blind VQA models (whose code is pub-

licly available). I used three performance metrics to quantify the performance of the different VQA models; PLCC, SROCC and RMSE. I applied non-linear regression to the predicted scores when computing PLCC and RMSE. as recommended in [147].

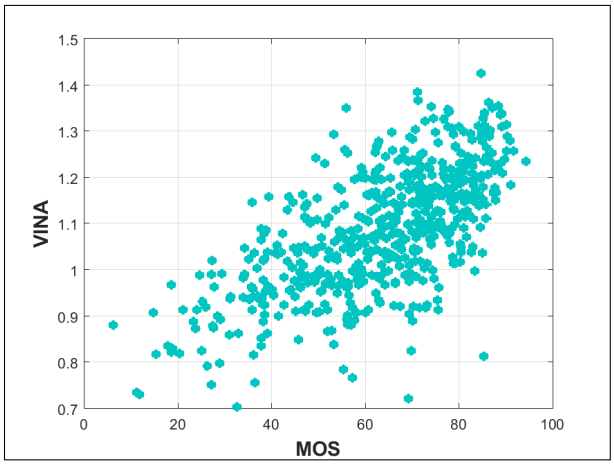


Figure 7.5: Distribution of VINA's scores vs MOS.

The models I considered are NIQE [8] model that the NR algorithms V-BLIINDS [72] and BRISQUE [7], both of which require training. The latter two models learned mappings from their feature spaces to ground truth MOS using a support vector regressor (SVR) [98], which has been successfully deployed in many prior image and video quality models. I used the LIBSVM package [110] to implement the SVR, to predict MOS using a radial basis function (RBF) kernel. I randomly divided the videos into two disjoint sets (80%-20%). I used the larger set for training and the other one for testing, after normalizing the features [98] to predict the MOS. I repeated train-test process 100 times, then computed the median PLCC, SROCC and RMSE

values between the predictions and MOS. V-BLIINDS [72] could not be run successfully on a few of the videos, especially at lower resolutions. Hence. the results using V-BLIINDS are reported on a subset of 553 of the 585 videos. I note that applying NIQE [8] and BRISQUE [7] on the same subset slightly improved their performances. The results reported for VINA, NIQE [8] and BRISQUE [7] were computed on the entire database. The NIQE [8] features were computed on non-overlapping blocks of size 96×96, then the computed NIQE distances were computed on each frame, then averaged over time, similar to its implementation in V-BLIINDS [72], which incorporates NIQE [8]. BRISQUE [7] was calculated on each frame, then averaged over time. A summary of the results obtained for all the models is given in Table 7.4. It may be observed that VINA outperformed the popular NIQE model which is also completely blind, while its performance was only beaten by V-BLIINDS, which requires the expensive computation of motion. It also requires training, which costs its generalizibility into some doubt. VINA is much more computationally efficient than V-BLIINDS, while nearly matching its prediction performance without training.

Table 7.4: Performance Metrics Measured on the Compared VQA Models.

	Training Videos	Testing Videos	PLCC	SROCC	RMSE
NIQE [8]	0	585	0.5832	0.5635	13.857
BRISQUE [7]	468	117	0.6456	0.6072	12.908
VINA	0	585	0.6808	0.6619	12.50
V-BLIINDS [72]	442	111	0.7196	0.7083	11.478

7.4 Concluding Remarks

I presented a new, completely blind VQA model, called VINA. VINA does not require any sort of training since it uses features that capture the loss of spatial and temporal naturalness in video which is degraded by the presence of distortions. Three basic aesthetic-related features that capture colorfulness, luminance and contrast are used, along with a directional space-time naturalness model. I exploited the idea that pristine or undistorted videos obey statistical regularities that are violated by distortions. VINA is computationally efficient, while providing prediction results that outperform other completely blind video quality predictors.

Chapter 8

Conclusion and Future Directions

In this dissertation, I presented a new closed form bivariate spatial correlation model of bandpass and normalized image samples. The model was developed on high-quality naturalistic photographs, and was shown to hold as well for the case of distorted images, however its parameters change consistently based on the type and amount of distortion introduced. Provided this important property, I exploited it in order to build a build IQA model and a model for predicting 3D visual discomfort.

