The Dissertation Committee for Phani Deepti Ghadiyaram
certifies that this is the approved version of the following dissertation:

Perceptual Quality Assessment of Real-World Images and Videos

Committee:

Alan C. Bovik, Supervisor

Kristen Grauman

Donald Fussell

Anne Aaron

Perceptual Quality Assessment of Real-World Images and Videos

by

Phani Deepti Ghadiyaram

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2017

To my dearest Peddamma Late Dr. B. K. Ranga who is my lifelong role model of compassion, courage, ingenuity, and humility.

Acknowledgments

I am very fortunate to have had splendid teachers and advisors, friends, and family who continue to motivate and support me throughout my life.

First and foremost, I want to extend my sincerest acknowledgement to my advisor Prof. Al Bovik. He has deeply inspired me with his brilliant teaching abilities and his strive for excellence, and motivated me to choose research as a career. I am thankful to him for believing in me and entrusting me with crucial responsibilities long before I believed in myself and that I was capable of making it through this journey. If I had transformed from a very shy and silent graduate student to an adequately confident researcher, I owe it all to Dr. Bovik. He has made me a better researcher and a better individual through his intellect, patience, upbeat attitude, and considerateness. Every time I felt that I had met the high bar he had set for me, he continued to raise it even further, thereby motivating me to only get better. That, I believe, is the true hallmark of a perfect advisor. He is incredibly hardworking and passionate about research and sets an example of what it takes to aim for and achieve the absolute best in everything. I will continue to be inspired from him.

I am fortunate to have had the wonderful opportunity to be a graduate student at UT Austin. The university's openness and top-class education gave me the opportunity to collaborate with the bright minds from diverse backgrounds and also offered a great platform to do pioneering research. I am thankful to the

Department of Computer Science for offering me various scholarships during my entire graduate program, especially in my earlier years at UT.

During my stay at UT, I had the incredible opportunity to interact with some of the brightest mentors. My heartfelt appreciation goes to Prof. Kristen Grauman, who has been a role model to me even before I joined UT and I continue to greatly admire her for her guidance, intellect, and her superb teaching abilities. I constantly draw inspiration from her incredible passion and dedication to research. I want to thank Prof. Donald Fussell for his patience, kindness, and especially for recommending me to Prof. Al Bovik during my first semester at UT. I am awe-inspired by the unbelievable quest and passion of Prof. Wilson Geisler, who has engaged my curiosity to understand more about neuroscience despite my engineering background. I want to thank Anne Aaron for being a great committee member and Prof. Joydeep Ghosh for his general guidance throughout this program.

I want to extend my warmest appreciation to my absolutely lovable and truly inspiring LIVE lab members – Janice, Zeina, Praful, Todd, Christos, Leo, Shubham, Jerry, and Lark for their cheerful companionship, lively discussions, and constructive collaborations. As we were all learning to navigate through this lonesome doctoral course in our own ways, I feel so lucky to have found such a wonderful group of friends who have always encouraged and inspired each other. A special warm acknowledgement goes to the terrific Janice with whom I have collaborated the most during the course of my PhD, and rewardingly so. She has inspired me to have a positive outlook on all aspects of life, to stay motivated, and to give one's best shot at everything. Another very special acknowledgment goes to my dearest friend Zeina, who with her upbeat attitude helped all the lab members in having a sense of balance between our demanding professional and personal lives by organizing numerous fun and social activities.

I also want to thank all my teachers throughout my primary and secondary

education and my professors at my alma mater IIIT-Hyderabad for constantly challenging my intellectual ability and inspiring me. The stimulating academic environment in IIIT-Hyderabad vetted my inquisitive appetite and motivated me to pursue research in a much bigger way at UT Austin, and I am very fortunate to have been a part of such a wonderful environment.

My deepest admiration and gratitude goes to my parents, Ravi Prasad Ghadiyaram and B. K Rama Devi for their unconditional love and unabated support throughout my life. All during my childhood, they have taught me that it is important to be inquisitive and to give one's absolute best to every task one takes upon with utmost dedication and perseverance. They have helped me build a tenacious, independent character through their incredible encouragement and love. Amma and Nanna, completing this degree would not have been possible without your love and encouragement and I am forever grateful to you for everything I have in my life.

I have lifelong role models in my late Peddamma Dr. B. K. Ranga and my Peddananna Dr. M. Ramanujacharyulu, who with their philosophical thoughts have imbibed in me, among other things, that education is the true wealth one could possess and that constantly seeking knowledge is the ultimate goal of a human's life. I will continue to carry their invaluable lessons with me as I navigate my own adult life. During the course of my doctoral journey, my grandfather B. K. Ramanujacharulu had time and again convinced and consoled my mom that encouraging my dream is the right thing to do and I am indebted to his wisdom and love.

Words cannot express my love and gratitude towards my dearest brother Raghu Kalyan for always looking out for me and shielding me from any distractions that may come my way during this journey and I truly appreciate his continuous love and unwavering support.

My heartfelt appreciation goes to my best friend and my husband Rohit Gernapudi, who from the onset of this doctoral journey, with his optimism, compas-

sion, and an unwavering trust in my abilities, cheered me towards completion. From proof-reading my publication drafts in the earlier stages of my PhD to attentively listening to all my talks before every conference and interview, and the numerous stimulating technical discussions, Rohit has been passionately involved in this entire journey. With his high-spirited nature and a big smile, Rohit always reminded me to enjoy every moment of what I initially perceived to be a daunting trail and I am forever thankful for his positive perspective on all things.

I am deeply grateful to my parents, my dearest brother, and to Rohit, who, aside from Prof. Al Bovik, were the ones who have watched me very closely as I navigated through this challenging course and shared my worries and victories with an open heart.

I am lucky to have truly wonderful parents-in-law Nagesh and Charanmayee Gernapudi and I want to thank them from the bottom of my heart for always loving and encouraging me. I would like to thank my dearest aunt and uncle Sri Vidya and Srinivas and my amazing siblings-in-law Neha, Ram, and Lakshmi, for their constant love and support. I would also like to thank my beloved friends Abhinaya, Jeevitha, Anoushaka, Dinesh, Prem, and Akhila for all the countless hours of conversations and for making my six years of stay in Austin so wonderfully lovely and fun.

Finally, I would like to reiterate that my earnest appreciation to everybody in my life goes beyond what I could express in these few words.

Phani Deepti Ghadiyaram

The University of Texas at Austin
August 2017

Perceptual Quality Assessment of Real-World Images and Videos

Publication No. _____

Phani Deepti Ghadiyaram, Ph.D.
The University of Texas at Austin, 2017

Supervisor: Alan C. Bovik

The development of online social-media venues and rapid advances in technology by camera and mobile device manufacturers have led to the creation and consumption of a seemingly limitless supply of visual content. However, a vast majority of these digital images and videos are often afflicted with annoying artifacts during acquisition, subsequent storage, and transmission over the network. All these factors impact the quality of the visual media as perceived by a human observer, thereby compromising their quality of experience (QoE).

This dissertation focuses on constructing datasets that are representative of real-world image and video distortions as well as on designing algorithms that accurately predict the perceptual quality of images and videos. The primary goal

of this research is to design and demonstrate automatic image and continuous-time video quality predictors that can effectively tackle the widely diverse authentic spatial, temporal, and network-induced distortions – contrary to all present-day algorithms that operate on single, synthetic visual distortions and predict a single overall quality score for a given video.

I introduce an image quality database which contains a large number of images captured using a representative variety of modern mobile devices and afflicted with a widely diverse authentic image distortions. I will also describe the design of an online crowdsourcing system which aided a very large-scale image quality assessment subjective study. This data collection facilitated the design of a new image quality predictor that is founded on the principles of natural scene statistics of images in different color spaces and transform domains. This new quality method is capable of assessing the quality of images with complex mixtures of distortions and yields high correlation with human perception.

Pertaining to videos, this dissertation describes a video quality database created to understand the impact of network-induced distortions on an end user's quality of experience. I present the details of a large-scale subjective study that I conducted to gather continuous-time ground truth QoE scores on a collection of 180 videos afflicted with diverse stalling events. I also present my analysis of the temporal variations in the perceived QoE due to the time-varying video quality and present insights on the impact of relevant human cognitive aspects such as long-term and short-term memory and recency on quality perception. Next, I present a continuous-time objective QoE predicting model that effectively captures the complex interactions between the aforementioned human cognitive elements, spatial and temporal distortions, properties of stalling events, and models the state of any given client-side network buffer. I also show how the proposed framework can be extended by further supplementing with any number of additional inputs (or by eliminating

any ineffective ones), based on the information available at the content providers during the design of adaptive stream-switching algorithms. This QoE predictor supports future research in the design of quality-aware stream-switching algorithms which could control the position, location, and length of stalls, given a network bandwidth budget and the end user's device information, such that the end user's QoE is maximized.

Contents

List of Tables

List of Figures

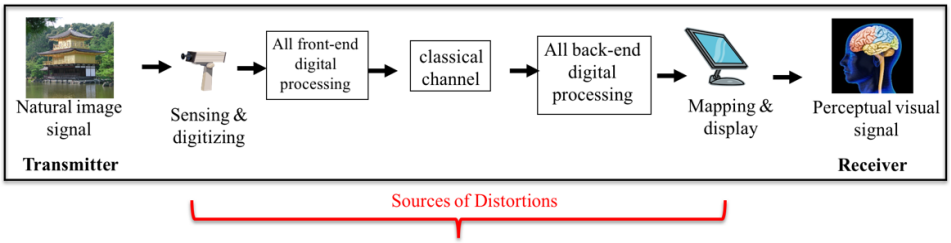
Chapter 1

Introduction

1.1 Perceptual Visual Quality Assessment

We live in a world obsessed with taking pictures, and recording and streaming videos. Visual media is increasingly pervasive everywhere; entertainment, social networks, and news reports are all available at the touch of a button, and the online film industry and social media websites supply the populace with an almost bottomless supply of visual content. The Internet offers many venues such as Instagram, Facebook, YouTube, Vimeo, and Vine for publishing pictures and videos that reflect the individualized creativity of the increasingly knowledgeable consumer base. Such ubiquitousness of visual media has led to rapid, synergistic advances in technology by camera and mobile device manufacturers, allowing consumers to efficiently capture, share, and store high-resolution pictures and videos.

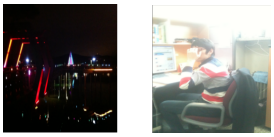
The fundamental problem I address in this dissertation is to effectively predict the *perceptual quality* of pictures and videos. The word "quality" is touted as a measure of excellence, but there are several distinct connotations to it in the image and video processing literature. For instance, "perceptual quality" refers to the subjective quality of a visual stimuli as perceived by a human observer. By



(a) The Image/Video Channel

(b) Amatuer camera users

(c) Varied camera devices



(d) Ambient lighting conditions

(e) Network congestion

Figure 1.1: Examples of sources of image and video distortions. To design robust quality predictors, the algorithms must tackle an extraordinary amount of variation introduced in the visual content by these different distortion sources.

contrast, "aesthetic quality" refers to the appreciation of beauty and possibly the artistic quality of the visual stimuli. In this dissertation, however, I exclusively deal with perceptual image and video quality. In this first chapter, I will discuss the factors that affect visual quality, motivate the need for accurate quality methods, and overview my contributions towards tackling this challenging problem.

1.1.1 Sources of Visual Distortions

A vast majority of the digital pictures (and videos) are captured by amateur photographers whose unsure hands and eyes could potentially introduce annoying artifacts during the capture process. Furthermore, every digital image and video passes through various stages during its acquisition, storage, and transmission – as illus-

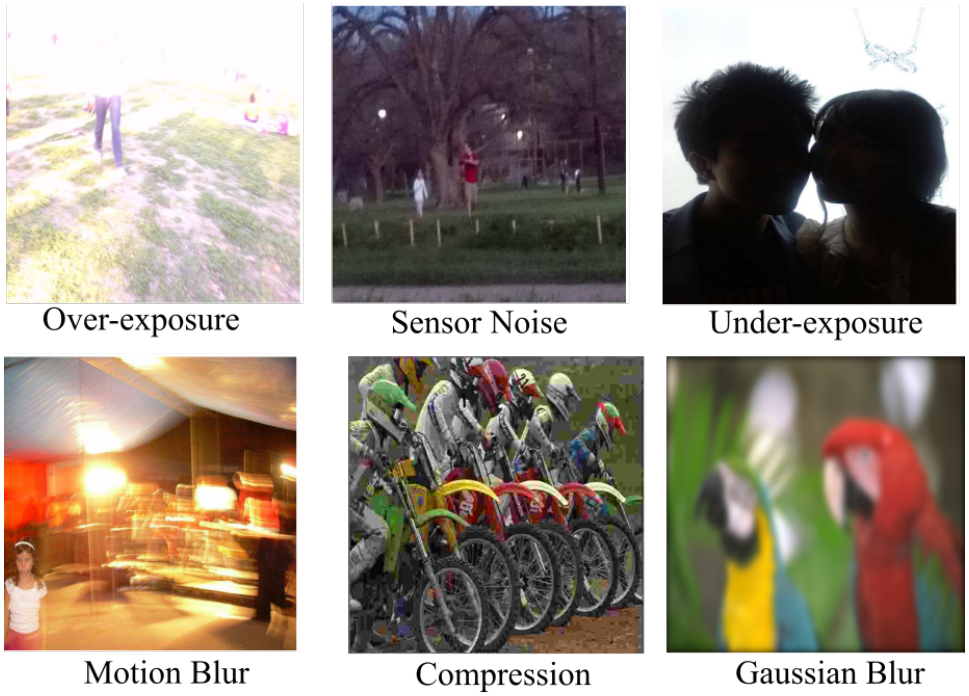


Figure 1.2: Sample images that illustrate few spatial distortions introduced during the capture or storage process.

trated in Figure 1.1, and thus could suffer from a wide-variety of spatial, temporal, and network-induced distortions.

This leads to large numbers of images and videos of unsatisfactory perceptual quality being captured and stored along with more desirable ones. Pertaining to over-the-top (OTT) video streaming, network impairments or bandwidth limitations can cause volatile network conditions, resulting in rebuffering or stalling events and bitrate fluctuations, which interrupt a video's playback and cause user annoyance.

Blocking, overexposure, underexposure, ringing, noise, compression, and blurring are some of the examples of commonly-occurring spatial distortions whereas compression, distortions due to transmission loss, ghosting, smearing, and flickering are some examples of temporal distortions that typically afflict videos. Figures 1.2 and 1.3 show a sample of typical distortions that afflict visual media.

Figure 1.3: Sample video frames that illustrate some commonly occurring distortions (a) MPEG-2 compressed frame (b) H.264 compressed frame (c) IP loss simulated frame (d) Wireless loss simulated frame.

1.1.2 Advantages and Challenges of Quality Assessment

The increasing demand for visual data necessitates the development of accurate metrics that understand and even estimate its quality in light of the evolving standards and devices. Furthermore, given that the ultimate receiver of any visual media is the human eye, and that an increasingly knowledgeable base of consumer users are demanding better quality image and video display services, accounting for an end user's quality of experience (QoE) has also become very important. QoE refers to a viewer's holistic perception and satisfaction while viewing any visual media. The presence of poor quality images and videos on any multimedia service impacts a

viewer's QoE with that service. With regards to OTT video streaming, network-induced stalling events can negatively impact a viewer's satisfaction with content and network services and could lead to user attrition for those services.

Efficiently predicting the perceptual quality of images and videos and QoE has several important advantages [12]. They can assist in identifying and eliminating poor quality pictures and videos that are captured and stored along with good quality videos and pictures on any digital device. They can also assist the design of automatic image and video enhancement algorithms by identifying the type and severity of the potential distortions afflicting the image, thereby guiding the choice of the appropriate enhancement techniques. These algorithms can be adapted to design perceptually optimized digital cameras and lenses which can potentially maximize the quality of the pictures (videos) during their capture. QoE prediction models can control and monitor the quality of streaming video content and assist in the reduction of network operational costs by encouraging the design and deployment of efficient "quality-aware" network solutions. Such strategies could help ensure that end users have a satisfactory quality of experience (QoE).

However, both perceptual visual quality and an end user's quality of experience are highly subjective in nature and are the result of a combined effect produced by factors such as diverse distortions, visual content, an individual's sensitivity to distortions, aesthetics, visual foveation, and so on. More importantly, how the visual brain plausibly perceives the affect of simultaneously-occurring distortions in a picture (or a video) and effectively carries out various recognition tasks nevertheless is not well-understood. Therefore, designing a quality predictor that accounts for all these subjective factors but still correlates well with human opinion scores is highly challenging.

1.2 Thesis Overview

1.2.1 Concepts in Quality Assessment

Given that the ultimate receivers of most images and videos are humans, the only reliable way to understand and predict the effect of distortions on a typical person's viewing experience is to capture opinions from a large sample of human subjects, which is termed *subjective visual quality assessment*. These subjective studies are vital for understanding human perception of visual quality and aid in gathering authentic data. Subjective quality is measured by displaying images or videos to human observers. The subject then indicates a quality score on a numerical or qualitative scale. To account for human variability and to assert statistical confidence, multiple subjects are required to view each image/video, and a Mean Opinion Score (MOS) is computed. While subjective quality assessment is the only completely reliable method, these studies are cumbersome, expensive, given the tremendous surge in the volume of visual media content across the Internet. Nevertheless, the data gathered from the subjective studies is crucial for designing, evaluating, and benchmarking *objective quality models* to measure their degree of consistency with subjective human evaluations.

An objective *full-reference quality assessment* algorithm assumes that a pristine signal is available to it, thus allowing a full comparison between pristine and distorted signals. Requiring the reference signal is favorable for explicitly interpreting the fidelity of the distorted signal by measuring the mutual information of the two signals. However, the availability of a pristine signal along with its distorted version is impractical in many real-world scenarios, which poses serious limitations on the applicability of full-reference quality assessment models to several practical applications and analysis. *No-reference (NR) or blind QA* techniques lie on the other end of this spectrum of information availability as they are not based on any additional information except for the distorted signal whose quality needs to be

ascertained. Blind quality assessment is certainly the most challenging as well as the most interesting problem with a potential for being integrated into various real-world applications. In this dissertation, I focus exclusively on no-reference quality predictors for images and videos.

A large number of image and video quality assessment databases have been designed in the past decade. These legacy image and video quality databases have played an important role in advancing the field of quality prediction. Existing benchmark databases have been designed to contain a small set of high quality real-world photographs (or videos), each corrupted by only one of a few synthetically introduced distortions, e.g., images corrupted by simulated camera sensor noise, Gaussian blur, or H.264/MPEG-2 compressed videos. Current top-performing IQA/VQA models are designed, trained, and evaluated based only on the statistical perturbations observed on such 'singly' distorted datasets. This might result in quality prediction models that inadvertently assume that every image/video has a single distortion that most objective viewers could agree upon.

However, the unsure eyes and hands of most amateur photographers frequently lead to occurrences of annoying visual artifacts, which are usually mixtures of several possible distortions. The unrepresentativeness of the legacy benchmark databases challenges the robustness, scalability, and applicability of the current quality assessment models in several user-centric visual media applications. It is thus desirable to design challenging databases containing a large number of authentically distorted images and videos of different quality "types," mixtures, and distortion severities, and a wide variety of visual content. It is also crucial to design efficient no-reference quality prediction algorithms that have better prediction capability on real-world image and video distortions.

1.2.2 Contributions

This dissertation is divided into two major topics. The first deals with no-reference image quality assessment for authentically distorted pictures, while the second topic is about predicting an end user's quality of experience in streaming videos under constrained network environments. In both these scenarios, I describe my approach on attacking the difficult problem of objective quality assessment from the ground up and summarize my contributions below:

No-Reference Image Quality Assessment

1. **A distortion-representative Image Database:** I will first describe a challenging blind image quality database that I created that contains images that were captured using numerous individual mobile devices, including tablets and smart-phones on real scenes in the U.S and Korea. Each picture was collected without artificially introducing any distortions beyond those occurring during capture, processing, and storage by a user's device. These images are affected by unknown mixtures of single or more commonly occurring multiple interacting authentic distortions of diverse severities. I will introduce the content and characteristics of the new LIVE In the Wild Image Quality Challenge Database, which contains 1162 *authentically* distorted images captured from many diverse mobile devices in Chapter 3.

2. **Crowdsourcing Framework for Subjective Quality Assessment:** With an aim to gather very rich human data on the aforementioned authentic picture collection, I designed and implemented an extensive online subjective study by leveraging Amazon's crowdsourcing system, the Mechanical Turk. This substantial effort helped in gathering over $350,000$ human opinion scores from more than $8,100$ unique subjects, making it the world's largest, most comprehensive study of perceptual image quality ever conducted. I will de-

scribe the design and infrastructure of this online crowdsourcing system and how it was used to conduct a very large-scale, multi-month image quality assessment subjective study, wherein a wide range of diverse observers recorded their judgments of image quality in Chapter 3.

3. **Objective, Automatic Image Quality Prediction:** A significant pragmatic contribution that I make is a potent new blind image quality assessment model called **F**eature maps based **R**eferenceless **I**mage **QU**ality **E**valuation **E**ngine (FRIQUEE), which more accurately predicts the perceptual quality of authentically distorted images than state-of-the-art NR IQA models. FRIQUEE is based on the principles of natural scene statistics of images (more in Chapter 2). FRIQUEE combines a larger and more diverse collection of perceptually relevant statistical features across multiple transform domains and color spaces, that is able to generalize over many different authentic distortion types, mixtures, and severities. These features avoid assumptions about the type of distortion(s) contained in an image and focus instead on capturing consistencies, or departures therefrom, of the statistics of real world images. I will also present the prediction performance of FRIQUEE along with current top-performing image quality predictors on six different image quality databases in Chapter 4.

No-Reference, Continuous-time Video QoE Prediction

Given the increasing demand for over-the-top (OTT) video content, my overarching goal was to thoroughly study and understand the influence of the effect of several quality-degrading factors caused due to volatile network conditions on quality of experience (QoE) and design generalizable QoE models for mobile videos. Below, I summarize my efforts towards realizing this goal:

1. **A video collection modeling playback interruptions:** I designed a new mobile video database that accurately represents the diverse stalling events and startup delays typically encountered while streaming videos. The database contains 180 distorted videos that were generated by simulating 26 unique stalling events and startup delays on 24 high-quality reference videos. These 26 stall patterns varied in the position, frequency, and length of video stalling events. A large-scale subjective study was conducted on these videos, hence each video has an associated per-frame continuous-time as well as an overall QoE subjective score. I will describe the way I simulated the diverse stalling events to create a corpus of distorted videos and the details of the human study in Chapter 5. I will also present the outcomes of my comprehensive analysis of the impact of several factors that influence subjective quality of experience (QoE) in Chapter 5.

2. **A Continuous-Time Video QoE Predictor** With a goal to assist the design of "quality-aware" stream-switching algorithms, I developed a model that can accurately predict viewers' instantaneous subjective QoE for streaming video in the wild under volatile network conditions. This model, called the **T**ime-**V**arying **QoE** (TV-QoE) Indexer, accounts for the interactions between stalling events, analyzes the spatial and temporal content of a video, predicts the perceptual video quality, models the state of the client-side network buffer, and consequently predicts continuous-time quality scores that agree quite well with human opinion scores. TV-QoE also embeds the impact of relevant human cognitive factors, such as memory and recency, and their complex interactions with the video content being viewed. I present the details of this very simple and easily extensible quality predictor in Chapter 6.

   I conclude this dissertation with a discussion of avenues for future work. Throughout my dissertation, I provide extensive evaluation on challenging

image and video datasets and also compare with many state-of-the-art methods and other baselines, thereby validating the strengths of the proposed algorithms. The outcomes across several different experiments show that the proposed quality prediction algorithms significantly outperform all prior IQA and QoE models and are of great pragmatic value.

Chapter 2

Background and Prior Work

In this chapter, I first discuss the fundamentals of human visual system and provide some background on visual neuroscience in Section 2.1.1. I then proceed to introduce the remarkable statistical regularities observed in natural images and videos and their relation to the design of objective quality assessment models. I organize my discussion of previous work on image and video quality assessment into four main subtopics: approaches for constructing image quality databases (Section 2.2), objective quality assessment models (Section 2.3), approaches for designing video databases that reflect constrained network environments (Section 2.4), and techniques for predicting temporal quality of streaming videos. Throughout, I identify the unaddressed issues in the existing approaches and propose my solutions to the same. I also draw comparisons and differences between the existing approaches and this work.

2.1 Human Visual Processing and Natural Scene Statistics

2.1.1 Human Visual System (HVS)

The human visual system (HVS) processes all of the information incident upon an eye and renders it into an efficient form amenable for the human brain to conduct high-level cognitive tasks. Substantial strides have been made towards understanding and modeling low-level visual processing in the human visual system [13] (Figure 2.1). As light from the outside world falls onto the retina of a human eye, multiple photoreceptors contained in the retina produce local responses to the visual signal. First, the bipolar cells near the surface of the retina relay the gradient potentials from the photoreceptors to the ganglion cells. There is considerable evidence that local center-surround excitatory-inhibitory processes occur at the receptive fields of the ganglion cells, thus providing a bandpass response to the input luminance (Fig. 2.2). These ganglion cells are interconnected and together, they compute a retinal contrast signal that can simply be approximated as

$$A(\mathbf{x}, t) = \frac{f(\mathbf{x}, t) - \overline{f}(\mathbf{x}, t)}{\overline{f}(\mathbf{x}, t)} \tag{2.1}$$

This spatial contrast signal is directed to the Lateral Geniculate Nucleus (LGN) (depicted in Fig. 2.1). LGN decomposes these two center-surround contrast response signals from each retina, and is primarily responsible for the temporal decorrelation of the responses using a set of difference of temporal gamma filters. This operation yields a set of lagged and unlagged temporal responses. Thus, the retina decorrelates the spatial signal while the LGN decorrelates the retinal signal temporally. These four response signals provide a simple and complete "dictionary" necessary to describe the visual input signal [14] for later parts of the visual pathway. Even though the model assumes separability and linearity, it closely predicts

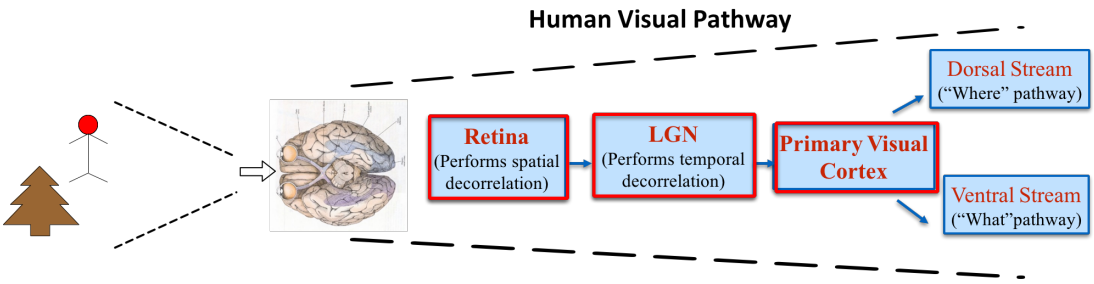


Figure 2.1: The big picture of the visual pathway. The ganglion cells in the retina spatially decorrelate the incoming visual input, and the cells in the Lateral Geniculate Nucleus (LGN) temporally decorrelate the resulting spatial signal. This spatio-temporally decorrelated signal is transmitted to area V1 for further processing. After V1, the two streams split to perform two main categories of processing (popularly known as the *what* and *where* pathways) in the Human Visual System.

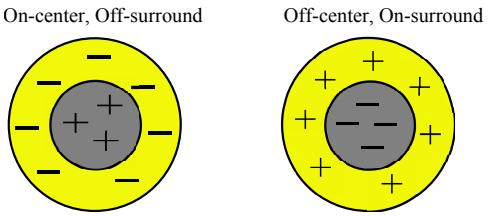


Figure 2.2: Depiction of On-center and Off-center excitatory-inhibitory responses of ganglion cells in the retina.

the behavior observed in human visual system, thus reinforcing the "efficient coding" hypothesis, which we describe later in Section 2.1.2. A bridge connecting LGN to V1 transmits these four response signals to the *simple and complex cells* in V1. The simple cells in area V1 can be modeled as collectively providing a large bank of quadrature pairs of log-gabor type responses. *Complex cells* can be simply modeled as adding the local half-square rectified responses of simple cells, and further normalizing them. The collective output of complex cells mimics a filter-bank of log-gabor filters tuned for various spatio-temporal orientations of visual stimuli across scales.

After area V1, the flow of information along the visual pathway is often

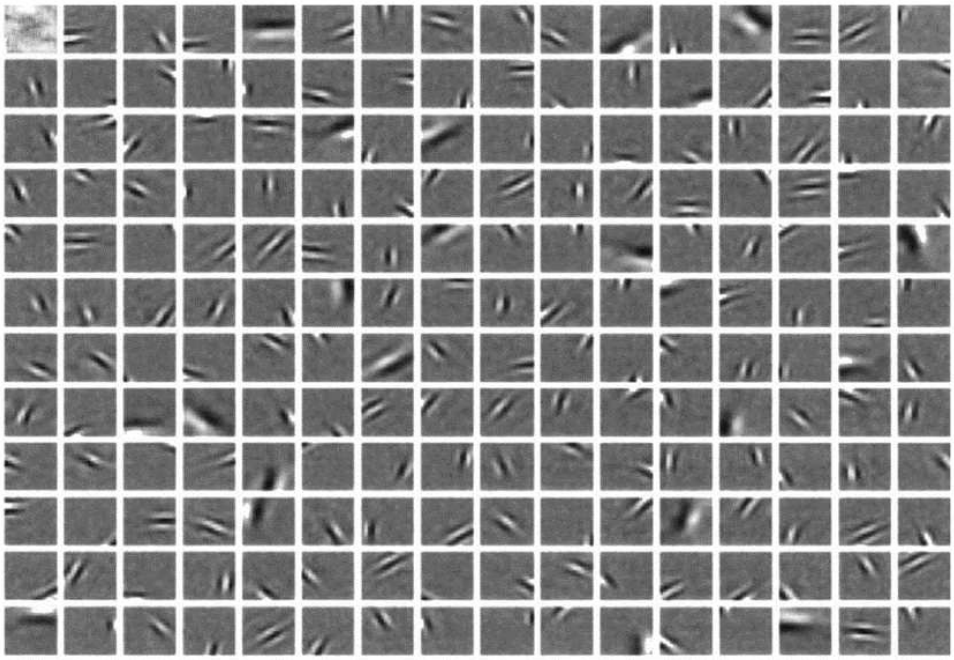


Figure 2.3: Illustration of a few sparse spatial codes derived on natural image data. These codes strongly resemble the receptive field profiles (2D impulses responses) of 2D simple cells in primary visual cortex Figure reproduced from [10] with permission.

broadly modeled as being split into *ventral* and *dorsal* streams. The ventral stream ("What Pathway") mostly follows the pathway from V1 to the temporal lobe via V2 and V4, and corresponds to object recognition and shape representation. The dorsal stream ("Where Pathway") follows the pathway from V1 to MT via V2 and corresponds to motion computation of object locations and trajectories including the control of eyes and arms.

Much has been understood about the functionalities of many neurons in the human visual system, however a lot remains unknown [15]. However, it is clear that the spatial-visual signal is decomposed over multiple orientations and scales/frequency bands in area V1 [16]. A number of low-level image processing and computer vision algorithms are based on this model [17, 18, 19, 20, 21, 22].

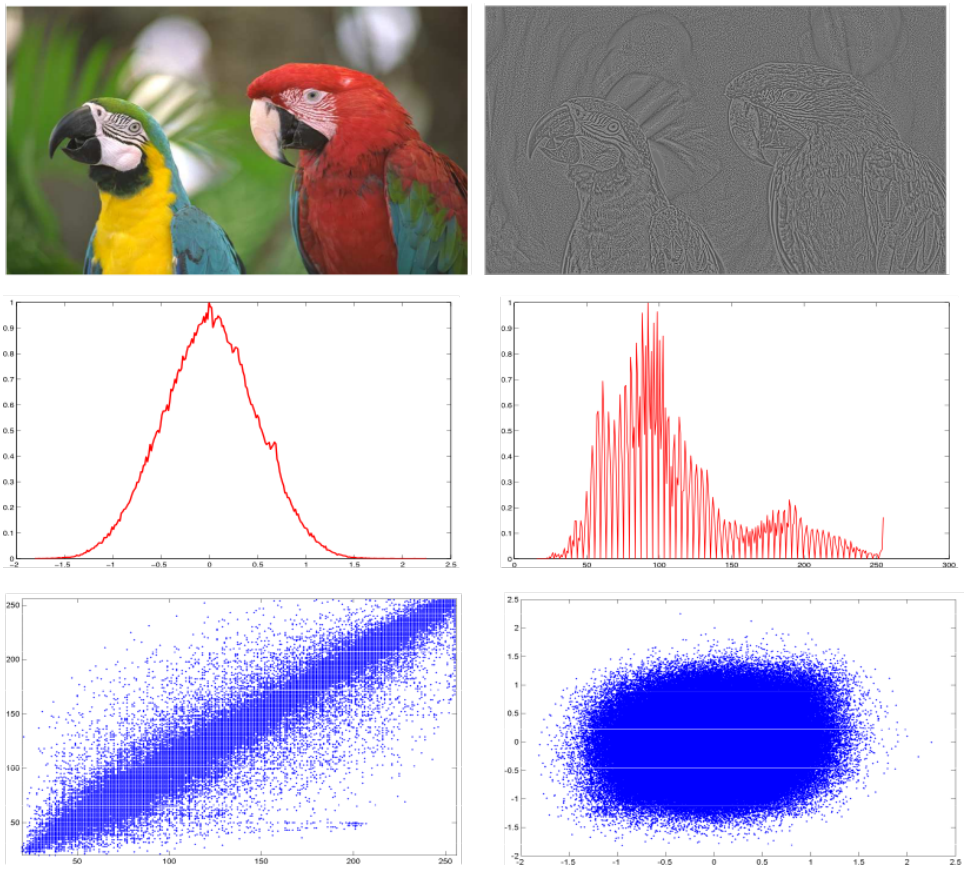


Figure 2.4: A natural undistorted image (shown in the upper left) when processed by applying the debiasing and normalization produces a decorrelated NLC map (shown in upper right). The histogram of the intensity values of the NLC map (middle right) follows a Gaussian distribution. The scatter plots (lower row) contrast the highly correlated natural image and the nearly decorrelated NLC map.

2.1.2 Natural Scene Statistics (NSS)

The development of human visual system is strongly dependent on early visual stimulation and the statistics of the surrounding visual environment. Although statistically analyzing the image environment could lead to a deeper understanding of human visual processing, there is no way to collect enough data to fully characterize the same.

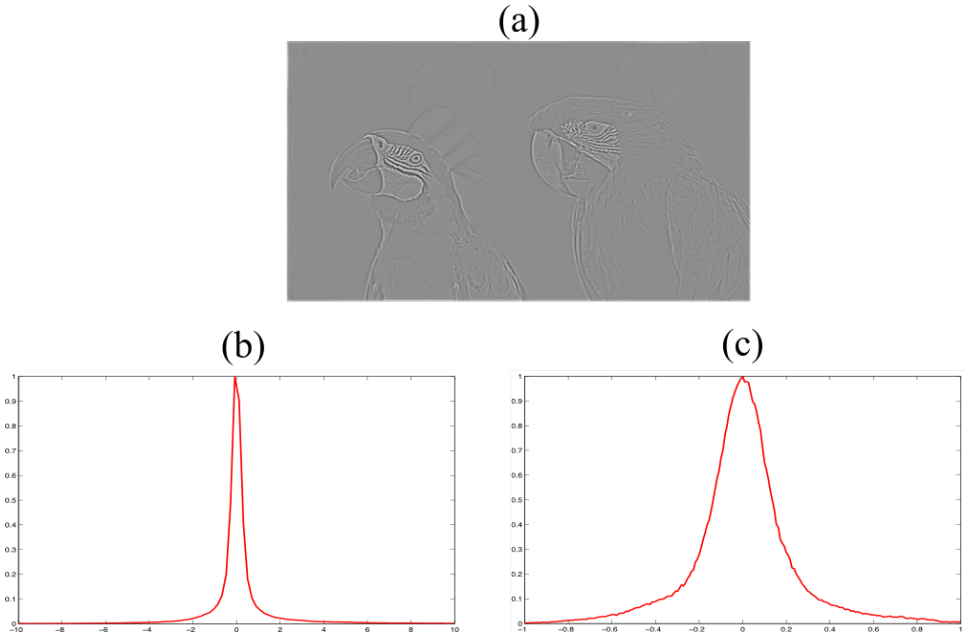


Figure 2.5: (Left) The residual when the Difference of Gaussian filter is applied on a natural image shown in Fig. 2.4 (a). (Middle) The histogram of the DoG residual. (Right) The debiased and normalized residual also closely follows a Gaussian distribution.

Nevertheless, the statistics of real-world natural images, generally referred to as *Natural Scene Statistics* (NSS), have been deeply studied for the past several years. NSS models are based on the principled observation that good quality real-world photographic images exhibit certain perceptually relevant statistical regularities. Despite the tremendous diversity of natural images in terms of content and capture processes, these statistical regularities are remarkably consistent. For instance, the amplitude spectra of the spatial Fourier transforms of natural images obey an approximate reciprocal law [23, 24]. This is a statistically self-similar phenomenon and is invariant to the scale of the natural images.

Another powerful statistical regularity is founded on the thesis that the firings of sensory neurons along the visual pathways carry efficient representations of the

visual information. Olshausen and Field [10] thus conjectured that natural images also have such an efficient and *sparse* representation that can be exploited by the visual system. The 'sparse codes' derived from natural images provide minimal reconstruction error while preserving information. They strongly resemble Gabor filters, or the receptive field profiles of 2D simple cells in primary visual cortex. Another analysis showed that the principal and independent spatial (and spatio-temporal) components of natural time-varying images strongly resemble the simple cell responses in the visual cortex [25, 26].

Another particularly useful statistical regularity of natural images surfaces when subjected to a spatial linear bandpass filter such as a Difference of Gaussian (DoG) or a predictive coding filter. This filter attenuates the low spatial frequencies (such as smoothly varying content in the image) and the resulting filtered responses can be reliably modeled using a Gaussian probability distribution [11].

For example, given an image's intensity map I of size $M \times N$, a divisive normalization operation [11] yields a normalized luminance coefficients (NLC) map:

$$NLC(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + 1}, \tag{2.2}$$

where

$$\mu(i,j) = \sum_{k=-3}^{3} \sum_{l=-3}^{3} w_{k,l} I_{k,l}(i,j) \tag{2.3}$$

and

$$\sigma(i,j) = \sqrt{\sum_{k=-3}^{3} \sum_{l=-3}^{3} w_{k,l} \left[I_{k,l}(i,j) - \mu(i,j)\right]^2}, \tag{2.4}$$

where $i \in 1, 2..M, j \in 1, 2..N$ are spatial indices and w is a 2D circularly-symmetric Gaussian weighting function. This *debiasing* and *divisive normalization* process mimics the normalization operation performed by complex cells.

Similarly, a good filter to model the center-surround excitatory-inhibitory

processes that occur at various stages of visual processing (as described earlier in Sec. 2.1.1) is the 2D difference of isotropic Gaussian filters (DoG):

$$DoG = \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sigma_1} e^{\frac{-(x^2+y^2)}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{\frac{-(x^2+y^2)}{2\sigma_2^2}} \right), \tag{2.5}$$

where $\sigma_2 = 1.5\sigma_1$.

Figure 2.4 depicts a pristine natural image, its NLC map, their corresponding histograms of intensity values along with the scatter plots of horizontally adjacent pixels. The white noise like scatter plot of the NLC map of I is indicative of the decorrelation of the pixels which contrasts from the near linear correlation between the plot of I. As illustrated in Fig. 2.5 the empirical probability density function of $I''(x, y)$ (obtained from applying DoG on I) of natural images also closely follow a Gaussian-like distribution.

