The Dissertation Committee for Debarati Kundu
certifies that this is the approved version of the following dissertation:

Subjective and Objective Quality Evaluation of Synthetic and High Dynamic Range Images

Committee:

Brian L. Evans, Supervisor

Alan C. Bovik

Donald S. Fussell

Wilson S. Geisler III

Joydeep Ghosh

Subjective and Objective Quality Evaluation of Synthetic and High Dynamic Range Images

by

Debarati Kundu, B.E., M.S.E.

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

May 2016

To the bard *Rabindranath Tagore*, who ensured that a song is readily available whenever words fail to express my emotions.

Acknowledgments

I decided to write the acknowledgment section at the very end because somehow it always felt as if writing it at the beginning would jinx my graduation plans. The years at UT have taught me that even though you may not be sure of how your dissertation would end, pretty early in the graduate program you mentally start drafting a list of names that would find their place in the acknowledgment section. It gives me immense pleasure to think that finally after my PhD defense, I have got the opportunity to express my gratitude to all the people who helped in shaping up my dissertation.

First and foremost, I would like to thank my adviser Prof. Brian Evans for his constant support and encouragement throughout the course of my graduate studies. Under his mentorship, I learned to conduct independent research and hone my problem solving skills. I am also grateful to him for instilling in me his sense of discipline, his focus on successful technology transfers to actual prototypes and his motto of 'We are engineers, we learn things by doing them.' Apart from research, I am indebted to Prof. Evans for teaching me how to teach, by setting a remarkable example himself. I consider myself to be extremely lucky to have got the opportunity to work on my PhD at his research group.

I would also like to thank my dissertation committee members: Prof. Alan Bovik, Prof. Don Fussell, Prof. Bill Geisler, and Prof. Joydeep Ghosh. I am immensely thankful to Prof. Bovik for the concepts that I learned in course of collaborating with him on my PhD research. His highly interesting courses on image and video processing lay the

cornerstone of my PhD contributions and I am honored to have him as a co-author on my paper. I am also indebted to Prof. Fussell for introducing me to the fascinating world of computer graphics and the insightful discussions on my research. Prof. Geisler is an amazing teacher who made me interested in the study of the human visual system. My first brush with data mining occurred when I took Prof. Ghosh's course during my first year of graduate studies, and that became an important tool for my research on blind image quality evaluation.

I am also thankful to Prof. Sanjay Shakkottai, Prof. Sujay Sanghavi and Prof. Constantine Caramanis for offering extremely interesting courses on probability, machine learning and optimization that have significantly strengthened my mathematical foundation and encouraged me to think analytically.

I would like to extend my gratitude towards Dr. Vinayak Nagpal and Dr. Klaus Doppler for mentoring my internship at Nokia Research Center and to Dr. Pawan Baheti and Dr. Rajiv Soundararajan for a fulfilling internship experience at Qualcomm Research. Special thanks to Prof. Mark Fairchild at the Rochester Institute of Technology for providing some of his source high dynamic range images.

My graduate days would have been incomplete without the support of ESPLers past and present: Dr. Aditya Chopra, Dr. Yousof Mortazavi, Dr. Marcel Nassar, Dr. Jing Lin, Dr. Kyle Wesson, Dr. Chao Jia, Dr. Karl Nieman, Ghadi Sebaali, Jinseok Choi, Junmo Sung, Scott Johnston, Jeremy Gin, Hugo Andrade, Marcus DeYoung. Special thanks to Chao for helping me learn the ropes of computer vision and also extending his helpful hand when I acted as a teaching assistant for the first time. I would also like to thank Prof. Hazem Hajj and Gilbert Badaro of the American University of Beirut for

their helpful comments during my presentations.

I am indebted to the students of Prof. Bovik's LIVE research group: Dr. Lark Kwon Choi, Deepti Ghadiyaram, Leo Isikdogan, Christos Bampis, Todd Goodall, Janice Pan and Zeina Sinno not only for the extremely helpful academic brainstorming sessions but also for the frequent social outings that paved the way for fluid cultural exchanges and in turn made my stay in the US a truly international experience. Lark and Deepti deserve special mention for their meticulous help on subjective image quality assessment. I would also like to extend my gratitude towards Dr. Anush Moorthy, Dr. Michele Saad, and Dr. Anish Mittal, whose work inspired me to focus on image quality assessment. I would like to thank my colleagues: Suriya, Subhashini, Abhishek G., Divya, Hardik, Joyce, Natalia, Rajat, Soumya, Abhishek S., Harpreet, Sarabjyot, Amin, Murat, Derya, Mandar, Kiran, Jianhua, Preeti, Marcus T. for an enriching experience at WNCG.