A second direction in this dissertation was to tackle to blind video quality prediction problem. To do so, I constructed a new video quality database containing 585 videos of unique contents, impaired by authentic distortion combinations, captured by 80 users around the globe using 43 different device models. To gather quality labels, I built a framework to crowdsource more than 205000 online opinion scores. The subjects who participated in my study were working under widely differing viewing conditions and resources (bandwidth, hardware...). So my design took into account all these factors which can affect the quality ratings. I demonstrated that the design is robust against the many variables affecting the video rating process.

Provided the importance of the bivariate NSS in tackling several image quality related applications, I studied the relationship between shifted frame differences and developed a completely blind model VQA model, VINA. VINA utilizes measurements of the physical statistics of spatial frames as well as directional bandpass statistical space-time features to predict quality. I was able to demonstrate that VINA is the best performing completely blind VQA model although it is computationally efficient as it does not require any motion related computations.

I believe that VINA can be extended further. An important avenue for improving this predictor would be by extending directional naturalness to include more motion directions, greater spatial and temporal displacements, and multiple space-time scales. I also believe that even larger and more diverse VQA databases targeting user generated content would be of significant importance. Being able to predict the quality of videos without any training and in the absence of additional information is an important endeavor for a variety of invested practitioners, such as camera designers, cloud engineers, and users who could be directed to recapture videos of poor quality. Videos provided by users are very diverse in terms of content, style, encountered distortions, resolution, capture devices, compression protocols, and so on, hence building larger databases that better represent all these factors would enable the design of better VQA models, and better model verification and algorithm comparisons. However, such databases are not yet available, since their creation would require very substantial time and expense.

Lastly, while I believe in the great potential of today's deep neural models which given adequate subjective data, could improve on current VQA algorithms, there is also the need for effective lightweight models that can be massively rolled out. VINA falls in this category. Light weight models can be expected to dominate streaming VQA applications for several years (at least), given the dearth of large datasets needed to create effective deep models, and the substantial hardware requirements and hardware compatibility problems needed to implement them at scale.

Bibliography

[1] D. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am.*, vol. 4, no. 12, pp. 2379–2394, 1987.

[2] A. D. D'Antona, J. S. Perry, and W. S. Geisler, "Humans make efficient use of natural image statistics when performing spatial interpolation," *J. Vision*, vol. 13, no. 14, p. 11, 2013.

[3] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[4] M. Clark and A. C. Bovik, "Experiments in segmenting texton patterns using localized spatial filters," *Patt. Recog.*, vol. 22, no. 6, pp. 707–717, 1989.

[5] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.

[6] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[7] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. on Imag. Proc.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[8] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Sign. Process, Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

[9] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, 2011.

[10] C. Su, L. K. Cormack, and A. C. Bovik, "Color and depth priors in natural images," *IEEE Trans. on Imag. Proc.*, vol. 22, no. 6, pp. 2259–2274, 2013.

[11] C. C. Su, L. Cormack, and A. C. Bovik, "Closed-form correlation model of oriented bandpass natural images," *IEEE Sign. Process. Lett.*, vol. 22, no. 1, pp. 21–25, Jan 2015.

[12] "Cisco visual networking index: Forecast and methodology, 20162021," *Cisco*, Sep 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[13] S. Bergman, "We spend a billion hours a day on YouTube, more than Netflix and Facebook video combined," *Forbes*, Feb

2017, accessed on Jan. 2 2018. [Online]. Available: https://www.forbes.com/sites/sirenabergman/2017/02/28/we-spend-a-billion-hours-a-day-on-youtube-more-than-netflix-and-facebook-video-combined/#38c001ba5ebd

[14] C. Goodrow, "You know whats cool? A billion hours," Feb 2017, accessed on Jan. 2 2018. [Online]. Available: https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html

[15] S. Fiegrman, "Netflix adds 9 million paying subscribers, but stock falls," Jan 2019, accessed on Jan. 17 2019. [Online]. Available: https://www.cnn.com/2019/01/17/media/netflix-earnings-q4/index.html

[16] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," *SPIE Int'l Symp. Opt. Sci., Eng., Instrum.*, pp. 188–195, 1999.