Applying divisive normalization using the neighboring coefficient energies in a wavelet or other bandpass transform domain yields a similar result of reduction of statistical dependencies and Gaussianization of the data. Divisive normalization or contrast-gain-control [27] accounts for specific measured nonlinear interactions between neighboring neurons. It models the response of a neuron as governed by the responses of a pool of neurons surrounding it. Further, divisive normalization models partially account for contrast masking [28] – when a signal reduces or eliminates the visibility of another signal, typically of similar frequency, orientation, motion, color, or other attribute.

I will revisit the concepts of natural scene statistics and divisive normalization and present their significance and relevance to the design of accurate perceptual quality prediction models in Section 2.3 and also in Chapter 4. In the following section, I describe the current approaches of designing image quality assessment databases.

2.2 Benchmark Image Quality Databases

2.2.1 Image content

As mentioned in Section 1.2.1, subjective quality assessment is the only reliable way to truly understand the impact of distortions on an end user's quality of experience. Several subjective studies have been conducted in the past decade which has led to the design of a large number of quality assessment models (Figure 2.6). Most of the top-performing IQA models (full, reduced, and no-reference) [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39] [40] [41] [42] [43] [44] [45] [46] [47] have been designed and extensively evaluated on two popular benchmark databases: the LIVE IQA Database [3] which was designed in 2005 and the TID2008 Database [48], designed and released in 2008. The LIVE IQA Database, one of the first comprehensive IQA databases, consists of 779 images, much larger than the small databases that existed at the time of its introduction [49] [50] [51]. This legacy database contains 29 pristine reference images and models five distortion types - jp2k, jpeg, Gaussian blur, white noise, and fast fading noise [3]. The TID2008 Database is larger, consisting of 25 reference and 1700 distorted images over 17 distortion categories. TID2013 [5] is a recently introduced image quality database with an end goal to include the peculiarities of color distortions in addition to the 17 simulated spatial distortions included in TID2008. It consists of 3000 images and includes seven new types of distortions, thus modeling a total of 24 distortions. More details on the categories and severities of image distortions contained in these database can be found in [3] [48] [5]. Aside from these three databases, there exist a few other smaller databases [6] [52] all modeling single, synthetic distortions.

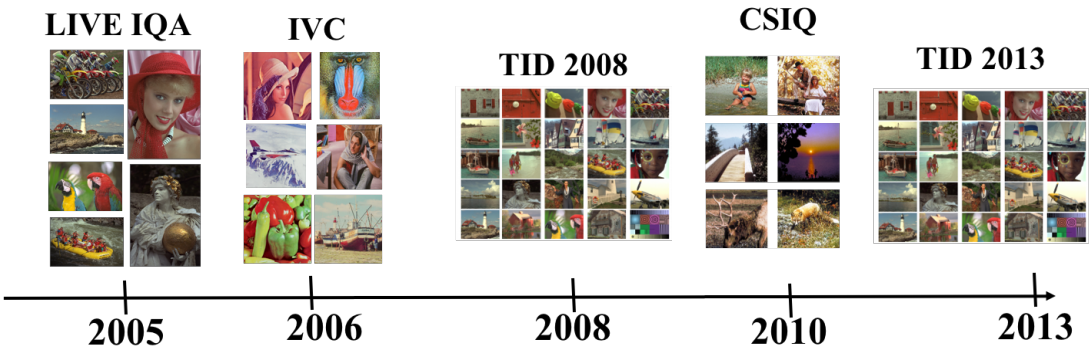


Figure 2.6: A few popular legacy Image Quality Assessment Databases designed in the past decade.

Limitations - Inauthentic Distortions:

The aforementioned legacy IQA databases, have been designed to contain images corrupted by only one of a few synthetically introduced distortions. Specifically, all of these databases have been developed beginning with a small set of high-quality pristine images (29 distinct image contents in [3] and 25 in [48] [5]), which are subsequently distorted. The distortions are introduced in a controlled manner by the database architects (Figure 2.7) and these distortion databases have three key properties. First, the distortion severities / parameter settings are carefully (but artificially) selected, typically for psychometric reasons, such as mandating a wide range of distortions, or dictating an observed degree of perceptual separation between images distorted by the same process. Second, these distortions are introduced by computing them from an idealized distortion model. Third, the pristine images are of very high quality, and are usually distorted by one of several single distortions. These databases therefore contain images that have been impaired by one of a few synthetically introduced distortion types, at a level of perceptual distortion chosen by image quality scientists.

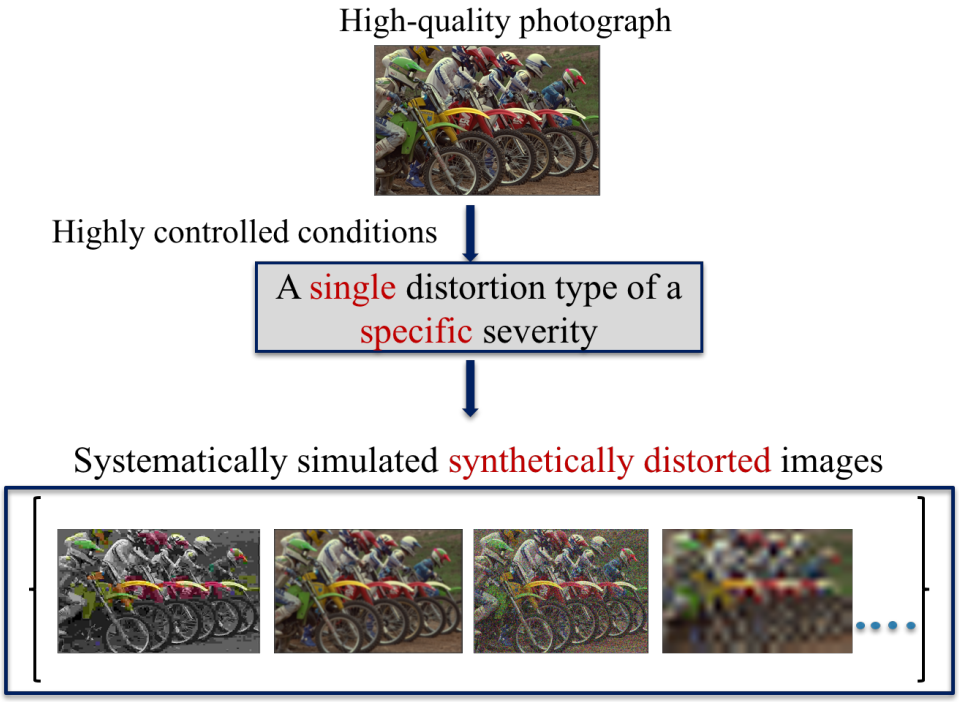


Figure 2.7: Outline of the standard procedure followed by most legacy image database originators.

Though the existing legacy image quality databases have played an important role in advancing the field of image quality prediction and facilitated the study of the effects of distortion-specific parameters on human perception, I contend that determining image quality databases such that the distorted images are derived from a set of high quality source images and by simulating image impairments on them is much too limiting. In particular, traditional databases fail to account for difficult mixtures of distortions that are inherently introduced during image acquisition and subsequent processing and transmission. For instance, consider the images shown in Fig. 2.8(a) - Fig. 2.8(d). Figure 2.8(d) was captured using a mobile device and can be observed to be distorted by both low-light noise and compression errors. Figure 2.8(b) and (c) are from the legacy LIVE IQA Database [3] where JPEG compression

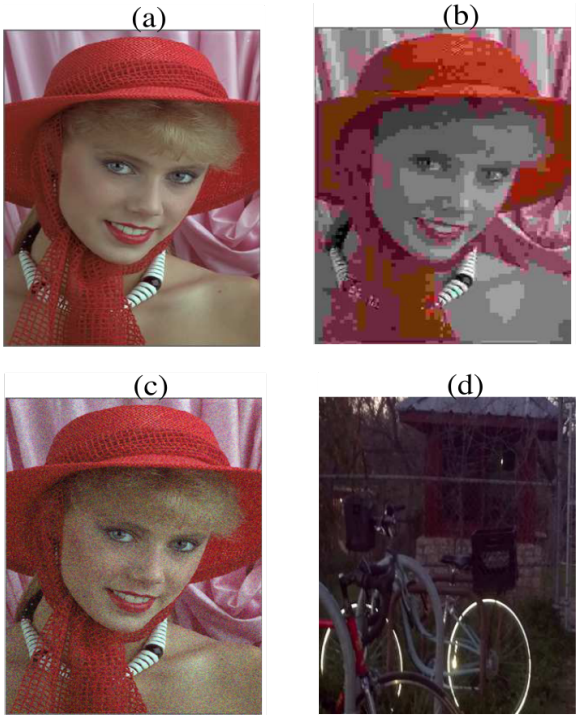


Figure 2.8: (a) A pristine image from the legacy LIVE Image Quality Database [3] (b) JPEG compression distortion artificially applied to (a). (c) White noise added to (a). (d) A blurry image also distorted with low-light noise from the new LIVE In the Wild Image Quality Challenge Database.

and Gaussian blur distortions were synthetically introduced on a pristine image (Fig. 2.8(a)).

Since cameras on mobile devices make it extremely easy to snap images spontaneously under varied conditions, the complex mixtures of image distortions that occur are not well-represented by the distorted image content in either of these legacy image databases. This limitation is especially problematic for *blind* IQA models which have great potential to be employed in large-scale user-centric visual media applications. Designing, training, and evaluating IQA models based only on the statistical perturbations observed on these restrictive and non-representative datasets might result in quality prediction models that inadvertently assume that every image

has a "single" distortion that most objective viewers could agree upon. Although top-performing algorithms perform exceedingly well on these legacy databases (e.g., the median Spearmann correlation of 0.94 on the legacy LIVE IQA Database [3] reported by BRISQUE [29] and 0.96 reported by Tang *et. al* in [34]), their performance is questionable when tested on naturally distorted images that are normally captured using mobile devices under highly variable illumination conditions.

In this dissertation, I address this limitation by designing a new image quality database that models naturally occurring, authentic image distortions. I describe the content and characteristics of this unique data collection in Chapter 3.

2.2.2 Traditional subjective study methodologies

All of the benchmark image quality assessment databases have human opinion scores captured from a large sample of human subjects. These human opinion scores in most of the legacy datasets [3, 48, 5] were collected by conducting subjective studies in laboratory settings with stringent controls on the experimental environments. The TID2008 opinion scores were obtained from 838 observers by conducting batches of large scale subjective studies, whereby a total of $256,000$ comparisons of the visual quality of distorted images were performed. Although this is a large database, some of the test methodologies that were adopted do not abide by the ITU recommendations. For instance, the authors followed a *swiss competition principle* and presented three images, wherein two of them are the distorted versions of the third one. A subject was asked to choose one image of superior quality amongst the two distorted images. I believe that this kind of presentation does not accurately reflect the experience of viewing and assessing distorted images in the most common (e.g. mobile) viewing scenarios. Furthermore, in each experiment, a subject would view and compare 306 instances of the same reference image containing multiple types and degrees of distortions, introducing the significant possibility of serious *hysteresis*

effects that are not accounted for when processing the individual opinion scores.

In pairwise comparison studies, the method for calculating preferential ranking of the data can often dictate the reliability of the results. Certain probabilistic choice model-based ranking approaches [53, 54, 55] offer sophisticated ways to accurately generate quality rankings of images. However, the opinion scores in the TID2008 database were obtained by first accumulating the points "won" by each image. These points are driven by the preferential choices of different observers during the comparative study. The mean values of the winning points on each image were computed in the range $[0-9]$ and are referred to as mean opinion scores. This method of gathering opinion scores, which diverges from accepted practice, is in our view questionable.

The LIVE IQA Database was created by following a single-stimulus methodology. Both the reference images as well as their distorted versions were evaluated by each subject during each session. Thus, *quality difference scores* which address user biases were derived for all the distorted images and for all the subjects. The creators of the LIVE IQA Database used two 21-inch CRT monitors with display resolutions of 1024×768 pixels in a normally lit room, which the subjects viewed from a viewing distance of 2 - 2.5 screen heights. Although the LIVE test methodology and subject rejection method adheres to the ITU recommendations, the test sessions were designed to present a subject with a set of images, all afflicted by the same type of distortion (for instance, all the images in a given session consisted of different degrees of JPEG 2000 distortion) that were artificially added to different reference images. It is possible that this could have led to over-learning of each distortion type by the subjects as the study session progressed.

Limitations with traditional study setups:

As mentioned earlier, the above subjective studies were conducted in a controlled laboratory environment, where images were displayed on a single device with a fixed display resolution and which the subjects viewed from a fixed distance, involving small, non-representative subject samples (typically graduate and undergraduate university students). Additionally, each subjective study setup has a few shortcomings with regards to the order in which the pictures are presented during the study as mentioned above.

However, significant advances in technology made by camera and mobile device manufacturers now allow users to efficiently access visual media over wired and wireless networks. Thus, the subjective image quality opinions gathered under artificially controlled settings do not necessarily mirror the picture quality perceived on widely used portable display devices having varied resolutions. Gathering representative subjective opinions by simulating different viewing conditions would be exceedingly time-consuming, cumbersome, and would require substantial manual effort.

2.2.3 Online Subjective Studies

The highly variable ambient conditions and the wide array of display devices on which a user might potentially view images will have a considerable influence on her perception of picture quality. This greatly motivated my interest in conducting IQA studies on the Internet, which can allow access to a much larger and more diverse subject pool while allowing for more flexible study conditions. A few studies have recently been reported that used web-based image, video, or audio rating platforms [56] [57] [58] [59] [60] [61] [62] [63]. Some of these studies employed pairwise comparisons followed by ranking techniques [53] [54] [55] to derive quality scores, while others adopted the single stimulus technique and an absolute category rating

(ACR) scale. Since performing a complete set of paired comparisons (and ranking) is time-consuming and monetarily expensive when applied on a large scale, Xu *et al.* [64] [65] introduced the HodgeRank on Random Graphs (HRRG) test, where random sampling methods based on Erdös-Rényi random graphs were used to sample pairs and the HodgeRank [66] was used to recover the underlying quality scores from the incomplete and imbalanced set of paired comparisons. More recently, an active sampling method [67] was proposed that actively constructs a set of queries consisting of single and pair-wise tests based on the expected information gain provided by each test with a goal to reduce the number of tests required to achieve a target accuracy.

However, all of these studies were conducted on small sets of images taken from publicly available databases of synthetically distorted images [3], mostly to study the reliability and quality of the opinion scores obtained via the Internet testing methodology. In most cases, the subjective data from these online studies is publicly unavailable. By contrast, in this dissertation, I present the design of a web-based crowdsourced subjective study which enabled the collection of high-quality subjective scores on about 1200 pictures by engaging over 8100 unique subjects. Furthermore, as I mention in Chapter 3, this image collection as well as the subjective scores are made publicly available [68].

2.3 No-reference Image Quality Assessment Models

As already mentioned in Section 2.1.2, the statistics of real-world natural[1] images, generally referred to as *Natural Scene Statistics* (NSS), have been deeply studied for the past several years [11] [69]. NSS models are based on the principled observation

[1]Natural images are not necessarily images of natural environments such as trees or skies. Any natural visible-light image that is captured by an optical camera and is not subjected to artificial processing on a computer is regarded here as a natural image including photographs of man-made objects.

that good quality real-world photographic images exhibit certain perceptually relevant statistical regularities. Despite the tremendous diversity of natural images in terms of content and capture processes, these statistical regularities are remarkably consistent. Wainwright *et al.* [27], building on Ruderman's work [11], empirically determined that applying a non-linear *divisive normalization* operation, similar to the non-linear behavior of certain cortical neurons, wherein the rectified linear responses are divided by a weighted sum of rectified neighboring responses, greatly reduce such observed statistical dependencies. Furthermore, the empirical probability distributions of these filtered responses can be reliably modeled using a Gaussian probability distribution as described in Section 2.1.2. To further illustrate this well-studied phenomenal regularity, I processed 29 pristine images from the legacy LIVE IQA Database [3] which vary greatly in their image content and plotted the collective histogram of the normalized coefficients of all 29 images in Figure 2.9. Specifically, I concatenated the normalized coefficients of all the images into a single vector and plotted its histogram.

Most efficient quality assessment algorithms to date are perceptual-based and are founded on the basis of natural scene statistics (NSS). It has been observed by vision scientists and image engineers that certain perceptually relevant statistical laws obeyed by natural scenes are violated by the presence of common distortions. Pertaining to images, if they are singly distorted, that is, if images contain only one of the few synthetically introduced distortions, then the natural statistics of such distorted images make it possible to determine the presence and identify the type of distortion as well. Effectively quantifying these deviations is crucial for being able to predict the perceptual quality of that image.

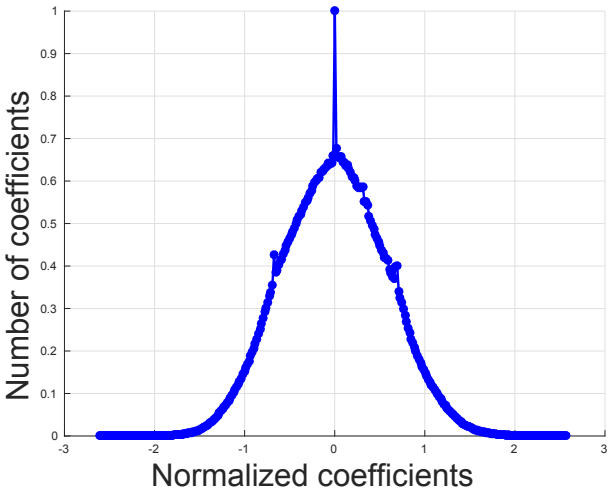


Figure 2.9: Histogram of normalized luminance coefficients of all 29 pristine images contained in the legacy LIVE IQA Database [3]. Notice how irrespective of the wide-variety of image content of the 29 pristine images, their collective normalized coefficients follow a Gaussian distribution (Estimated GGD shape parameter = 2.15.)

2.3.1 Overview of the Existing Approaches

Perceptual-based Techniques

State-of-the-art NSS-based NR IQA models [29] [30] [31] [32] [35] [33] [70] [71] exploit the statistical perturbations of these statistics by first extracting image features in a normalized bandpass space in different transform domains, then learning a kernel function that maps these features to ground truth subjective quality scores. These models do not make *a priori* assumptions on the contained distortion or image content. Tang *et al.* [32] proposed an approach combining NSS features along with texture, blur, and noise statistics. The DIIVINE Index [30] deploys summary statistics under an NSS wavelet coefficient model. Another model, BLIINDS-II [31] extracts a small number of NSS features in the DCT domain. BRISQUE [29] trains an SVR on a small set of spatial NSS features. CORNIA [36], which is not an NSS-

based model, builds distortion-specific code words to compute image quality. NIQE [33] is an unsupervised technique driven by spatial NSS-based features, requiring no exposure to distorted images at all.

Deep Learning based techniques

The authors of [72] use a convolutional neural network (CNN), divide an input image to be assessed into 32×32 non-overlapping patches, and assign each patch a quality score equal to its source image's ground truth score during training. The CNN is trained on these locally normalized image patches and the associated quality scores. In the test phase, an average of the predicted quality scores is reported. This data augmentation and quality assignment strategy could be acceptable in their work [72] since their model is trained and tested on legacy benchmark datasets containing single homogeneous distortions [3, 48]. However, real-world pictures suffer from non-homogeneous, authentic distortions, i.e., they different types of distortions affect different parts of images with varied severities. Thus, the CNN model and the quality assignment strategy in the training phase and the predicted score pooling strategy in the test phase, as used in [72] cannot be directly extended to real-world images. Similarly, the authors of [34] use a deep belief network (DBN) combined with a Gaussian process regressor to train a model on quality features proposed in their earlier work [73]. This model also has not been evaluated on real-world image distortions.

2.3.2 Limitations of the state-of-the-art IQA models:

All of these models (other than NIQE) were trained on synthetic, and usually singly distorted images contained in benchmark databases [3] [48]. They are also evaluated on the same data challenging their extensibility on images containing complex mixtures of authentic distortions such as those found in the real-world pictures that

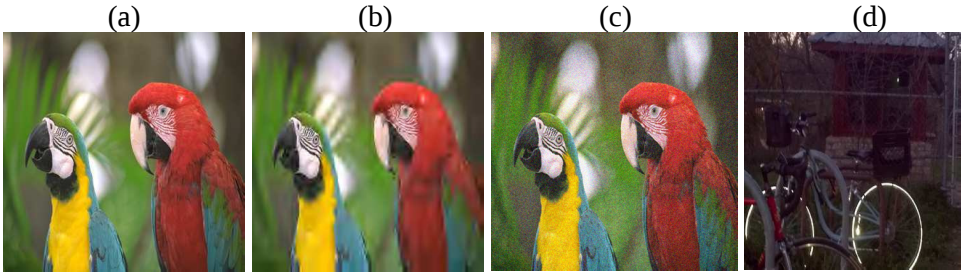


Figure 2.10: (a) A pristine image from the legacy LIVE Image Quality Database [3] (b) JP2K compression distortion artificially added to (a). (c) White noise added to (a). (d) A blurry image also distorted with low-light noise from the new LIVE In the Wild Image Quality Challenge Database [1, 2].

are captured using mobile devices.

Consider the images in Fig. 2.10. These images were transformed by a band-pass *debiasing* and *divisive normalization* operation [11]. This normalization process reduces spatial dependencies in natural images. The empirical probability density function (histogram) of the resulting normalized luminance coefficient (NLC) map of the pristine image in Fig. 2.10(a) is quite Gaussian-like (Fig. 2.11). I deployed a generalized Gaussian distribution (GGD) model [74] and estimated its parameters - shape (α) and variance (σ^2) (more details in Chapter 4). I found that the value of α for Fig. 2.10(a) is 2.09, in accordance with the Gaussian model of the histogram of its NLC map. It should be noted that the family of generalized Gaussian distributions include the normal distribution when $\alpha = 2$ and the Laplacian distribution when $\alpha = 1$. This property is not specific to Fig 2.10(a), but is generally characteristic of *all* natural images (as already described in Section 2.1.2).

The same property is not held by the distorted images shown in Fig. 2.10(b) and (c). The estimated shape parameter values computed on those images was 1.12 and 3.02 respectively. This deviation from Gaussianity of images containing single distortions has been observed and established in numerous studies on large comprehensive datasets of distorted images, irrespective of the image content. Quantifying

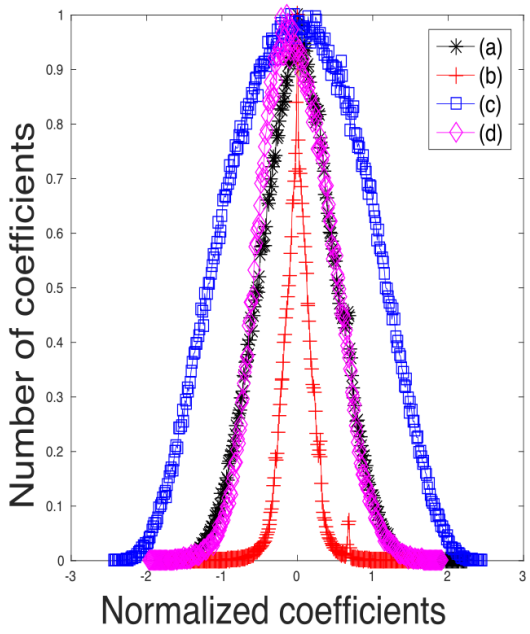


Figure 2.11: Histogram of normalized luminance coefficients of the images in Figures 2.10(a) - (d). Notice how each single, unmixed distortion affects the statistics in a characteristic way, but when mixtures of authentic distortions afflict an image, the histogram resembles that of a pristine image. (Best viewed in color).

these kinds of statistical deviations as learned from databases of annotated distorted images is the underlying principle behind several state-of-the-art objective blind IQA models [29, 31, 30, 73, 35, 33, 70, 71].

While this sample anecdotal evidence suggests that the statistical deviations of distorted images may be reliably modeled, consider Fig. 2.10(d), from the new LIVE In the Wild Image Quality Challenge Database [2]. This image contains an apparent mixture of blur, sensor noise, illumination, and possibly other distortions, all nonlinear and difficult to model. Some distortion arises from compositions of these, which are harder to understand or model. The empirical distribution of its NLC (Fig. 2.11) also follows a Gaussian-like distribution and the estimated shape

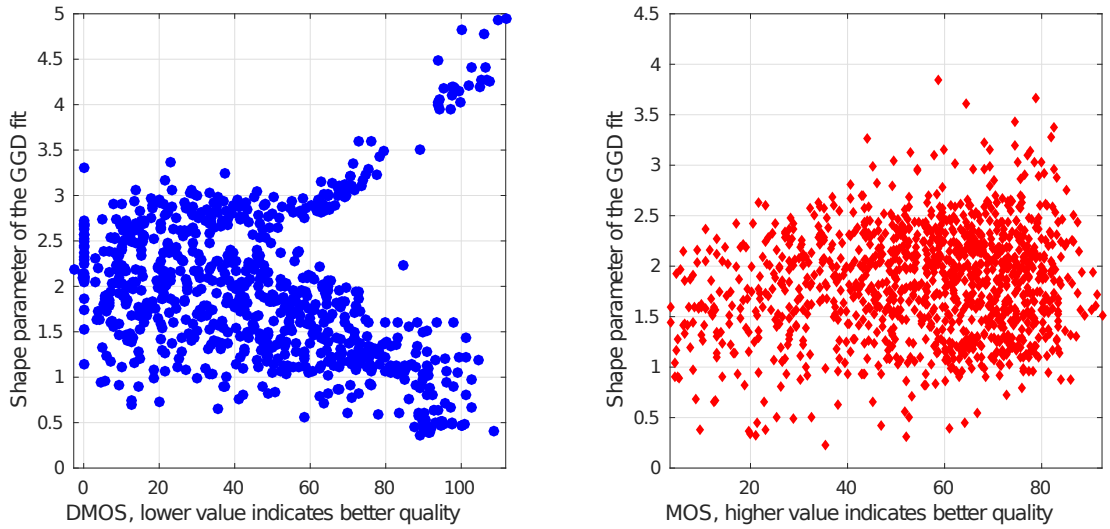


Figure 2.12: 2D scatter plots of subjective quality scores against estimated shape parameters (α) obtained by fitting a generalized Gaussian distribution to the histograms of normalized luminance coefficients (NLC) of all the images in (a) the legacy LIVE Database [3] and (b) the LIVE Challenge Database [1, 2].

parameter value (α) is 2.12, despite the presence of multiple severe and interacting distortions. As a way of visualizing this problem, I show scatter plots of subjective quality scores against the α values of the best GGD fits to NLC maps of all the images (including the pristine images) in the legacy LIVE IQA Database (of synthetically distorted pictures) [3] in Fig. 2.12(a) and for all the authentically distorted images in the LIVE Challenge Database in Fig. 2.12(b). From Fig. 2.12(a), it can be seen that most of the images in the LIVE legacy IQA Database that have high human subjective quality scores (i.e., low Difference of Mean Opinion Scores (DMOS)) associated with them (including the pristine images) have estimated α values close to 2.0, while pictures having low quality scores (i.e., high DMOS), take different α values, thus are statistically distinguishable from high-quality images. However, Fig. 2.12(b) shows that authentically distorted images from the new

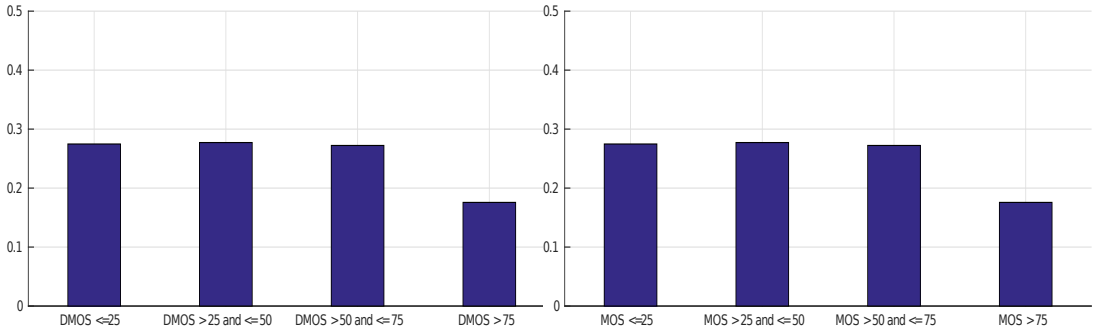


Figure 2.13: Bar plots illustrating the distribution of the fraction of images from (Left) the legacy LIVE IQA Database and (Right) the LIVE Challenge Database belonging to 4 different DMOS and MOS categories respectively. These histograms demonstrate that the distorted images span the entire quality range in both the databases.

LIVE Challenge Database may be associated with α values close to 2.0, even on heavily distorted pictures (i.e., with low Mean Opinion Scores (MOS)). Figure 2.13 plots the distribution of the fraction of all the images in the database that fall into four discrete MOS and DMOS categories. It should be noted that legacy LIVE IQA Database provides DMOS scores while the LIVE Challenge Database contains MOS scores. These histograms show that the distorted images span the entire quality range in both databases and that there is no noticeable skew of distortion severity in either databases that could have affected the results in Fig. 2.12 and Fig. 2.14.

Figure 2.14 also illustrates our observation that authentic and inauthentic distortions affect scene statistics differently. In the case of single inauthentic distortions, it may be observed that pristine and distorted images occupy different regions of this parameter space. For example, images with lower DMOS (higher quality) are more separated from the distorted image collection in this parameter space, making it easier to predict their quality. There is a great degree of overlap in the parameter space among images belonging to the categories 'DMOS <= 25' and 'DMOS > 25 and <= 50', while heavily distorted pictures belonging to the other two DMOS

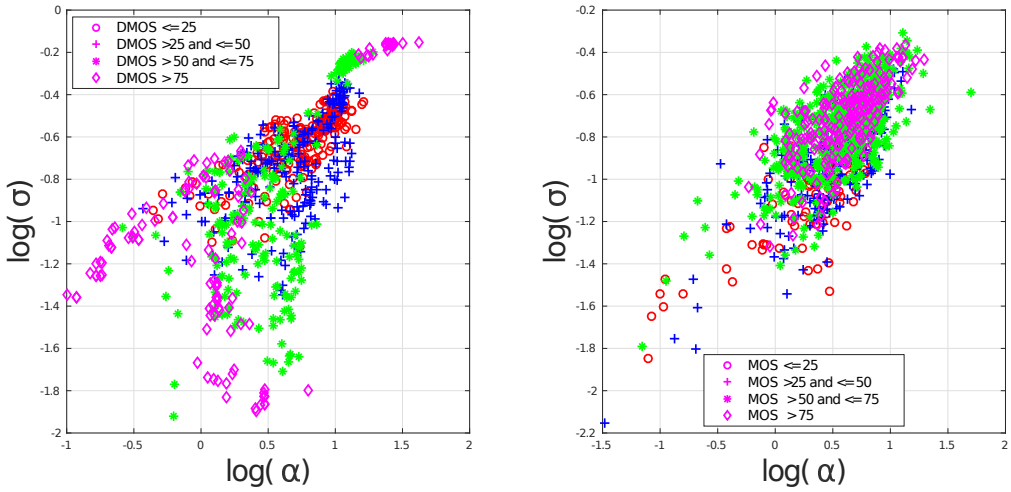


Figure 2.14: 2D scatter plots of the estimated shape and scale parameters obtained by fitting a generalized Gaussian distribution to the histograms of normalized luminance coefficients (NLC) of all the images in (a) the legacy LIVE Database [3] and (b) the LIVE Challenge Database [1, 2]. Best viewed in color.

categories are separated in the parameter space. On the other hand, all the images from the LIVE Challenge Database, which contain authentic, often agglomerated distortions overlap to a great extent in this parameter space despite the wide spread of their quality distributions.

Although the above visualizations in Figs. 2.12 and 2.14 were performed in a lower-dimensional space of parameters, it is possible that authentically distorted images could exhibit better separation if modeled in a higher dimensional space of perceptually relevant features. It is clear, however that mixtures of authentic distortions may affect the statistics of images distorted by single, synthetic distortions quite differently. Figures 2.12 and 2.14 also suggest that although the distortion-informative image features used in several state-of-the-art IQA models are highly predictive of the perceived quality of inauthentically distorted images contained in legacy databases [3, 48], these features are insufficient to produce accurate predic-

tions of quality on real-world authentically distorted images.

In this dissertation, I address these limitations by capturing other, more diverse statistical image features that improve the quality prediction power on authentically distorted images. Specifically, in Chapter 4, I present the perceptually relevant natural scene statistics of authentically distorted images, in different color spaces and transform domains. I propose a bag of *feature-maps* approach which avoids assumptions about the *type of distortion(s)* contained in an image and focuses instead on capturing consistencies, or departures therefrom, of the statistics of real world images and achieves standout performance.

2.4 Stalling Events in Mobile Streaming Videos

In this section, I will frame the setting for quality assessment in the context of streaming videos. I first provide some background of over-the-top adaptive streaming protocols adapted by a number of media streaming services and motivate the need for high-quality subjective QoE data and accurate objective predictors.

2.4.1 Quality of Experience and HTTP-based Adaptive Bitrate Streaming Protocols

Most digital content goes through several stages of processing, which can degrade quality, before ultimately being delivered to viewers. One of these stages is the transmission of videos over wired or wireless networks. The limits of network capacities, network fluctuations, and bandwidth limitations can cause volatile network conditions, that, at the client, can result in *rebuffering* or *stalling events*, which interrupt video playback.

An example of a stalling event in a video is illustrated in Fig. 2.15. Such network-induced stalling events, along with bitrate fluctuations, can negatively impact a viewer's degree of satisfaction with the experience of the delivered video

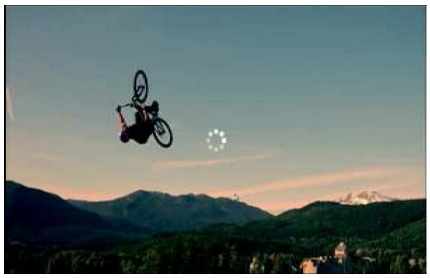


Figure 2.15: A sample stalled video sequence.

content or of the network service, which can lead to user attrition and even viewer abandonment.

As mentioned in Chapter 1, Quality of Experience (QoE) refers to a viewer's holistic perception and satisfaction with a given content, communication network, or content-providing service. As a consequence of viewers' demand for higher-quality video content, cast against the increasing competition among an expanding crowd of content and network providers (e.g., Netflix, HBO, T-Mobile, AT&T), accounting for and improving an end user's QoE has become an essential goal of content, network, and cellular services.

Media streaming services, such as YouTube and Netflix, typically leverage HTTP-based adaptive streaming protocols such as Dynamic Adaptive Streaming over HTTP (DASH) [75] and HTTP Live Streaming (HLS) [76] to make video delivery scalable and adaptable to the available network bandwidth. Under such protocols, videos are typically divided into *segments* (of fixed duration), where each video segment is encoded at multiple bitrates and resolutions (also called *video levels*). A stream-switching controller designed either at the server-side [77] or the client-side [78, 79, 80, 81] *adaptively* predicts (and then requests) an "optimal" video level depending on the requester's device, network buffer occupancy, network conditions, or other factors. These algorithms aim to minimize the number of rebuffering events and bitrate switches as a way to lessen user annoyance.

The main drawback of these algorithms is that the end user's perceived QoE is not being objectively measured (or maximized). Though reducing the number of stalls and bitrate switches is a reasonable approach to reduce viewer annoyance, it does not capture a viewer's holistic perception of quality or guarantee the best QoE. A user's perceived QoE is greatly influenced by the complex interplay of video content, number of rebuffering events, rebuffering lengths, rebuffering locations within a video, and so on. Therefore, having a fast and accurate *objective* QoE predictor that automatically predicts quality scores that correlate well with perceived QoE can serve as feedback to improve the performance of stream-switching algorithms (at either the client or the server side).

2.4.2 Subjective Assessment of Viewer's QoE

A key ingredient towards designing accurate QoE predictors is the availability of realistic, representative training data, i.e., videos containing stalling events, that are annotated with human opinion scores. These human opinion scores obtained from conducting subjective studies help to thoroughly understand the specific factors regarding video stream quality that effect viewers' QoE can help researchers better understand how increases in network video quality affect viewer behavior. This understanding can lead to design choices that make more efficient use of network resources. These studies are also critical for designing reliable models for objective evaluation of QoE that account for stalling events in a way that is consistent with subjective human evaluation, regardless of video content or the type and strength of rebuffering.

Video Quality Assessment is a thriving area of research, and a number of popular, public-domain video quality databases have been designed in the past decade [82, 83, 84, 85, 86, 87]. The videos in these data collections are about 10–15 seconds long and model different post-capture and in-capture spatial and

temporal distortions, such as compression, transmission errors, frame freezes, artifacts due to exposure and lens limitations, focus distortions, and color aberrations. Though these databases have undoubtedly guided the development of VQA algorithms [88, 89, 90, 91, 92, 93, 94], they do not model network-induced distortions, such as start-up delays and stalling events.

A few video quality studies have been conducted in the recent past to analyze the effects of network streaming quality on QoE [95, 96, 97, 98, 99]. The focus of these studies has been to investigate the influence of simple factors such as startup delays and total stall length on an end user's QoE. Certain general conclusions, e.g., that longer start-up delays are more annoying than shorter ones, have been reported in these studies. However, the datasets and the subjective scores used in these studies are not publicly available. Other recent studies have focused specifically on stall, or as they refer to it – *pause*, features, such as position and duration, in Transmission Control Protocol (TCP) applications [100, 101, 102]. However, these works draw very general conclusions about the correlations between pause features and subjective QoE, and again, none of the source content or subjective data is publicly available. Recent works have also focused on how HTTP Adaptive Streaming (HAS), specifically, impacts QoE [103, 104, 105, 106, 107, 105, 106, 107], but because the goal of HAS is to limit the chance of a stall occurrence, these studies focus primarily of spatial quality fluctuations.

The authors of the publicly available Waterloo Quality-of-Experience database [8] combine bitrate variations and stalling events in their videos. However, their stall patterns are unvaried, as all of the stalls are of fixed duration (5 seconds) and are added at fixed locations (start and middle) in every video in their collection. Such fixed distortion patterns do not reflect typical video streaming situations, and may even bias the subjects to have expectations of subsequent videos after initial viewings. Repetitive patterns may also cause even shallow learners to overfit to the data.

Further, none of these studies attempt to gather continuous-time human judgments of quality, and they only record overall subjective QoE scores. Real-time QoE measurements are of far more interest, because when streaming video to a client, bitrate decisions must be made to maximize QoE, so continuous-time video quality must be balanced against the likelihood of stall events.

The study presented in [9] is similar to this work, as their database was also designed by taking realistic network bandwidth usage into account. However, their database is not publicly available in its entirety (only 24 out of 112 videos are publicly available). While I seek to deeply understand the effect of the complex interplay of different stall-specific parameters by looking at 26 diverse stall patterns, [9] only considers 8 distortion patterns that combine bitrate variations and rebuffering events in very specific ways.

Thus, in summary, the methods used in previous studies do not adequately advance my goals, as they suffer from one or more of the following problems:

1. Small, insignificant sizes of the video collections.
2. Insufficient number of subjective judgments.
3. Unknown video sources with limited variability in content.
4. Lack of public availability of the databases.
5. Lack of fine-grained and *continuous-time* subjective ratings.
6. Lack of a variety in the distortion severities and patterns that would broadly reflect the different bandwidth limitations that need to be tackled by video-streaming services such as Netflix and YouTube.

In this dissertation, I sought to address all of the above shortcomings of the existing data collections. Moreover, as I will discuss in Chapter 5, every key aspect of the proposed database construction, such as the choice of reference videos, the design of

distortion patterns, and the subjective study, were tailored to be as close as possible to the real world scenario of streaming online videos on mobile devices. Furthermore, I also present details of a novel subject rejection strategy for continuous-time data and present my analysis of continuous-time subjective responses.