My graduate life would have been far less enjoyable without the amazing friends at Austin. I cannot begin to describe in words how much I am grateful to my friend Avik Ray for supporting me untiringly through all these years ever since we boarded the same flight from Kolkata to Austin for the first time. Thanks to Somsubhra Barik for having the patience to listen to my one-sided conservations and trying to teach me (in vain) how to play the piano. Words would fall short to describe the way Anamika Dubey, Swagata Bhattacharjee, and Shrijata Chattopadhyay have acted as pillars of encouragement. I am also grateful to Ayan Acharya, Urmimala Roy, Subhendu Roy, Abhik Bhattacharya, Arindam Sanyal, Avhishek Chatterjee, Bhavatharini Sridharan, Sundeep Korrapati, Aparajita Sant, and Suma Jothibasu for the fun time we had in Austin.

The present and past staff at UT-ECE, Melanie Gulick, Barry Levitch, Melody

Singleton, Janet Preuss, Jennifer Graham, Lauren Bringle and Karen Little, have been very supportive in helping me deal with administrative issues. Melanie Gulick deserves special mention, who got bombarded with my questions about the graduate program even before I had made up my mind to come to UT. Thanks to Gabriel Hernandez and Ricardo Gomez for their prompt help in dealing with the system administration issues that I faced.

This dissertation perhaps would never have seen the light of the day without the companionship and words of encouragement of my friends from Kolkata, namely Kaushik Suresh, Sattam Kumar Datta, Sambuddha Kumar, and Pummy Mukherjee, who provided their shoulder of support whenever I missed my home in India and made sure to rejoice at each and every one of my achievements. I would also like to express my gratitude towards Prof. Swagatam Das and Prof. Amit Konar at the Department of Electronics and Telecommunications Engineering, Jadavpur University, for encouraging me to pursue a PhD in the first place. Special thanks go to my photographer friends Chiranjib Saha, Debdipta Goswami, Soumya Goswami and Sangram Biswas for providing me samples of high dynamic range images for my subjective study and discussing at great length different image processing software.

My family. My parents have been the bedrock of strength upon which the entire foundation of my academic career has been built. Without their unflinching support I could not have carried on my education. I am grateful to my grandmother Mrs. Sova Rani Paul and my grandfather, the late Mr. Amulya Chandra Paul for instilling in me the core value of 'Education is the magic key that opens the door of unimaginable potential'. My extended family, comprising of uncles, aunts, and cousins was also just a phone call away whenever I needed them.

Subjective and Objective Quality Evaluation of Synthetic and High Dynamic Range Images

Publication No. _____

Debarati Kundu, Ph.D.
The University of Texas at Austin, 2016

Supervisor: Brian L. Evans

Recent years have seen a huge growth in the acquisition, transmission, and storage of videos. The visual data consists of both natural scenes as well as synthetic scenes, such as animated movies, cartoons and video games. In all these cases, the ultimate goal is to provide the viewers with a satisfactory quality-of-experience. In addition to the traditional 8-bit images, high dynamic range imaging is also becoming popular because of its ability to represent the real world luminances more realistically. Coming up with objective image quality assessment algorithms for these applications is an interesting research problem.

In this work, I have developed a synthetic image quality database by introducing varying degrees of different types of distortions and conducted a subjective experiment in order to obtain the ground-truth data. I evaluated the performance of state-of-the-art image quality assessment algorithms (typically meant for natural images) on this database, especially no-reference algorithms that have not been applied to the domain of computer graphics images before. I identified the top-performing algorithms along with

analyzing the types of distortions on which the present algorithms show a less impressive performance.

For high dynamic range(HDR) images, I have designed two new full-reference image quality assessment algorithms to judge the quality of tonemapped HDR images using statistical features extracted from them. I have also conducted a massive online crowd-sourced subjective test for HDR image artifacts arising from tonemapping, multiple-exposure fusion and post processing. To the best of our knowledge, presently this is the largest HDR image database in the world involving the largest number of source images and most number of human evaluations. Based on the subjective evaluations obtained, I have also proposed machine learning based no-reference image quality assessment algorithms to predict the perceptual quality of HDR images.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent years have seen tremendous growth in the acquisition, transmission, and storage of digital visual data[1]. With the proliferation of hand-held smart devices, the exponential increase in the amount of mobile image/video traffic will likely continue in the upcoming years. Some of the popular applications of visual data are streaming websites like YouTube, High Definition TVs, Video-on-demand services like Hulu and Netflix, Digital Cinema etc. On an average 350M photos are uploaded to Facebook every year and YouTube has over a billion users, roughly one-third of all internet users. Apart from the images and videos captured by optical cameras, the visual data traffic also comprises of computer graphics generated content, such as those in animated movies and video games. The genre of massively multi-player online gaming has 23.4M subscribers worldwide. In addition, fusion of natural and synthetic content is becoming increasing popular due to the widespread use of augmented reality applications (such as Google Glass).