[17] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Process.*, vol. 10, no. 11, pp. 1647–1658, 2001.

[18] L. Sendur and I. W. Selesnick, "Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency," *IEEE Trans. Signal Process.*, vol. 50, no. 11, pp. 2744–2756, 2002.

[19] D.-Y. Po and M. N. Do, "Directional multiscale modeling of images using the contourlet transform," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1610–1620, 2006.

[20] D. Mumford and B. Gidas, "Stochastic models for generic images," *Quarterly App. Math.*, vol. 59, no. 1, pp. 85–112, 2001.

[21] A. B. Lee, D. Mumford, and J. Huang, "Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model," *Int. J. Comput. Vision*, vol. 41, no. 1-2, pp. 35–59, 2001.

[22] C. C. Su, L. K. Cormack, and A. C. Bovik, "Bivariate statistical modeling of color and range in natural scenes," *Proc. SPIE, Human Vis. Electron. Imag. XIX*, vol. 9014, Feb. 2014.

[23] C. Su, L. Cormack, and A. Bovik, "Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1685–1699, May 2015.

[24] Z. Sinno and A. C. Bovik, "Generalizing a closed-form correlation model of oriented bandpass natural images," in *IEEE Glob. Conf. Sig. Inf. Process. (GlobalSIP)*, 2015, pp. 373–377.

[25] Z. Sinno and A. C. Bovik, "Relating spatial and spectral models of oriented bandpass natural images," in *IEEE South. Symp. Image Anal. Interpret.(SSIAI)*, 2016, pp. 89–92.

[26] D. J. Tolhurst, Y. Tadmor, and T. Chao, "Amplitude spectra of natural images," *Opth. Phys. Opt.*, vol. 12, no. 2, pp. 229–232, 1992.

[27] J. B. Johnson, "Johnson and 1/f noise," *Nature*, vol. 119, p. 50, 1927.

[28] J. M. Halley, "Ecology, evolution and 1f-noise," *Trends in Ecology & Evolution*, vol. 11, no. 1, pp. 33–37, 1996.

[29] A. E. Cohen, L. J. H. Gonzalez, A., O. L. Petchey, D. Wildman, and J. E. Cohen, "A novel experimental apparatus to study the impact of white noise and 1/f noise on animal populations," *Proc. of the Roy. Soc. of London B: Bio. Sci.*, vol. 265, no. 1390, pp. 11–15, 1998.

[30] R. T. Baillie, "Long memory processes and fractional integration in econometrics," *J. Econometrics*, vol. 73, no. 1, pp. 5–59, 1996.

[31] D. L. Gilden, T. Thornton, and M. W. Mallon, "1/f noise in human cognition," *Science*, vol. 267, no. 5205, pp. 1837–1839, 1995.

[32] R. F. Voss and J. Clarke, "1/f noise in music: Music from 1/f noise," *J. Acoust. Soc. of Amer.*, vol. 63, no. 1, pp. 258–263, 1978.

[33] A. H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional brownian motion," *IEEE Trans. Info. Theo.*, vol. 38, no. 2, pp. 904–909, 1992.

[34] B. B. Mandelbrot and J. W. Van Ness, "Fractional brownian motions, fractional noises and applications," *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.

[35] B. Mandelbrot, *Fractals: Form, Chance, and Dimension*, 1977.

[36] M. A. Riley and G. C. Van Orden, "Tutorials in contemporary nonlinear methods for the behavioral sciences," p. 2005, 2005.

[37] C. Carlson, "Thresholds for perceived image sharpness," *Photog. Sci. Engng.*, vol. 22, pp. 69–71, 1982.

[38] M. Carandini, D. J. Heeger, and A. J. Movshon, "Linearity and normalization in simple cells of the macaque primary visual cortex," *J. Neurosci.*, vol. 17, no. 21, pp. 8621–8644, 1997.