2.4.3 Automatic QoE Predictors

As mentioned in Section 2.4.1, automatic QoE models have the potential to motivate the design of solutions that strike a balance between reducing network operational costs while delivering video with the highest possible quality to customers. Top-performing video quality predictors [88, 89, 90, 91, 93] that have been developed in the past decade deal with post-processing distortions but not network-induced impairments. Although the impact of bitrate fluctuations on viewer QoE can be understood by employing these VQA models, the impact of stalling events interspersed with bitrate variations cannot be captured using them, thus creating a demand for reliable QoE predictors for analyzing streaming videos. A handful of objective QoE predictors have been designed [108, 109, 98] that derive global video statistics based on the total stall length and the number of random video stalls. The DQS model [110] also considers global stall statistics and a linear model to predict a continuous-time QoE score. Specifically, this model defines three events: start-up delay, first rebuffering, and multiple rebuffering (explicitly) based on empirical observations on the final QoE scores of the LIVE Mobile Stall Video Database-I [111]. The underlying assumption of the DQS model is that an end user's QoE is driven by these predefined events. Different model parameters are chosen for each event that drives the DQS model's quality prediction. Specifically, the mean and variance of the opinion scores of videos afflicted only by startup delays, or by a startup delay followed by a single (or multiple) rebuffering event(s) are computed, and the parameters of the DQS model are determined. Thus, the generalizability of the DQS model to more

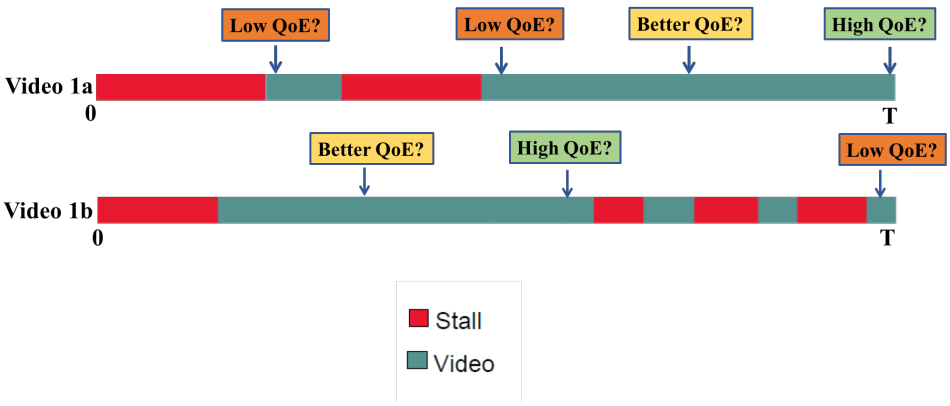


Figure 2.16: Illustrating the affect of hysteresis on perceived quality of experience. A video content afflicted with two different stalling patterns (video 1a and 1b) with equal total stall length time. Stalls in a video are illustrated in red and the video playback is illustrated in blue. Despite their common attributes, these two videos will be perceived very differently by a viewer.

diverse stall patterns is questionable. The recently proposed SQI model [8] combines perceptual video presentation quality and simple stalling event-based features to predict QoE.

Limitations of the existing approaches:

One key limitation of most models is that they are based only on global statistics and cannot capture the time-varying levels of satisfaction experienced when viewing streaming videos. Furthermore, QoE also depends on a behavioral hysteresis or recency "after effect," whereby a user's QoE at a particular moment also depends on their viewing experiences preceding that moment. For example, the memory of an early unpleasant viewing experience caused by a stalling event may negatively impact future QoE and, thus, may also negatively impact the overall QoE. A long initial delay (of length L, for example) at the beginning of a video sequence may more likely to lead to viewer abandonment, than when viewing the same video content containing multiple stalls whose total length equals L. Additionally, a stalling event

occurring towards the end of a video sequence could have a more negative impact on the final overall perception of video quality, than a stall of the same length occurring at an earlier position in the same video. This dependency on previous viewing experiences is generally nonlinear and can be crucial in determining both the overall as well as the instantaneous QoE of viewers, but this information is not currently being exploited by contemporary QoE prediction models.

I illustrate the affect of memory on perceived quality in Figure 2.16. Consider a given video content that is afflicted with two different stalling patterns, such that the sum of the individual stalls in both the patterns is the same. in both the cases. It is reasonable to assume that a user is more likely to be annoyed during and immediately after a stall, but would probably recover from this bad experience after some period of uninterrupted video playback. Thus, these two video contents will be perceived very differently by an end user. This behavioral response is not specific to stalls and has been observed in videos with bitrate specific distortions.

Thus, the memory of the past poor quality is retained in the memory for some non-negligible amount of time even after the video has returned to acceptable levels of quality. Clearly such a complex property of our HVS is not effectively modeled when global features such as the sum of the stall length and their number is taken into consideration.

In this dissertation, I propose a perceptually-driven objective predictor that predicts the time-varying QoE by modeling the non-linearities and the hysteresis properties of the human visual system. As I describe in Chapter 6, I also model the client-side network buffer state, measure the spatial and temporal video complexity, and the perceptual video quality and effectively combine them to create a continuous-time QoE predictor.

Chapter 3

Crowdsourced Study of Subjective Picture Quality

As mentioned earlier, I approached the problem of blind image quality assessment on authentically distorted images from the ground up and first constructed a novel image quality database of real-world distortions. In this chapter, I will introduce this database and present the details of the subjective study I conducted to obtain a very large number of human opinion scores. This chapter is organized as follows:

1. First, I introduce the content and characteristics of the new **LIVE In the Wild Image Quality Challenge Database**, which contains 1162 authentically distorted images captured from many diverse mobile devices. Each image was collected without artificially introducing any distortions beyond those occurring during capture, processing, and storage by a user's device.

2. Next, I aimed to gather very rich human data and designed and implemented an extensive online subjective study by leveraging Amazon's crowdsourcing system, the Mechanical Turk. I will describe the design and infrastructure of my online crowdsourcing system and how I used it to conduct a very large-

scale, multi-month image quality assessment subjective study, wherein a wide range of diverse observers recorded their judgments of image quality.

3. I also discuss the critical factors that are involved in successfully crowdsourcing human IQA judgments, such as the overall system design of the online study, methods for subject validation and rejection, task remuneration, influence of the subjective study conditions on end users' assessment of perceptual quality, and so on.

3.1 LIVE In the Wild Image Quality Challenge Database

As already mentioned in Section 2.2.1, current IQA models have been designed, trained, and evaluated on benchmark databases such as the LIVE Image Quality Database [3], the TID databases [48, 5], the CSIQ database [6], and a few other small databases [52], all of which model single, *inauthentic* distortions. These image distortions fixed by a database designer for the purpose of ensuring a statistically significant set of human responses are not the same as real-world distortions introduced by highly diverse cameras in the hands of real-world users. In my work, I refer to the latter distorted images obtained as *authentically distorted*.

Some important characteristics of real-world, authentically distorted images captured by naïve users of consumer camera devices, is that the pictures obtained generally cannot be accurately described by a simple generative model, nor as suffering from single, statistically separable distortions. For example, a picture captured using a mobile camera under low-light conditions is likely to be under-exposed, in addition to being afflicted by low-light noise and blur. Subsequent processes of saving and/or transmitting the picture over a wireless channel, will generally introduce further compression and transmission artifacts. Further, the characteristics of the overall distortion "load" of an image will depend on the device used for capture

and on the camera-handling behavior of the user. Consumer-grade digital cameras differ widely in their lens configurations, levels of noise sensitivity and acquisition speed, and in post-acquisition in-camera processing. Camera users differ in their shot selection preferences, hand steadiness, and situational awareness. Overall, our understanding of true, authentic image distortions is quite murky. Such complex, unpredictable, and currently un-modeled mixtures of distortions are characteristic of real-world pictures that are authentically distorted. There currently isn't any known way to categorize, characterize, or model such complex and uncontrolled distortion mixtures, and it is certainly unreasonable to expect an image quality scientist to be able to excogitate a protocol for creating authentically distorted images in the laboratory, by synthetically combining controlled, programmed distortions into what must ultimately be regarded as highly authentically distorted images.

There is a way to create databases of authentically distorted images, which is by acquiring images taken by many casual camera users. Normally, inexpert camera users will acquire pictures under highly varied and often suboptimal illuminations conditions, with unsteady hands, and with unpredictable behavior on the part of the photographic subjects. Such real-world, authentically distorted images exhibit a broad spectrum of authentic quality "types," mixtures, and distortion severities, that defy attempts at accurate modeling or precise description. Authentic mixtures of distortions are even more difficult to model when they interact, creating new agglomerated distortions not resembling any of the constituent distortions. A simple example would be a noisy image that is heavily compressed, where the noise presence heavily affects the quantization process at high frequencies, yielding hard-to-describe, visible compressed noise artifacts. Users of mobile cameras will be familiar with this kind of spatially-varying, hard-to-describe distortion amalgamation.

As mentioned in Section 2.2.1, the lack of content diversity and mixtures of bonafide distortions in existing, widely-used image quality databases [48, 3] is a con-

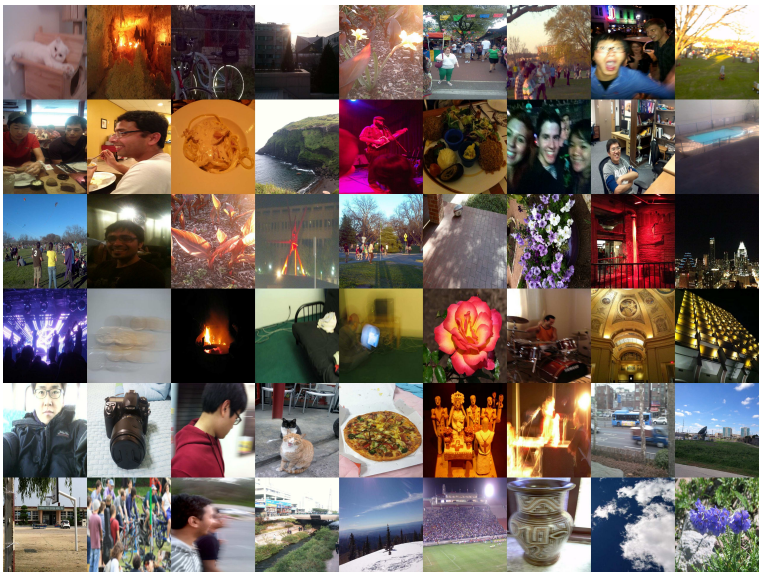


Figure 3.1: Sample images from the LIVE In the Wild Image Quality Challenge Database. These images include pictures of faces, people, animals, close-up shots, wide-angle shots, nature scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest.

tinuing barrier to the development of better IQA models and prediction algorithms of the perception of real-world image distortions. To overcome these limitations and towards creating a holistic resource for designing the next generation of robust, perceptually-aware image assessment models, I designed and created the **LIVE In the Wild Image Quality Challenge Database**, containing images afflicted by diverse authentic distortion mixtures on a variety of commercial devices. Figure 3.1 presents a few images from this database. The images in the database were captured using a wide variety of mobile device cameras as shown in Fig. 3.2. The images include pictures of faces, people, animals, close-up shots, wide-angle shots, nature scenes, man-made objects, images with distinct foreground/background configurations, and images without any specific object of interest. Some images contain high luminance and/or color activity, while some are mostly smooth. Since these images

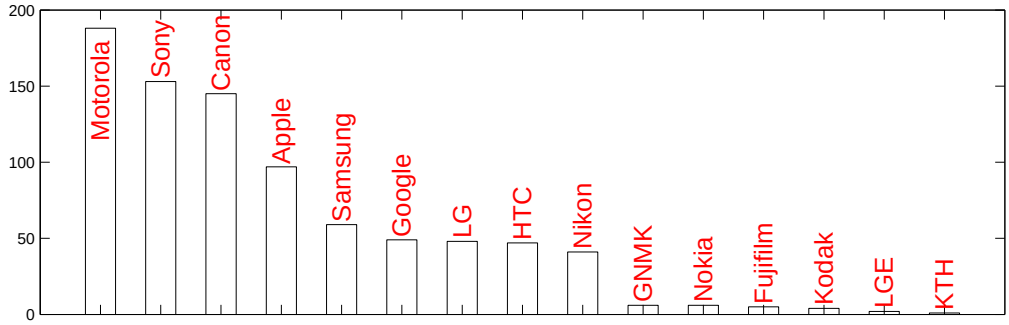


Figure 3.2: Distribution of different manufacturers of the cameras that were used to capture a sample of images contained in my database.

are naturally distorted as opposed to being artificially distorted post-acquisition pristine reference images, they often contain mixtures of multiple distortions creating an even broader spectrum of perceivable impairments.

3.2 Crowdsourced framework for gathering subjective scores

With a goal to gather a large number of human opinion scores on the image collection detailed in Section 3.1, I designed and implemented an online crowdsourcing system which I used to gather more than $350,000$ human ratings of image quality which amounts to about 175 ratings on each image in the new LIVE Challenge Database. In this section, I will briefly describe the details of the online study framework and the subjective study details.

Crowdsourcing systems like Amazon Mechanical Turk (AMT), Crowd Flower [112], and so on, have emerged as effective, human-powered platforms that make it feasible to gather a large number of opinions from a diverse, distributed populace over the web. On these platforms, "requesters" broadcast their task to a selected pool of registered "workers" in the form of an open call for data collection. Workers

who select the task are motivated primarily by the monetary compensation offered by the requesters and also by the enjoyment they experience through participation.

3.2.1 Instructions, Training, and Testing

The data collection tasks on AMT are packaged as HITs (Human Intelligence Tasks) by requesters and are presented to workers, who first visit an instructions page which explains the details of the task. If the worker understands and likes the task, she needs to click the "Accept HIT" button which then directs her to the actual task page at the end of which, she clicks a "Submit Results" button for the requester to capture the data.

Crowdsourcing has been extensively and successfully used on several object identification tasks [113, 114] to gather segmented objects and their labels. However, the task of labeling objects is often more clearly defined and fairly straightforward to perform, by contrast with the more subtle, challenging, and highly subjective task of gathering opinion scores on the perceived quality of images. The generally naive level of experience of the workers with respect to understanding the concept of image quality and their geographical diversity made it important that detailed instructions be provided to assist them in understanding how to undertake the task without biasing their perceptual scores. Thus, every unique participating subject on AMT that selects this HIT was first provided with detailed instructions to help them assimilate the task. A screenshot of this web page is shown in Fig. 3.3. Specifically, after defining the objective of the study, a few sample images were presented which are broadly representative of the kinds of distortions contained in the database, to help draw the attention of the workers to the study and help them understand the task at hand. A screenshot of the rating interface was also given on the instructions page, to better inform the workers of the task and to help them decide if they would like to proceed with it.

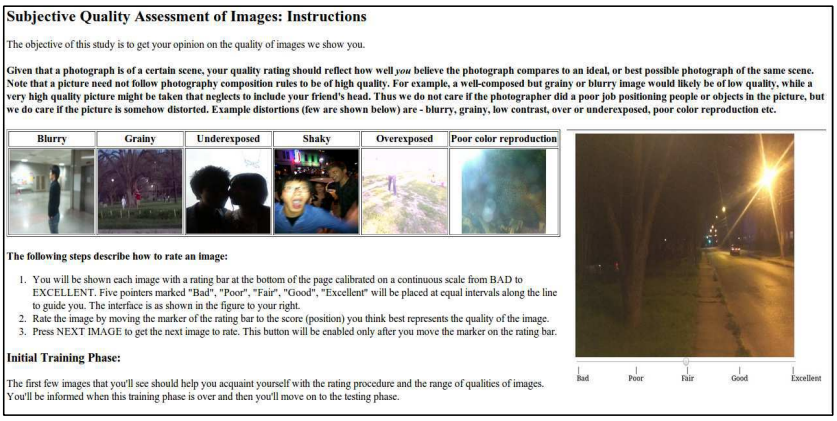


Figure 3.3: Instructions page shown before the worker accepts the task on AMT.

Ensuring unique participants:

After reading the instructions, if a worker accepted the task, and did so for the first time, a rating interface was displayed that contains a slider by which opinion scores could be interactively provided. A screenshot of this interface is also shown in Fig. 3.4. In the event that this worker had already picked this task earlier, I informed the worker that the study requires unique participants and this worker was not allowed to proceed beyond the instructions page. Only workers with a confidence value[1] greater than 0.75 were allowed to participate. Even with such stringent subject criteria, I gathered more than 350,000 ratings overall.

Study framework:

I adopted a single stimulus continuous procedure [115] to obtain quality ratings on images where subjects reported their quality judgments by dragging the slider located below the image on the rating interface. This continuous rating bar is divided into five equal portions, which are labeled "bad," "poor," "fair," "good,"

[1]AMT assigns a confidence score in the range of 0-1 to each worker, based on the accuracy of their responses across all the HITs they have accepted thus far. The higher this number, the more trustworthy a worker is.

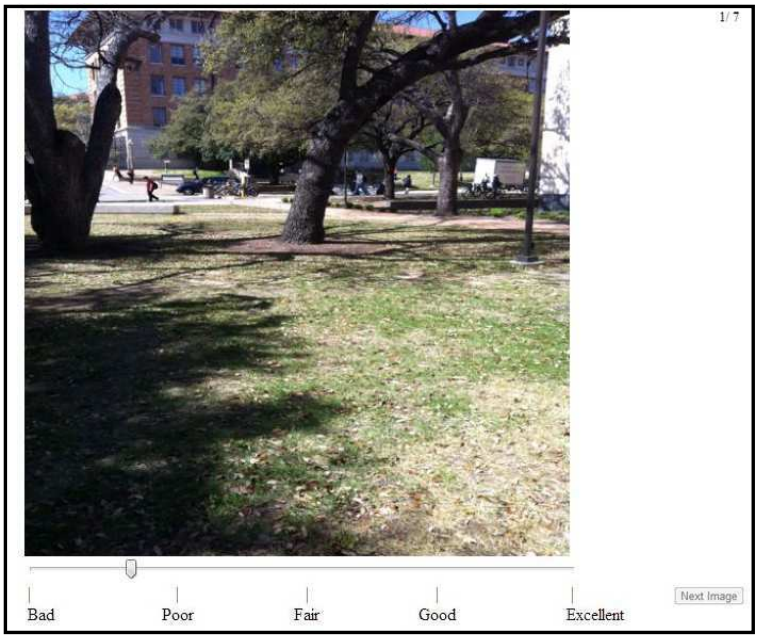


Figure 3.4: The rating interface presented to every subject on which they can provide opinion scores on images.

and "excellent." After the subject moved the slider to rate an image and pressed the *Next Image* button, the position of the slider was converted to an integer quality score in the range $1 - 100$, then the next image was presented. Before the actual study began, each participant is first presented with 7 images that were selected by us as being reasonably representative of the approximate range of image qualities and distortion types that might be encountered. I call this the **training phase**. Next, in the **testing phase**, the subject is presented with 43 images in a random order where the randomization is different for each subject. This is followed by a quick survey session which involves the subject answering a few questions. Thus, each HIT involves rating a total of 50 images and the subject receives a remuneration of 30 cents for the task. Figure 3.5 illustrates the detailed design of the HIT on IQA and Fig. 3.6 illustrates how I package the task of rating images as a HIT and

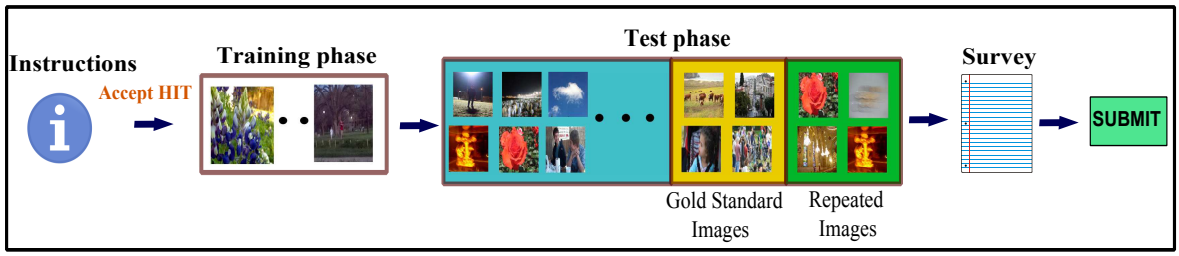


Figure 3.5: Illustrating the design of the HIT. Once a worker clicked the "Accept HIT" button and did so for the first time, I directed her to the training phase which was followed by a test phase. A worker who had already participated once in my study and attempted to participate again was not allowed to proceed beyond the instructions page. For the purpose of illustration, I show gold standard and repeated images in exclusion. In reality, the pool of 43 test images was presented in a random order.

effectively disperse it online via AMT to gather thousands of human opinion scores.

3.2.2 Subject Reliability and Rejection Strategies

Crowdsourcing has empowered us to efficiently collect large amounts of ratings. However, it raises interesting issues such as dealing with noisy ratings and addressing the reliability of the AMT workers.

Intrinsic metric To gather high quality ratings, only those workers on AMT with a confidence value greater than 75% were allowed to select my task. Also, in order to not bias the ratings due to a single worker picking my HIT multiple times, I imposed a restriction that each worker could select my task no more than once.

Repeated images 5 of each group of 43 test images were randomly presented twice to each subject in the testing phase. If the difference between the two ratings that a subject provided to the same image each time it was presented exceeded a threshold on at least 3 of the 5 images, then that subject was rejected. This served to eliminate workers that were providing unreliable, "random" scores. Prior to the full-fledged

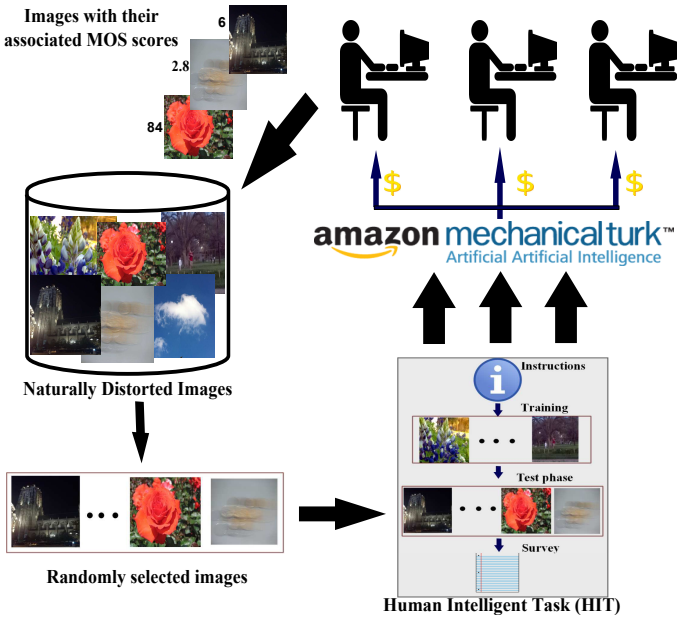


Figure 3.6: Illustrating how the system I designed packages the task of rating images as a HIT and disperses it on Mechanical Turk.

study, I conducted an initial subjective study and obtained ratings from 300 unique workers. I then computed the average standard deviation of these ratings on all the images. Rounding this value to the closest integer yielded 20 which I then used as my threshold for subject rejection.

Gold Standard Data 5 of the remaining 38 test images were drawn from the LIVE Multiply Distorted Image Quality Database [4] to supply a control. These images along with their corresponding MOS from that database were treated as a *gold standard*. The mean of the Spearman's rank ordered correlation values computed between the MOS obtained from the workers on the gold standard images and the corresponding ground truth MOS values from the database was found to be **0.9851**. The mean of the absolute difference between the MOS values obtained from my crowdsourced study and the ground truth MOS values of the gold standard images was found to be **4.65**. Furthermore, I conducted a paired-sampled t-test and

observed that this difference between gold standard and crowdsourced MOS values is not statistically significant. This high degree of agreement between the scores gathered in a traditional laboratory setting and those gathered via an uncontrolled online platform with several noise parameters is critical to us. Although the uncontrolled test settings of an online subjective study could be perceived as a challenge to the authenticity of the obtained opinion scores, this high correlation value indicates a high degree of reliability of the scores that are being collected by us using AMT, reaffirming the efficacy of my approach of gathering opinion scores and the high quality of the obtained subject data.

3.2.3 Subject-Consistency Analysis

In addition to measuring correlations against the gold standard image data as discussed above, I further analyzed the subjective scores in the following two ways:

Inter-Subject consistency To evaluate subject consistency, I split the ratings obtained on an image into two disjoint equal sets, and computed two MOS values on every image, one from each set. When repeated over 25 random splits, an average Spearman's rank ordered correlation between the mean opinion scores between the two sets was found to be **0.9896**.

Intra-Subject consistency Evaluating intra-subject reliability is a way to understand the degree of consistency of the ratings provided by individual subjects [116]. I thus measured the Spearman's rank ordered correlation (SROCC) between the individual opinion scores and the MOS values of the gold standard images. A median SROCC of **0.8721** was obtained over all of the subjects.

All of these additional experiments further highlight the high degree of reliability and consistency of the gathered subjective scores and of my test framework.

3.2.4 Analysis of the Subjective Scores

The database currently comprises of more than $350,000$ ratings obtained from more than $8,100$ unique subjects (after rejecting unreliable subjects). Enforcing the aforementioned rejection strategies led us to reject 134 participants who had accepted my HIT. Each image was viewed and rated by an average of 175 unique subjects, while the minimum and maximum number of ratings obtained per image were 137 and 213, respectively. While computing these statistics, I excluded the 7 images used in the training phase and the 5 gold standard images as they were viewed and rated by all of the participating subjects. Workers took a median duration of 4.37 minutes to view and rate all 50 images presented to them. The Mean Opinion Scores (MOS) after subject rejection was computed for each image by averaging the individual opinion scores from multiple workers. MOS is representative of the *perceived viewing experience* of each image. The MOS values range between $[3.42 - 92.43]$. Figure 3.7 is a scatter plot of the MOS computed from the individual scores I have collected. In order to compare the MOS values with single opinion scores (SOS), I computed the standard deviation of the subjective scores obtained on every image and obtained an average standard deviation of 19.2721.

The uncontrolled online test environment poses certain unique challenges: a test subject of any gender or age may be viewing the image content on any kind of a display, under any sort of lighting, from an unknown distance, and an unknown level of concentration, each of which can affect her choice of quality score. Figures 3.9 (a) and 3.9 (b) illustrate the demographic details of the unique subjects who have participated in my study[2]. Most of them reported in the final survey that they are inexperienced with image quality assessment but do get annoyed by image impairments they come across on the Internet. Since I did not test the subjects for vision problems, they were instructed to wear corrective lenses during

[2]Gathering demographic details of the workers is a common practice on Mechanical Turk. None of the workers expressed any concerns when providing us with these details.

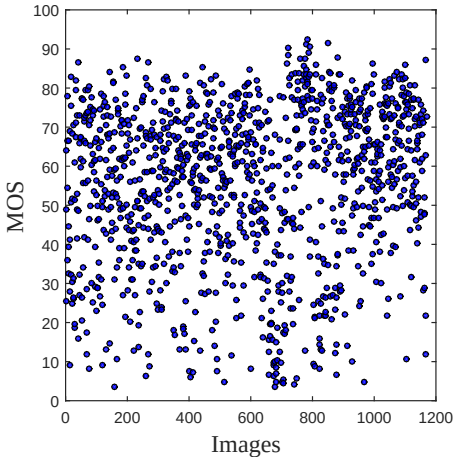


Figure 3.7: Scatter plot of the MOS scores obtained on all the images in the database.

the study if they do so in their day-to-day life. Later in the survey, the subjects were asked if they usually wore corrective lenses and whether they wore the lenses while participating in the study. The ratings given by those subjects who were not wearing their corrective lenses they were otherwise supposed to wear were rejected. Figures 3.9 (c) and 3.9 (d) illustrate the distribution of the distances from which workers have viewed the images and the broad classes of different display devices used by them. These four plots illustrate the highly varied testing conditions that exist during the online study and also highlight the diversity of the subjects. Figure 3.8 (a) illustrates the distribution of the types of consumer image capture devices that are preferred by the users. It is evident from this plot that most of the workers reported that they prefer using mobile devices to capture photographs in their daily use. One of the questions I posed to the subjects in the survey was whether the poor quality of pictures that they encounter on the Internet bothers them. Subjects chose between the following four options - "Yes," "No," "I don't really care," and "I don't know." The distribution of the responses to this question is plotted in Fig.3.8 (b) which clearly indicates that a large population of the workers are bothered by

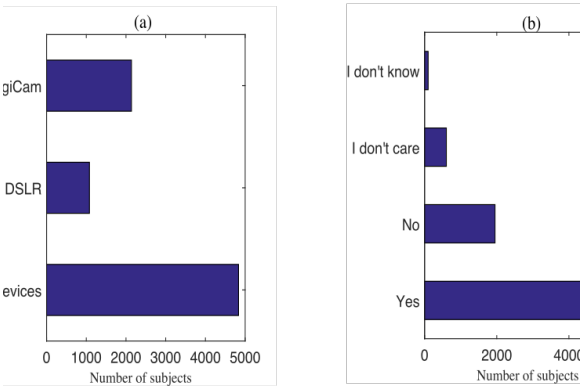


Figure 3.8: Illustrating (a) the kind of consumer image capturing devices preferred by users and (b) their sensitivity to perceived distortions in digital pictures viewed on the Internet.

poor quality Internet pictures.

I next present my analysis of the influence of several factors such as age, gender, and display devices on user's perceptual quality. In all cases, I study the effect of each factor independently while fixing the values of the rest of the factors. I believe this strategy helped to closely study the influence of each factor independently and to help avoid combined effects caused by the interplay of several factors on a user's perceptual quality. Note that the results presented in the following sections are consistent irrespective of the specific values that were fixed for the factors.

3.2.5 Gender

To understand to what extent gender had an affect on the quality scores, I separately analyzed the ratings obtained from male and female workers on five randomly chosen images (Figures 3.10(a)-(e)) while maintaining all the other factors constant. Specifically, I separately captured the opinion scores of male and female subjects who are between $20-30$ years old, and reported in the survey to be using a desktop and sitting about $15-30$ inches from the screen. Under this setting and on the

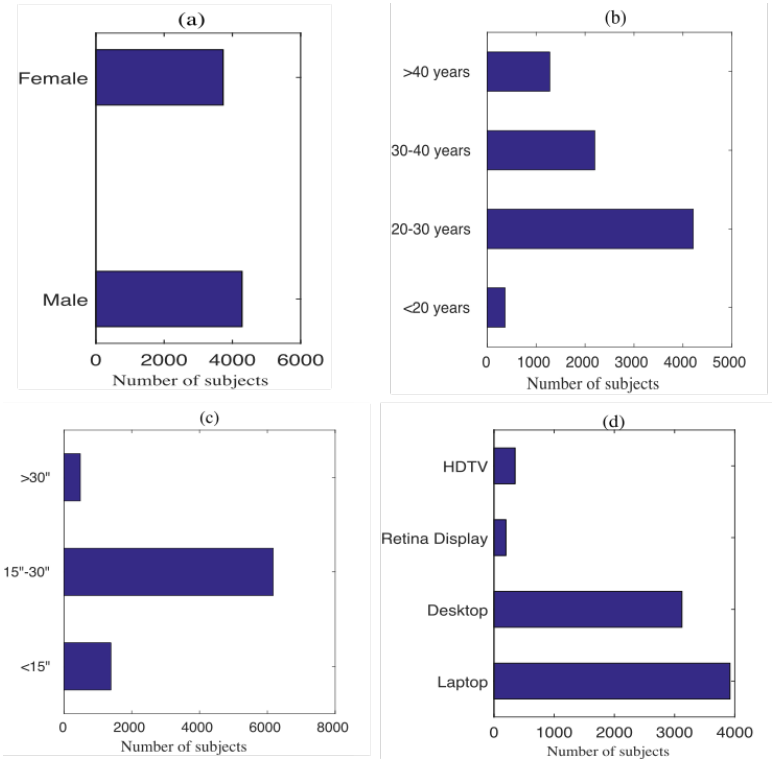


Figure 3.9: Demographics of the participants (a) gender (b) age (c) approximate distance between the subject and the viewing screen (d) different categories of display devices used by the workers to participate in the study.

chosen set of images, both male and female workers appeared to have rated the images in a similar manner. This is illustrated in Figure 3.11(a).

3.2.6 Age

Next, I considered both male and female workers who reported using a laptop during the study and were sitting about $15 - 30$ inches away from their display screen. I grouped their individual ratings on these 5 images (Fig. 3.10) according to their age and computed the MOS of each group and plotted them in Fig 3.11(b). For the images under consideration, again, subjects belonging to different *age categories*

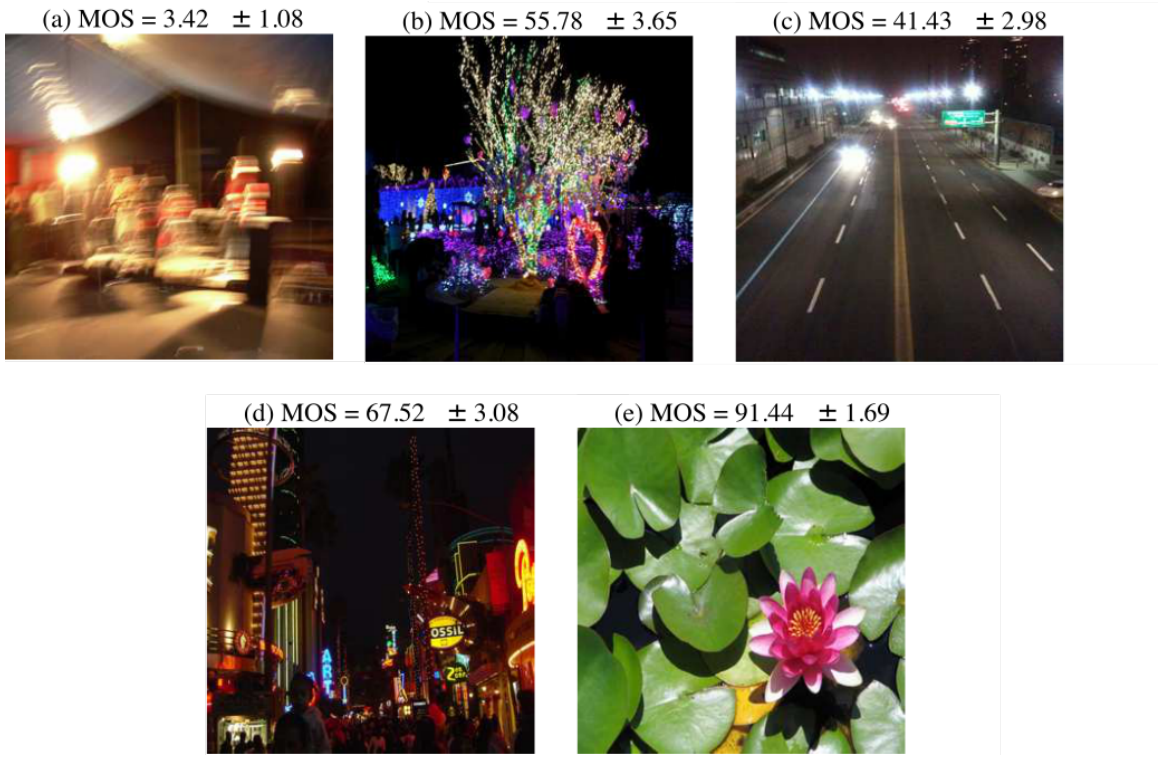


Figure 3.10: A few randomly chosen images from the LIVE In the Wild Image Quality Challenge Database that are used to illustrate the influence of various parameters on the QoE of the study participants. The upper caption of each image gives the image MOS values and the associated 95% confidence intervals.

appeared to have rated them in a similar manner.

Although gender and age did not seem to significantly affect the ratings gathered on the randomly chosen images discussed above, I believe that other factors such as the content in the image can play a significant role in being appealing to one group more than to another. A systematic study focused exclusively on understanding the interplay of image content, gender, and age using this database might help better understand the impact of each of these factors on perceptual quality.

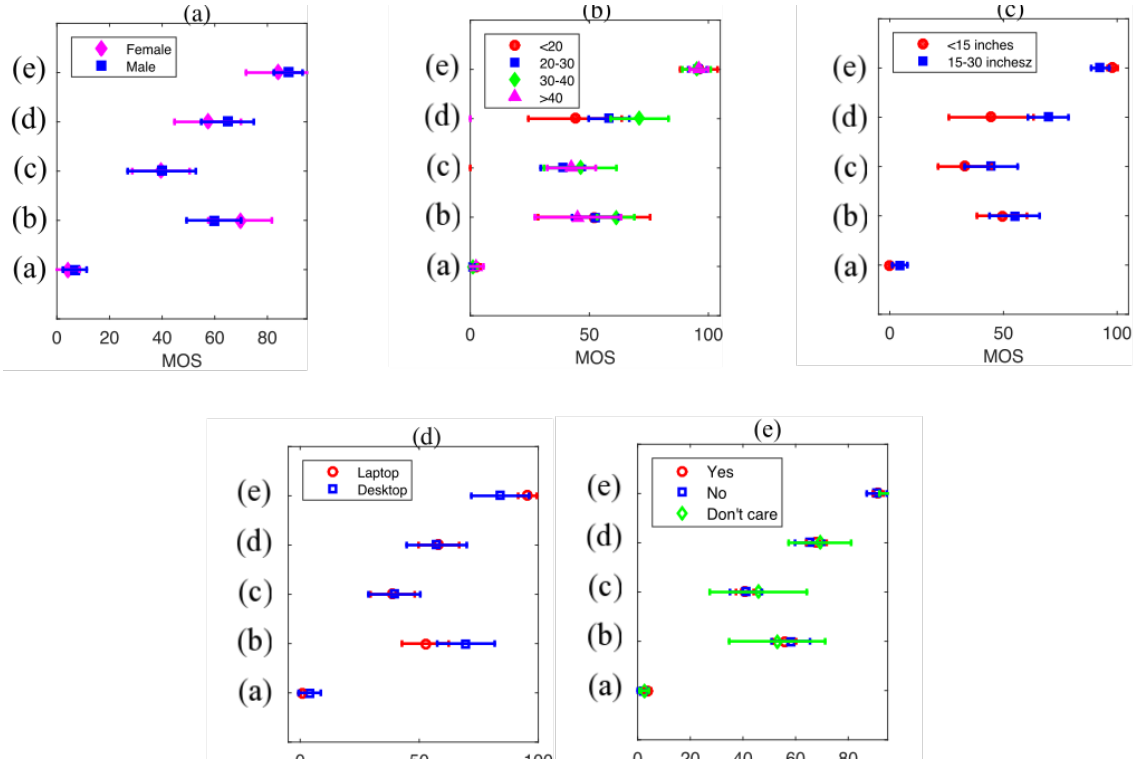


Figure 3.11: Plots showing the influence of a variety of factors on a user's perception of picture quality. The factors are: (a) gender (b) age (c) approximate distance between the subject and the viewing screen and (d) types of display devices used by the workers to participate in the study. Plot (e) shows the influence of users' distortion sensitivity on their quality ratings. The plots detail the range of obtained MOS values and the associated 95% confidence intervals.

3.2.7 Distance from the Screen

I next explored the influence of the distance between a subject and her monitor, on the perception of quality. One of the questions in the survey asked the subjects to report which of the three *distance categories* best described a subject's location relative to the viewing screen - "less than 15 inches," "between 15 to 30 inches," and "greater than 30 inches."

I gathered the ratings of subjects who reported to be between $30-40$ years old

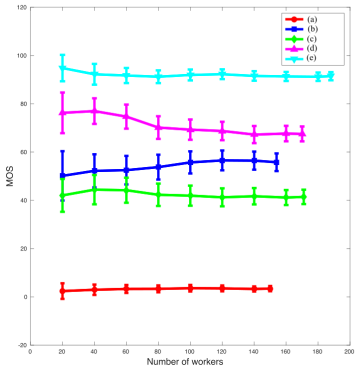


Figure 3.12: MOS plotted against the number of workers who viewed and rated the images shown in Fig. 3.10.

and were participating in the study using their desktop computer. I grouped their ratings[3] on the five test images (Fig. 3.10) according to these distance categories and report the results in Fig. 3.11(c). It may be noticed that the difference between the mean of the ratings obtained on the same image when viewed from a closer distance as compared to when the same image was viewed by subjects from a greater distance is not statistically significant. However, I do not rule out the possible influences that viewing distance may have on distortion perception from an analysis of 5 random images. The observed indifference to viewing distance could be due to an interplay of the resolution of the display devices, image content, and viewing distances which is a broad topic worthy of future study.