[39] B. A. Olshausen and D. J. Field, "How close are we to understanding v1?" *Neural Comp.*, vol. 17, no. 8, pp. 1665–1699, 2005.

[40] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," *Adv. in Neur. Infor. Proc. Sys.*, pp. 873–880, 2008.

[41] M. S. Keshner, "1/f noise," *Proc. IEEE*, vol. 70, no. 3, pp. 212–218, 1982.

[42] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.

[43] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H. 264/AVC video database for the evaluation of quality metrics," *Int. Conf. Acous. Sp. Sign. Process. (ICASSP)*, pp. 2430–2433, 2010.

[44] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, and A. Bovik, "Modeling the time varying subjective quality of HTTP video streams with rate adaptations," *IEEE Trans. on Image Process.*, vol. 23, no. 5, pp. 2206–2221, 2014.

[45] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Select. Topics Sign. Process.*, vol. 6, no. 6, pp. 652–671, 2012.

[46] C. Keimel, J. Habigt, T. Habigt, M. Rothbucher, and K. Diepold, "Visual quality of current coding technologies at high definition IPTV bitrates," *IEEE Int'l Wkshp. Multidim. Sign. Process.*, pp. 390–393, 2010.

[47] C. Keimel, A. Redl, and K. Diepold, "The TUM high definition video datasets," *Int'l Wkshp. Qual. Multim. Exper.*, pp. 97–102, 2012.

[48] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Repres.*, vol. 30, pp. 1–9, 2015.

[49] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang *et al.*, "Videoset: A large-scale compressed video quality dataset based on jnd measurement," *J. of Vis. Comm. Im. Rep.*, vol. 46, pp. 292–302, 2017.

[50] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circ. Syst. Video Technol.*, 2018.

[51] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014 a database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, 2016.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comp. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[53] J. A. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel, "Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal," *ACM Int. Wkshp Crowds. Multim.*, pp. 29–34, 2013.

[54] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2016.

[55] K. T. Chen, C. J. Chang, C. C. Wu, Y. C. Chang, and C. L. Lei, "Quadrant of euphoria: A crowdsourcing platform for QoE assessment," *IEEE Net.*, vol. 24, no. 2, 2010.

[56] M. Seufert and T. Hoßfeld, "One shot crowdtesting: Approaching the extremes of crowdsourced subjective quality testing," *Wkshp Perc. Qual. Sys. (PQS 2016)*, pp. 122–126, 2016.

[57] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd a framework for crowd-based quality evaluation," *Pict. Cod. Symp. (PCS), 2012*, pp. 245–248, 2012.

[58] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," *Qual. Mult. Exp. (QoMEX)*, pp. 1–6, 2017.

[59] "Methodology for the subjective assessment of the quality of television pictures." ITU-R Rec. BT. 500-13, 2012.

[60] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multim.*, vol. 16, no. 2, pp. 541–558, 2014.

[61] Ó. Figuerola Salas, V. Adzic, A. Shah, and H. Kalva, "Assessing internet video quality using crowdsourcing," *Proc. ACM Int'; Wkshp. Crowd. Multim.*, pp. 23–28, 2013.

[62] M. Shahid, J. Søgaard, J. Pokhrel, K. Brunnström, K. Wang, S. Tavakoli, and N. Gracia, "Crowdsourcing based subjective quality assessment of

adaptive video streaming," *Wkshp. Qual. Multim. Exper.*, pp. 53–54, 2014.

[63] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Comm. Surv. Tutorials*, vol. 17, no. 2, pp. 1126–1165, 2015.

[64] B. Rainer and C. Timmerer, "Quality of experience of web-based adaptive HTTP streaming clients in real-world environments using crowdsourcing," *ACM Wkshp. Desig. Quality Deployment Adapt. Video Streaming*, pp. 19–24, 2014.

[65] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Imag. Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[66] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conf. Sign. Sys. Comp.*, vol. 2. Ieee, 2003, pp. 1398–1402.

[67] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Jun 2016, accessed on Jan. 3 2019. [Online]. Available: https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652

[68] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," *First Int. Worksh. Vid. Proc. Quall. Met.*

Cons. Electr., pp. 23–25, 2005.

[69] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, 2010.

[70] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circ. Sys. Vid. Techn.*, vol. 23, no. 4, pp. 684–694, 2013.

[71] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transa. Broadcast.*, vol. 50, no. 3, pp. 312–322, 2004.

[72] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

[73] S. J. Daly, "Visible differences predictor: an algorithm for the assessment of image fidelity," *SPIE Proc. Hum. Vis. Vis. Process. Dig. Disp.*, vol. 1666, pp. 2–16, 1992.

[74] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inform. Th.*, vol. 20, no. 4, pp. 525–536, 1974.

[75] J. Lubin and D. Fibush, "Sarnoff jnd vision model," 1997.

[76] C. J. Van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Dig. Vid. Comp. Alg. Tech.*, vol. 2668. International Society for Optics and Photonics, 1996, pp. 450–462.

[77] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, 2007.

[78] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *IEEE Int. Conf. Image Process.* IEEE, 2011, pp. 2505–2508.

[79] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2016.

[80] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, 2016.

[81] H. Men, H. Lin, and D. Saupe, "Empirical evaluation of no-reference vqa methods on a natural video quality database," *Wkshp. Qual. Multim. Exper.*, pp. 1–3, 2017.

[82] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE*

Trans. Circ. Sys. Vid. Techn., vol. 25, no. 4, pp. 533–546, 2015.

[83] C. Wang, L. Su, and Q. Huang, "Cnn-mr for no reference video quality assessment," *Inform. Sci. Contr. Eng. (ICISCE)*, pp. 224–228, 2017.

[84] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov 2006.

[85] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Machine Intell.*, no. 9, pp. 891–906, 1991.

[86] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," *Proc. IEEE Int. Conf. Imag. Process.*, vol. 3, pp. 3444–3444, 1995.

[87] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Phys. Rev. Lett.*, vol. 73, no. 6, p. 814, 1994.

[88] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, "Parameter estimation for multivariate generalized gaussian distributions," *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5960–5971, 2013.

[89] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *App. Comp. Harm. Anal.*, vol. 11, no. 1, pp. 89–123, 2001.

[90] M. Fisz, "The limiting distribution of a function of two independent random variables and its statistical application," in *Colloq. Math.*, vol. 2, no. 3, 1955, pp. 138–146.

[91] P. Fryzlewicz and G. P. Nason, "A haar-fisz algorithm for poisson intensity estimation," *J. Comp. Graph. Stat.*, vol. 13, no. 3, pp. 621–638, 2004.

[92] B. A. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[93] P. E. Gill and W. Murray, "Quasi-newton methods for unconstrained optimization," *IMA J. of App. Math.*, vol. 9, no. 1, pp. 91–108, 1972.

[94] E. Larson and D. Chandler, "Categorical image quality CSIQ database 2009." [Online]. Available: http://vision.okstate.edu/csiq

[95] Y. Horita, Y. Kawayoke, and Z. M. Parvez Sazzad, "Image quality evaluation database."

[96] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, 2001.

[97] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, and F. Battisti, "Image database

tid2013: Peculiarities, results and perspectives," *Sign. Process. Im. Comm.*, vol. 30, pp. 57–77, 2015.

[98] S. A. J. W. R. C. Schölkopf, B. and P. L. Bartlett, "New support vector algorithms," *Neur. Comp.*, vol. 12, no. 5, pp. 1207–1245, 2000.

[99] S. Heyman, "Photos, photos everywhere," Jul 2015, accessed on Jan. 2 2018. [Online]. Available: https://www.nytimes.com/2015/07/23/arts/\international/photos-photos-everywhere.html

[100] J. Chen, Y. Zhang, L. Liang, S. Ma, R. Wang, and W. Gao, "A no-reference blocking artifacts metric using selective gradient and plainness measures," pp. 894–897, 2008.