3.2.8 Display Device

To better understand the influence of display devices on QoE, I focused on workers between $20-30$ years old and who reported to be $15-30$ inches away from the screen while participating in the study. I grouped the ratings of these subjects on the five

[3]I received very few ratings from subjects who reported to be sitting greater than 30 inches away from their display screen and hence excluded those ratings from this segment of analysis.

images in Fig. 3.10 according to the display device that the subjects reported to have used while participating in the study.

As illustrated in Fig. 3.11(d), the influence of the specific display device that was used for the study appears to have had little effect on the recorded subjective ratings. Of course, I am not suggesting that the perceptual quality of images is unaffected by the display devices on which they are viewed. It is possible that more fine-grained detail regarding the type of display device used by the study participants (e.g., screen resolution, display technology involved, shape of the screen etc.) could deepen our understanding of the dependency between display device and perceptual image quality. However, I chose to focus as much of each participants' effort on the visual tasks as reasonable, and so did not poll them on these details, leaving it for future studies.

3.2.9 Annoyance of Low Image Quality

As mentioned earlier, one of the questions posed to the subjects in the survey was whether the quality of pictures they encounter on the Internet bothers them (distribution of the responses in Fig. 3.8 (b)). When I grouped the ratings according to these three answers, I noticed that the subjects from each of these three response categories were almost equally sensitive to the visual distortions present in the images from my dataset. This is illustrated in Figure 3.11 (e).

Figure 3.12 illustrates how MOS values flatten out with increases in the number of subjects rating the images. It is interesting to note that there is much more consistency on images with very high and very low MOS values than on intermediate-quality images. Of course, the opinion scores of subjects are affected by several external factors such as the order in which images are presented, a subject's viewing conditions, and so on, and the MOS thus exhibit variability with respect to the number of workers who have rated them.

Table 3.1: Summary of my analysis of the different QoE influencing factors on the perception of image distortions.

Influencing factor	**Factors that were held constant**	**Observation**
Gender	Age: 20-30 years, Display device: Desktop, Distance from the screen: 15-30 inches	Both male and female workers appeared to rate images in a similar manner.
Age	Gender: Both male and female workers, Display device: Desktop, Distance from the screen: 15-30 inches	Very little difference was noticed in the ratings of the subjects of different age groups.
Subject's distance from the display screen	Gender: Both male and female workers, Age: 30-40 years, Display device: Desktop	Little effect on the subjective ratings.
Display device	Gender: Both male and female workers, Age: 20-30 years, Distance from the screen: 15-30 inches.	Little effect on subjective ratings.
Subject's general sensitivity to perceptual quality	None	People who claimed to differ in their level of annoyance in response to image distortions appeared to rate images in a similar manner.

I summarize all the factors whose influence I studied and presented in this section (by controlling the other factors) in Table 3.1.

3.2.10 Limitations of the current study

Crowdsourcing is a relatively new tool with considerable potential to help in the production of highly valuable and generalized subjective databases representative of human judgments of perceptual quality. However, the approach involves many complexities and potential pitfalls which could affect the veracity of the subject results. A good summary and analysis of these concerns may be found in [116].

For example, while I have a high degree of faith in the subject results, it is

based on a deep analysis of them rather than simply because the participants were screened to have high AMT confidence values. As mentioned earlier, the confidence values of the workers computed by AMT is an aggregate that is measured over all the tasks in which a worker has participated. This metric thus is not necessarily an indicator of reliability with regards to any specific task and should be accompanied by rigorous, task-specific subject reliability methods. Future studies would benefit by a more detailed data collection and analysis of the details of workers' display devices [116] and viewing conditions. While the current philosophy, even in laboratory studies, is to not screen the subjects for visual problems, given the newness of the crowdsourcing modality, it might be argued that visual tests could be used to improve subject reliability checks. Many other environmental details could be useful, such as reports of the time spent by a worker in viewing and rating images, to further measure worker reliability.

Conclusion

The crowdsourcing image quality study allowed diverse subjects to participate at their convenience, and in diverse, uncontrolled viewing circumstances, enhancing my ability to investigate the effects of each of these factors on perceived picture quality. The results of my analysis of the factors affecting data reliability, and my observations of the high correlations of the objective quality scores against the MOS values of the gold standard images that were obtained under controlled laboratory conditions, both strongly support the efficacy of my online crowdsourcing system for gathering large scale, reliable data.

Chapter 4

Objective Automatic Quality Prediction of Images in the Wild

In this chapter, I will describe the details of a new image quality predictor called **F**eature maps based **R**eferenceless **I**mage **QU**ality **E**valuation **E**ngine (FRIQUEE). I designed FRIQUEE with a goal to tackle authentic distortions such as those captured in LIVE In the Wild Image Quality Challenge Database [117]. There is an extensive prior work on statistical modeling of normalized coefficients that currently drives top-performing blind IQA models as detailed in Section 2.3.1. However, as illustrated in Figures. 2.11 - 2.14 in Section 2.3.2, complex mixtures of authentic image distortions modify the image statistics in ways not easily predicted by these models. They exhibit large, hard to predict statistical variations as compared to synthetically distorted images. Thus, I devised an approach that leverages the idea that different perceptual image representations may distinguish different aspects of the loss of perceived image quality. Specifically, given an image, I first construct several *feature maps* in multiple color spaces and transform domains, then extract

individual and collective scene statistics from each of these maps.

I have described the perceptual significance and the operational details of the divisive normalization operator in Section 2.3. Most of the feature maps I construct as part of extracting the proposed bag of features are processed using divisive normalization. Before I describe the types of feature maps that I compute in this work, I first introduce the general statistical modeling techniques that I employ to derive and extract features from any given (divisively normalized) feature map.

4.1 Statistical Modeling of Normalized Coefficients

4.1.1 Generalized Gaussian Distributions

My approach builds on the idea exemplified by observations like those depicted in Fig. 2.11, viz., that the normalized luminance or bandpass/wavelet coefficients of a given image have characteristic statistical properties that are predictably modified by the presence of distortions. Effectively quantifying these deviations is crucial to be able to make predictions regarding the perceptual quality of images. A basic modeling tool that I use throughout is the generalized Gaussian distribution (GGD), which effectively models a broad spectrum of (singly) distorted image statistics, which are often characterized by changes in the tail behavior of the empirical coefficient distributions [74]. A GGD with zero mean is given by:

$$f(x;\alpha,\sigma^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right), \tag{4.1}$$

where

$$\beta = \sigma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \tag{4.2}$$

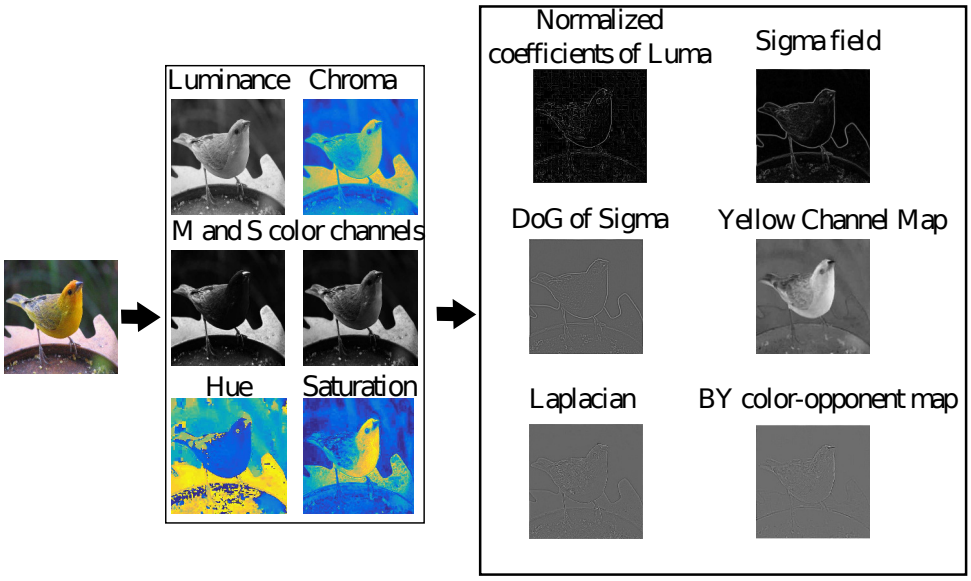


Figure 4.1: Given any image, the proposed feature maps based model first constructs channel maps in different color spaces and then constructs several feature maps in multiple transform domains on each of these channel maps (only a few feature maps are illustrated here). Parametric scene statistic features are extracted from the feature maps after performing perceptually significant divisive normalization [11] on them. The design of each feature map is described in detail in later sections.

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

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt \quad a > 0. \tag{4.3}$$

A GGD is characterized by two parameters: the parameter α controls the 'shape' of the distribution and σ^2 controls its variance. A zero mean distribution is appropriate for modeling NLC distributions since they are (generally) symmetric. These parameters are commonly estimated using an efficient moment-matching based approach [74] [29].

4.1.2 Asymmetric Generalized Gaussian Distribution Model

Additionally, some of the normalized distributions derived from the feature maps are skewed, and are better modeled as following an asymmetric Generalized Gaussian distribution (AGGD) [118]. An AGGD with zero mode is given by:

$$f(x;\nu,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{\nu}{(\beta_l+\beta_r)\Gamma(1/\nu)} exp\left(-\left(\frac{-x}{\beta_l}\right)^{\nu}\right) & x<0 \\ \frac{\nu}{(\beta_l+\beta_r)\Gamma(1/\nu)} exp\left(-\left(\frac{x}{\beta_r}\right)^{\nu}\right) & x>0, \end{cases} \tag{4.4}$$

where

$$\beta_l = \sigma_l \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \tag{4.5}$$

$$\beta_r = \sigma_r \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}, \tag{4.6}$$

where η is given by:

$$\eta = (\beta_r - \beta_l)\,\frac{\Gamma(2/\nu)}{\Gamma(1/\nu)}. \tag{4.7}$$

An AGGD is characterized by four parameters: the parameter ν controls the 'shape' of the distribution, η is the mean of the distribution, and σ_l^2, σ_r^2 are scale parameters that control the spread on the left and right sides of the mode, respectively. The AGGD further generalizes the GGD [74] and subsumes it by allowing for asymmetry in the distribution. The skew of the distribution is a function of the left and right scale parameters. If $\sigma_l^2 = \sigma_r^2$, then the AGGD reduces to a GGD. All the parameters of the AGGD may be efficiently estimated using the moment-matching-based approach proposed in [118].

Although pristine images produce normalized coefficients that reliably follow a Gaussian distribution, this behavior is altered by the presence of image distortions. The model parameters, such as the shape and variance of either a GGD or an AGGD fit to the NLC maps of distorted images aptly capture this non-Gaussianity and

hence are extensively utilized in my work. Additionally, sample statistics such as kurtosis, skewness, and goodness of the GGD fit, have been empirically observed to also be predictive of perceived image quality and are also considered here. Thus, I deploy either a GGD or an AGGD to fit the empirical NLC distributions computed on different feature maps of each image encountered in any given data collection.

Images are naturally multi-scale, and distortions affect image structures across scales. Existing research on quality assessment has demonstrated that incorporating multi-scale information when assessing quality produces QA algorithms that perform better in terms of correlation with human perception [31, 40]. Hence, I extract these features from many of the feature maps at two scales - the original image scale, and at a reduced resolution (low pass filtered and downsampled by a factor of 2). It is possible that using more scales could be beneficial, but I did not find this to be the case on this large dataset, hence only report scores using two scales.

4.2 Feature Maps

My approach to feature map generation is decidedly a "Bag of Features" approach, as is highly popular in the development of a wide variety of computer vision algorithms that accomplish tasks such as object recognition [119, 120]. However, while my approach uses a large collection of highly heterogeneous features, as mentioned earlier, all of them either have a basis in current models of perceptual processing and/or perceptually relevant models of natural picture statistics, or are defined using perceptually-plausible parametric or sample statistic features computed on the empirical probability distributions (histograms) of simple biologically and/or statistically relevant image features.

I also deploy these kinds of features on a diverse variety of color space representations. Currently, our understanding of color image distortions is quite limited.

By using the "Bag of Features" approach on a variety of color representations, I aim to capture aspects of distortion perception that are possibly distributed over the different spaces. Figure 4.1 schematically describes some of the feature maps that are built into my model, while Fig. 4.2 shows the flow of statistical feature extraction from these feature maps. Further, I will use the images illustrated in Figure 4.3 (a) - (d) in the below sections to illustrate the proposed feature maps and the statistical variations that occur in the presence of distortions.

4.2.1 Luminance Feature Maps

Next I describe the feature maps derived from the luminance component of any image considered.

a. Luminance Map

There is considerable evidence that local center-surround excitatory-inhibitory processes occur at several types of retinal neurons [121, 122], thus providing a bandpass response to the visual signal's luminance. It is common to also model the local divisive normalization of these non-oriented bandpass retinal responses, as in [29].

Thus, given an $M \times N \times 3$ image I in RGB color space, its luminance component is first extracted, which I refer to as the *Luma* map. A normalized luminance coefficient (NLC) map as defined in Equation (2.2) is then computed on it by applying a divisive normalization operation on it [11]. A slight variation from the usual retinal "contrast signal" model is the use of divisive normalization by the standard deviation (as defined in Equation (2.4)) of the local responses rather than by the local mean response. The best-fitting GGD model to the empirical distribution of the NLC map is found [29]. Two parameters, (α, σ^2) are estimated and two sample statistics are computed (kurtosis, skewness) from the empirical distribution over two scales, yielding a total of 8 features. The features may be regarded as essential NSS

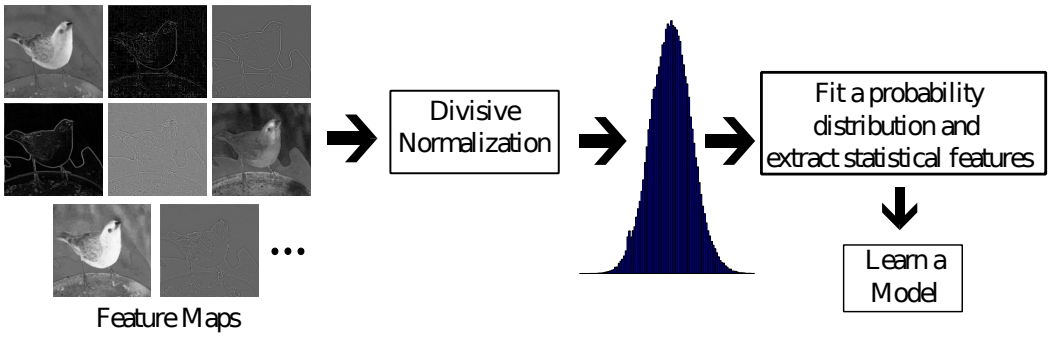


Figure 4.2: The proposed model processes a variety of perceptually relevant feature maps by modeling the distribution of their coefficients (divisively normalized in some cases) using either one of GGD (in real or complex domain), AGGD, or wrapped Cauchy distribution, and by extracting perceptually relevant statistical features that are used to train a quality predictor.

features related to classical models of retinal processing.

b. Neighboring Paired Products

The statistical relationships between neighborhood pixels of an NLC map are captured by computing four product maps that serve as simple estimates of local correlation. These four maps are defined at each coordinate (i, j) by taking the product of $NLC(i, j)$ with each of its directional neighbors $NLC(i, j + 1)$, $NLC(i + 1, j)$, $NLC(i + 1, j + 1)$, and $NLC(i + 1, j - 1)$. These maps have been shown to reliably obey an AGGD in the absence of distortion [29]. A total of 24 parameters (4 AGGD parameters per product map and two sample statistics - kurtosis, skewness) are computed. These features are computed on two scales yielding 48 additional features. These features use the same NSS/retinal model to account for local spatial correlations.

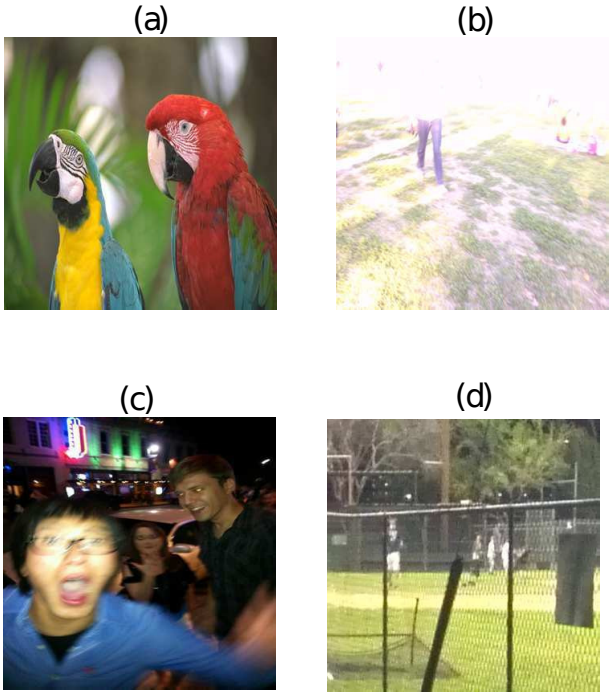


Figure 4.3: (a) A high-quality image and (b) - (d) a few distorted images from the LIVE Challenge Database [1, 2].

c. Sigma Map

The designers of existing NSS-based blind IQA models, have largely ignored the predictive power of the sigma field Equation (2.4) present in the classic Ruderman model. However, the sigma field of a pristine image also exhibits a regular structure which is disturbed by the presence of distortion. I extract the sample kurtosis, skewness, and the arithmetic mean of the sigma field at 2 scales to efficiently capture structural anomalies that may arise from distortion. While this feature map has not been used before for visual modeling, it derives from the same NSS/retinal model and is statistically regular.

d. Difference of Gaussian (DoG) of Sigma Map

Center-surround processes are known to occur at various stages of visual processing, including the multi-scale receptive fields of retinal ganglion cells [123]. A good model is the 2D difference of isotropic Gaussian filters [124, 125]:

$$DoG = \frac{1}{\sqrt{2\pi}} \left(\frac{1}{\sigma_1} e^{\frac{-(x^2+y^2)}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{\frac{-(x^2+y^2)}{2\sigma_2^2}} \right), \tag{4.8}$$

where $\sigma_2 = 1.5\sigma_1$. The value of σ_1 in my implementation was 1.16. The mean subtracted and divisively normalized coefficients of the DoG of the sigma field (obtained by applying Equation (2.4) on the DoG of the sigma field, denoted henceforth as $\mathbf{DoG_{sigma}}$) of the luminance map of a pristine image exhibits a regular structure that deviates in the presence of some kinds of distortion (Fig. 4.4(a)). Features that are useful for capturing a broad spectrum of distortion behavior include the estimated shape, standard deviation, sample skewness and kurtosis. The DoG of the sigma field can highlight conspicuous, 'stand-out' statistical features that may particularly affect the visibility of distortions.

I next extract the sigma field of $\mathbf{DoG_{sigma}}$ and denote its mean subtracted and divisively normalized coefficients as $\mathbf{DoG'_{sigma}}$. The sigma field of $\mathbf{DoG_{sigma}}$ is obtained by applying Equation (2.4) on $\mathbf{DoG_{sigma}}$. I found that $\mathbf{DoG'_{sigma}}$ also exhibit statistical regularities disrupted by the presence of distortions (Fig. 4.4(b)). The sample kurtosis and skewness of these normalized coefficients are part of the list of features that are fed to the regressor.

e. Laplacian of the Luminance Map

A Laplacian image is computed as the downsampled difference between an image and a low-pass filtered version of it. The Laplacian of the luminance map of a pristine image is well-modeled as AGGD, but this property is disrupted by image

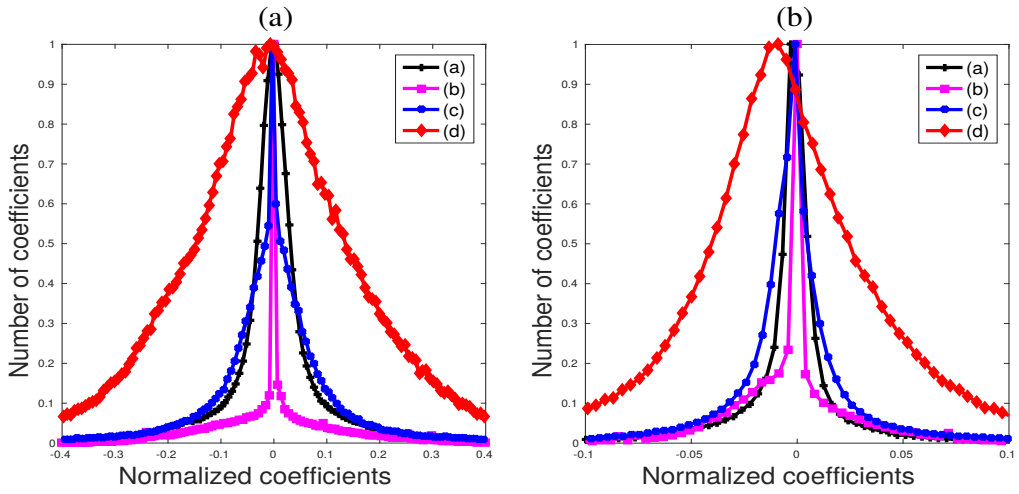


Figure 4.4: Histogram of normalized coefficients of a) DoG_{sigma} and (b) $DoG^{'}_{sigma}$ of the luminance components of Figures 4.3 (a) - (d).

distortions [126]. I therefore compute the Laplacian of each image's luminance map (*Luma*) and model it using an AGGD. This is also a bandpass retinal NSS model, but without normalization. The estimated model parameters $(\nu, \sigma_l^2, \sigma_r^2)$ of this fit are used as features along with this feature map's sample kurtosis and skewness.

f. Features extracted in the wavelet domain

The next set of feature maps are extracted from a complex steerable pyramid wavelet transform of an image's luminance map. This could also be accomplished using Gabor filters [127] but the steerable pyramid has been deployed quite successfully in the past on NSS-based problems [3, 30, 27, 45]. The features drawn from this decomposition are strongly multi-scale and multi-orientation, unlike the other features. C-DIIVINE [35] is a complex extension of the NSS-based DIIVINE IQA model [30] which uses a complex steerable pyramid. Features computed from it enable changes in local magnitude and phase statistics induced by distortions to be effectively captured. One of the underlying parametric probability models used by C-DIIVINE is

the wrapped Cauchy distribution. Given an image whose quality needs to be assessed, 82 statistical C-DIIVINE features are extracted on its luminance map using 3 scales and 6 orientations. These features are also used by the learner.

4.2.2 Chroma Feature Maps

Feature maps are also defined on the *Chroma* map defined in the perceptually relevant CIELAB color space of one luminance (L*) and two chrominance (a* and b*) components [128]. The coordinate L* of the CIELAB space represents color lightness, a* is its position relative to red/magenta and green, and b* is its position relative to yellow and blue. Moreover, the nonlinear relationships between L*, a*, and b* mimic the nonlinear responses of the L, M, and S cone cells in the retina and are designed to uniformly quantify perceptual color differences. *Chroma*, on the other hand, captures the perceived intensity of a specific color, and is defined as follows:

$$C^*_{ab} = \sqrt{a^{*2} + b^{*2}} \tag{4.9}$$

where a^* and b^* refer to the two chrominance components of any given image in the LAB color space. The chrominance channels contained in the chroma map are entropy-reduced representations similar to the responses of color-differencing retinal ganglion cells.

g. Chroma Map:

The mean subtracted and divisively normalized coefficients of the *Chroma* map Equation (4.9) of a pristine image follow a Gaussian-like distribution, which is perturbed by the presence of distortions (Fig. 4.5 (a)) and thus, a GGD model is apt to capture these statistical deviations. I extract two model parameters – shape and standard deviation and two sample statistics – kurtosis and skewness at two scales to serve as image features.

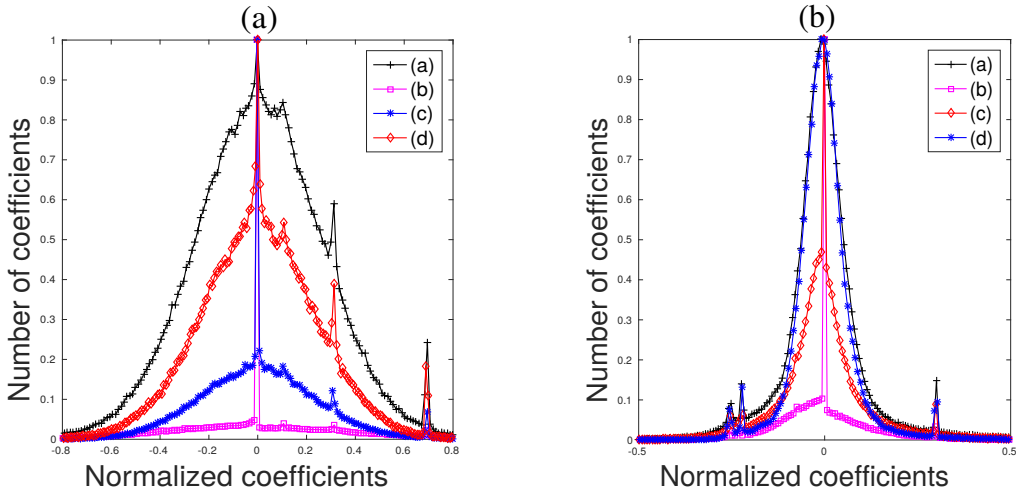


Figure 4.5: Histogram of normalized coefficients of (a) the *Chroma* map and (b) $Chroma_{sigma}$ of Fig. 4.3 (a) - (d).

h. Sigma field of the Chroma Map:

I next compute a sigma map (as defined in Equation (2.4)) of *Chroma* (henceforth referred to as $Chroma_{sigma}$). The mean subtracted and divisively normalized coefficients of $Chroma_{sigma}$ of pristine images also obey a unit Gaussian-like distribution which is violated in the presence of distortions (Fig. 4.5(b)). I again use a GGD to model these statistical deviations, estimate the model parameters (shape and standard deviation), and compute the sample kurtosis and skewness at two scales. All of these are used as features deployed by the learner.

Furthermore, as was done on the luminance component's sigma field in the above section, I compute the sample mean, kurtosis, and skewness of $Chroma_{sigma}$. I also process the normalized coefficients of the *Chroma* map and generate four neighboring pair product maps, the Laplacian, $\mathbf{DOG_{sigma}}$, and $\mathbf{DOG'_{sigma}}$ maps, and extract the model parameters and sample statistics from them. C-DIIVINE features on the *Chroma* map of each image are also extracted to be used later by the learner.

4.2.3 LMS Feature Maps

The LMS color space mimics the responses of the three types of cones in the human retina. Hurvich and Jameson [129] suggested that the retina contains three types of cone photoreceptors, selectively sensitive to different color mixtures of Long, Medium, and Short wavelengths. They also postulated that each photo-receptor pair has two *physiologically opponent* color members: *red-green, yellow-blue*, in addition to an *achromatic white-black*. Later, Ruderman *et al.* [130] later experimentally gathered cone response statistics and found robust orthogonal decorrelation of the (logarithmic) data along three principal axes, corresponding to one achromatic ($\hat{l}$) and two chromatic-opponent responses (RG and BY).

Denoting L, M, and S as the three components of LMS color space, the three chromatic-opponent axes are:

$$\hat{l} = \frac{1}{\sqrt{3}}\left(\hat{L} + \hat{M} + \hat{S}\right), \tag{4.10}$$

$$BY = \frac{1}{\sqrt{6}}\left(\hat{L} + \hat{M} - 2\hat{S}\right), \tag{4.11}$$

$$RG = \frac{1}{\sqrt{2}}\left(\hat{L} - \hat{M}\right), \tag{4.12}$$

where $\hat{L}, \hat{M}$, and $\hat{S}$ are the NLCs Equation (2.2) of the logarithmic signals of the L, M, and S components respectively, i.e.,

$$\hat{L}(i,j) = \frac{logL(i,j) - \mu_L(i,j)}{\sigma_L(i,j) + 1}, \tag{4.13}$$

where $\mu_L(i,j)$ is the mean and $\sigma_L(i,j)$ is the standard deviation of $logL$, similar to those defined in Equations (2.3) and (2.4) for L. $\hat{M}(i,j)$ and $\hat{S}(i,j)$ are defined in the same manner as Equation (4.13) from $logM(i,j)$ and $logS(i,j)$ respectively.

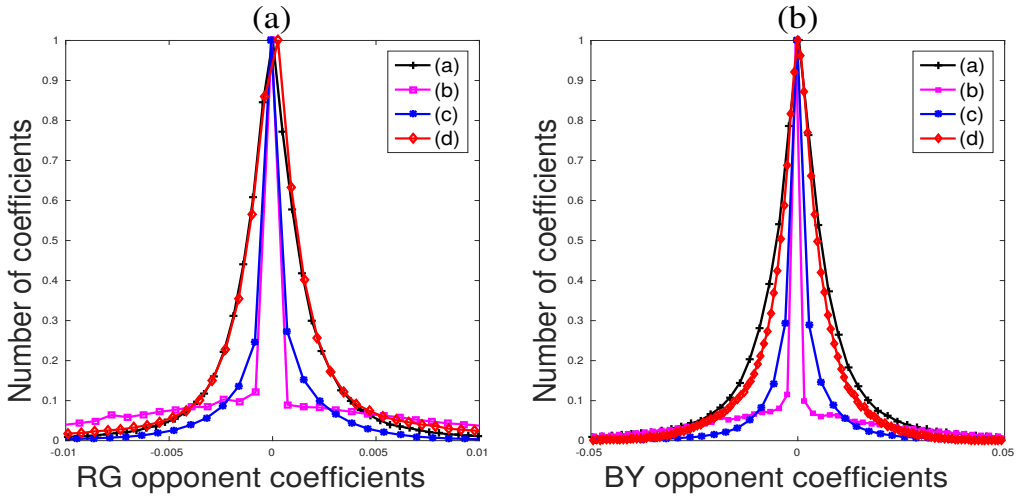


Figure 4.6: Histogram of color opponent maps of (a) red-green channel (RG). (b) blue-yellow channel (BY).

i. Blue-Yellow (BY) and Red-Green (RG) color-opponent maps:

The marginal distributions of image data projected along each opponent axis follow a Gaussian distribution (Fig. 4.6). In the presence of distortion, this statistical regularity is perturbed along all three axes Equations (4.10) - (4.12). By projecting each image along the two color opponent axes RG and BY, then fitting them with an AGGD model, I captured additional distortion-sensitive features in the form of the model parameters $(\nu, \sigma_l^2, \sigma_r^2)$. I also compute the sample kurtosis and skewness of the color opponent maps RG and BY.

j. M and S Channel Maps:

After transforming an image into LMS color space, the M and S components are processed as in the previous section and their normalized coefficients are modeled along with their sigma field. I also generate the Laplacian, $\mathbf{DOG_{sigma}}$, and $\mathbf{DOG'_{sigma}}$ feature maps from both M and S channels, and extract model parameters and sample statistics from them. C-DIIVINE features at 3 scales and 6 orientations are also

Table 4.1: Summary of different feature maps and the features extracted from them in all three color spaces. The last three columns refer to feature counts from each feature map in each color space and the number in their headings refer to the total number of features in those color spaces.

Feature map	**Color Channels or Spaces**	**Model parameters (derived from GGD (real and complex), AGGD, wrapped Cauchy)**	**Sample statistics**	**Luma** (155)	**Chroma** (163)	**LMS** (240)
Yellow color map and its sigma field	RGB		goodness of GGD fit	2	0	0
Red-Green Equation (4.12) and Blue-Yellow Equation (4.11) color opponent maps	LMS	shape, left and right standard deviation	kurtosis, skewness	0	0	10
Neighboring pair product map	Luminance, Chroma (from LAB space)	shape, mean, left and right variance	kurtosis, skewness	48	48	0
Debiased and normalized coefficients	Luminance, Chroma (from LAB space), M and S (from LMS space)	shape, variance	kurtosis, skewness	8	8	16
Sigma field	Luminance, Chroma (from LAB space), M and S (from LMS space)	shape, variance	kurtosis, skewness, mean	6	14	28
DOG_{sigma} and DOG'_{sigma}	Luminance, Chroma (from LAB space), M and S (from LMS space)	shape, standard deviation	kurtosis, skewness	6	6	12
First Laplacian	Luminance, Chroma (from LAB space), M and S (from LMS space)	shape, left and right standard deviations	kurtosis, skewness	5	5	10
Complex steerable decomposition	Luminance, Chroma (from LAB space), M and S (from LMS space)	Model parameters from magnitude and phase coefficients (See [35])	-	82	82	164

computed on both the channel maps and added to the final list of features.

4.2.4 Statistics from the Hue and Saturation Components

I also extract the hue and saturation components of every image in the HSI (hue, saturation, intensity) color space and compute the arithmetic mean and standard deviation of these two components. These four features are also added to the list of features to be considered by the learner. I examined the bandpass distributions of the HS components, but found that they were redundant with respect to those of other color channels in regards to distortion. Thus, in order to avoid redundancy in my final feature collection, I exclude these from the final feature list.

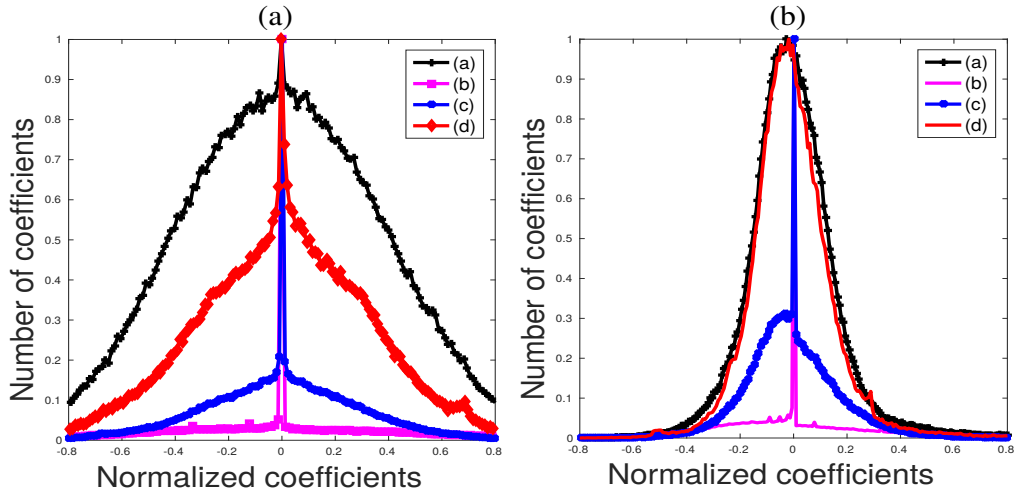


Figure 4.7: Histogram of the normalized coefficients of (a) Y and (b) Y_{sigma} of Fig. 4.3 (a) - (d).

4.2.5 Yellow Color Channel Map

Similar to the design of saliency-related color channels in [131], I constructed a yellow color channel map of an RGB image I, which is defined as follows:

$$Y = \frac{R+G}{2} - \frac{|R-G|}{2} - B, \tag{4.14}$$

where R, G, and B refer to the red, green, and blue channels respectively. My motivation for using the yellow channel is simply to provide the learner with direct yellow-light information rather than just B-Y color opponency, which might be relevant to distortion perception, especially on sunlit scenes.

Divisive normalization of $\mathbf{Y}$ computed on a pristine image yields coefficients which, as illustrated in Fig. 4.7(a), exhibit Gaussian-like behavior on good quality images. Furthermore, the normalized coefficients of the sigma map of $\mathbf{Y}$ (denoted henceforth as $\mathbf{Y_{sigma}}$) also display Gaussian behavior on pristine images (Fig. 4.7(b)). This behavior is often not observed on distorted images. Thus, the good-

ness of generalized Gaussian fit of both the normalized coefficients of $\mathbf{Y}$ and $\mathbf{Y_{sigma}}$ at the original scale of the image are also extracted and added as features used in this model. As discussed in the next section, features drawn from the yellow color channel map were able to efficiently capture a few distortions that were not captured by the luminance component alone.

4.3 Advantages of the proposed Feature Maps

As an example to illustrate the advantages of the proposed feature maps, consider the four images presented in Fig. 2.10. To reiterate, Fig. 2.10(a) is a pristine image from the legacy LIVE Image Quality Database [3] while Fig. 2.10 (b) and (c) are JPEG2000 compression and additive white noise distortions (respectively) artificially applied to Fig. 2.10 (a). On the other hand, Fig. 2.10 (d) is a blurry image distorted by low-light noise and presumably compression, drawn from the LIVE In the Wild Image Quality Challenge Database [1].

I processed these four images using three different operations - (a) the mean subtraction, divisive normalization operation used in [29] on singly distorted images, (b) the yellow color channel map Equation (4.14), and (c) the DOG_{sigma} map on the luminance map as defined in earlier sections. It may be observed that, though the histograms of the singly distorted images differ greatly from those of the pristine image in Fig. 4.8(a), the distribution of an authentically distorted image containing noise, blur, and compression artifacts closely resembles the distribution of the pristine image. However, when the normalized coefficients of the proposed yellow color channel and the DOG_{sigma} feature maps are observed in Fig. 4.8(b)-(c), it is clear that these distributions are useful for distinguishing between the pristine image and both singly and authentically distorted images. I have observed the usefulness of all of the proposed feature maps on a large and comprehensive collection of images contained in the LIVE Challenge Database.

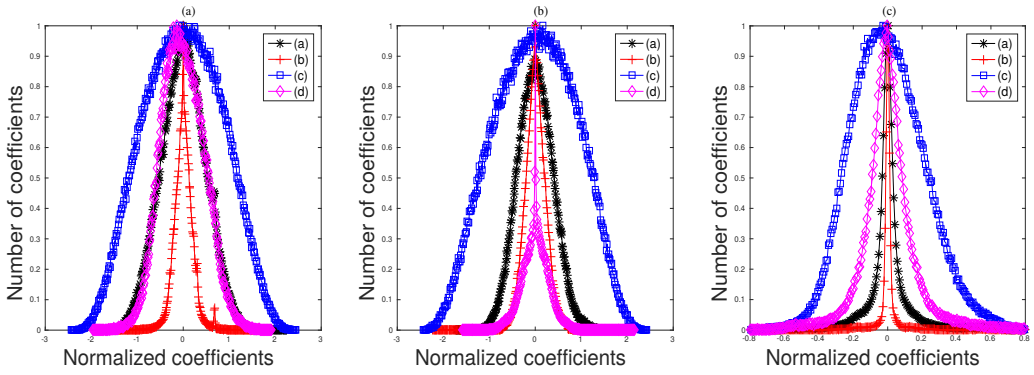


Figure 4.8: (a) Histogram of the normalized coefficients of the images in Figures 2.10(a) - (d) when processed using (a) BRISQUE-like normalization defined in Equation (2.2), (b) yellow color channel maps Equation (4.14), and (c) DoG_sigma computed on the luminance map. Notice how for the authentically distorted image Fig. 2.10 (d), the corresponding histogram in (a) resembles that of a pristine image. But in the case of the two feature maps - yellow color map and DoG_{sigma}, the histograms of pristine vs. authentically distorted images vary. (Best viewed in color).

I have thus far described a series of statistical features that I extract from a set of feature maps and also how each of these statistics are affected by the presence of image distortions (summarized in Table 4.1). I note that in Table 4.1, I did not show the 4 features extracted in the HSI space due to space constraints. Also, the number of features shown in the LMS column refer to the sum of the number of features extracted on the M and S channel maps. Predicting the perceptual severity of authentic distortions is recondite, and a 'bag of features' approach is a powerful way to approach the problem.

4.4 Regression

These perceptually relevant image features, along with the corresponding real-valued MOS of the training set, are used to train a support vector regressor (SVR). SVR is the most common tool for learning a non-linear mapping between image features

and a single label (quality score) among IQA and VQA algorithms [29, 35, 30, 31, 132, 133]. Given an input image (represented by a feature vector), SVM maps the high-dimensional feature vector into a visual quality score [134, 135]. While the database is large, it is not large enough to motivate deep learning methods. The SVM classifier and regressor is widely used in many disciplines due to its high accuracy, ability to deal with high-dimensional data, and flexibility in modeling diverse sources of data [134].

In this algorithm, I used an SVR with a radial basis kernel function. Following this, given any test image's image features as input to the trained SVR, a final quality score may be predicted. The optimal model parameters of the learner were found via cross-validation. The choice of the model parameters was driven by the obvious aim of minimizing the learner's fitting error to the validation data.