[101] S. A. Golestaneh and D. M. Chandler, "No-reference quality assessment of *JPEG* images via a quality relevance map," *IEEE Sign. Process. Lett.*, vol. 21, no. 2, pp. 155–158, 2014.

[102] R. Barland and A. Saadane, "Reference free quality metric using a region-based attention model for *JPEG*-2000 compressed images," *Proc. SPIE Electronic Imaging 2006*, p. 605905, 2006.

[103] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," *IEEE Conf. on Comput. Vis. Patt. Rec.*, pp. 305–312, 2011.

[104] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Trans. Imag. Proc.*, vol. 21, no. 7, pp. 3129–3138, 2012.

[105] A. Saha and Q. M. J. Wu, "Utilizing image scales towards totally training free blind image quality assessment," *IEEE Trans. on Imag. Proc.*, vol. 24, no. 6, pp. 1879–1892, 2015.

[106] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE Trans. on Imag. Proc.*, vol. 21, no. 8, pp. 3339–3352, 2012.

[107] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Trans. Circ. Sys. Vid. Techn.*, vol. 5, no. 1, pp. 52–56, 1995.

[108] Z. Sinno, C. Bampis, and A. C. Bovik, "Detecting, localizing and correcting exposure-staurated regions using a natural image statistics model," *Ann. Meet. of Vis. Sci. Soc.*, 2017.

[109] N. E. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *IEEE Int. Conf. Image Process. (ICIP)*, 2009, pp. 2281–2284.

[110] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Sys. Tech. (TIST)*, vol. 2, no. 3, p. 27, 2011.

[111] D. Ghadiyaram and A. C. Bovik, "Scene statistics of authentically distorted images in perceptually relevant color spaces for blind image quality assessment," *IEEE Int. Conf. Imag. Process.*, pp. 3851–3855, 2015.

[112] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Sig. Proc.: Image Commun.*, vol. 29, no. 7, pp. 725–747, 2014.

[113] Z. Sinno, C. Caramanis, and A. C. Bovik, "Towards a closed form second-order natural scene statistics model," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3194–3209, 2018.

[114] Z. Sinno and A. C. Bovik, "On the natural statistics of chromatic images," in *IEEE South. Symp. Image Anal. Interpret. (SSIAI)*, 2018, pp. 81–84.

[115] Z. Sinno, C. Caramanis, and A. C. Bovik, "Second order natural scene statistics model of blind image quality assessment," *IEEE Conf. Acous. Spee. Sig. Proc. (ICASSP)*, Apr. 2018.

[116] R. K. Jones and D. N. Lee, "Why two eyes are better than one: the two views of binocular vision." *J. Exp. Psych. Hum. Perc. Perf.*, vol. 7, no. 1, p. 30, 1981.

[117] A. R. Fielder and M. J. Moseley, "Does stereopsis matter in humans?" *Eye*, vol. 10, no. 2, p. 233, 1996.

[118] F. L. Kooi and M. Lucassen, "Visual comfort of binocular and 3D displays."

[119] S. Pastoor, "Human factors of 3d imaging: results of recent research at heinrich-hertz-institut berlin," in *Proc. 2nd International Display Workshop IDW'95*, vol. 3, 1995, pp. 69–72.

[120] M. Emoto, T. Niida, and F. Okano, "Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television," *Journal of display technology*, vol. 1, no. 2, p. 328, 2005.

[121] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence–accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of vision*, vol. 8, no. 3, pp. 33–33, 2008.

[122] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks, "Visual discomfort with stereo displays: effects of viewing distance and direction of vergence-accommodation conflict," *SPIE Proc. Ster. Disp. App. XXII*, vol. 7863, 2011.

[123] M. T. Lambooij, W. A. IJsselsteijn, and I. Heynderickx, "Visual discomfort in stereoscopic displays: a review," *SPIE Proc. Ster. Disp. App. XIV*, vol. 6490, p. 64900I, 2007.

[124] Y. Nojiri, H. Yamanoue, A. Hanazato, and F. Okano, "Measurement of parallax distribution and its application to the analysis of visual comfort

for stereoscopic HDTV," *SPIE Proc. Ster. Disp. App. X*, vol. 5006, pp. 195–206, 2003.