4.5 Experiments

As described, FRIQUEE combines a large, diverse collection of perceptually relevant statistical features across multiple domains, which are used to train a regressor that is able to conduct blind image quality prediction. Variations called FRIQUEE-Luma, FRIQUEE-Chroma, FRIQUEE-LMS, and FRIQUEE-ALL are developed according to the subset of overall features considered. Thus FRIQUEE-Luma uses feature maps a. - f., FRIQUEE-Chroma uses feature maps g. - h., FRIQUEE-LMS uses feature maps i.-j., while FRIQUEE-ALL uses all feature maps as well as the two HSI color space feature maps and the yellow color channel map.

In all of the experiments I describe below, the model (initialized with the optimal parameters) was trained from scratch on a random sample of 80% training images and tested on the remaining non-overlapping 20% test data. To mitigate any bias due to the division of data, the process of randomly splitting the dataset was repeated 50 times. Spearman's rank ordered correlation coefficient (SROCC) and

Pearson's correlation coefficient (PLCC) between the predicted and the ground truth quality scores were computed at the end of each iteration. The median correlation over these 50 iterations is reported. A higher value of each of these metrics indicates better performance both in terms of correlation with human opinion as well as the performance of the learner. I also report the outlier ratio (OR) [136] which is the fraction of the number of predictions lying outside the range of $\pm$ 2 times the standard deviations of the ground truth MOS. A lower value of the outlier ratio indicates better performance of a given model.

4.5.1 Comparing Different IQA Techniques

I trained several other well-known NR IQA models (whose code was publicly available) on the LIVE In the Wild Image Quality Challenge Database, using identical train/test settings and the same cross-validation procedure over multiple random trials. In the case of DIIVINE [30] and C-DIIVINE [35] which are two-step models, I skipped the first step of identifying the probability of an image belonging to one of the five distortion classes present in the legacy LIVE IQA Database as it doesn't apply to the newly proposed database. Instead, after extracting the features as proposed in their work, I learned a regressor on the training data. An SVR with a RBF (radial basis function) kernel was trained using FRIQUEE features and I denote this model as FRIQUEE-ALL. The median and the standard deviations of the correlations and the mean of the outlier ratios across the 50 train-test iterations is reported in Table 4.2. I note that the NIQE [33] score is a measure of how far a given image is from 'naturalness,' which is different from the subjective MOS values. Since it is not trained on MOS values, I do not compute an outlier ratio on the predicted NIQE scores. I conclude from this table that the performance of the proposed model on unseen test data is significantly better than that of current top-performing state-of-the-art NR IQA models on the LIVE Challenge Database

[1, 2].

To justify the design choice of an SVR with an RBF kernel, I also trained a linear SVR (FRIQUEE-LSVR) on FRIQUEE features extracted from the images in the LIVE Challenge Database. The training was performed under the same setting (on 50 random train/test splits). The median correlations across 50 iterations are reported in Table 4.2. From this table I conclude that the performance of FRIQUEE-ALL is better than the other learners. Also, comparing the median correlation scores of FRIQUEE-ALL with those of top-performing IQA models such as C-DIIVINE, BRISQUE, and DIIVINE, all of which also use an SVR as a learning engine, reveals that the perceptually-driven FRIQUEE NSS features perform better than the features designed in the other top-performing IQA models.

The high internal statistical consistency and reliability of the subjective scores gathered in the crowdsource study make it possible to consider the MOS values obtained from the online study as ground truth quality scores of the images [2]. Moreover, the poor correlation scores reported by most algorithms suggests that the LIVE Challenge Database is a difficult test of the generalizability of those models.

4.5.2 Statistical Significance and Hypothesis Testing

Although there exist apparent differences in the median correlations between the different algorithms (Table 4.2), I evaluated the statistical significance of the performance of each of the algorithms considered. Thus, I performed hypothesis testing based on the paired t-test [137] on the 50 SROCC values obtained from the 50 train-test trials. The results are tabulated in Table 4.3. The null hypothesis is that the mean of the two paired samples is equal, i.e., *the mean correlation for the (row) algorithm is equal to the mean correlation for the (column) algorithm with a confidence of 95%*. The alternative hypothesis is that *the mean correlation of the row algo-*

Table 4.2: Median PLCC and SROCC, and mean OR of several no-reference IQA metrics across 50 train-test combinations on the LIVE Challenge Database [1, 2]. FRIQUEE-ALL refers to the scenario where the proposed learning engine, i.e.,SVR with an RBF was used. The IQA algorithm that achieves top-performance is indicated in bold font.

	PLCC	SROCC	OR
FRIQUEE-ALL	$\mathbf{0.72 \pm 0.04}$	$\mathbf{0.72 \pm 0.04}$	**0.04**
FRIQUEE-LSVR	0.65 ± 0.04	0.62 ± 0.04	0.04
BRISQUE [29]	0.61 ± 0.06	0.58 ± 0.05	0.07
DIIVINE [30]	0.59 ± 0.05	0.56 ± 0.05	0.06
BLIINDS-II [31]	0.45 ± 0.05	0.40 ± 0.05	0.09
NIQE [33]	0.48 ± 0.05	0.42 ± 0.05	–
C-DIIVINE [35]	0.66 ± 0.04	0.63 ± 0.04	0.05

rithm is greater than or lesser than the mean correlation of the column algorithm. A value of '1' in the table indicates that the row algorithm is statically superior to the column algorithm, whereas a '-1' indicates that the row is statistically worse than the column. A value of '0' indicates that the row and column are statistically indistinguishable (or equivalent), i.e., I could not reject the null hypothesis at the 95% confidence level.

From Table 4.3 I conclude that FRIQUEE-ALL is statistically superior to all of the no-reference algorithms that I evaluated, when trained and tested on the LIVE Challenge Database.

4.5.3 Contribution of Features from Each Color Space

I next evaluated the performance of FRIQUEE-Luma, FRIQUEE-Chroma, and FRIQUEE-LMS. I trained three separate SVRs with features extracted from each color space serving as an input to each SVR and report the median correlation values across 50 random train/test splits in Table 4.4. These values justify my choice of

Table 4.3: Results of the paired one-sided t-test performed between SROCC values generated by different measures. '1,' '0,' '-1' indicate that the NR IQA algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column.

	DIIVINE	BRISQUE	NIQE	C-DIIVINE	BLIINDS-II	FRIQUEE-LSVR	FRIQUEE-ALL
DIIVINE	0	-1	1	-1	1	-1	-1
BRISQUE	1	0	1	-1	1	-1	-1
NIQE	-1	-1	0	-1	0	-1	-1
C-DIIVINE	1	1	1	0	1	0	-1
BLIINDS-II	-1	-1	0	-1	0	-1	-1
FRIQUEE-LSVR	1	1	1	0	1	0	-1
FRIQUEE-ALL	1	1	1	1	1	1	0

Table 4.4: Median PLCC and Median SROCC across 50 train-test combinations on [1, 2] when FRIQUEE features from each color space were independently used to train an SVR.

	PLCC	SROCC
FRIQUEE-Luma	0.64 ± 0.04	0.61 ± 0.04
FRIQUEE-LMS	0.63 ± 0.04	0.60 ± 0.04
FRIQUEE-Chroma	0.36 ± 0.05	0.34 ± 0.05

different color spaces, all of which play a significant role in enhancing image quality prediction.

4.5.4 Contribution of Different Feature Maps

To better understand the relationship between the proposed feature set and perceptual quality, I trained separate learners (SVR with radial basis kernel functions) on the statistical features extracted from each feature map on 50 random, non-overlapping train and test splits. I report the median Spearman rank ordered correlation scores over these 50 iterations in Fig. 4.9. This plot illustrates the degree to which each of these features accurately predict perceived quality, while also justifying the choice of the feature set. I want to point out that I included the Yellow Map under FRIQUEE-Luma in Fig. 4.9 purely for the purpose of illustration. It is

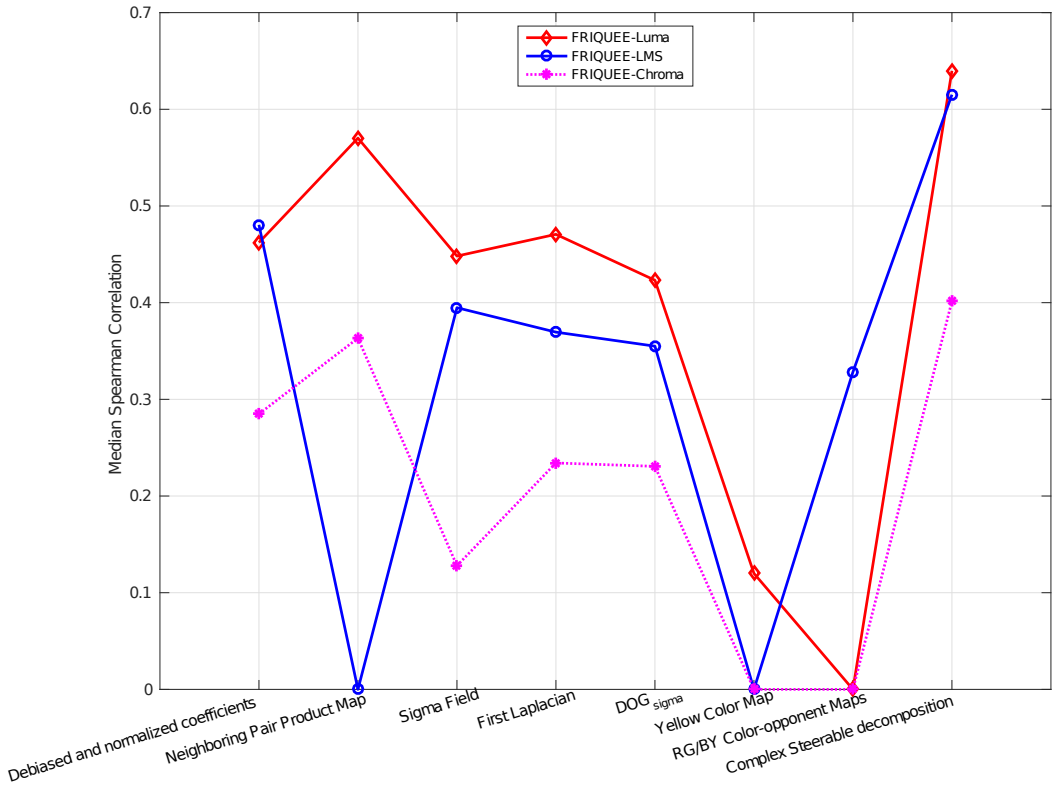


Figure 4.9: Contribution of different *types* of features that are extracted in different color spaces. A correlation of 0 in a color space indicates that that specific feature map was not extracted in that color space.

not extracted from the luminance component of an image but is a color feature as described earlier.

4.5.5 Evaluating the Robustness of Different IQA Techniques

The goal of this experiment was to study the efficacy of training IQA models on the synthetically distorted images contained in current benchmark databases relative to training on authentically distorted images. Some of the current top-performing IQA learning models have been made publicly available (i.e., the model parameter

Table 4.5: Median PLCC and Median SROCC across 50 train-test combinations of a few NR-IQA models on [1, 2] when models trained on the LIVE IQA Database are used to predict the quality of the images from the LIVE Challenge Database. The IQA algorithm that achieves top-performance is indicated in bold font.

	PLCC	SROCC
FRIQUEE-ALL	**0.6289±0.0425**	**0.6303±0.0405**
BRISQUE [29]	0.3296 ± 0.0505	0.2650 ± 0.0505
DIIVINE [30]	0.3667 ± 0.0504	0.3328 ± 0.0536
BLIINDS-II [31]	0.1791 ± 0.0713	0.1259 ± 0.0704
C-DIIVINE [35]	0.4705 ± 0.0549	0.4589 ± 0.0515

values used by their SVRs) after being trained on the images on the legacy LIVE IQA Database. I sought to understand the performance of these publicly available models when they are used in real-world scenarios, to predict the quality of real-world images captured using mobile devices. I used the publicly available model BRISQUE [29] trained on the legacy LIVE Database. With regards to the other blind algorithms, I extracted image features from each image in the LIVE IQA Database following the same procedure as was originally presented in their work and separately trained SVRs for each model on these image features.

Next, I used the 50 randomly generated test splits and evaluated each learned engine (trained on the LIVE IQA Database) on the 50 test splits. I report the median of the correlations of the predicted scores with human judgments of visual quality across the 50 test splits in Table 4.5. This analysis provides an idea of how well state-of-the-art quality predictors generalize with respect to image content and real-world distortions. As can be seen from the results reported in Table 4.5, although FRIQUEE performed better than all of the algorithms, the performance of all the models suffer when they are trained only on images containing synthetic, inauthentic distortions.

Table 4.6: Performance on legacy LIVE IQA Database [3]. Italics indicate NR-IQA models. -NA- indicates data not reported in the corresponding paper.

	SROCC	PLCC
PSNR [138]	0.8636	0.8592
SSIM [39]	0.9129	0.9066
MS-SSIM [40]	0.9535	0.9511
CBIQ [36]	*0.8954*	*0.8955*
LBIQ [73]	*0.9063*	*0.9087*
DIIVINE [30]	*0.9250*	*0.9270*
BLIINDS-II [31]	*0.9124*	*0.9164*
BRISQUE [29]	0.9395	*0.9424*
NIQE [33]	0.9135	0.9147
C-DIIVINE [35]	0.9444	0.9474
FRIQUEE-ALL	***0.9477*** ±0.0250	***0.9620*** ±0.0223

4.5.6 Evaluating IQA models on Legacy LIVE Database

I next compared the performance of the proposed model against several other top-performing blind IQA models on the older standard benchmark LIVE IQA Database [3]. Regarding FRIQUEE-ALL, 560 features were extracted on all the images of the LIVE IQA Database and the image data was divided into training and test subsets, with no overlap in content. This process was repeated 1000 times and I report the median correlation values in Table 4.6. With regards to the other models, I report the median correlation scores as reported in their papers. I note that [34] report an SROCC value of 0.9650 on the legacy LIVE IQA Database, but this result is not verifiable since the authors do not make the code publicly available. Since I cannot validate their claim, I do not include it in Table 4.6.

Comparing the correlation scores reported in Table 4.2 with those in Table 4.6, I observe that several other blind IQA models are not robust to authentic dis-

Table 4.7: Median PLCC and Median SROCC across 100 train-test combinations of a few NR-IQA models on LIVE-Multiply Database - Part I [4]. The IQA algorithm that achieves top-performance is indicated in bold font.

	PLCC	SROCC
FRIQUEE-ALL	**0.9667**	**0.9591**
BRISQUE [29]	0.9391	0.9238
DIIVINE [30]	0.9424	0.9327
NIQE [33]	0.9075	0.8614
C-DIIVINE [35]	0.9336	0.9179

Table 4.8: Median PLCC and Median SROCC across 100 train-test combinations of a few NR-IQA models on LIVE-Multiply Database - Part II [4]. The IQA algorithm that achieves top-performance is indicated in bold font.

	PLCC	SROCC
FRIQUEE-ALL	**0.9664**	**0.9632**
BRISQUE [29]	0.9070	0.8748
DIIVINE [30]	0.8956	0.8677
NIQE [33]	0.8316	0.7762
C-DIIVINE [35]	0.8837	0.8772

tortions, since while they achieve superior performance on the legacy LIVE IQA Database, they fail to accurately predict the quality of authentically distorted images. On the other hand, it may be observed that FRIQUEE not only performs well on the LIVE Challenge Database (Table 4.2), but also competes very favorably with all the other blind IQA models as well as with full-reference IQA models on the legacy LIVE Database. It reaches and exceeds the 'saturation level' of performance achieved on this long-standing synthetic distortion database by the tested prior models. This supports my contention that a combination of semantically rich, perceptually informative image features feeding a highly discriminative learning model

Table 4.9: Median PLCC and Median SROCC across 100 train-test combinations of a few NR-IQA models on TID2013 Database [5].. The IQA algorithm that achieves top-performance is indicated in bold font.

	PLCC	SROCC
FRIQUEE-ALL	**0.9287**	**0.9138**
BRISQUE [29]	0.7781	0.7515
DIIVINE [30]	0.8066	0.7644
NIQE [33]	0.3592	0.3137
C-DIIVINE [35]	0.7319	0.6602

Table 4.10: Median PLCC and Median SROCC across 100 train-test combinations of a few NR-IQA models on CSIQ Database [6]. The IQA algorithm that achieves top-performance is indicated in bold font.

	PLCC	SROCC
FRIQUEE-ALL	**0.9622**	**0.9627**
BRISQUE [29]	0.8926	0.8823
DIIVINE [30]	0.9171	0.9282
NIQE [33]	0.6943	0.6142
C-DIIVINE [35]	0.8660	0.8611

is a powerful way to automatically predict the perceptual quality of images afflicted by both authentic and synthetic distortions.

4.5.7 Evaluating IQA models on other legacy databases

Although my primary focus was to evaluate the performance of the proposed algorithm on the LIVE In the Wild Challenge Database (since I wanted to benchmark the superior performance of FRIQUEE on authentically distorted images), I understand that some readers may find performance on the legacy databases to be relevant. Therefore, I evaluated FRIQUEE and a few other top-performing NR

IQA algorithms on other legacy databases such as TID2013 [5] and CSIQ [6] both of which contain single, synthetic distortions and LIVE-Multiply Database [4], which contains Gaussian blur followed by JPEG compression distortions (in Part I) and Gaussian blur followed by additive white noise distortions (in Part II). The images in all of these datasets were divided into non-overlapping training and test sets and this process was repeated 100 times. For each IQA algorithm on every database, optimal model parameters were chosen for an SVR with a radial basis kernel while training a model. In Tables 4.7 - 4.10, I report the median correlation values between ground truth and predicted quality scores across 100 iterations on all these databases.

Comparing the correlation scores, it may be observed that FRIQUEE features perform better than the features designed in all the other top-performing IQA models on synthetic distortions modeled in [5, 6, 4].

4.6 Conclusion

In this chapter, I have described a first effort towards the design of blind IQA models that are capable of predicting the perceptual quality of images corrupted by complex mixtures of authentic distortions. Its success encourages the feasibility of adapting the proposed model for application to real-world problems such as perceptual optimization of digital camera capture, perceptual image enhancement, and so on.

Chapter 5

A Subjective Study of Stalling Events in Mobile Streaming Videos

As mentioned in Section 2.4, over-the-top mobile video streaming is invariably influenced by volatile network conditions which cause playback interruptions (stalling events), thereby impairing users' quality of experience (QoE). Numerous subjective studies have been conducted in the past with an aim of better understanding the effects of volatile networks on an end user's QoE. However, all existing databases have certain limitations as described in Section 2.4.2 and do not support subjective and objective instantaneous QoE assessment of videos in-the-wild.

In this chapter, I present the details of a subjective study that I recently conducted on a novel video collection to address the limitations of all existing video QoE datasets. This chapter is organized as follows:

1. I first provide details of the new **LIVE Mobile Stall Video Database-II**, which consists of 174 videos that model 26 different patterns of rebuffering

events and startup delays. I also describe the distribution of stalling patterns, selection of reference videos, and the process of generating the 174 distorted videos in Section 5.1.

2. I next discuss the subjective study I conducted to gather ground-truth human opinion scores on these videos. I detail the set-up, the subjects, and the testing methodology (Section 5.2).

3. I present the results of my thorough analysis of the subjects' continuous-time subjective behavior to better understand the temporal variations in the perceived QoE due to the influence of factors such as stall positions, number of stalls, length of the stalls, and varied video content (Section 5.4).

5.1 Construction of the database

A few previous studies have gathered network and client-side media analytics by capturing the network traces of viewers streaming videos from services such as Akamai [96]. The authors of [98] manually varied the client-side network bandwidth in their study set-up, thereby causing bottlenecks in the network and stalling events in streaming videos. Videos with these types of network-induced stalling events were then viewed and rated by the subjects. Though these studies have provided useful insights on viewers' QoE, such uncontrolled network settings can often lead to non-reproducible stalling patterns. Videos and subjective data captured in this manner, with no preset goals and an insufficient number of subjective scores per distortion pattern, limits researchers' and system designers' abilities to dissect the effects of specific stall types on an end user's QoE.

Thus, in this dissertation, I chose to first obtain a small set of high-quality videos, into which I systematically inserted a wide variety of predefined—yet realistic—stalling patterns in a controlled manner, thereby generating distorted video content.

Figure 5.1: Sample frames of the reference video contents contained in the LIVE Mobile Stall Video Database-II.

To be able to evaluate the influence of different stall parameters on subjective behavior, the new database was designed to meet the following requirements:

1. The reference video content used to generate the distorted videos should be interesting and representative of what a typical viewer may stream on a mobile device.

2. The distortion severities and the stall pattern parameter settings should model a wide range of realistic distortions.

3. The distorted video contents should model adequate perceptual separability, i.e., they should reasonably span the quality spectrum, from low quality to high quality. Otherwise, a large number of videos may cluster too tightly at high and/or low quality, making it difficult to distinguish the performances of

different quality assessment models.

4. There should be an overlap of video content across different distortion patterns, enabling us to analyze the interplay of video content and different stall patterns on user's QoE.

5. There should be a reasonable number of data samples per distortion pattern, enabling us to reasonably analyze the influence of these patterns on subjective behavior.

6. The videos should not contain any post-capture distortions (such as compression artifacts or frame drops), since I wish to focus exclusively on network-induced distortions.

5.1.1 Source Sequences

I selected 24 High Definition (HD) creative commons licensed videos (with audio) from YouTube and Vimeo. These public-domain videos have different original resolutions: 1280×720, 1280×640, 480×360, 484×360, 490×360, 540×360, and 640×360. All videos are of 30 fps. Any visual distortions due to aliasing were deemed minimal or invisible. In order to focus exclusively on network impairments, I excluded any jittered or delayed videos, and thus each of the 24 selected video sequences contained minimal spatial distortions or abrupt camera shakes. From each of these video sequences, I chose a video segment that was semantically and temporally coherent and long enough to be meaningful on its own. The lengths (after adding rebuffering impairments) of these video segments range between 29 and 134 seconds. Though streaming services typically deliver longer video sequences, in order to test a variety of content and stall patterns while maintaining reasonable study session durations, I had to limit the sequence lengths.

In order to understand the impact of video content type in the presence of

Table 5.1: Number of videos classified into broad content categories.

Sports	Talk shows / Documentaries	Music	Advertisements	Newscasts
9	8	2	3	2

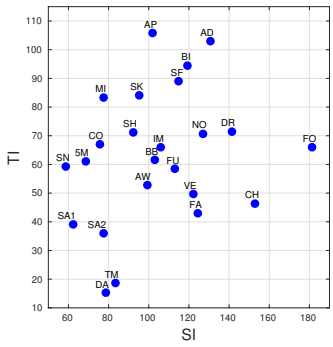


Figure 5.2: Spatial Information (SI) against Temporal Information (TI) for the 24 video contents in the database.

network delays on QoE, I selected video sequences of varied content categories that a typical video viewer is likely to encounter on the Internet. The 24 video contents were categorized into five broad categories as detailed in Table 5.1.

As noted in [139], measuring the amount of spatial and temporal information or activity in the reference videos can help to broadly characterize their span in the spatio-temporal plane. An appropriate set of test scenes should span a wide range of spatial and temporal information to be representative of the wide variety of content typically viewed over the Internet. Let F_n denote the luminance component of a video frame at instant n, and let (i, j) denote the spatial coordinates of within the frame. A frame filtered with the spatial Sobel operator is denoted as $Sobel(F_n)$. A frame difference operation can be defined as $M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$. As formulated in [139], the spatial perceptual information (SI) and temporal perceptual

information (TI) measurements are given by

$$SI = \max_{time}\Big\{ std_{space}\Big[Sobel(F_n(i,j))\Big]\Big\} \tag{5.1}$$

$$TI = \max_{time}\Big\{ std_{space}\Big[M_n(i,j)\Big]\Big\}, \tag{5.2}$$

where $\max_{time}$ denotes the maximum value computed over all the frames and std_{space} denotes the standard deviation over all the pixels of a given image (F_n or M_n). As shown in Fig. 5.2, the reference video content of the database widely spans the spatio-temporal space.

5.1.2 Distortion Patterns

The studies in [98] and [8] model scenarios where network congestion occurs at a constant rate, leading to periodic stalling patterns wherein each stall is of a fixed duration. This, however, is not realistic, as stall patterns could occur arbitrarily.

I chose to construct different stalling patterns by varying four defining features: start-up delay length, stall lengths, stall positions, and the number of stalls. Table 5.2 provides information about the range of values these four parameters take for different stalling patterns. Figure 5.3 illustrates the different positions in a video sequence where stalls were placed. Different settings for any single stall pattern parameter were not combined, i.e., if the stall position for a given video is chosen as beginning (B), and the stall length was chosen as medium, then all the stalls would be of medium length and occur only in the first half of that video. By varying the four stall pattern parameters, I designed a comprehensive set of distorted videos with the following constraints in mind:

1. Videos without any rebuffering should also be presented to subjects as they would serve as a baseline.

2. To understand the impact of start-up delays, a subset of videos should only contain start-up delay events (of varied lengths), with no additional rebuffering

events.

3. To understand the impact of inherent memory biases on a user's overall QoE, a subset of videos should have stall events of varied lengths in the latter half of the video.

4. To understand the recency, or the hysteresis effect [140] on quality perception, a subset of videos should have stall events in the beginning and middle regions of a video.

5. In order to model a variety of real-world network bandwidth capacities, a subset of videos should have many (or longer) stalls while a disjoint subset of videos should have very limited stalling events.

6. To tease out the most dominant factor impacting viewer QoE, the value of each stall parameter should be sufficiently varied within the set of modeled distortion patterns.

The 26 unique stall patterns designed with the above constraints are listed in Table 5.3. Notice that a given stall pattern (defined in terms of length and number of stalls) was introduced at multiple positions in a video (denoted by x). This was done to help us understand the influence of the position of stalling events on QoE. Further, since start-up delays are a very common and frequently experienced phenomenon, all of the distorted videos in the collection have start-up delays of varied lengths. Figure 5.4 illustrates example stall patterns. The videos with the fewest number of stalls, excluding the reference videos, contain only an initial delay, whereas the largest number of stalls that were added to any video was 7.

5.1.3 Distortion Simulation Process

A different randomly-selected set of stall patterns was introduced on each of the 24 reference videos, with the constraint that each of the distortion patterns may be

Table 5.2: Description of the four stall parameters (left column) and the different values of these parameters considered constructing the stalling patterns in the following database. L refers to the total length of a given video.

Number of stalls	Few ($1-3$ stalls)	Many ($4-7$ stalls)		
Stall length	Short ($2-4$ sec.)	Medium ($5-9$ sec.)	Long ($10-15$ sec.)	
Position of the stalls	Beginning (between $0-L/2$ sec.)	Middle (between $L/4-3L/4$ sec.)	End (between $L/2-L$ sec.)	Uniformly throughout (between $0-L$ sec.)
Startup delay	Short ($0-7$ sec.)	Long ($8-20$ sec.)		

Table 5.3: Summary of different simulated stall patterns. The prefix **x** refers to the position where the pattern is introduced and takes values {B, M, E, U} as defined in Fig. 5.3. '#' refers to the count of videos in each column.

Stalling patterns	**# videos with x = B**	**# videos with x = M**	**# videos with x = E**	**# videos with x = U**	**Total # videos**
Only short initial delays (**shortInitial**)	-	-	-	-	5
Only long initial delays (**longInitial**)	-	-	-	-	4
Short initial + **f**ew **m**edium (**x_sfm**)	6	6	8	-	20
Short initial + **f**ew **l**ong (**x_sfl**)	6	6	6	-	18
Short initial + **m**any **m**edium (**x_smm**)	4	-	4	6	14
Short initial + **m**any **s**hort (**x_sms**)	6	-	6	6	18
Long initial + **f**ew **m**edium (**x_lfm**)	6	6	6	-	18
Long initial + **f**ew **l**ong (**x_lfl**)	6	6	4	-	16
Long initial + **m**any **m**edium (**x_lmm**)	4	-	4	5	13
Long initial + **m**any **s**hort (**x_lms**)	6	-	6	4	16

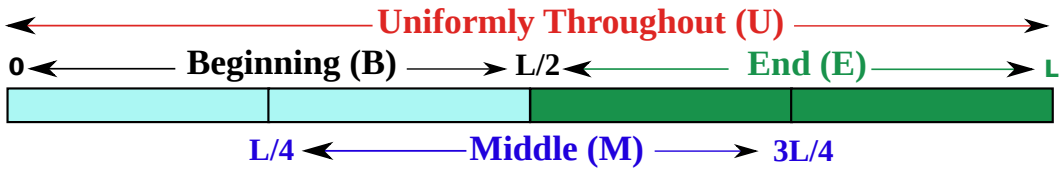


Figure 5.3: Illustrating different positions of stalls in a video of length L represented by {B, M, E, U}.

Figure 5.4: Illustrating (a) the stall pattern B_sfl (**s**hort initial delay followed by a **f**ew **l**ong stalls in the **b**eginning.) (b) stall pattern E_lmm (**l**ong initial delay with **m**any **m**edium stalls towards the **e**nd.) as defined in Table 5.3 for any video sequence.

added to a minimum of 4 and a maximum of 8 reference videos. For each video content, a stall pattern file was generated containing the stall location(s) and the length(s) in seconds ($< stall\_loc,\ stall\_length >$). Using the FFmpeg tool [141], the source .mp4 sequences were decoded yielding uncompressed raw .yuv files. At every *stall_loc* contained in the stall pattern file of a given video, the corresponding .yuv file was split into two chunks. The last frame of the chunk preceding *stall_loc* was copied for *stall_length* seconds to simulate the scenario where the client-side network buffer has emptied, thereby forcing the last frame in the buffer to be displayed on the screen until the next frame is streamed, thus generating the rebuffering video chunk. Further, a loading or buffering icon was overlaid at the center of each frame of the rebuffering video chunk to realistically simulate the network impairment as encountered by a viewer when streaming over-the-top video content.

5.2 Subjective Study

5.2.1 Subjects and Study Set-up

Due to the large amount of video data in the LIVE Mobile Stall Video Database-II, it was difficult to obtain a sufficient amount of subjective data on each video, and having each subject view every video in the database would be be asking subjects to volunteer approximately four hours of their time, which would likely discourage participation. Therefore, as was done in [142], I divided the videos into two groups, A and B, containing 88 and 86 videos respectively. Each set of videos was further equally divided over three study sessions of approximately 40 minutes each, to avoid mentally and visually fatiguing the subjects. The duration of all three sessions were roughly equivalent for both groups of subjects. Subjects were randomly assigned to either the A or the B group, and each viewed the same two training videos prior to the start of their first session. These training videos are not included in the database and were simply used to introduce subjects to the rebuffering events they would encounter in the study and to allow them time to practice providing continuous-time feedback and using the ratings bar. The study was conducted over a two-week period, and in total, 54 undergraduate and graduate students from The University of Texas at Austin participated. Both the A and B groups contained 27 subjects after random assignments.

Before participating in the study, each subject read instructions informing them that they should provide both continuous-time and overall opinion scores based on their viewing experience. I then verbalized instructions to reiterate the subject's tasks and to ensure that the subjects understood that they were not to judge the content of a video. The monitor position was also adjusted to ensure a viewing distance between 2 and 2.5 feet, which was considered a comfortable viewing distance. The subjects were informed not change their viewing distance and position too much throughout the study to maintain viewing conditions. The subjects were

allowed to take breaks during their session if they felt visually fatigued, but they were not allowed to take breaks while the video was playing or while the interface was prompting for submission of a rating. I did not test subjects for vision problems, because I wanted to capture the impressions of a realistic sampling of human vision systems. I did, however, ask them to wear corrective lenses during the study, if they usually wore them during for daily activities.

5.2.2 Study Interface

The study was conducted on a PC using a GUI I designed using the XGL toolbox [143] with MATLAB 2015b. I chose to use a PC over a smartphone with a touch screen to make it easy for subjects to provide accurate continuous-time ratings. The XGL toolbox was developed for the presentation of psychophysical stimuli to human observers, and I encountered no display issues while using it. Each video sequence was stored as raw YUV 4:2:0 frames, and to avoid additional playback interruptions, I loaded each video in its entirety into memory before presenting to subjects. Corresponding audio files were played with each video and no latency was experienced with audio playback. The PC contained an ATI Radeon X600 graphics card, and the attached monitor was an ASUS VG248QE. Videos with original resolutions of 1280×720 and 1280×640 were converted to 1024×576 and 1024×512 respectively, using FFmpeg [141] to resize raw YUV frames. These were the highest resolutions that were supported by the monitor and that did not cause the video load times to exceed three seconds, which was a cutoff I set to avoid annoying subjects. All other videos were displayed in their native resolutions. I believed that showing the videos in high definition, or even 1080p, was not necessary, as the focus of the current work is on *mobile* streaming QoE and monitoring its trends over time. Showing lower-resolution videos on a typical laptop-sized monitor allowed us to gather continuous scores with minimal distractions to the subjects, as I was not trying to simulate

watching a movie in HD at home. Prior to the subjective tests, the display monitor was color-calibrated using the Datacolor Spyder5PRO Display Calibration System [144], which measures ambient lighting conditions and provides on-screen guidance to change the monitor settings so that colors are authentically presented. Finally, I processed each video sequence to have a 30-Hz frame rate, set the monitor refresh rate to be 60 Hz, and displayed each frame for two monitor refresh cycles for avoid adding flicker artifacts.

5.2.3 Testing Methodology

I used a single stimulus continuous quality evaluation (SSCQE) procedure [145] to obtain both continuous-time and overall quality ratings on video sequences. All videos were shown in random order without repeated content shown successively. Subjects used a continuous-scale ratings bar with a neutral initial cursor position, i.e., qualitatively *Fair*, to minimize additional bias. The qualitative range of the ratings bar started at *Bad* (far left) and reached a maximum at *Excellent* (far right), with *Poor* and *Good* equally spaced between *Bad*, *Fair*, and *Excellent* to reflect the ITU-R Absolute Category Rating (ACR) scale [145]. Subjects used a mouse to adjust their desired ratings on the quality scale. Figure 5.5 shows a screenshot of the GUI during video playback with the continuous-scale ratings bar at the bottom of the screen. Following the presentation of each video, subjects were shown a screen displaying the same continuous-scale ratings bar with instructions to provide a single overall quality score. A screenshot of this screen is shown in Fig. 5.6.

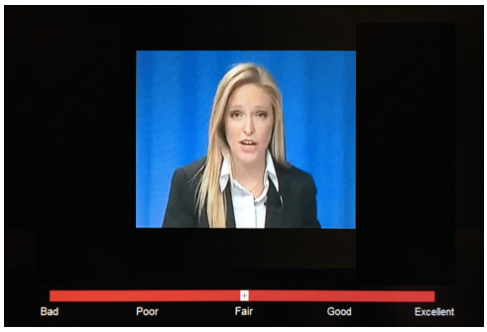


Figure 5.5: Screenshot of the GUI showing the continuous ratings bar placed below the video sequence for gathering continuous-time scores during playback.

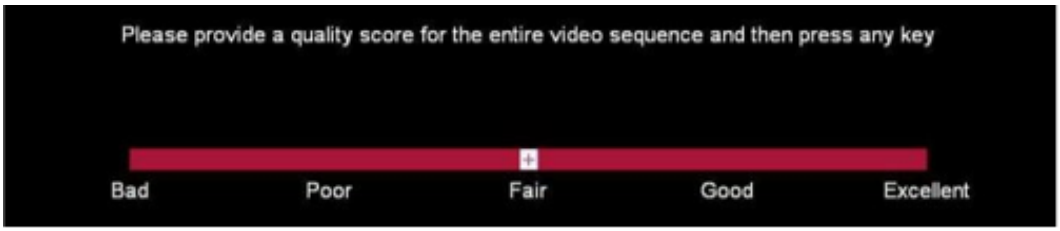


Figure 5.6: Screenshot of instructions and ratings bar for gathering an overall QoE score at the end of each video's playback.

5.3 Processing of the Subjective Scores

5.3.1 Accounting for Intra-Subject Variability

As there are inevitable variations between each subject's use of the quality scale, possibly also across sessions, I compute overall Z-scores [146] for continuous-time and overall QoE scores which I describe next. If I let s_{ijk} denote the final QoE score assigned by subject i to video j during session k,

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} s_{ijk}, \tag{5.3}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik} - 1} \sum_{j=1}^{N_{ik}} (s_{ijk} - \mu_{ik})^2} \tag{5.4}$$

$$z_{ijk} = \frac{s_{ijk} - \mu_{ik}}{\sigma_{ik}}. \tag{5.5}$$

Here, N_{ik} is the number of test videos seen by subject i in session k, $k \in \{1, 2, 3\}$.

To compute continuous-time Z-scores, let s_{ijkl} denote the score assigned by subject i to the l^{th} frame of a video j during the session k. The per-frame Z-scores are then computed as:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{N_{ik}} N_j} \sum_{j=1}^{N_{ik}} \sum_{l=1}^{N_j} s_{ijkl}, \tag{5.6}$$

$$\sigma_{ik} = \sqrt{\frac{1}{\sum_{j=1}^{N_{ik}} N_j - 1} \sum_{j=1}^{N_{ik}} \sum_{l=1}^{N_j} (s_{ijkl} - \mu_{ik})^2}, \tag{5.7}$$

$$z_{ijkl} = \frac{s_{ijkl} - \mu_{ik}}{\sigma_{ik}}. \tag{5.8}$$

In this case, N_{ik} is the total number of test videos that were seen by subject i in session k, and N_j is the total number of frames in video j. The per-frame Z-scores z_{ijkl} are thus computed using the scores the subject assigned to all frames $l \in [1, N_j], j \in [1, N_{ik}]$ over the entire session.

5.3.2 Subject Rejection Methodology for Continuous-Time Scores

Perceptual video quality is inherently subjective, and the task of providing continuous-time feedback is even more likely to induce variability across subjects, so I wanted to filter out data from subjects who were unable to provide consistent ratings. There is no standard acceptable method for subject rejection that can be applied to continuous-time quality monitoring. Moreover, adopting the rejection strategies recommended for single overall scores per video by discarding the rich, temporal subjective data is inefficient.

Thus, with a goal to identify and eliminate inconsistent continuous-time subjective data, I used a subject rejection method based on the Dynamic Time Warping

(DTW) technique [147], similar to what is described in [9]. Dynamic Time Warping (DTW) is a helpful technique for aligning time series data and computing a distance measure between the best aligned data. Specifically, in the DTW method, given a pair of temporal sequences (seq_1, seq_2), their time axis are warped (stretched or compressed) to achieve a reasonable alignment between them, by minimizing an aggregated Euclidean distance between corresponding points of the two sequences. Computing a DTW distance between (seq_1, seq_2) yields a measurement of dissimilarity between the two sequences.

I chose to apply DTW on pairs of sequences with a *locality restriction* that the maximum window for matching temporally-corresponding points of any two sequences was 5 seconds. This was done to account for the variation in typical human response time to an event (in this case, a quality variation, which could be the start or end of a stalling event), including the motor activity required to adjust opinions on the rating bar in response to a change in video quality [148]. I tested subject rejection using warping windows of $\{2, ..., 12, \infty\}$. By manually inspecting the instances where different windows rejected different subjects, I determined that a window of 5 seconds allowed for reasonably associating corresponding data points and waveform patterns without matching completely unrelated temporal data points. Further, unlike the piece-wise distortion-localized sequences constructed in [9], I chose to consider complete temporal subjective data waveforms, excluding data from frames associated with start-up delays, as inputs to the DTW algorithm. I only removed start-up delay data to identify inconsistent subjective responses, because I found that subjects were generally not bothered by start-up delays (more in Sec. 5.4.1).

I conjectured that the reliability of the continuous-time opinion scores can depend on the following two scenarios [145]:

- **Scenario 1: Systematic shifts:** If a subject is too negligent, optimistic or pessimistic, or has misunderstood the voting procedures, a series of continuous-

time ratings could be systematically shifted away from the general consensus, if not completely out of range.

- **Scenario 2: Local inversions:** A subject can sometimes vote without taking too much care in watching and tracking the variation of the quality of the sequence displayed. In this case, the opinion curve can be relatively similar to the general consensus curve, but local discrepancies can be observed.

These two undesirable behavioral patterns need to be identified and the corresponding ratings have to be discarded. I thus followed a two-step subject rejection strategy; the first step was devoted to detecting and discarding observers exhibiting a strong shift of votes compared to the consensus behavior (Scenario 1), while the second step was used for detecting inconsistent observers (Scenario 2). I then rejected subjects identified as outliers in either of the two scenarios.

To assist the identification of unreliable subjects due to Scenario 1, I computed a *mean QoE waveform* per frame for each video, by averaging the continuous-time waveforms obtained from the individual subjects. Figure 5.7 illustrates the individual continuous QoE waveforms along with the mean QoE waveforms for a few videos. The standard deviations of the mean QoE waveforms and the stall patterns afflicting the videos are also illustrated. It may be observed that, despite the local temporal fluctuations in the continuous-time scores of individual subjects, in general, they followed similar trends which were effectively captured by the mean QoE waveforms. However, I note that the individual continuous-time subjective scores were not temporally aligned to account for each subject's motor response delays prior to computing the mean QoE waveforms. To identify the unreliable subjects due to Scenario 2, I compared individual waveforms of subjects in a pairwise manner to identify responses that were highly inconsistent with other individual responses. I now describe the proposed two-step subject rejection strategy in detail:

1. From each continuous-time subjective score waveform, I removed the data

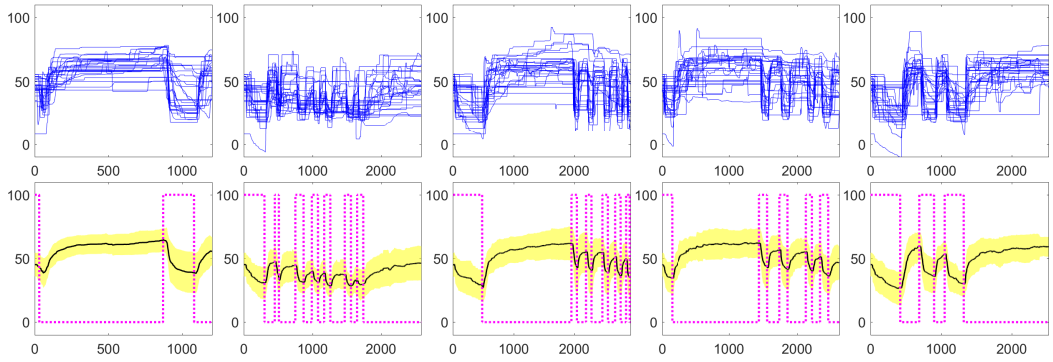


Figure 5.7: In each top-bottom pair of plots, the top shows all the responses from individual subjects plotted together, and the bottom shows the average response and standard deviation (in yellow) around the mean (in black) along with the stall pattern (in magenta). Here, values of 100 indicates stall frames, and values of 0 indicates playback frames.

points associated with the start-up delay for each video with a start-up delay.

2. I computed the mean QoE waveform for each video by averaging the continuous-time waveforms from each subject per frame.

3. For each video:

    a. For each subject, I computed:

        i. their DTW distance from the mean QoE waveform;

        ii. their average DTW distance from all other subjects who viewed that video.

    b. I used the distribution of DTW distances from the mean QoE waveforms to determine inconsistent subjects (Scenario 1).

    c. I used the distribution of average pairwise DTW distances to determine additional inconsistent subjects (Scenario 2).

Note that I rejected subjects on a *per-video* basis. A large distance value in either Step 3(a)-i or Step 3(a)-ii for a given subject suggests that the subject was

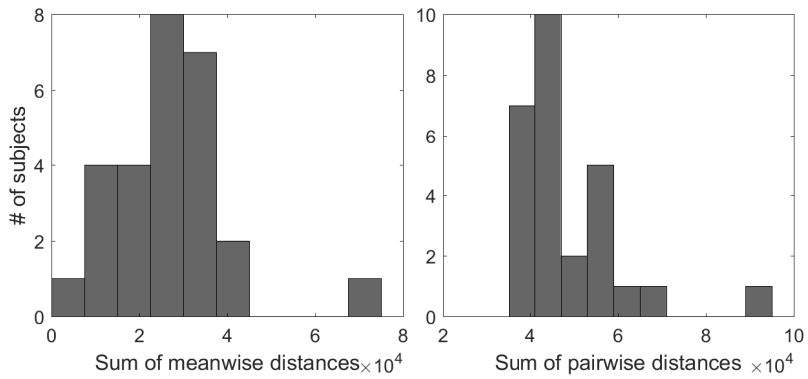


Figure 5.8: Example histograms of accumulated DTW distances when using meanwise (left) and pairwise (right) comparisons.

misaligned from other subjects or the consensus and thus could be an outlier. To identify the outlier subjective responses, in Steps 3(b) and 3(c), I used the adjusted boxplot method for skewed distributions detailed in [149] and implemented in a MATLAB toolbox called LIBRA [150]. I used their default parameter selections and rejected only subjects having very high aggregated DTW values. Consider Fig. 5.8, which shows histograms of accumulated DTW distances for two example videos. On the left, the DTW distances to the mean QoE response are summed (meanwise distances), and on the right, the DTW distances to other individual responses are summed (pairwise distances). In both cases, the rightmost subject gets rejected as an outlier. Specifically, given a set of DTW distances (from Step 3(b) or 3(c)) for a video, I reject any subject having DTW distances (either meanwise or pairwise) greater than:

$$Q_3 + h_u(\text{MC})\ \text{IQR}, \tag{5.9}$$

where Q_3 is the third quartile cutoff, IQR is the interquartile range $Q_3 - Q_1$, and $h_u(\text{MC}) = 1.5e^{b\text{MC}}$, with $b = 3$, as found in [149] to produce a robust outlier detection model. The medcouple (MC) is defined as the scaled median difference between the left and right halves of a univariate distribution and is used here to

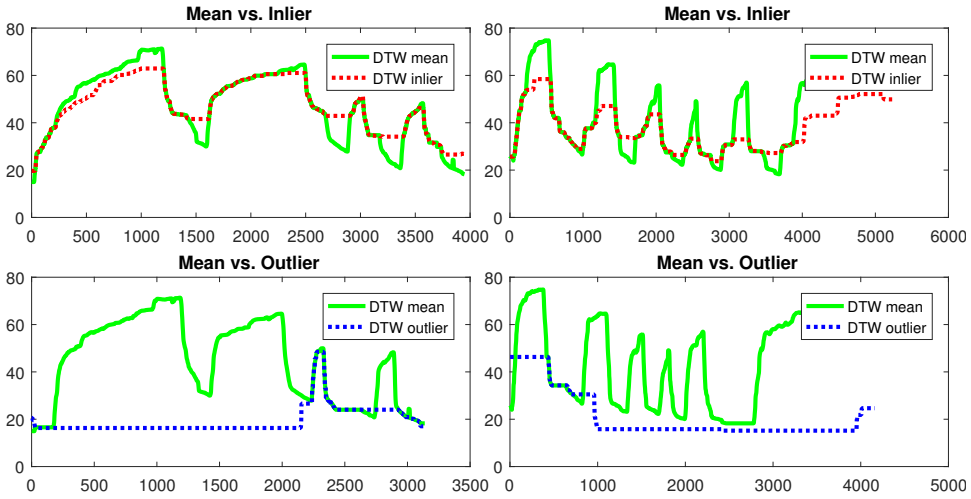


Figure 5.9: Examples of inliers and outliers using meanwise DTW subject rejection. Each column represents a different video. The top row shows the dynamic time warped mean waveform and an example inlier waveform. The second row shows the dynamic time warped mean and an example outlier waveforms.

measure the distribution of skewness in the DTW distances. In my experiments I found that most of the time, $|\text{MC}_{pairwise}| > |\text{MC}_{meanwise}|$, meaning that histograms of accumulated pairwise DTW distances tended to be more skewed than those of accumulated DTW distance to the mean waveforms. However, there were many cases in which the opposite inequality was true. Thus, pairwise and meanwise subject rejection methods may handle border cases differently, and by combining them, I could enforce stricter subject rejection cutoffs. Figures 5.9 and 5.10 show inlier and outlier examples using both the meanwise and pairwise DTW subject rejection steps. With either method, it is clear the outlier waveforms have much larger DTW distances to either the mean (Fig. 5.9) or an identified inlier (Fig. 5.10) than the inliers do. On average, the proposed subject rejection method identified only **0.7** subjects per video to be rejected.

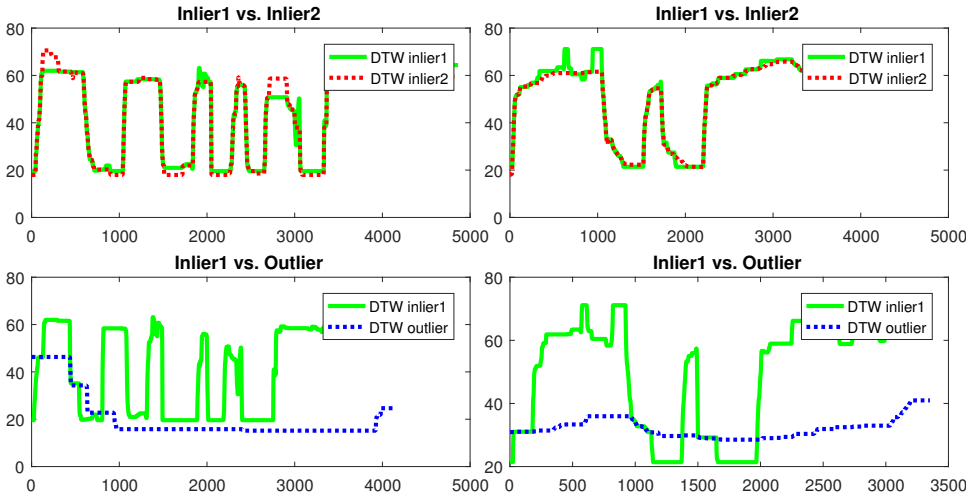


Figure 5.10: Examples of inliers and outliers using pairwise DTW subject rejection. Each column represents a different video. In the top row, two dynamic time warped inlier responses are plotted against one another. The second row shows dynamic time warped inlier and outlier waveforms.

5.3.3 Subject Rejection Methodology for Overall QoE Scores

As mentioned earlier, each subject also provided an overall QoE score for each video, upon watching the entire sequence. In order to be able to use these scores, I wanted to retain only those subjects who were able to rate the videos consistently. I followed a different subject rejection procedure, detailed in the ITU-R BT.500-11 recommendation [139], and used *overall* QoE Z-scores. By following the recommended steps, I rejected 6 subjects out of total 54 (3 from each set) and retained 24 subjects per set for analysis.

As was done in [82], the Z-scores were linearly rescaled so that the scores lay in the range [0,100]:

$$z'_{ij} = \frac{100(z_{ij} + 3)}{6}. \tag{5.10}$$

I can scale the scores in this way, because if the Z-scores are normally distributed, then $> 99\%$ of the scores will lie in the range $[-3, 3]$, which I found to be the case in my subjective data.

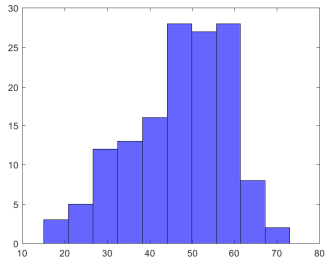


Figure 5.11: Histogram of the overall MOS of the distorted videos.

Finally, I computed a Mean Opinion Score (MOS) for each video by averaging the rescaled Z-scores:

$$\text{MOS}_j = \frac{1}{M}\sum_{i=1}^{M} z'_{ij}, \tag{5.11}$$

where $M = 24$ subjects after subject rejection in both sets A and B. Figure 5.11 illustrates the mean overall QoE scores across all the distorted test video sequences, after subject rejection. These scores were found to lie in the range [19.12, 75.82], and the mean of the standard deviations of the Z-scores obtained from all subjects across all the videos from both the groups was 16.62.

5.3.4 Temporally Pooling Continuous-Time QoE Scores

I sought to derive a single quality value per video from the continuous data in order to analyze and compare them with the overall subjective QoE values. I achieved this by adopting the following two temporal pooling strategies: (a) a per-frame average of the continuous-time scores (henceforth referred to as average-pooled QoE) and (b) VQ Pooling as described in [151]. VQ Pooling technique combines the scores giving lower-quality frames more influence on the overall quality score, supporting the concept that the worst part of a video tends to attract more attention of the viewers, thus dominating the overall quality perception of the video [151]. I achieved a Spearman rank ordered correlation of 0.9128 between the average-pooled QoE and overall QoE and 0.8685 between the VQ-pooled QoE and the overall QoE.

This analysis is helpful in comparing *overall* continuous-time data to reported overall scores, which are the most commonly collected data points in QoE subjective studies [82, 83, 84, 85, 86, 87]. Given that a simple arithmetic average weighs each per-frame score equally, I found that it did not often effectively capture temporal changes in viewer QoE, which depends on stall features, such as their absolute and relative locations, lengths, and so on, which do factor into the final overall subjective QoE scores. Thus, for the subjective data analysis which I describe next, I chose to rely on the mean overall QoE scores defined in (5.11).

5.4 Analysis of the Subjective Data

Continuous-time subjective data is a valuable resource that could help us better understand the aspects of rebuffering events (such as their length and frequency of occurrence) that impact viewer behavior. Such analysis could assist the design of quality-aware stream-switching algorithms which could influence the occurrence and lengths of the stalls, for a given network bandwidth budget, such that the end user's QoE is maximized. However, when trying to understand the effects of stall *patterns*, rather than a single stall event, on continuous-time QoE, I use a single representative QoE score per video. In this section, I utilize the mean overall QoE scores to represent the continuous-time waveforms to analyze the effect of the four stall parameters—stall length, number, position, and the startup delay—on an end user's overall QoE. Also, wherever feasible, I report the statistical significance of the observed differences in QoE by conducting a paired, two-sided Wilcoxon signed rank test [152] (at significance level $\alpha = 0.05$). However, as described in Sec. 5.1, the videos in my collection are of varied lengths, and not all stall patterns were added to all video contents (Table 5.3). This constraint prevented us from comparing all of the distortion patterns for each video content or aggregating the continuous-time scores of all the videos for each distortion pattern. Thus, for the analysis I present

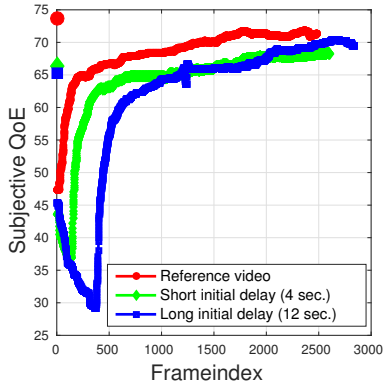


Figure 5.12: Temporal subjective ratings of a reference video and its corresponding distorted variants afflicted with short and long initial delays. At $x = 0$, I plot their overall QoE scores.

in the following sections, I illustrate the results using only those video contents that model the specific stall patterns under investigation. I acknowledge that the design of the test sequences limits the conclusions I can draw from holding video lengths constant, however, my goal is to develop a general VQA model that can predict QoE regardless of sequence length.

5.4.1 Effect of Start-up Delay on QoE

To understand the impact of the length of the start-up delay on QoE, I compared the overall QoE values of videos with only short or long initial delays with the overall QoE of the reference videos, which do not contain any stalling events (delay lengths are defined in Table 5.2, and the number of distorted videos per stall pattern are listed in Table 5.3). I found that the observed difference in the overall QoE scores of the 4 videos with and without the long initial delays was not statistically significant. Similar results were obtained for the 5 videos with and without short initial delays. The duration of the 9 reference videos used to generate the distorted sequences under consideration were in the range $[29, 82]$ seconds in length with an average of about 60 seconds. Figure 5.12 illustrates the temporal subjective ratings for a

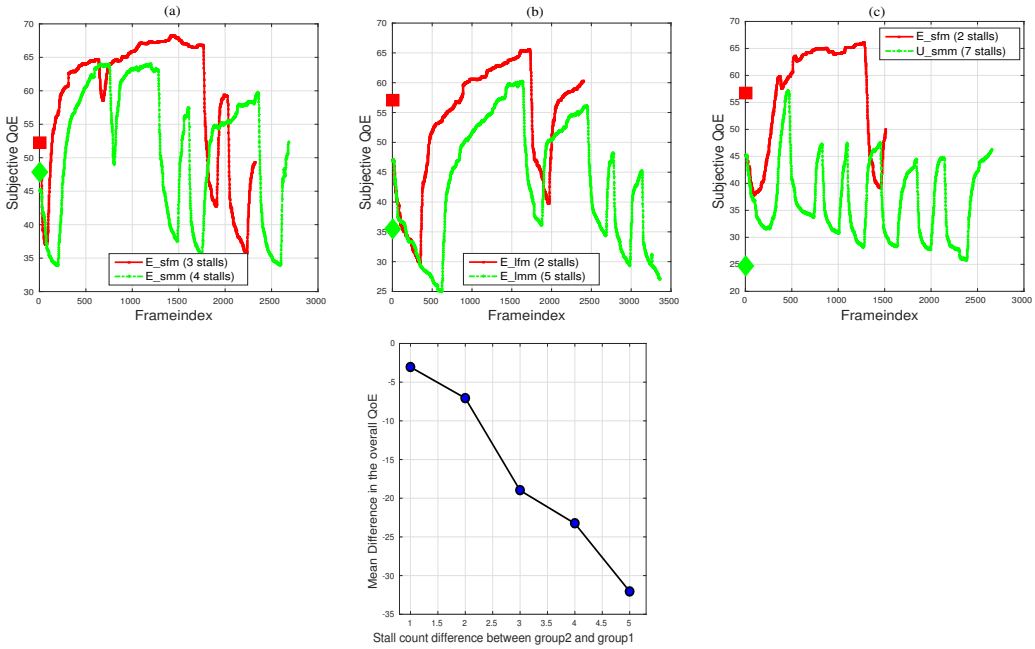


Figure 5.13: Temporal QoE ratings and the overall QoE scores (presented at $x = 0$) of three different video contents (a) - (c) modeling few (V_f in red) and many (V_m in green) stalling events of similar lengths. It may be observed that videos with many stalls consistently score lower, and the gap between the overall scores increases with the increase in the difference between the many and few stall counts, as illustrated in (d). Best viewed in color.

reference video[1] (of length 82 sec.) and its corresponding distorted variants. It may be observed that for this video, the subjects were not too frustrated with the initial delay, and thus the overall opinion scores (presented at $x = 0$) did not seem to vary greatly around this factor.

Start-up delays are currently very commonly experienced in OTT video streaming, which could have led to viewers having a higher level of tolerance to stalls placed in this particular position, even in short video sequences. It could also be attributed to the hysteresis (recency) effect [140] experienced by subjects, which

[1]This was the only reference sequence that was used to model both shortInitial and longInitial stall patterns.

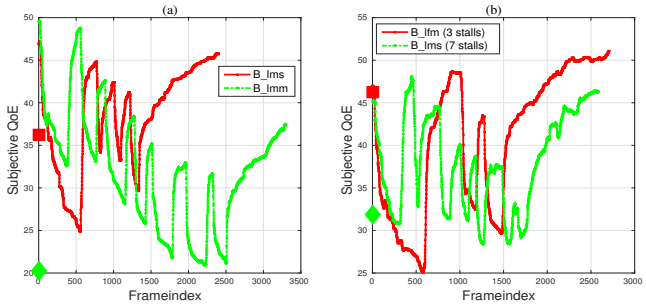


Figure 5.14: Temporal QoE ratings and the overall QoE scores (presented at $x = 0$) of two different video contents (a) modeling shorter ('V_l,' in red) and longer ('V_h,' in green) stalling events.

supports the phenomenon that stalls that occur early in a video tend to cause less viewer annoyance. I study this more closely in Sec. 5.4.4.

5.4.2 Effect of the Number of Stalls on QoE

In order to understand the effect of the number of stalls on QoE (i.e., few vs. many as defined in Table 5.2), I conducted the following experiment. From each of the 24 unique video contents, I selected contents that were used to model *both* $s\mathbf{f}m$ and $s\mathbf{m}m$ stall patterns or those that model *both* $l\mathbf{f}m$ and $l\mathbf{m}m$ (refer Table 5.3). Identifying such video sequences can help us understand the effect of the number of stalls while keeping stall lengths and start-up delay lengths fixed for a given video content. I then grouped all the videos containing few distortions in set V_f and their corresponding videos containing $many$ distortions in set V_m. Some sample video sequences are illustrated in Figure 5.13. From Fig. 5.13(a) to 5.13(c), I notice that irrespective of the stall location and the start-up delay lengths, videos with $many$ stalls annoyed viewers more than videos with few stalls. Further, I found that the observed difference in the overall QoE scores between the 18 videos that belonged to sets V_f and V_m was statistically significant. The average duration of these 18 videos was 79.2 seconds.

To effectively quantify my observation that annoyance increases with increases in the number of stalls, I computed the difference between the overall QoE scores between the video contents belonging to V_m and V_f and aggregated them based on the *difference between the number of stalls* in V_m and V_f[2]. In Fig. 5.13(d), along the x-axis, I plot the difference in the number of stalls between videos belonging to V_f and V_m, and along the y-axis I plot the aggregated average difference between the overall QoE scores. It is obvious in Fig. 5.13(d) that as the difference in the number of stalls increases, user annoyance increases monotonically.

5.4.3 Effect of the Lengths of the Stalls on QoE

Next, I wanted to understand the effect of stall lengths on viewers' QoE, i.e., compare short vs. medium vs. long length stalls as defined in Table 5.2. Towards this end, out of the 24 unique video contents, I selected contents that were used to model at least one of the following pairs of distortions (also refer Table 5.3):

1. $sf\mathbf{m}$ *and* $sf\mathbf{l}$
2. $sm\mathbf{s}$ *and* $sm\mathbf{m}$
3. $lf\mathbf{m}$ *and* $lf\mathbf{l}$
4. $lm\mathbf{s}$ *and* $lm\mathbf{m}$

As was done earlier, I constructed two sets V_l and V_h, of videos with *low* and *high* stall lengths respectively. V_l consists of distorted videos with shorter stalls relative to videos in V_h. For example, V_l may consist of all videos containing sfm or sms stall patterns, while V_h may consists of the corresponding video contents containing sfl or smm stall patterns, respectively. Identifying such video sequences can help us isolate and understand the effect of the length of stalls, keeping the number of stalls

[2]The videos with *few* stalls had $1-3$ stalls and videos with *many* stalls had $4-7$ stalls. Therefore, the difference in the number of stalls between videos belonging to V_f and V_m ranges between $1-6$

and the length of start-up delays fixed for a given video content. However, I note that a video with few stalling events can have anywhere between $1-3$ stalls, while videos with *many* stalling events can have anywhere between $4-7$ stalls. Thus, I denote the total number of stalls for each video content belonging to V_l and V_h as ns_l and ns_h respectively.

By comparing the overall QoE scores from both the groups, I observed that:

1. For the 25 video contents where the total number of stalls in V_l were less than those in V_h (i.e., $ns_l \leq ns_h$), I found that the observed difference in the overall QoE was statistically significant.

2. However, for the 11 video contents where $ns_l > ns_h$, the observed difference in overall QoE was not statistically significant.

3. Irrespective of the number of stalls, during a stall event, viewers' QoE values dropped more steeply to lower values for longer stalls than for short (or medium) length stalls. Figure 5.14 illustrates two examples.

4. Irrespective of the number of stalls, after a stall event, viewers' QoE seems to recover quickly. However, it appears to recover to a lower value each time due to the occurrence of the stall.

This analysis indicates that, given two stall patterns, if the number of stalls are equal, then the stall length exerts more influence on an end users' QoE. Also, an increase in the number of stalls causes more viewer annoyance than do the lengths of the stalls. To more closely study this observation, I chose to compare video contents that modeled *few medium* length stalls, i.e., sfm or lfm (set V_{fm}) against those containing *many short* length stalls, i.e., sms or lms (set V_{ms}). A video content belonging to both sets would have stalls occurring at the same location (B, M, E, or U). For each of the 7 videos contents that belonged to both video sets, V_{fm} and V_{ms}, I found that videos with many short stalls were consistently rated lower (Fig. 5.14

(b)). Further, I also found that this observed difference in perceived QoE between these two stall patterns was statistically significant.

5.4.4 Effect of the Position of the Stalls on QoE

As mentioned earlier in Sec. 5.1.2, all of the stall patterns designed in the current study were placed in one of the four positions: beginning, middle, end, or distributed uniformly throughout a video (refer Fig. 5.3). This was done to enable us to study the effect of the position of the stalls on viewers' QoE. Therefore, as was done earlier, out of the 24 unique video contents, I selected the distorted contents that consisted of similar distortion patterns occurring both at an earlier and at a later point in the video, and constructed two sets: V_p and V_s. This was done for all distortion patterns listed in Table 5.3 except for *shortInitial* and *longInitial*. I considered the following order of positions: $B \prec M$, $M \prec E$, $B \prec E$.[3] Thus, the *predecessor* belonged to V_p and the *successor* to V_s. For example, I identified distorted videos afflicted with B_smm and E_smm, for each of the 24 unique video contents. The distorted video sequence with B_smm belonged to V_p and the one with E_smm belonged to V_s. Further, I denote the number of stalls occurring in a video belonging to either V_p or V_s as ns_p or ns_s respectively. Identifying such video sequences helps us isolate and understand the effect of stall position, while keeping the stall lengths and the start-up delay lengths fixed for a given video content. By comparing the overall QoE scores from both sets, I observed the following:

1. For the 9 videos where $ns_p \leq ns_s$, in most cases, videos belonging to V_p had higher overall QoE when compared to their corresponding video contents in V_s. This phenomenon of distortions that occur early on in a video sequence having less effect on a viewer's QoE is referred to as the 'subjective hysteresis' effect [140], since the memory of poor-quality elements in the recent past

[3]Position U (uniformly throughout) was not considered in this analysis.

causes subjects to provide lower quality scores immediately following the event. However, this observed difference was not statistically significant in the case of these 9 videos. The average duration of these 9 videos was 71 seconds.

2. For the 3 videos where $ns_p > ns_s$, I found that the overall QoE values of videos belonging to V_p was less than the overall QoE of those belonging to V_s. This suggests that videos with more frequent interruptions annoy a viewer the most, even if the interruptions occur at the beginning of a video playback. The average length of these 3 videos was 60 seconds.

It is not the case that the perceived overall QoE is unaffected by stall position. Note that I was limited by the number of available videos from conducting a more detailed analysis. In this direction, an even more focused study using videos of longer duration, with stalls occurring at various locations and sufficient data samples per location could further deepen my understanding of the dependency between the stall positions and QoE.

5.4.5 Recency, Primacy, and Repetition Priming

Since subjective overall QoE scores were gathered at the end of each video's playback, I wanted to evaluate the *extent* to which the 'recency effect' influenced these scores. I observed evidence supporting the influence of recency, detailed in Sec. 5.4.4, where videos with stalls at the beginning of video playback were perceived to be of higher quality than those with the exact same stall pattern occurring at the end of playback. However, I wanted to understand the extent of this effect and the complex interplay of recency and the number and length of stalls in more depth. To further my understanding, I once again compared video contents that modeled either of the following pairs of distortions (also refer to Table 5.3):

1. B_smm vs. E_sms

2. B_smm vs. E_sfl

3. B_smm vs. E_sfm

4. B_sms vs. E_sfl

I then grouped all the videos afflicted with either of the above listed distortions into V_b or V_e, depending on whether the stalls were located towards the beginning or end, respectively, of the video sequence. I found that in the 5 video contents that modeled any of the above pairs of distortions, the number of stalls had a more prevailing effect on viewer perception, as illustrated in Fig. 5.15. The average length of these 5 videos was 88 seconds.

Many factors other than recency could contribute to this subjective behavior. The 'primacy effect' [153] refers to the phenomenon of being able to recollect (stalling) events that occurred at or near the beginning of a temporal sequence more clearly than those that occurred around the middle. 'Repetition priming,' [154, 155] caused by the repetition of stalling events at the beginning of a video's playback, could lead to the encoding of this experience in memory. Thus, when providing overall QoE scores, a viewer could be influenced by recollections of unpleasant viewing events related to repetition priming, primacy, and recency.

In the real world scenario of video streaming on mobile devices, frequent rebuffering events—either long or short—at the beginning of a video could lead to viewer abandonment [96]. My analysis, albeit on a small video collection, suggests that, under a given network bandwidth budget, introducing stalling events early on in a video sequence by the stream-switch controllers can lead to severe viewer annoyance and should probably be avoided.

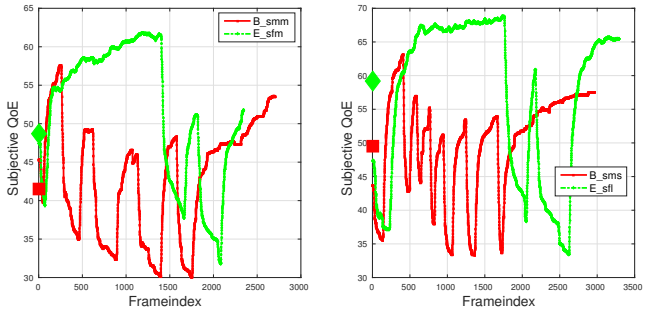


Figure 5.15: Temporal QoE ratings and the overall QoE scores (presented at $x = 0$) of two different video contents (a) modeling stalls in the beginning ('V_b,' in red) and in the end ('V_e,' in green).

Table 5.4: Questionnaire Responses

Survey question	**% of subjects responding Yes**	**% of subjects responding No**
Did you find it hard to rate your viewing experience?	40%	60%
Would you have preferred fewer videos per session and more sessions?	42%	58%
Would you have preferred fewer sessions and more videos per session?	30%	70%
Would you have preferred longer videos?	23%	77%
Did you feel any visual fatigue during the course of study?	38%	62%
Do you think you would be less fatigued and more engaged if there were fewer stalling events?	72%	28%

5.4.6 Summary of the Analysis

In this section, with the help of this subjective data, I focused on gaining insights into the way viewers respond to different stall parameters. I found that start-up delays did not significantly impact overall QoE, whereas the number of stalls did. I also found that stall length negatively impacts overall QoE, but the frequency of stall events had a more significant effect. Additionally, when analyzing evidence of the recency effect, I found that frequent poor-quality experiences in distant memory can also have a significant impact on overall QoE.

5.5 Analyzing the Survey Responses

As mentioned earlier, all of the subjects participated in a survey at the end of the subjective study. I summarize all of their responses in this section.

Level of concentration of the subjects

I asked the subjects from both A and B groups to rate their level of concentration throughout the study on a scale of 1 to 5, with 1 indicating that the subject was *very distracted* and 5 indicating that the subject was *very concentrated* throughout the study. I found that the average concentration level of all the subjects was 3.99 and the average standard deviations of the reported concentration levels was 0.78.

Level of interest in the test video contents

I also asked all of the subjects to rate their interest level in the test video content on a scale of 1 to 5, with 1 indicating *very boring content* and 5 indicating *very interesting content*. Subjects reported an average interest level of 3.18.

Repetition of the video content

As described in Sec. 5.1.3, each stall pattern was added to $4-8$ videos. This design choice was partly to assist us in studying the effects of different stall patterns on any given content. However, about 76% of subjects expressed that the content was too repetitive. When asked what they believed to be an acceptable number of times a given content could be repeated, their average response was 3.7 videos per session, with a standard deviation of 1.5.

Table 5.4 presents the responses to a few more questions asked in the survey session. I wanted to understand if the highly subjective and subtle task of rating QoE was understandable to the subjects. The questions were designed to gather feedback on the test set-up, number of sessions, and length of videos preferred by

the subjects. From the responses presented in Table 5.4, it can be understood that subjects would prefer watching fewer videos distorted with stalling events, as they are visually fatiguing.

5.6 Conclusion

In this chapter, I described the details of a new database called the LIVE Mobile Stall Video Database-II designed to gain a deeper understanding of viewer behavior when exposed to videos with realistic network-induced rebuffering events. I also discussed a subjective study I conducted using this database to collect continuous-time subjective QoE scores. Further, with the lack of publicly-available video databases that model such distortions, I hope that this new video database can be of use to researchers developing stall-dependent QoE models.

This subjective data and the insights gained by the analysis presented in Section 5.4 proved to be valuable when analyzing the aspects of stalling events that most impact viewer experience, and assisted in designing automatic QoE predictors which I describe in the next chapter.

Chapter 6

A Continuous-Time Streaming Video QoE Model

Using the analysis of the subjective study results presented in Chapter 5, I next aimed to tackle the problem of developing perceptually-accurate QoE prediction models that perform well on video contents afflicted with several different aspects of stalling events, in addition to variations in their spatio-temporal quality (due to compression, upscaling, frame rate distortions, and so on). As mentioned in Section 2.4.2, a number of QoE prediction algorithms exist in the literature that predict a global quality score for an entire video sequence. However, current streaming technologies are almost exclusively based on DASH or HLS, which adaptively select the optimal video levels. Thus, having an accurate measure of video QoE at any given instant would be much more valuable than obtaining a single overall QoE prediction at the end of the video.

In this chapter, I will describe an objective, no-reference, continuous-time QoE predictor called the **T**ime-**V**arying **QoE** (TV-QoE) Indexer for processing streaming videos afflicted by stalling events and quality variations [156]. Towards solving this problem, I sought to solve several sub-problems simultaneously, includ-

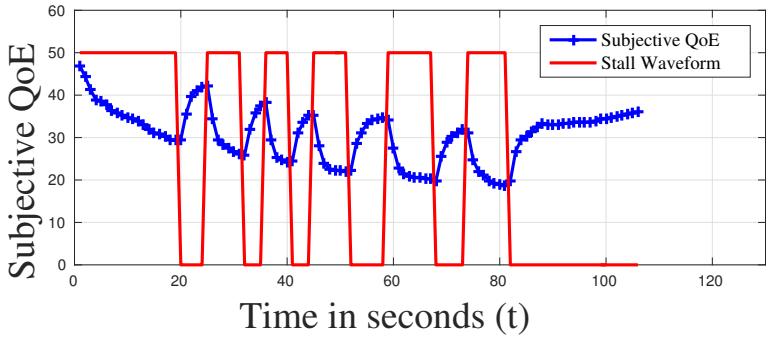


Figure 6.1: An example video sequence with 6 stalling events from the LIVE Mobile Stall Video Database-II [7], where the stall waveform (in red) is overlaid on the average of the temporal subjective QoE scores from each subject (in blue). For the purpose of illustration, a value of 0 in the stall waveform indicates normal video playback, while a value of 50 in the stall waveform indicates a stalling event.

ing how factors such as stalling event properties, the network buffer state, and video content impact an end user's QoE. This chapter is organized as follows:

1. First, I describe the comprehensive set of continuous-time, stall-informative, video content-informative, and perceptual quality-informative inputs that I derive from distorted videos. These inputs contain useful evidence descriptive of the effects of stalls and quality degradations on the time-varying QoE of a streaming video (Section 6.1.1).

2. In Section 6.1.2, I mathematically model the dynamics of a client-side network buffer that takes into account the variations in the bitrates at which the streaming video segments are encoded, as well as the instantaneous network throughput. This dynamic model serves as a valuable indicator of the fluctuations in perceived QoE due to stalling events.

3. I employ a Hammerstein-Wiener model that effectively captures the hysteresis effects that contribute to QoE using a *linear filter*. Further, it also accounts for the nonlinearity of human behavioral responses using nonlinear functions

at the input of the linear filter (Fig. 6.7). Each distortion-informative input is independently used to train a HW model with memory, resulting in an ensemble of HW models. The details of these approaches is discussed in Section 6.2.

4. I fuse the predictions of these individual HW models by employing and comparing two different strategies: (a) a multi-stage approach, in which the Hammerstein-Wiener models are concatenated, such that the predictions from the learners at one stage are supplied as inputs to another HW Model at a subsequent stage and (b) a multi-learner approach, in which the predictions of the individual HW models are used to train a different learner, called the *meta-learner* [157]. These two predictors, described in detail in Section 6.2, are independently trained to predict continuous-time ground truth QoE scores.

5. To address the side problem of predicting the overall perception of the quality of experience after viewing a video, I derive useful global statistics from the proposed comprehensive set of continuous-time inputs, and I also design a global overall QoE predictor (Section 6.3).

6. In Section 6.4, I present the results obtained from evaluating state-of-the-art models and TV-QoE predictors (multi-stage, multi-learner, and global) on three different video QoE databases. I also present my analysis on the performances of the proposed and existing models when different amounts and types of information about the test video are available.

6.1 Modeling Continuous-Time Inputs for QoE Prediction

The subjective data that I gathered on the LIVE Mobile Stall Video Database-II [7] (Chapter 5) proved to be a valuable resource to better understand various aspects of rebuffering events, such as their length, density, and location within a video, on an end users' QoE. With a goal to model the effects of stalls as well as video content, distortion, and other factors on QoE, I designed a number of stall-informative and content-informative input channels that I describe next.

6.1.1 Video Stall-Driven Inputs

Stall Length

One of the inputs ($u_1[t]$) is designed to capture the impact of stall lengths on QoE. Given a video, if $s_1[t]$ denotes the length of a stall at a discrete time instance t, then let

$$u_1[t] = e^{\alpha_1 s_1[t]} - 1, \tag{6.1}$$

where α_1 is a scalar chosen via cross-validation (Sec. 6.4). The choice of a nonlinear exponential function to express the influence of stall lengths on predicted QoE is motivated by the basic observation that viewer annoyance increases with rebuffering length [7]. Using a parameterized exponential makes it possible for TV-QoE to learn the steepness of the stall-length / annoyance relationship.

Total number of stalls

I also found from my analysis of the subjective data in [7] that, as the number of stalls increases, user annoyance increases monotonically, irrespective of the video content or duration. Further, as may be observed in the example in Fig. 6.1, perceived QoE tends to decrease with every stall occurrence. To capture the impact

of the number of stalls on QoE, I defined another dynamic input

$$u_2[t] = e^{\alpha_2 s_2[t]} - 1, \tag{6.2}$$

where $s_2[t]$ is the total number of stalls up to a discrete time instance t. Again, using an exponential model makes it possible to capture a viewer's annoyance against the number of stalls. The parameter α_2 is also a scalar chosen via cross-validation (Sec. 6.4).

Time since the previous stall

The next continuous-time input targets recency. Viewers generally react sharply to a stall occurrence, and as the period of time following the end of a stall increases, the viewer's perceived QoE may reflect improved satisfaction with the streaming video quality. However, immediately (and for some period of time) following a stall, the viewer's perceived QoE generally reflects heightened annoyance with the streaming quality. I found clear evidence for this behavior in the continuous-time subjective data obtained on the LIVE Mobile Stall Video Database-II [7]. Figure 6.1 also illustrates this behavior.

Thus, the third input to the TV-QoE model is the *time since the preceding rebuffering event* at every discrete instant t, with values of zero representing times during stalls. If $[T_{i,end}, T_{i+1,begin}]$ denotes the discrete time interval between the stall event (s_i) ending at time $T_{i,end}$ and the next stall event (s_{i+1}) starting at time $T_{i+1,begin}$[1], then

$$u_3[t] = \begin{cases} t - T_{i,end} & if \quad [T_{i,end} \leq t < T_{i+1,begin}] \\ 0 & if \quad [T_{i,begin} \leq t < T_{i,end}] \end{cases}. \tag{6.3}$$

[1]If there does not exist a stall event s_{i+1}, then $T_{i+1,begin}$ denotes the end of the video.