[125] S. Ide, H. Yamanoue, M. Okui, F. Okano, M. Bitou, and N. Terashima, "Parallax distribution for ease of viewing in stereoscopic hdtv," *SPIE Proc. Ster. Disp. App. IX*, vol. 4660, pp. 38–46, 2002.

[126] S. Yano, S. Ide, T. Mitsuhashi, and H. Thwaites, "A study of visual fatigue and visual comfort for 3D HDTV/HDTV images," *Displays*, vol. 23, no. 4, pp. 191–201, 2002.

[127] J. Choi, D. Kim, S. Choi, and K. Sohn, "Visual fatigue modeling and analysis for stereoscopic video," *Optical Engineering*, vol. 51, no. 1, p. 017206, 2012.

[128] D. Kim and K. Sohn, "Visual fatigue prediction for stereoscopic image," *IEEE Trans. Circ. Sys. Vid. Techn.*, vol. 21, no. 2, pp. 231–236, 2011.

[129] J. Park, H. Oh, S. Lee, and A. C. Bovik, "3D visual discomfort predictor: Analysis of disparity and neural activity statistics," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1101–1114, 2015.

[130] T. Kim, S. Lee, and A. C. Bovik, "Transfer function model of physiological mechanisms underlying temporal visual discomfort experienced when viewing stereoscopic 3D images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4335–4347, 2015.

[131] H. Oh, S. Lee, and A. C. Bovik, "Stereoscopic 3D visual discomfort prediction: A dynamic accommodation and vergence interaction model," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 615–629, 2016.

[132] J. Park, S. Lee, and A. C. Bovik, "3D visual discomfort prediction: Vergence, foveation, and the physiological optics of accommodation." *IEEE J. Select. Topics Sign. Process.*, vol. 8, no. 3, pp. 415–427, 2014.

[133] H. Oh, S. Ahn, S. Lee, and A. C. Bovik, "Deep visual discomfort predictor for stereoscopic 3d images," *IEEE Trans. Image Process.*, 2018.

[134] S. Li, "Stereoscopic (3-D imaging) Database," Standard for the Quality Assessment of Three Dimensional (3D) Displays, 3D Contents and 3D Devices based on Human Factors http://grouper.ieee.org/groups/3dhf/, (Accessed: 07 July 2018).

[135] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *IEEE Conf. Comp. Vis. Patt. Recog. (CVPR).* IEEE, 2010, pp. 2432–2439.

[136] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *Int. J. Comp. Vis.*, vol. 106, no. 2, pp. 115–137, 2014.

[137] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, 2019.

[138] Z. Sinno and A. C. Bovik, "Live video quality challenge (VQC) database," accessed on Aug. 11 2018. [Online]. Available: http://live.ece.utexas.edu/research/LIVEVQC/index.html

[139] Z. Sinno and A. C. Bovik, "Large scale subjective video quality study," *IEEE Int. Conf. Image Process. (ICIP)*, pp. 276–280, 2018.

[140] L. Gausduff and A. A. Forni, "Gartner says worldwide sales of smartphones grew 7 percent in the fourth quarter of 2016," Feb 2017, accessed on Jan. 2 2018. [Online]. Available: https://www.gartner.com/newsroom/id/3609817l

[141] "Browser market share worldwide November 2016−December 2017," Dec 2017, accessed on Jan. 2 2018. [Online]. Available: http://gs.statcounter.com/

[142] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vision*, vol. 101, no. 1, pp. 184–204, 2013.

[143] S. Van Leuven, "Measuring perceived video quality on mobile devices," *Twitter's Engineering Blog*, Jul 2018, accessed on Aug. 9 2018. [Online]. Available: https://blog.twitter.com/engineering/en_us/topics/insights/2018/videoqualityonmobile.html

[144] I. Abramov, J. Gordon, O. Feldman, and A. Chavarga, "Sex and vision II: color appearance of monochromatic lights," *Biol. Sex Diff.*, vol. 3,

no. 1, p. 21, 2012.