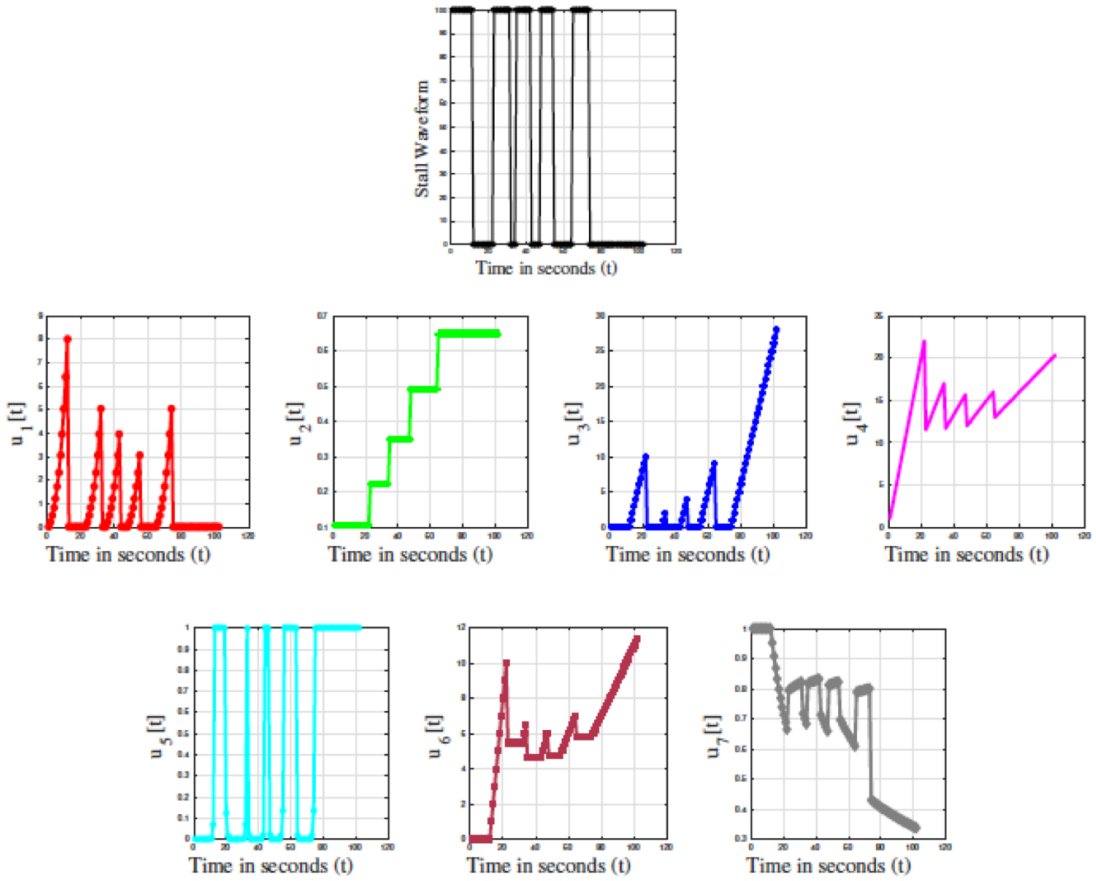


Figure 6.2: (Top row) A sample test video impaired by intermittent stalling events. (Bottom two rows) Stall-descriptive continuous input waveforms computed from a video sequence as described in Sec. 6.1.1. The vertical axis labels the type of input. Best viewed in color.

Inverse Stall Density

In addition to the various individual stall properties, the length of the video sequence containing stalls also affects viewer QoE. For instance, two videos of different lengths, one 20 seconds long and another 100 seconds long, but having the same stall pattern (i.e., one 5 second stall towards the end of the video) are likely to be perceived very differently by viewers. Thus, the next dynamic input I design is the reciprocal stall density at discrete time t, which provides a way of accounting for the length of the video (that includes the rebuffering events) relative to the total number of stall events that have occurred at any given instant. The inverse stall density is simply a discrete time t divided by the number of rebuffering events in the video up to time t:

$$u_4[t] = \frac{t}{s_2[t]}. \tag{6.4}$$

The inverse stall density adds a normalized time factor to the model and incorporates the combined effect of video length and stall positions.

Frequency of stalling events

Next, I sought to define a model input that excludes the effects of stall lengths, instead capturing the interplay between the number of stalling events and the length of *video playback time up to a given time instant* ($p[t]$). Therefore, the next input captures the effects of the density of the stalls on QoE relative to the current moment. The frequency of stalling events ($u_5[t]$) at a given time t is given by

$$u_5[t] = \frac{p[t]}{s_2[t]}, \tag{6.5}$$

where $s_2[t]$ is the total number of stalls up to time t.

Rebuffering rate

While the inverse stall density and frequency of stalling events inputs capture important interactions between video length, playback time, and the number of stalls, the next input to the model focuses exclusively on the interplay between stall lengths and the length of playback up to a given time instant ($p[t]$). To motivate the construction of this input, consider a single, very long stalling event in a video of length 90 seconds. This event may impact QoE differently than would a relatively short stalling event in a 20-second video. To effectively account for this hypothesis, I define the rate of rebuffering events as:

$$u_6[t] = \frac{r[t]}{r[t] + p[t]}, \tag{6.6}$$

where $r[t]$ is the total sum of stall lengths up to time t, and $p[t]$ is the playback time up to time instant t.

6.1.2 Modeling the Dynamics of the Client-Side Network Buffer

As previously mentioned, OTT services employ adaptive bitrate streaming algorithms, wherein the end-to-end network conditions are constantly monitored, and the bitrates of future video segments are chosen based on the current network buffer status of the client's media player, with a goal to minimize the occurrences of stalling events. However, under constrained network conditions, the state of the network buffer varies dynamically, and a stream-switching controller constantly chooses either to request a lower bitrate video segment or to risk the possibility of stall occurrence. Thus, the dynamics of a network buffer have a direct impact on streaming video quality but are not being modeled in any existing QoE models [110, 8, 158].

Existing publicly available QoE databases have been constructed by systematically inserting a variety of predefined stalling patterns in a controlled manner into

a small set of high-quality videos, thereby generating distorted video content [7, 8]. Thus each distorted video is not accompanied by the associated ground truth dynamic network buffer capacity that could have caused the underlying impairments. To overcome this challenge, I designed a simple model for a client-side network buffer, which I describe next. Note that, in the event that a media-streaming service or a QoE database can make available the actual network buffer capacity trace, then it could be directly plugged into the proposed QoE learner without needing to explicitly model the network buffer as I do below.

Assumptions and Notation: I make the following assumptions about the client-side network buffer, which are reflected in my mathematical model.

1. That a client-side network buffer is of a fixed size with a capacity measured in seconds. $BUFF\_MAX\_CAPACITY$ is defined as the maximum amount of video content that can be stored in a buffer.

2. Without loss of generality, that each video segment that is being adaptively transmitted from the media server is 1 second long.

3. That the buffer capacity builds and depletes exponentially. However, a different function can be easily applied in place of the exponential function.

4. That $BUFF\_MIN\_CAPACITY = 1$, which is the minimum amount of video (in seconds) that should be present in the network buffer for it to be able to handle input bitrate variations. This quantity can also be understood as requiring the network buffer to contain at least one second's worth of video content in order to continue playback on the client's media player. If the buffer state does not satisfy this minimum requirement, the result on the client side would be the occurrence of a rebuffering event.

5. $O[t]$ is the rate at which the video content leaves the buffer at time t. It can

take one of the two possible values

$$O[t] = \begin{cases} 1, & \text{during playback} \\ 0, & \text{during stall} \end{cases} \tag{6.7}$$

i.e., one second of video content leaves the network buffer during each second of playback, and no video content leaves during a playback interruption.

6. Let $B[t]$ be the amount of buffer that is occupied with video content and $\Delta B[t]$ be the rate of change of the buffer occupancy at a given discrete time instant t.

7. Let $L[t]$ denote the bitrate at which the incoming video segment is encoded, and $I[t]$ be the network throughput at time t.

At a given discrete time instant t, the rate of change of the buffer occupancy can be defined as follows [77]:

$$\Delta B[t] = \frac{I[t]}{L[t]} - O[t], \tag{6.8}$$

i.e., $\Delta B[t]$ is the difference between the amount of video (in seconds) that is entering the buffer and the amount of video (in seconds) that is leaving the buffer. Thus, variations in the buffer occupancy can be introduced due to changes in $I[t]$ or $L[t]$. When videos are encoded under a constant bitrate (CBR) regime, then $L[t]$ is fixed over the entire duration of the video sequence being streamed, which would not be the case for videos encoded under a variable bitrate (VBR) regime.

In the proposed model, the client-side network buffer can exist in one of the following three phases:

1. A **steady-state phase** where there is sufficient content in the network buffer to be transmitted to the client's media player at time t, i.e., $B[t] \geq BUFF\_MIN\_CAPACITY$.

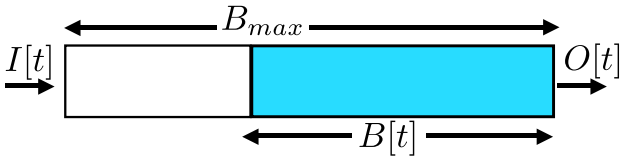


Figure 6.3: Illustration of a possible client-side network buffer state.

2. A **build-up phase**, where the buffer builds up until it reaches $BUFF\_MIN\_CAPACITY$, i.e., until $B[t] < BUFF\_MIN\_CAPACITY$. In this phase, set $O[t] = 0$ until $I[t]/L[t] \approx BUFF\_MIN\_CAPACITY$. In other words, until the buffer contains at least one second of content, no amount of video content can leave the buffer, and thus, a rebuffering event occurs.

3. A **depletion phase** that occurs when $I[t]/L[t] < O[t]$. In this phase, the amount of data leaving the buffer is greater than the amount of data entering the buffer, which causes the buffer to slowly deplete until there is no more data to transmit.

I will now illustrate how a network buffer might transition from one state to another through a general scenario. Consider a video sequence $v[t]$ such that

1. there is a stall event between times t_1 and t_2;

2. there is smooth continuous playback between times t_2 and t_3;

3. and there is another stall event that occurs after t_3.

A possible network buffer scenario can be described as follows:

1. At a discrete time instant $t = t_1$, the buffer starts off empty, but it *must* enter the **build-up phase** before $t = t_2$ for playback to begin.

2. Next, the buffer *must* be in the **steady-state phase** for some time (t_s) between t_2 and t_3.

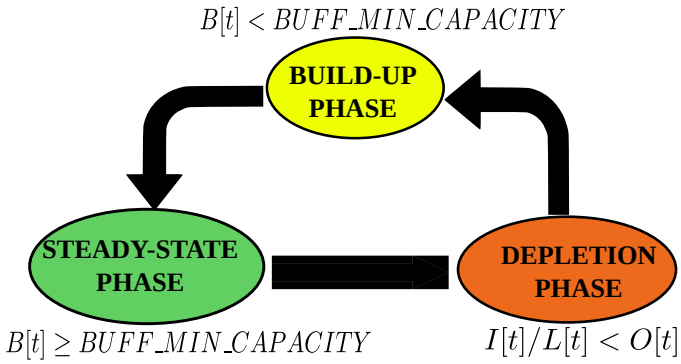


Figure 6.4: Possible states of the client-side network buffer model.

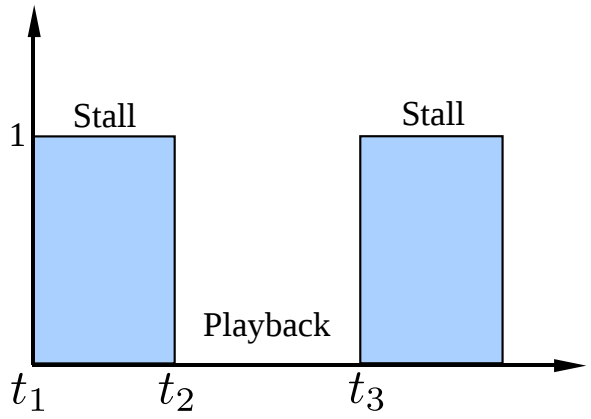


Figure 6.5: Example video sequence with one stall between t_1 and t_2 and another stall at t_3. A value of 0 in this waveform indicates successful video playback, while a value of 1 indicates a stall event.

3. The buffer should next enter the **depletion phase** and *must* be completely empty at $t = t_3$, for a stall to occur at $t = t_3$.

I effectively model the different phases of the buffer as follows:

1. **Modeling the buildup phase (between t_1 and t_2):**
   - I *randomly* sample a discrete time instant t_b between times t_1 and t_2.
   - I fit an exponential function between data points (t_b, 0) and (t_2, $BUFF\_MIN\_CAPACITY$).
2. **Modeling the depletion phase (between t_2 and t_3):**
   - I *randomly* sample a discrete time instance t_d between times t_2 and t_3.
   - I fit an exponential function between data points (t_d, $BUFF\_MIN\_CAPACITY$) and (t_3, 0).

During smooth playback (steady-state phase), the buffer capacity is not necessarily always at $BUFF\_MIN\_CAPACITY$, but instead can fall anywhere in the range $[BUFF\_MIN\_CAPACITY, BUFF\_MAX\_CAPACITY]$. However, I chose not to model the steady-state phase of the network buffer, because it does not cause any quality degradations in the streaming video, and therefore does not influence viewer QoE.

6.1.3 Video Content-Driven Inputs

As mentioned earlier, in addition to stalling events, a viewer's QoE can further be affected by the interplay of other factors such as video quality (due to the presence of distortions), and the spatial and temporal complexities of the video. During the subjective study in [7], I instructed subjects to not judge a video based on their interest in the content, but I did not provide instructions regarding the audio or the video presentation quality. To deepen my understanding in these regards, I sought to study the contributions of these aspects on QoE.

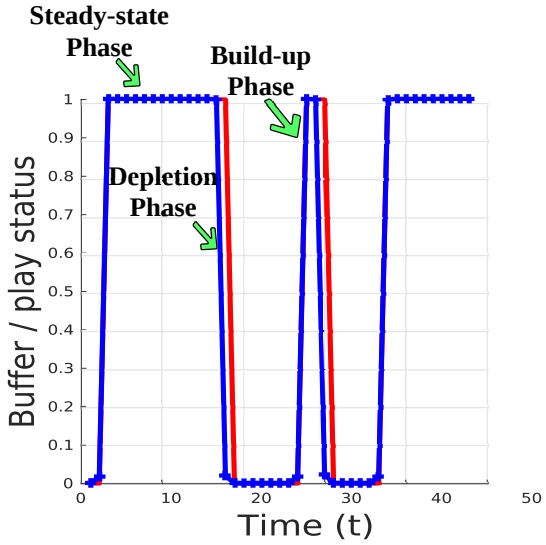


Figure 6.6: Illustrating a possible client-side buffer state (in blue) for a given playback state (in red). Best viewed in color.

Perceptual Video Quality

Perceptual video quality can be defined as the quality of a digital video as *perceived* by human observers, as a reaction to the presence of different forms of spatial and temporal distortions. Rebuffering events, while a form of distortion, do not fall in this category. Bitrate variations and rebuffering events co-occur in streaming videos, and although rebuffering events are more likely to dominate a viewer's QoE, rapid bitrate variations can also significantly impact an end user's dynamic viewing experience and must be accounted for when designing a QoE predictor.

Towards this end, I incorporate either a full-reference, a reduced-reference, or a no-reference video quality assessment (VQA) algorithm [91, 33, 39] in the TV-QoE model, depending on the application scenario. Given the information provided by an objective VQA algorithm, I compute a perceptual VQA score at every second, which provides a continuous-time waveform of perceptual quality. This serves as another continuous-time input to the proposed QoE predictor.

Video Space-Time Perceptual Measurement

Videos contain highly diverse spatial and temporal complexities, and different video contents may be retained differently in memory [159, 160]. These may both interact with past memories of unsatisfactory viewing experiences, e.g., caused by rebuffering events or bitrate drops. The perceptual video quality input designed in Sec. 6.1.3 is not sufficient to capture this aspect, so I chose to use a variant of a spatial-temporal metric called *scene criticality* [161]. Let F_n denote the luminance component of a video frame at instant n, and (i, j) denote spatial coordinates within the frame. A frame filtered with the spatial Sobel operator [162] is denoted as $Sobel(F_n)$. Also define the frame difference operation $M_n(i, j) = F_n(i, j) - F_{n-1}(i, j)$. As formulated in [139], spatial perceptual information (SI) and temporal perceptual information (TI) measurements are computed as

$$SI[n] = STD_{space}\Big[Sobel(F_n(i, j))\Big], \tag{6.9}$$

$$TI[n] = STD_{space}\Big[M_n(i, j)\Big], \tag{6.10}$$

where STD_{space} denotes the standard deviation computed over all the pixels of a given image (F_n or M_n). These are simple, widely used measurements of video activity [139].

By combining these quantities, a continuous-time scene criticality input at every n is arrived at:

$$Criticality[n] = log_{10}\Big[SI[n] + TI[n]\Big]. \tag{6.11}$$

I study the efficacies of each of these content-driven inputs on continuous-time QoE prediction in Sec. 6.4.

6.2 Training a Continuous-Time QoE Predictor

6.2.1 Hammerstein-Wiener Model

When designing a dynamic model that can accurately predict perceived QoE, structural simplicity and computational efficiency are highly desirable. Moreover, the dynamic model should also crucially account for the affects of subjective hysteresis and memory on viewers' QoE [140, 163]. While a simple linear system model would be desirable, human visual responses contain numerous nonlinearities [164, 165, 166], which should also be modeled.

Thus, towards simultaneously capturing the nonlinearities in human visual responses and the hysteresis effect, I employ a classical nonlinear temporal system called the Hammerstein-Wiener (HW) model [167]. The core of the HW model is a linear filter with memory [167] to capture the hysteresis effect, with an input point nonlinearity of a very general form to allow the model to learn nonlinearities. This simple design makes it possible to capture both linear and nonlinear aspects of human behavioral responses. The output linear scaling block simply scales the output of the linear filter to continuous-time quality scores. Figure 6.7 shows a block diagram of the single-input single-output (SISO) Hammerstein-Wiener model that I use.

The linear filter block of the proposed model (See Fig. 6.7) has the following form:

$$\begin{aligned} x[t] &= \sum_{d=0}^{n_b} b_d w[t-d] + \sum_{d=1}^{n_f} f_d x[t-d] \\ &= \mathbf{b}^T (w)_{t-n_b:t} + \mathbf{f}^T (x)_{t-n_f:t-1}, \end{aligned} \tag{6.12}$$

where $w[t]$ is the output of the nonlinear input block at time t. The parameters n_b and n_f define the model order, while the coefficients $\boldsymbol{b} = (b_1, ..., b_{n_b})^T$ and $\boldsymbol{f} = (f_1, ..., f_{n_f})^T$ are learned.

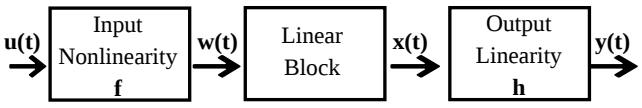


Figure 6.7: Block diagram representing the structure of a Hammerstein-Wiener model.

At the input, I process the signal with a generalized sigmoid function of the form

$$w[t] = \beta_3 + \beta_4 \frac{1}{1 + \exp(-(\beta_1 u[t] + \beta_2))}. \tag{6.13}$$

The output block, which scales the output of the linear IIR filter to a continuous-time QoE prediction, is a simple linear function of the form

$$y[t] = \gamma_1 x[t] + \gamma_2, \tag{6.14}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_4)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T$ are also learned.

6.2.2 An Ensemble of Hammerstein-Wiener Models

The HW model is the building block of the proposed continuous-time QoE predictor. Each of the distortion-informative continuous-time inputs (detailed in Sec. 6.1) is independently used to train a HW model, thereby leading to an ensemble of M HW models. My next task is to accurately combine them to jointly model the interactions between these factors. Formally, if $Y_i \ \forall i = 1, 2, ..M$ are the continuous-time outputs predicted from each HW model (HW_i), then

$$Y_{combined} = \mathbf{\Phi}(Y_1, Y_2, Y_3, ..Y_M), \tag{6.15}$$

where $\mathbf{\Phi}$ is a function that maps the individual outputs to a combined desired output $Y_{combined}$. In this case, I have a total of 9 inputs (7 stall-derived and 2 content-derived) to design an ensemble of $M = 9$ SISO HW models. I chose to

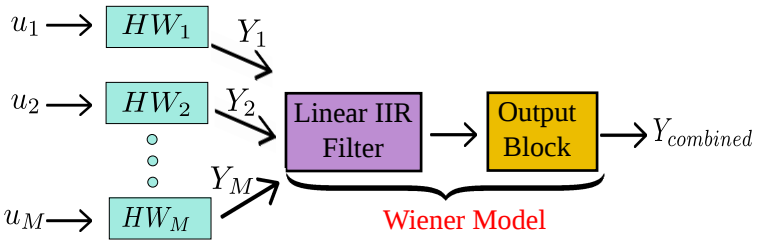


Figure 6.8: The multi-stage framework for predicting dynamic QoE.

strategically combine these models by learning $\mathbf{\Phi}$ via the following two alternative approaches:

Multi-Stage Approach

In this approach, I utilize the predictions from each HW model (HW_i) to train another Wiener model[2] (Fig. 6.8). Specifically, each individual prediction serves as input to another linear filter (in the second stage), followed by an output linearity block. Thus, the model in the second stage is a MISO (Multiple Input, Single Output) Wiener model. Given a test video's distortion-informative inputs, I use the trained multi-stage framework to directly derive the final continuous-time QoE prediction $Y_{combined}$.

Although I utilized a two-stage model here, I note that this can be easily extended to more stages if desired. For example, by training separate MISO Wiener models for stall-derived and content-derived inputs, then fusing them using another MISO model, a third stage could be added to the framework, and so on.

Multi-Learner Approach

I denote the continuous-time output of a HW_i model for a given video content of length V seconds as $Y_i = [y_{i1}, y_{i2}, y_{i3}, ...y_{iV}]$. In this approach, I first construct a set of instance-label pairs for each video content, $(\bar{\boldsymbol{y}}_{\boldsymbol{n}}^{\boldsymbol{I}}, y_n^L)\ \forall\ n = 1, 2, ..V$, where y_n^L is

[2] A Hammerstein-Wiener model without an input non-linearity block is a Wiener model [168].

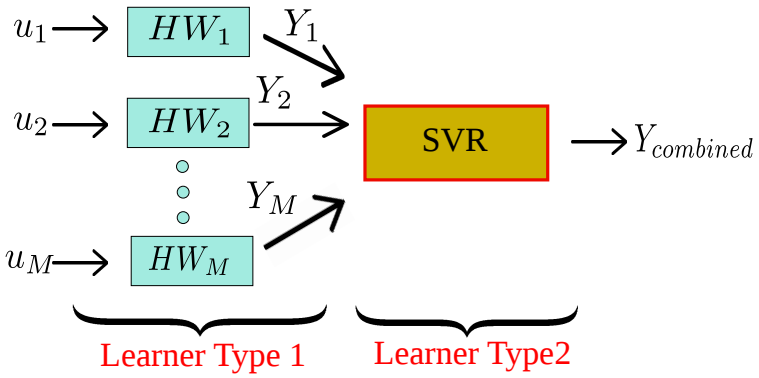


Figure 6.9: The multi-learner framework for predicting dynamic QoE.

the ground truth subjective QoE at time instant n and $\bar{\boldsymbol{y}}_{\boldsymbol{n}}^{\boldsymbol{I}} = [y_{1n}, y_{2n}, ...y_{Mn}]$ are the predictions from each of the M HW models at time instant n. Using the instant-label pairs of all the video contents in the training set, I train a support vector regressor (SVR) to learn the mapping function $\boldsymbol{\Phi}$. I illustrate this learning framework in Fig. 6.9. Thus, in this approach, I use multiple learners: Hammerstein-Wiener models that predict the continuous-time outputs Y_i, and an SVR that learns $\boldsymbol{\Phi}$. Given a test video's distortion-informative inputs, I use the pre-trained HW models and $\boldsymbol{\Phi}$ to directly derive $Y_{combined}$ using an SVR. Since the SVR is trained on the predictions of other learners, it is a *meta-learner* [157]. Other learners (random forests, multilayer perceptron, etc...) could also be used in place of the SVR.

6.2.3 Advantages of the Proposed Dynamic Frameworks:

- **Structural Flexibility:** The proposed ensemble framework is extremely flexible, since it can be further supplemented with any number of additional inputs (or by eliminating any ineffective ones), without changing the general structure of the model.

- **Computational Efficiency:** Each of the SISO HW models are extremely fast (the average training time on a video of average length 86 seconds was

0.54 seconds[3]). The models can also be trained in parallel to improve the overall computation time on a test video.

- **Modeling the effects of Memory:** The long-term and short-term effects of memory on viewing experience could be easily modeled by adding more SISO HW models to the ensemble using the same kinds of distortion-informative dynamic inputs, but with varied amounts of memory parameters (n_b and n_f defined in Equation (6.12)).

Note that a single MISO HW model could potentially be designed instead of constructing an ensemble of SISO *HW* models. However, this approach has several disadvantages: 1) the train and test times would be very high when training on multiple nonlinearly transformed inputs. 2) a MISO HW model cannot be trained on different inputs in parallel, and 3) jointly learning multiple input non-linear functions requires a very large amount of training data which is not available in any of the existing QoE databases.

6.3 An Overall QoE Predictor with Global Video Features

Although continuous-time QoE predictors are valuable, there is also a need for accurate, computationally efficient overall (end-of-video) QoE predictors that could be used when the resources of the stream-switching controllers are limited or when a different analysis is desired. Thus, I also trained an overall QoE predictor by designing comprehensive global features (listed in Table 6.1) which are derived by effectively encapsulating the aforementioned continuous-time inputs. With regards to the perceptual quality score feature, as I describe in Sec. 6.4.4, I tested different

[3]These runtimes were obtained using MATLAB's implementation of the Hammerstein Wiener model [168] when executed on Ubuntu 14.04 OS with an Intel i7 CPU (single processor) and 32 GB of RAM.

Table 6.1: Description of the proposed global video QoE features.

Type of the feature	Definition
Number of stalls	-
Sum of the lengths of all stalls	-
Rebuffering Rate	$\frac{TotalPlaytime}{TotalVideoLength}$
Frequency of stalling events	$\frac{TotalPlaytime}{NumberofStalls}$
Time since the end of last quality impairment	-
Perceptual Quality Score	-

pooling strategies and different objective VQA algorithms in regards to their ability to derive a single, effective, representative quality score to be used as a feature to feed the proposed global model.

A non-linear mapping was learned between these global features and the corresponding real-valued overall QoE scores of the training videos, using an SVR with a radial basis kernel function. Given any test video's features as input to the trained SVR, a final QoE score may be predicted. The optimal model parameters of the learner were found via cross-validation. My choice of the model parameters was driven by the obvious aim of minimizing the learner's fitting error to the validation data (details in Sec. 6.4).

6.4 Experiments

I evaluated the proposed TV-QoE model and all other currently known continuous-time QoE and global QoE predictors on three different databases: the LIVE Mobile Stall Video Database-II [7], the Waterloo QoE Database [8], and the recent LIVE-Netflix Video QoE Database [9]. Every distorted video in the LIVE Mobile Stall

Video Database-II is afflicted by at least one stalling event. However, 60 of the 180 distorted videos in the Waterloo QoE Database, and 56 of the 112 videos of the LIVE-Netflix Video QoE Database are afflicted only by compression artifacts. Since stall-based inputs are not applicable to videos having only compression artifacts, I constructed two disjoint video sets: V_s and V_c, comprising videos afflicted with only compression artifacts and videos afflicted with combinations of stalling events and compression artifacts (if any), respectively[4]. In each of the experiments I describe below, I evaluated the performance of the various predictors on both of these disjoint video collections, wherever applicable.[5]

For every experiment, each database (and video set) was partitioned into training and testing data (80/20 split) with non-overlapping content. To mitigate any bias due to the division of data, the process of randomly splitting each dataset was repeated 50 times. Since global TV-QoE and one of the compared models (V-ATLAS [169]) are learning-based, in each iteration, a model was trained from scratch on the 80% of the data that was set aside for training, then evaluated on the remaining 20% of the test data. FTW [170] and the Streaming QoE Index (SQI) [8] are training-free algorithms, but for a fair comparison with the learning-based models, I report their performance on the test data alone. For each test split, depending on the type of predictor being evaluated (continuous-time or global), I computed three different metrics as described below:

1. **Continuous-time performance** was evaluated by computing the median of the per-frame correlation and root mean square error (RMSE) between the subjective and the estimated continuous QoE for each distorted test video. The median of these per-video correlations and errors was computed as a performance indicator of the given split.

[4]Skipping stall-based inputs is the same as setting all stall-based inputs to zero, provided that these instances are carefully handled in the feature normalization step.

[5]Note that V_c is the empty set $\emptyset$ for LIVE Mobile Stall Video Database-II.

2. **Overall QoE performance** of a global QoE predictor for a given split was evaluated by computing the correlation and RMSE between the predicted overall QoE and the ground truth overall QoE of the test videos.

For continuous-time as well as global predictors, I report the median Pearson Linear Correlation Coefficient (PLCC), median Spearman Rank-Order Correlation Coefficient (SROCC), and the median RMSE across the 50 test splits. Higher correlation values indicate better performance of a QoE prediction model with better monotonicity and linear accuracy, while lower RMSE values indicate better accuracy of the model. Since SQI and FTW are training-free algorithms, their predictions were passed through a logistic non-linearity [3] mapping them to the ground truth QoE scores before computing PLCC. Furthermore, since the Waterloo QoE Database [8] does not contain ground truth continuous-time subjective scores, I was only able to evaluate global QoE models on that database. The continuous-time TV-QoE predictors were superior to all the compared models on all databases with statistical significance. I report the results of the statistical significance tests that I conducted on the results of every experiment described below in Section 6.4.5.

Parameter selection: To find the optimal parameters for each individual Hammerstein-Wiener QoE prediction model in the ensemble, I determined the model order parameters (n_b, n_f, $\mathbf{b}$, $\mathbf{f}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$), and the input nonlinearities via cross-validation on the LIVE Mobile Stall Video Database-II [7]. Performing a simple grid-search resulted in the values $n_b = 4$ and $n_f = 3$ being chosen as the final model parameters for each of the SISO Hammerstein-Wiener models in the ensemble. I also determined the values of the weights α_1 in (6.1) and α_2 in (6.2) using cross-validation. Specifically, I performed a grid search varying both scalars between 0.1 and 0.7 in steps of 0.1, trained a series of models using the training data, and evaluated the performance of each on the validation data. I found that the models with $\alpha_1 \approx 0.2$ and $\alpha_2 \approx 0.1$ yielded maximum correlation scores, and thus, these

values were used in all the experiments on all the databases.

The parameters of both the Wiener model in the multi-stage framework (Sec. 6.2.2) and the SVR in the multi-learner framework (Sec. 6.2.2) were also determined via cross-validation. I found that the value of $n_b = 1$ and a simple linear output block yielded maximum correlation scores on all the databases.

6.4.1 Performance of Continuous-Time Predictors on LIVE Mobile Stall Video Database-II

First, I evaluated the performance of continuous-time QoE models on the distorted videos of the LIVE Mobile Stall Video Database-II. The results are reported in Table 6.2. Since I proposed two different ways of combining the ensemble of Hammerstein-Wiener models (the multi-learner will be referred to as TV-QoE-1 while the multi-stage learner will be referred to as TV-QoE-2), I report the performance of both of these models. SQI, which is the only other existing continuous-time QoE predictor, uses a per-frame quality metric to compute the spatial quality on each frame. Specifically, the instantaneous QoE (Q_n) at each frame n is computed as the sum of a video presentation quality (P_n), i.e., spatial quality, and a stall-dependent experience quality (S_n). Since the LIVE Mobile Stall Video Database-II does not have reference videos, I relied on a popular per-frame NR-IQA metric, NIQE [33], to compute the continuous-time SQI scores.

It may be observed from Table 6.2 that the proposed set of dynamic inputs and learners significantly outperform SQI. It may also be observed that the proposed multi-learner approach (Sec. 6.2.2), which uses an SVR with a nonlinear radial basis kernel function performs better than the multi-stage approach (Sec. 6.2.2), for every given input combination. I will show next that NIQE scores [33] are poor indicators of instantaneous QoE, so including NIQE as an input slightly deteriorates the performance of TV-QoE. Figure 6.10 illustrates a few examples of the ground

Table 6.2: Performance of continuous-time QoE predictors on the LIVE Mobile Stall Video Database-II. Note that the per-frame QoE values lie in the range $[0, 100]$. The best performing model is indicated in bold font.

Learner Type		PLCC	SROCC	RMSE
Multi-learner (TV-QoE-1)	Stall only Inputs	0.9599	**0.9474**	4.6305
Multi-learner	Stall only Inputs + Scene Criticality	**0.9601**	0.9444	**4.4241**
Multi-learner	Stall only Inputs + Scene Criticality + NIQE [33]	0.9297	0.9262	5.5052
Multi-stage (TV-QoE-2)	Stall only Inputs	0.9394	0.9378	5.3155
Multi-stage	Stall only Inputs + Scene Criticality	0.9429	0.9330	5.0244
Multi-stage	Stall only Inputs + Scene Criticality + NIQE [33]	0.9348	0.9162	5.2517
	SQI + NIQE [8]	0.8348	0.6988	4.4901

truth and the predicted continuous-time QoE waveforms of a few test videos from the proposed approach (using the multi-learner approach and the stall-based inputs in isolation). It may be observed that the proposed model does not overfit to the existing dataset, but instead attempts to accurately predict the varying trends in each dynamic QoE prediction. In some of the examples, it may be observed that the QoE predictions occasionally fall outside of the 95% confidence interval, despite maintaining a strong monotonic relationship with the ground truth dynamic QoE.

6.4.2 Intrinsic Analysis of the Individual Dynamic Inputs

To better understand the relationship between the proposed input set and the dynamic QoE, I trained separate Hammerstein-Wiener Models on each input on the same 50 random, non-overlapping train and test splits of the LIVE Mobile Stall Video Database-II, as were used in Sec. 6.4.1. I report the median SROCC and PLCC scores over these 50 iterations in Table 6.4. These results illustrate the degree to which each of these inputs accurately predict perceived QoE, while also justifying the choice of the proposed inputs. It may also be observed that the per-second

Table 6.3: Performance of continuous QoE predictors on the video set V_c of the LIVE-Netflix Video QoE Database. Note that the per-frame QoE values lie in the range $[-2.26, 1.52]$. The best performing model is indicated in bold font.

Learner Type		Quality Predictor	PLCC	SROCC	RMSE
Multi-learner (TV-QoE-1)	Scene Criticality	NIQE [33]	0.2412	0.1711	0.5462
Multi-learner	Scene Criticality	SSIM [39]	0.6314	0.3998	0.4723
Multi-learner	**Scene Criticality**	**STRRED [91]**	**0.6733**	**0.5776**	0.3965
Multi-stage (TV-QoE-2)	Scene Criticality	NIQE [33]	0.2512	0.1594	0.5609
Multi-stage	Scene Criticality	SSIM [39]	0.6387	0.3786	0.4732
Multi-stage	Scene Criticality	STRRED [91]	0.6728	0.5715	0.3969
	SQI	NIQE [33]	0.2123	0.1408	**0.3102**
	SQI	SSIM [39]	0.2392	0.0934	0.3136
	SQI	STRRED [91]	0.1984	0.1917	0.3214

NIQE scores[33] performed rather poorly at predicting QoE scores when videos were afflicted by stalling events. Thus including this input when conducting continuous-time QoE prediction degrades performance (Sec. 6.4.1). Of course, the NIQE model utilizes only spatial information and does not benefit from any reference signal or training process.

6.4.3 Performance of Continuous-Time Predictors on the LIVE-Netflix Video QoE Database

As mentioned, I divided the entire collection of 112 videos in the LIVE-Netflix Video QoE Database into two disjoint video sets: V_c and V_s. Videos belonging to V_s contain both compression and stalling artifacts, while those in V_c contain only compression artifacts. Hence, on the video set V_c, I did not use any stall-based inputs[6], relying instead only on the content-driven inputs. I report the performance

[6]This is same as setting stall-based inputs to zero.

Table 6.4: Contribution of the proposed stall and video content-based dynamic inputs towards continuous-time QoE on the 50 test splits of the LIVE Mobile Stall Video Database-II [7]. The video content-based inputs are italicized.

Dynamic Inputs	PLCC	SROCC
Stall position	0.6946	0.6962
Number of stalls	0.4399	0.4744
Time since previous stall	0.9109	0.8919
Stall density	0.6264	0.6945
Buffer model	0.7893	0.7829
Frequency	0.6812	0.7640
Rebuffering rate	0.5554	0.5495
Scene Criticality	*0.5701*	*0.4399*
NIQE [33]	*0.0758*	*0.0811*

of TV-QoE-1, TV-QoE-2, and SQI in Table 6.3. For videos in V_s, however, I used both stall-based as well as content-based inputs, and report the performance in Table 6.5. Furthermore, I also considered scenarios where either a FR, RR, or NR VQA model would be incorporated into the QoE predictor. It may be observed from these results that TV-QoE significantly outperforms SQI on both video sets, especially when videos were afflicted by both stalls and compression artifacts. Further, the multi-learner approach (TV-QoE-1) yielded better performance than the multi-stage approach (TV-QoE-2) on both video sets of the LIVE-Netflix Video QoE Database.

6.4.4 Performance of Global QoE Predictors

Next, I evaluated the performance of the proposed global features (Table 6.1) and other global QoE predictors under identical train/test settings on all three databases and report the results in Tables 6.6, 6.7, and 6.8. I computed the perceptual quality

Table 6.5: Performance of continuous QoE predictors on the video set V_s of the LIVE-Netflix Video QoE Database. Note that the per-frame QoE values lie in the range $[-2.26, 1.52]$. The best performing model is indicated in bold font.

Learner Type		Quality Predictor	PLCC	SROCC	RMSE
Multi-learner (TV-QoE-1)	**Stall only Inputs**	-	**0.9131**	**0.8579**	0.3536
Multi-learner	Stall only Inputs + Scene Criticality	NIQE [33]	0.9059	0.8306	0.3672
Multi-learner	Stall only Inputs + Scene Criticality	SSIM [39]	0.8694	0.7820	0.3151
Multi-learner	Stall only Inputs + Scene Criticality	STRRED [91]	0.8905	0.8061	**0.3004**
Multi-stage (TV-QoE-2)	Stall only Inputs		0.8800	0.7970	0.3851
Multi-stage	Stall only Inputs + Scene Criticality	NIQE [33]	0.8738	0.7775	0.4026
Multi-stage	Stall only Inputs + Scene Criticality	SSIM [39]	0.8345	0.7248	0.3584
Multi-stage	Stall only Inputs + Scene Criticality	STRRED [91]	0.8496	0.7471	0.3777
	SQI	NIQE [33]	0.6821	0.4281	0.3433
	SQI	SSIM [39]	0.6892	0.3793	0.3450
	SQI	STRRED [91]	0.6705	0.3275	0.3581

scores using various quality predictors and tested several pooling strategies to derive a single quality score from the per-frame perceptual quality score, to be used as a global feature. I only report the results obtained from the pooling strategy that yielded the best performance. Note that the LIVE Mobile Stall Video Database-II does not contain pristine videos, so I relied on the no-reference (NR) picture quality model NIQE [33] to supply VQA scores on this database. For the other two databases, the reference videos are available, so I report the performance using full-reference (SSIM [39]), reduced-reference (ST-RRED [91]), and no-reference VQA models (NIQE [33]). Note that when evaluating the proposed global QoE predictor on the video sets V_c of different databases, I utilized only the video content-based inputs, since the stall-informative global features do not capture any information. When evaluating the proposed global QoE predictor and V-ATLAS, I trained an SVR with a radial basis kernel function, by separately finding the optimal SVR

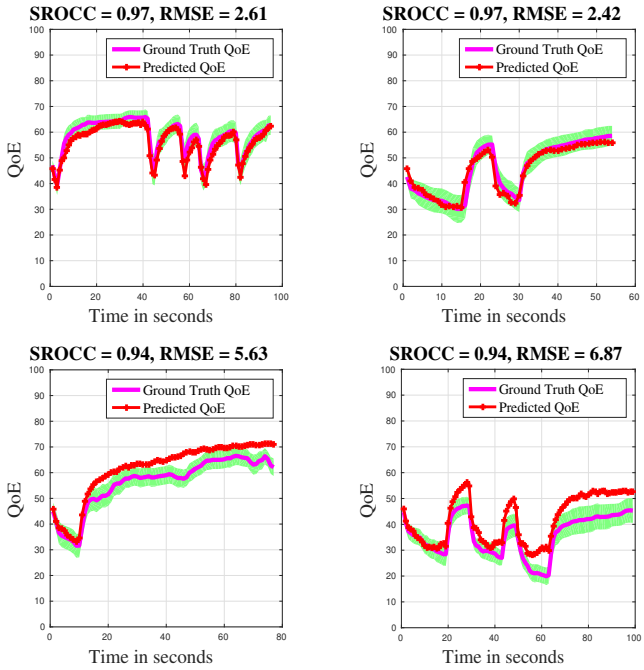


Figure 6.10: Some examples of the continuous-time predictions obtained from the proposed algorithm (indicated in red) on different test video sequences of the LIVE Mobile Stall Video Database-II. The ground truth dynamic QoE response is indicated in magenta and the associated 95% confidence interval derived from the responses from individual subjects is indicated in green. Spearman Rank Ordered Correlation (SROCC) and Root Mean Squared Error (RMSE) between the instantaneous predicted and ground truth QoE is also reported in each plot.

parameters via cross-validation on all three databases.

I found that the proposed global QoE predictor outperforms all existing QoE predictors on the LIVE Mobile Stall Video Database-II (Table 6.6). It is also clear from these results that including NIQE as a global perceptual quality metric benefits the QoE prediction. The scatter plots of the predicted and the ground truth QoE scores for one test split are illustrated in Fig. 6.11. With regards to the Waterloo QoE Database, there are a couple of oddities in the results arising from the database design. Each of the 120 videos in the Waterloo QoE Database belonging to V_s are of the same length and each contains one stalling event of

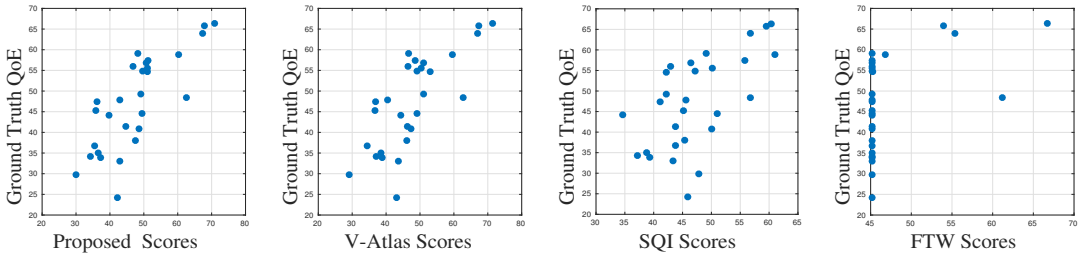


Figure 6.11: Scatter plots of the ground truth overall QoE scores and the predicted overall QoE scores obtained on a single test split from our different global QoE predictors on the LIVE Mobile Stall Video Database-II [7]. Global TV-QoE is statistically significant than all other global QoE predictors.

Table 6.6: Performance of global QoE models on the LIVE Mobile Stall Video Database-II [7]. Note that the final QoE values lie in the range $[0, 100]$. The best performing model is indicated in bold font.

	PLCC	SROCC	RMSE
TV-QoE Global Stall Features	0.7099	0.6836	8.7609
TV-QoE Global Stall Features + Max-Pooled NIQE	**0.7757**	**0.7797**	**7.7914**
SQI + NIQE	0.4828	0.4565	9.8512
V-ATLAS + Max-Pooled NIQE	0.7541	0.7572	8.1541
FTW	0.4411	0.6689	10.5074

duration 5 seconds. In this peculiar scenario, the otherwise different global TV-QoE and V-ATLAS features capture exactly the same information, thereby yielding identical performances (Table 6.7). Moreover, since the FTW model [170] is based on only two features (the number and the summed length of stalls), it predicts the same quality score on all video contents in the Waterloo QoE Database. On the LIVE-Netflix QoE Database, the global TV-QoE model competes very well with the performances of V-ATLAS and SQI (Table 6.8).

Table 6.7: Performance of global QoE predictors on the Waterloo QoE Database [8]. Note that the final QoE values lie in the range $[0, 100]$. The best performing model is indicated in bold font.

		V_s			V_c		
	Quality Predictor (Pool type = mean)	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Global TV-QoE Features	NIQE [33]	0.3200	0.3216	14.8681	0.2983	0.3356	19.9181
Global TV-QoE Features	SSIM [39]	**0.8660**	0.8604	8.5568	**0.8956**	**0.8531**	**12.2217**
SQI [8]	NIQE [33]	0.3046	0.4134	14.1765	0.2393	0.3357	18.6965
SQI [8]	SSIM [39]	0.8582	**0.8623**	**7.5603**	0.8910	0.8462	12.2412
V-ATLAS [169]	NIQE [33]	0.3200	0.3216	14.8681	0.2983	0.3356	19.9181
V-ATLAS [169]	SSIM [39]	**0.8660**	0.8604	8.5568	**0.8956**	**0.8531**	**12.2217**
FTW [170]		NaN	NaN	-	-NA-	-NA-	-NA-

6.4.5 Statistical Significance of Global and Dynamic QoE Predictors on Different QoE Databases

In this section, I report the results of the paired sample t-tests that I had conducted between SROCC values obtained from different global QoE predictors on three different databases in Tables 6.9 -6.13 and different continuous-time QoE predictors in Tables 6.14 - 6.16.

6.5 Conclusions

In this Chapter, I presented a continuous-time video QoE predictor that effectively captures the effects of a variety of QoE-influencing factors, and that models the client-side network buffer model, subjective hysteresis, and that is able to accurately predict viewers' instantaneous QoE. I have also designed a global QoE predictor that achieves top performance on all existing QoE databases. The success of these two models encourages the design of quality-aware stream-switching algorithms at either

Table 6.8: Performance of global QoE predictors on the LIVE-Netflix Video QoE Database [9]. Note that the final QoE values lie in the range $[-1.6, 1.6]$. The best performing model is indicated in bold font.

		V_s			V_c		
	Quality Predictor (Pool type = mean)	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
Global TV-QoE Features	NIQE [33]	0.6719	0.3318	0.4126	0.7676	0.4909	0.6616
Global TV-QoE Features	SSIM [39]	0.7821	0.7045	0.3475	0.9030	0.8000	0.7452
Global TV-QoE Features	STRRED [91]	0.8564	0.7591	0.3196	0.9246	0.8091	0.5663
SQI	NIQE [33]	0.1977	0.0272	0.4051	0.4895	0.3773	0.6630
SQI	SSIM [39]	0.6132	0.5500	0.3185	0.8262	0.8000	0.4596
SQI	STRRED [91]	**0.8597**	0.7500	**0.2151**	0.8061	0.8000	0.3820
V-ATLAS	NIQE [33]	0.8165	0.6091	0.3245	0.8170	0.6045	0.6250
V-ATLAS	SSIM [39]	0.7951	0.6591	0.3346	0.9406	0.8545	0.3902
V-ATLAS	STRRED [91]	0.8547	**0.7636**	0.3095	**0.9462**	**0.8591**	**0.3586**
FTW		0.2797	0.2778	0.3984	-NA-	-NA-	-NA-

the client or the server's end which could control the position, location, and length of stalls, given a network bandwidth budget and the end user's device information, such that the end user's QoE is maximized. Such models would greatly benefit both content and network providers.

Table 6.9: Results of the paired sample t-test performed between SROCC values generated by different global QoE predictors on LIVE Mobile Stall Video Database-II. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. Global TV-QoE is denoted as TV-QoE.

	TV-QoE (Stall only)	TV-QoE + NIQE	V-ATLAS + NIQE	SQI + NIQE	FTW
TV-QoE (Stall only)	0	-1	-1	1	1
TV-QoE + NIQE	1	0	1	1	1
V-ATLAS + NIQE	1	-1	0	1	1
SQI + NIQE	-1	-1	-1	0	-1
FTW	-1	-1	-1	1	0

Table 6.10: Results of the paired sample t-test performed between SROCC values generated by different global QoE predictors on the video set V_s of Waterloo QoE Database. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. Global TV-QoE is denoted as TV-QoE.

		TV-QoE		V-ATLAS		SQI	
		NIQE	SSIM	NIQE	SSIM	NIQE	SSIM
TV-QoE	NIQE	0	-1	0	-1	-1	-1
	SSIM	1	0	1	0	1	1
V-ATLAS	NIQE	0	-1	0	-1	-1	-1
	SSIM	1	0	1	0	1	1
SQI	NIQE	1	-1	1	-1	0	-1
	SSIM	1	-1	1	-1	1	0

Table 6.11: Results of the paired sample t-test performed between SROCC values generated by different global QoE predictors on the video set V_c of Waterloo QoE Database. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. Global TV-QoE is denoted as TV-QoE.

		TV-QoE		V-ATLAS		SQI	
		NIQE	SSIM	NIQE	SSIM	NIQE	SSIM
TV-QoE	NIQE	0	-1	0	-1	0	-1
	SSIM	1	0	1	0	1	0
V-ATLAS	NIQE	0	-1	0	-1	0	-1
	SSIM	1	0	1	0	1	0
SQI	NIQE	0	-1	0	-1	0	-1
	SSIM	1	0	1	0	1	0

Table 6.12: Results of the paired sample t-test performed between SROCC values generated by different global QoE predictors on the video set V_s of LIVE Netflix Video QoE Database. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. Global TV-QoE is denoted as TV-QoE.

		TV-QoE			V-ATLAS			SQI		
		NIQE	SSIM	ST-RRED	NIQE	SSIM	ST-RRED	NIQE	SSIM	ST-RRED
TV-QoE	NIQE	0	-1	-1	-1	-1	-1	1	-1	-1
	SSIM	1	0	-1	1	0	-1	1	1	-1
	ST-RRED	1	1	0	1	1	0	1	1	0
V-ATLAS	NIQE	1	-1	-1	0	-1	-1	1	0	-1
	SSIM	1	0	-1	1	0	-1	1	1	-1
	ST-RRED	1	1	0	1	1	0	1	1	0
SQI	NIQE	-1	-1	-1	-1	-1	-1	0	-1	-1
	SSIM	1	-1	-1	0	-1	-1	1	0	-1
	ST-RRED	1	1	0	1	1	0	1	1	0

Table 6.13: Results of the paired sample t-test performed between SROCC values generated by different global QoE predictors on the video set V_c of LIVE Netflix Video QoE Database. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. Global TV-QoE is denoted as TV-QoE.

		TV-QoE			V-ATLAS			SQI		
		NIQE	SSIM	ST-RRED	NIQE	SSIM	ST-RRED	NIQE	SSIM	ST-RRED
TV-QoE	NIQE	0	-1	-1	-1	-1	-1	1	-1	-1
	SSIM	1	0	-1	1	-1	-1	1	0	-1
	ST-RRED	1	1	0	1	-1	-1	1	1	0
V-ATLAS	NIQE	1	-1	-1	0	-1	-1	1	-1	-1
	SSIM	1	1	1	1	0	0	1	1	1
	ST-RRED	1	1	1	1	0	0	1	1	1
SQI	NIQE	-1	-1	-1	-1	-1	-1	0	-1	-1
	SSIM	1	0	-1	1	-1	-1	1	0	-1
	ST-RRED	1	1	0	1	-1	-1	1	1	0

Table 6.14: Results of the paired sample t-test performed between SROCC values generated by different continuous-time QoE predictors on LIVE Mobile Stall Video Database-II. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. TV-QoE-1 denotes the multi-learner approach and TV-QoE-2 denotes the multi-stage approach. The row to the left of NIQE represents the proposed QoE model with stall-derived inputs alone.

		TV-QoE-1		TV-QoE-2		SQI
		-	NIQE	-	NIQE	
TV-QoE-1	-	0	1	1	1	1
	NIQE	-1	0	-1	1	1
TV-QoE-2	-	-1	1	0	1	1
	NIQE	-1	-1	-1	0	1
SQI		-1	-1	-1	-1	0

Table 6.15: Results of the paired sample t-test performed between SROCC values generated by different continuous-time QoE predictors on the video set V_c of LIVE Netflix Video QoE Database. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. TV-QoE-1 denotes the multi-learner approach and TV-QoE-2 denotes the multi-stage approach.

		TV-QoE-1			TV-QoE-2			SQI		
		NIQE	SSIM	ST-RRED	NIQE	SSIM	ST-RRED	NIQE	SSIM	ST-RRED
TV-QoE-1	NIQE	0	-1	-1	0	-1	-1	0	1	0
	SSIM	1	0	-1	1	0	-1	1	1	1
	ST-RRED	1	1	0	1	1	0	1	1	1
TV-QoE-2	NIQE	0	-1	-1	0	-1	-1	0	1	0
	SSIM	1	0	-1	1	0	-1	1	1	1
	ST-RRED	1	1	0	1	1	0	1	1	1
SQI	NIQE	0	-1	-1	0	-1	-1	0	1	0
	SSIM	-1	-1	-1	-1	-1	-1	-1	0	-1
	ST-RRED	0	-1	-1	0	-1	-1	0	1	0

Table 6.16: Results of the paired sample t-test performed between SROCC values generated by different continuous-time QoE predictors on the video set V_s of LIVE Netflix Video QoE Database. '1,' '0,' '-1' indicate that the algorithm in the row is statistically superior, equivalent, or inferior to the algorithm in the column respectively. TV-QoE-1 denotes the multi-learner approach and TV-QoE-2 denotes the multi-stage approach. The row to the left of NIQE represents the proposed QoE model with stall-derived inputs alone.

		TV-QoE-1				TV-QoE-2				SQI		
		-	NIQE	SSIM	ST-RRED	-	NIQE	SSIM	ST-RRED	NIQE	SSIM	ST-RRED
TV-QoE-1	-	0	1	1	1	1	1	1	1	1	1	1
	NIQE	-1	0	1	0	1	1	1	1	1	1	1
	SSIM	-1	-1	0	-1	0	0	1	1	1	1	1
	ST-RRED	-1	0	1	0	1	1	1	1	1	1	1
TV-QoE-2	-	-1	-1	0	-1	0	1	1	1	1	1	1
	NIQE	-1	-1	0	-1	-1	0	1	1	1	1	1
	SSIM	-1	-1	-1	-1	-1	-1	0	-1	1	1	1
	ST-RRED	-1	-1	-1	-1	-1	-1	1	0	1	1	1
SQI	NIQE	-1	-1	-1	-1	-1	-1	-1	-1	0	1	1
	SSIM	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1
	ST-RRED	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0

Chapter 7

Conclusion

There is a growing awareness of the importance of understanding, predicting, and monitoring the perceptual quality of images and videos that are delivered to human viewers. The leaps of progress made thus far in vision science and image engineering continue to push the boundaries of achievable perceptual quality prediction. Existing quality prediction approaches only tackle a confined set of image and video distortions. With a motivation to address the scenarios encountered in real-world visual media applications, this dissertation focused on designing quality models that are robust to factors such as authentic mixtures of naturally occurring distortions in pictures and videos, varied visual content, network conditions, and display devices.

Specifically, I have introduced two image and video quality assessment databases and two separate image and video quality predictors that effectively predict the (temporal) quality of pictures and videos in-the-wild. I have demonstrated my algorithms successfully on a variety of datasets. These databases and automated tools could in turn have tremendous practical and industrial significance and could help in delivering the best possible visual content to end users.

I described an image database that models authentic picture distortions and an online study framework for subjective studies in Chapter 3. This database greatly

contrasts with existing benchmark IQA databases that are limited to single, synthetic distortions with subjective studies conducted in carefully calibrated laboratory settings. The online study framework is an important contribution as it an indispensable resource for conducting several other studies on tone-mapped HDR pictures [171], X-ray images, and mobile videos.

The automatic image quality predictor, FRIQUEE that I introduced in Chapter 4 is the first of its kind to address authentic image distortions. Such algorithms are invaluable for applications such as source inspection of user-uploaded pictures and videos on social-media applications such as Facebook, Instagram, and YouTube, and can guide content-based image/video compression strategies. It can also drive the next-generation mobile cameras that support on-device quality-adaptive picture and video capturing algorithms. These tools can serve as an additional factor for the image search ranking algorithms – identifying and culling low quality images in the top search results can greatly improve user experience.

With an ultimate goal to promptly and accurately predict an end user's instantaneous QoE, I conducted a subjective study on a video collection afflicted with simulated network impairments (detailed in Chapter 5). Contrary to previous studies that focused only on end-point subjective scores on limited video data collections, I obtained rich subjective per-frame data on a large video dataset and designed a novel subject rejection strategy for temporal data. Building on the insights obtained from a thorough analysis of the human behavioral responses to time-varying video quality, I designed a dynamic QoE predictor that requires only a video segment (to be transmitted), the instantaneous network conditions, and the end user's device information for accurate, instantaneous QoE prediction. Since existing approaches only predict end-point quality score for a given video with dynamic impairments, they do not offer sufficient benefit to the existing stream-switching algorithms. On the other hand, the TV-QoE model that I designed and described in Chapter 6

could easily have a direct and immediate impact on existing adaptive bitrate allocation protocols and stream-switching algorithms that are used in the client players of content providers such as YouTube and Netflix, thereby propelling user-centric mobile network planning and management.

In any case, given the current explosive growth rate of photos and videos on numerous data-driven services, finding effective and efficient ways to provide a high-quality of viewing experience is a pressing concern. The overarching goal of this work is to model image and video statistics by thoroughly understanding how naturally occurring distortions change these statistics thus designing low-level perceptual quality models. Though I have concentrated on applications for visible light pictures and videos, the perceptual quality models presented in this dissertation are absolutely suitable for applications involving alternative imaging modalities – infrared images, satellite images, and so on. I hope that this dissertation has laid a concrete foundation for exciting and practical prospective avenues of quality assessment research.

Bibliography

[1] D. Ghadiyaram and A. C. Bovik, "Crowdsourced study of subjective image quality," *Asilomar Conf. Signals, Syst. Comput.*, Nov 2014.

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

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

[4] D. Jayaraman, A. Mittal, A. K. Moorthy and A. C. Bovik, "Objective quality assessment of multiply distorted images," *Asilomar Conf. Signals, Syst. Comput.*, pp. 1693–1697, 2012.

[5] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, "Color image database tid2013: Peculiarities and preliminary results," *Proc. of 4th European Workshop on Visual Info. Proc.*, pp. 106–111, 2013.

[6] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, 2010.

[7] D. Ghadiyaram, J. Pan, and A.C. Bovik, "A subjective and objective study of stalling events in mobile streaming videos," *IEEE Trans. Circ. Syst. for Video Tech.*, 2017, (under review) [LINK].

[8] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality-of-experience prediction for streaming video," in *IEEE Int. Conf. on Multimedia and Expo*. IEEE, 2016, pp. 1–6.

[9] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," in *IEEE Trans. Img. Proc.*, (submitted).

[10] B. A. Olshausen and D. J. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, vol. 381, no. 6583, pp. 607, 1996.

[11] D. Ruderman, "The statistics of natural images," *Netw. Comput. Neural Syst.*, 1994.

[12] A. Moorthy and A. C. Bovik, "Visual quality assessment algorithms: what does the future hold?," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 675–696, 2011.

[13] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, J. David, Y. Dan, B. A. Olshausen, J. Gallant, N. C. Rust, "Do we know what the early visual system does?," *Journal of Neuroscience*, vol. 25, no. 46, pp. 10577–10597, 2005.

[14] D. W. Dong and J. J, Atick, "Temporal decorrelation: a theory of lagged and non-lagged responses in the lateral geniculate nucleus," *Network Computation in Neural Systems*, vol. 6, no. 2, pp. 159–178, 1995.

[15] B. A. Olshausen and D. J. Field, "How close are we to understanding V1?," *Neural computation*, vol. 17, no. 8, pp. 1665–1699, 2005.

[16] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer.*, vol. 2, no. 7, pp. 1160–1169, 1985.

[17] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, 1990.

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

[19] L. Wiskott, J. M. Fellous, N. Kruger, and C. V. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 19, no. 7, pp. 775–779, 1997.

[20] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 18, no. 8, pp. 837–842, 1996.

[21] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *IEEE Conf. Comp. Vision and Pattern Recog.*, 2005, vol. 2, pp. 994–1000.

[22] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area v2," in *Adv. in Neural Info. Proc. Syst.*, 2008, pp. 873–880.

[23] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Am., A.*, 1987.

[24] D. J. Tolhurst, Y. Tadmor, and T. Chao, "Amplitude spectra of natural images," *Ophthalmic and Physiological Optics*, vol. 12, no. 2, pp. 229–232, 1992.

[25] P. J. B. Hancock, R. J. Baddeley, and L. S. Smith, "The principal components of natural images," *Network: Computation in Neural Syst.*, vol. 3, no. 1, pp. 61–70, 1992.

[26] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters," *Vision Research*, 1997.

[27] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli, "Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons," *Statistical Theories of the Brain*, pp. 203–222, 2002.

[28] R. Sekuler and R. Blake, *Perception*, McGraw Hill, 2002.

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

[30] 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.

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

[32] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 305–312, 2011.

[33] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Sig. Proc. Letters.*, vol. 20, no. 3, pp. 209–212, 2012.

[34] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014.

[35] 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, 2014.

[36] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Int. Conf. Image Process.*, pp. 3129–3138, 2011.

[37] C. T. Vu, T. D. Phan, and D. M. Chandler, "S3: A Spectral and Spatial Measure of Local Perceived Sharpness in Natural Images," *IEEE Trans. on Image Proc.*, vol. 21, no. 3, pp. 934–945, 2012.

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

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

[40] Z. Wang, E. P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," *Proc. Asilomar Conf. Signals, Syst. Comput.*, vol. 2, pp. 1398–1402, 2003.

[41] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, 2009.

[42] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," *IEEE Int. Workshop Qual. Multimedia Experience*, pp. 87–91, 2009.

[43] S. Varadarajan and L. J. Karam, "An improved perception-based no-reference objective image sharpness metric using iterative edge refinement," *Proc. IEEE Int. Conf. Image Process.*, pp. 401–404, 2008.

[44] S. Varadarajan and L. J. Karam, "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling," *Proc. IEEE Int. Conf. Image Process.*, pp. 369–372, 2008.

[45] H. R . Sheikh, A. C. Bovik, and L. K. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, 2005.

[46] 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," *Proc. Pacific Rim Conf. Multimedia, Adv. Multimedia Inf. Process.*, pp. 894–897, 2008.

[47] R. Barland and A. Saadane, "Reference free quality metric using a region-based attention model for jpeg-2000 compressed images," *Proc. SPIE*, vol. 6059, pp. 605905–1 – 605905–10, 2006.

[48] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Adv. of Modern Radio Electron.*, vol. 10, no. 4, pp. 3045, 2009.

[49] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, 1995.

[50] I. Avcibas, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *J. Elec. Imag.*, vol. 11, no. 2, pp. 206–223, 2002.

[51] A. Mayache, T. Eude, and H. Cherifi, "A comparison of image quality models and metrics based on human visual sensitivity," *IEEE Int. Conf. Image Proc.*, pp. 409–413, 1998.

[52] P. Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database," [Online] Available: http://www.irccyn.ec-nantes.fr/ivcdb/.

[53] T. K. Huang, C. J. Lin, and R. C. Weng, "Ranking individuals by group comparisons," *23rd. Int. Conf. Machine Learn.*, pp. 425–432, 2006.

[54] M. Montague and J. A. Aslam, "Condorcet fusion for improved retrieval," *Proc. 11th Int. Conf. Inform. Knowledge Manage.*, pp. 538–548, 2008.

[55] O. Dykstra, "Rank analysis of incomplete block designs: A method of paired comparisons employing unequal repetitions on pairs," *Biometrics*, vol. 16, no. 2, pp. 176–188, 1960.

[56] K. T. Chen, C. C. Wu, Y. C. Chang, and C. L. Lei, "A crowdsourceable qoe evaluation framework for multimedia content," *Proc. of 17th Int. Conf. on Multimedia*, pp. 491–500, 2009.

[57] G. Qiu and A. Kheiri, "Social image quality," *Proc. IS&T/SPIE*, vol. 7867, pp. 78670S, 2011.

[58] D. R. Rasmussen, "The mobile image quality survey game," *Proc. SPIE*, vol. 8293, pp. 82930I–82930I–12, 2012.

[59] M. D. Harris, G. D. Finlayson, J. Tauber, and S. Farnand, "Web-based image preference," *J. of Imag. Science and Techn.*, vol. 57, no. 2, 2013.

[60] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," *Proc. of 18th IEEE Int. Conf. on Image Proc.*, pp. 3097–3100, 2011.

[61] T. Grzywalski, A. Luczak, and R. Stasinski, "Internet based subjective assessment of image quality experiment," *Proc. Int. Conf. Systems, Signals and Image Proc.*, pp. 1–4, 2011.

[62] C. Keimel, J. Habigt, C. Horch, and K. Diepold, T. Grzywalski, A. Luczak, and R. Stasinski, "Qualitycrowd – a framework for crowd-based quality evaluation," *Picture Coding Symp.*, pp. 245–248, 2012.

[63] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based qoe testing," *Int. Conf. Commun.*, pp. 245–248, 2014.

[64] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment," *ACM Int. Conf. on Multimedia*, pp. 359–368, 2012.

[65] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin, "Random partial paired comparison for subjective video quality assessment via hodgerank," *ACM Int. Conf. on Multimedia*, pp. 393–402, 2011.

[66] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial hodge theory," *Mathematical Programming*, vol. 127, no. 1, pp. 203–244, 2011.

[67] P. Ye and D. Doermann, "Active sampling for subjective image quality assessment," *Proc. IEEE Conf. on Comp. Vision and Pattern Recog.*, 2014.

[68] D. Ghadiyaram and A. C. Bovik, "LIVE In the Wild Image Quality Challenge Database," [Online] Available: http://live.ece.utexas.edu/research/ChallengeDB/index.html.

[69] B Mandelbrot, *The fractal geometry of nature*, Freeman, 1982.

[70] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal Processing: Image Communication*, vol. 29, no. 8, pp. 856–863, 2014.

[71] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.

[72] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," *Proc. Int. Conf. on Comp. Vision and Pattern Recog*, 2014.

[73] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," *Int. Conf. Comput. Vis. Pattern Recognit.*, pp. 305–312, 2011.

[74] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, 1995.

[75] MPEG Requirements Group, "ISO/IEC FCD 23001-6 Part 6: Dynamics adaptive streaming over HTTP (DASH)," 2011.

[76] R. Pantos and W. May, "HTTP live streaming," *IETF draft*, June, 2010, [Online] Available: https://tools.ietf.org/html/draft-pantos-http-live-streaming-04.

[77] L. De Cicco, S. Mascolo, and V. Palmisano, "Feedback control for adaptive live video streaming," *Proceedings of the second annual ACM conference on Multimedia systems*, pp. 145–156, 2011.

[78] L. De Cicco, V. Caldaralo, V. Palmisano, and S. Mascolo, "Elastic: a client-side controller for dynamic adaptive streaming over http (dash)," *International Packet Video Workshop*, pp. 1–8, 2013.

[79] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran, "Probe and adapt: Rate adaptation for http video streaming at scale," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 4, pp. 719–733, 2014.

[80] J. Jiang, V. Sekar, and H. Zhang, "Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive," *International Conference on Emerging Networking Experiments and Technology*, pp. 97–108, 2012.

[81] T. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 187–198, 2015.

[82] 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, June 2010.

[83] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AC video database for the evaluation of quality metrics," in *IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, March 2012, pp. 2430–2433.

[84] 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. Selected Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct 2012.

[85] VQEG HDTV Group, "VQEG HDTV Database. Video Quality Experts Group (VQEG)," [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx.

[86] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Hkkinen, "CVD2014;A Database for Evaluating No-Reference Video Quality Assessment Algorithms," *IEEE Trans. on Image Proc.*, vol. 25, no. 7, pp. 3073–3086, July 2016.

[87] D. Ghadiyaram, J. Pan A.C. Bovik, A. K. Moorthy, P. Panda, and K. C. Yang, "Subjective and objective quality assessment of mobile videos with in-capture distortions," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, 2017, (in print).

[88] Y. Tian and M. Zhu, "Analysis and modelling of no-reference video quality assessment," in *Computer and Automation Engineering, 2009. ICCAE '09. International Conference on*, March 2009, pp. 108–112.

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

[90] L. Ma, S. Li, and K. N. Ngan, "Reduced-reference video quality assessment of compressed video sequences," *IEEE Trans. on Circuits and Syst. for Video Tech.*, vol. 22, no. 10, pp. 1441–1456, Oct 2012.

[91] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. for Video Tech.*, vol. 23, no. 4, pp. 684–694, 2013.

[92] S. Li, L. Ma, and K. N. Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *IEEE Trans. on Circuits and Syst. for Video Tech.*, vol. 22, no. 7, pp. 1100–1112, 2012.

[93] 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, Feb 2010.

[94] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," *IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, April 2007.

[95] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," *ACM SIGCOMM Comp. Communication Review*, vol. 41, no. 4, pp. 362–373, 2011.

[96] S.S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs," *IEEE/ACM Trans. on Networking*, vol. 21, no. 6, pp. 2001–2014, 2013.

[97] M. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 129–134, 2014.

[98] T. Hoßfield, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via Crowdsourcing," *IEEE Int. Sym. On Multimedia*, pp. 494–499, 2011.

[99] N. Staelens, S. Moens, W.V. Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of iptv and video on demand services in real-life environments," *IEEE Trans. on Broadcasting*, vol. 56, no. 4, pp. 458–466, Dec 2010.

[100] T. Porter and X. H. Peng, "An objective approach to measuring video playback quality in lossy networks using tcp," *IEEE Communications Letters*, vol. 15, no. 1, pp. 76–78, 2011.

[101] D. Z. Rodriguez, J. Abrahao, D. C. Begazo, R. R. L. Rosa, and G. Bressan, "Quality metric to assess video streaming service over tcp considering temporal

location of pauses," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 985–992, 2012.

[102] R. Wang, Y. Geng, Y. Ding, Y. Yang, and W. Li, "Assessing the quality of experience of http video streaming considering the effects of pause position," in *16th Asia-Pacific Network Operations and Management Symposium*. IEEE, 2014, pp. 1–4.

[103] M. N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnström, and A. Raake, "Quality of experience and http adaptive streaming: A review of subjective studies," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 141–146.

[104] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hobfeld, and P. Tran-Gia, "A survey on quality of experience of http adaptive streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[105] T. Hoßfeld, M. Seufert, C. Sieber, T. Zinner, and P. Tran-Gia, "Identifying qoe optimal adaptation of http adaptive streaming based on subjective studies," *Computer Networks*, vol. 81, pp. 320–332, 2015.

[106] W. Robitza, M. N. Garcia, and A. Raake, "At home in the lab: Assessing audiovisual quality of http-based adaptive streaming with an immersive test paradigm," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[107] S. Egger, B. Gardlo, M. Seufert, and R. Schatz, "The impact of adaptation strategies on perceived quality of http adaptive streaming," in *Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming*. ACM, 2014, pp. 31–36.

[108] K. Watanabe, J. Okamoto, and T. Kurita, "Objective video quality assessment method for evaluating effects of freeze distortion in arbitrary video scenes," 2007, vol. 6494, pp. 64940P–64940P–8.

[109] R. K. P Mok, E. W. W Chan, and R. K. C Chang, "Measuring the quality of experience of http video streaming," in *IEEE Int. Symp. on Integrated Network Management*. IEEE, 2011, pp. 485–492.

[110] H. Yeganeh, R. Kordasiewicz, M. Gallant, D. Ghadiyaram, and A. C Bovik, "Delivery quality score model for internet video," in *IEEE Int. Conf. on Image Proc.*, 2014, pp. 2007–2011.

[111] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the Quality of Experience of mobile streaming videos," in *IEEE Global Conf. on Sig. and Info. Process.* IEEE, 2014, pp. 989–993.

[112] CrowdFlower, "Crowdflower — make your data useful," [Online] Available: https://crowdflower.com.

[113] B. Russell, A. Torralba, K. Murphy, and W.T. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comp. Vision*, vol. 77, pp. 157–173, 2008.

[114] L. Von Ahn and L. Dabbish, "Labeling images with a computer game," *Proc. SIGCHI Conf. on Human factors in Comp. Systems.*, pp. 319–326, 2004.

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

[116] T. Hoßfield, 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. Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[117] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 1–25, 2017.

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

[119] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *Int. Conf. Comp. Vision*, vol. 2, pp. 1458 – 1465, 2005.

[120] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *In Workshop on Statistical Learning in Computer Vision, ECCV*, vol. 1, no. 1–22, pp. 1–2, 2004.

[121] Stephen W Kuffler, "Discharge patterns and functional organization of mammalian retina," *J. neurophysiology*, vol. 16, no. 1, pp. 37–68, 1953.

[122] A. C Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, pp. 2008–2024, 2013.

[123] F. W. Campbell and J. G. Robson, "Application of fourier analysis to the visibility of gratings," *The Journal of Physiology*, vol. 197, no. 2, pp. 551, 1968.

[124] H. R. Wilson and J. R. Bergen, "A four mechanism model for threshold spatial vision," *Vision Research*, vol. 19, no. 1, pp. 19–32, 1979.

[125] R. W. Rodieck, "Quantitative analysis of cat retinal ganglion cell response to visual stimuli," *Vision Research*, vol. 5, no. 12, pp. 583–601, 1965.

[126] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. on Image Proc.*, vol. 24, no. 8, pp. 2579–2591, 2015.

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

[128] U. Rajashekar, Z. Wang, and E. P. Simoncelli, "Perceptual quality assessment of color images using adaptive signal representation," *SPIE*, vol. 31, no. 4, pp. 75271L–75271L, 2010.

[129] L. M. Hurvich and D. Jameson, "An opponent-process theory of color vision," *Psychological Review*, vol. 64, no. 6, pp. 384–404, 1957.

[130] D. L. Ruderman, T. W. Cronin, and C. C. Chiao, "Statistics of cone responses to natural images: Implications for visual coding," *JOSA A*, vol. 15, no. 8, pp. 2036 – 2045, 1998.

[131] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[132] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 043025–043025, 2013.

[133] T. Goodall, A. C. Bovik, and N. G. Paulter, "Tasking on natural statistics of infrared images," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 65–79, 2016.

[134] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Data mining techniques for the life sciences*, pp. 223–239, 2010.

[135] L. Xu, W. Lin, and C. C. J. Kuo, *Visual quality assessment by machine learning*, Springer, 2015.

[136] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Sig. Proc.: Img. Comm.*, vol. 19, no. 2, pp. 121–132, 2004.

[137] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, crc Press, 2004.

[138] Wikipedia, "Peak signal-to-noise ratio," 2017.

[139] Int. Telecommuncation Union Std., "ITU-T Recommendation P.910: Subjective video quality assessment methods for multimedia applications," 2008.

[140] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," *Int. Conf. on Acoust. Speech, Signal. Proc.*, pp. 1153–1156, May 2011.

[141] "FFmpeg," [Online] Available: https://ffmpeg.org/.

[142] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the Quality of Experience of mobile streaming videos," *IEEE Global Conf. on Signal and Information Proc.*, Dec. 2014.

[143] J. S. Perry, "XGL Toolbox," [Online]. Available: https://github.com/jeffsp/xgl, 2015.

[144] Spyder5PRO, "Datacolor: Spyder5PRO," [Online] Available: http://spyder.datacolor.com/portfolio-view/spyder5pro/.

[145] Int. Telecommun. Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures ITU-R Recommendation BT.500-11, Tech Rep.," 2002.

[146] A. M. van Dijk, J.-B. Martens, and A. B.Watson, "Quality asessment of coded images using numerical category scaling," *Proc. SPIE Advanced Image and Video Communications and Storage Technologies*, 1995.

[147] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)*, vol. 10, no. 16, 1994.

[148] S. Thorpe, D. Fize, , and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381.6582, pp. 520–522, 1996.

[149] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Computational statistics & data analysis*, vol. 52, no. 12, pp. 5186–5201, 2008.

[150] S. Verboven and M. Hubert, "LIBRA: A MATLAB Library for Robust Analysis," *Chemometrics and Intelligent Laboratory Systems*, pp. 127–136, 2005, [Online]. Available: http://wis.kuleuven.be/stat/robust/LIBRA/LIBRA-home.

[151] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, Feb 2013.

[152] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[153] A. J Greene, C. Prepscius, and W. B. Levy, "Primacy versus recency in a

quantitative model: activity is the critical distinction," *Learning & Memory*, vol. 7, no. 1, pp. 48–57, 2000.

[154] D. L Schacter and R. L. Buckner, "Priming and the brain," *Neuron*, vol. 20, no. 2, pp. 185–195, 1998.

[155] V. Maljkovic and K. Nakayama, "Priming of pop-out: I. role of features," *Memory & Cognition*, vol. 22, no. 6, pp. 657–672, 1994.

[156] D. Ghadiyaram, J. Pan, and A. C. Bovik, "Learning a continuous-time streaming video qoe model," in *IEEE Transactions of Image Processing*, (under review).

[157] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 1153–1156, 1992.

[158] D. Ghadiyaram, J. Pan, and A. C. Bovik, "A time-varying subjective quality model for mobile streaming videos with stalling events," in *SPIE Optical Engineering + Applications*. International Society for Optics and Photonics, 2015, p. 959911.

[159] A. Lang, K. Dhillon, and Q. Dong, "The effects of emotional arousal and valence on television viewers' cognitive capacity and memory," *J. of Broadcasting & Electronic Media*, vol. 39, no. 3, pp. 313–327, 1995.

[160] T. Sharot and E. A. Phelps, "How arousal modulates memory: Disentangling the effects of attention and retention," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 4, no. 3, pp. 294–306, 2004.

[161] C. Fenimore, J. Libert, and S. Wolf, "Perceptual effects of noise in digital video compression," *SMPTE journal*, vol. 109, no. 3, pp. 178–187, 2000.

[162] R. C Gonzalez and R. E. Woods, "Image processing," *Digital image processing*, vol. 2, 2007.

[163] D. E. Pearson, "Viewer response to time-varying video quality," *Proc. SPIE*, vol. 3299, pp. 16–25, July 1998.

[164] Watson, A. B. and Solomon, J. A., "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer. A*, vol. 14, no. 9, pp. 2379–2391, Sept. 1997.

[165] P. Teo and D. J. Heeger, "Perceptual image distortion," *Proc. IEEE ICIP*, vol. 2, pp. 982–986, Nov. 1994.

[166] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," *Digital Images Human Vis.*, vol. 1, pp. 179–206, June 1993.

[167] J. A. Nelder, "The fitting of a generalization of the logistic curve," *Biometrics*, vol. 17, no. 1, pp. 89–110, 1961.

[168] "Identifying Hammerstein Weiner Models," [Online]. Available: https://www.mathworks.com/help/ident/ug/identifying-hammerstein-wiener-models.html.

[169] C. G. Bampis and A. C. Bovik, "Learning to predict streaming video qoe: Distortions, rebuffering and memory," *IEEE Trans. Image Proc.*, (submitted).

[170] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience," *Data Traffic Monitoring and Analysis*, pp. 264–301, 2013.

[171] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. Evans, "Large-scale crowdsourced study for tone mapped hdr pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4725 – 4740, 2017.

Vita

Deepti Ghadiyaram received the B.Tech. degree in Computer Science from International Institute of Information Technology (IIIT), Hyderabad in 2009, and the M.S. degree from the University of Texas at Austin in 2013. She joined the Ph.D. program at the University of Texas at Austin in Fall 2013 under the guidance of Prof. Alan Bovik in the Laboratory for Image and Video Engineering (LIVE).

Her research interests include image and video processing, computer vision, and machine learning. Her Ph.D work focuses on perceptual image and video quality assessment, particularly on building quality prediction models for pictures and videos captured in the wild and understanding a viewers time-varying quality of experience while streaming videos. She was a recipient of the UT Austins Microelectronics and Computer Development (MCD) Fellowship from 2013 to 2014 and the Graduate Student Fellowship by the Department of Computer Science from 2013-2016.

Permanent Address: Flat No. 201, Laxmi Vani Residency
Kanta Reddy Nagar Colony
Attapur, Hyderabad - 500048
Telangana, India

This dissertation was typeset with LaTeX 2$_\varepsilon$[1] by the author.

[1] LaTeX 2$_\varepsilon$ is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.