[145] B. W. Rovner, R. J. Casten, and W. S. Tasman, "Effect of depression on vision function in age-related macular degeneration," *Arch. Opht.*, vol. 120, no. 8, pp. 1041–1044, 2002.

[146] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. Royal Stat. Soc. Ser. B (Methodol.)*, pp. 111–147, 1974.

[147] "Final report from the video quality experts group on the validation of objective models of video quality assessment." Video Quality Expert Group (VQEG), 2000.

[148] B. A. Wandell, *Foundations of Vision*. Sinauer Associates Sunderland, MA, 1995, vol. 8.

[149] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Sign. Process. Lett.*, vol. 17, no. 5, pp. 513–516, 2010.

[150] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *Asilomar Conf. Sign. Sys. Comp.* IEEE, 2011, pp. 723–727.

[151] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. John Wiley & Sons, 2005.

[152] A. E. Savakis, S. P. Etz, and A. C. Loui, "Evaluation of image appeal in consumer photography," *SPIE Proc. Hum. Vis. Electron. Imag. V*, vol. 3959, pp. 111–121, 2000.

[153] S. N. Yendrikhovski, F. J. Blommaert, and H. de Ridder, "Perceptually optimal color reproduction," *SPIE Proc. Hum. Vis. Electron. Imag. III*, vol. 3299, pp. 274–282, 1998.

[154] H. de Ridder, F. J. Blommaert, and E. A. Fedorovskaya, "Naturalness and image quality: chroma and hue variation in color images of natural scenes," *SPIE Proc. Hum. Vis. Electron. Imag. V*, vol. 2411, pp. 51–62, 1995.

[155] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *SPIE Proc. Hum. Vis. Electron. Imag. VIII*, vol. 5007. International Society for Optics and Photonics, 2003, pp. 87–96.

[156] A. A. Michelson, *Studies in optics*. Courier Corporation, 1995.

[157] G. Fechner, "Elements of psychophysics. vol. I." 1966.

[158] E. Peli, "Contrast in complex images," *J. Opt. Soc. Am.*, vol. 7, no. 10, pp. 2032–2040, 1990.

[159] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," in *Int. Conf. Acous. Sp. Sign. Process. (ICASSP)*, vol. 3. IEEE, 2004, pp. iii–709.

[160] R. Soundararajan and A. C. Bovik, "Rred indices: Reduced reference entropic differencing for image quality assessment," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 517–526, 2012.

[161] S. Roth and M. J. Black, "On the spatial statistics of optical flow," *Int. Conf. Comput. Vis.*, vol. 74, no. 1, pp. 33–50, 2007.

[162] D. Ghadiyaram, C. Chen, S. Inguva, and A. Kokaram, "A no-reference video quality predictor for compression and scaling artifacts," in *IEEE Int. Conf. Imag. Process.*, 2017, pp. 3445–3449.

[163] "Shutterstock," accessed on Dec. 26 2018. [Online]. Available: https://www.shutterstock.com/

[164] "Videezy," accessed on Dec. 26 2018. [Online]. Available: https://www.videezy.com/

[165] "Pixabay," accessed on Dec. 26 2018. [Online]. Available: https://pixabay.com/en/

[166] "Videvo," accessed on Dec. 26 2018. [Online]. Available: https://www.videvo.net/

[167] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acous, Sp., Sign. Process.*, vol. 29, no. 6, pp. 1153–1160, 1981.

[168] P. C. Mahalanobis, "On the generalized distance in statistics." National Institute of Science of India, 1936.

Vita

Zeina Sinno received her B.E. in Electrical and Computer Engineering with a minor in Mathematics from the American University of Beirut in 2013 with high distinction. She received her M.S. in Electrical and Computer Engineering from The University of Texas at Austin in 2015. She is a member of the Laboratory of Image and Video Engineering (LIVE) and Wireless Networking and Communications Group (WNCG). Her research interests include statistical modeling of images and videos, design of perceptual image and video quality assessment algorithms, statistical data analysis and machine learning.

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.