In addition to the standard dynamic range images (8bits/color/pixel), high dynamic range (HDR) images are also being captured by the users, either with high-end DSLRs or with hand-held smart-devices. For example, Qualcomm's Snapdragon S4 processors supports HDR capture. Also, video streaming services like Amazon Instant Video supports streaming of HDR videos and HDR displays for home entertainment are becom-

ing more popular (such as Samsung HDR TVs).

For all of these applications, the source image is subjected to a few processing stages, beginning with its capture (in case of natural scenes) or rendering (in case of computer graphics images). Properties of the capturing and display device, rendering GPUs, limited availability of transmission bandwidth may lead to loss of information in the source. Since humans are the final consumers of the visual data traffic, the ultimate goal is to provide a satisfactory quality-of-experience (QoE)[2]. The introduced distortions may or may not be visible to the human observers and if visible, they lead to varying degrees of annoyance[3]. Research in QoE deals with quantifying this visual annoyance and results in more perceptually optimized multimedia services, such as creating high-quality cinematic content, capturing better videos with hand-held smartphones, rendering photo-realistic computer-generated imagery in video games and animation movies, and video compression and transmission over bandwidth limited communication channels.

Conducting subjective experiments to ask for human opinion on multimedia content is the ultimate ground-truth of QoE evaluation, but these methods are time-consuming and expensive. Hence many research efforts in recent decades have focused on developing objective image quality assessment (IQA) algorithms which show a high degree of correlation with human judgment. The next section outlines the different types of Image Quality Assessment (IQA) methods.

1.1 Image Quality Assessment (IQA) Methods

Depending on the involvement of human subjects in evaluating the visual quality of images, the IQA algorithms may be divided into the following two categories:

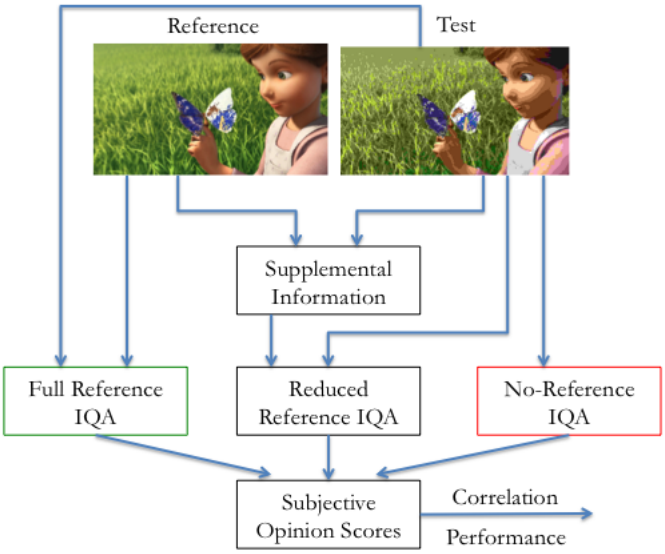


Figure 1.1: Different categories of objective image quality assessment algorithms. Full-reference methods need access to both the reference and the distorted image, reduced-reference methods need access to some supplementary information about the reference, no-reference methods evaluate the quality just by accessing the test image. The subjective scores obtained from the automated IQA algorithms are correlated with the ground subjective opinion scores obtained from human observers.

1.1.1 Subjective Quality Assessment

In case of subjective IQA, human subjects assess the visual quality to an image or video and assign a numerical score (say, on a scale of 1 to 100) based on the perceived quality. Due to the inherent variability among human subjects in judging the visual quality, in order to represent the population better, multiple subjects are required to judge and rate the quality of a corpus of images or videos. The subjective studies follow two paradigms:

- *Laboratory Experiments*: The studies are conducted in laboratory environments in order to control precisely the viewing conditions such as ambient illumination, viewing distance and proper calibration of the display device. The methodology for subjective testing and analyzing the data has been outlined in the recommendations provided by the The International Telecommunications Union (ITU). ITU-R Recommendation BT.500-11 citeitu outlines methodologies for the subjective assessment of the quality of television pictures. Also, for these subjective experiments, the distorted are images are synthetically created from high-quality pristine images by the introduction of graded simulated distortions.

- *Crowdsourced Experiments*: Studies conducted over crowdsourced platforms, mostly targeted at getting a larger and more varied source image corpus evaluated by a large group of subjects. The images considered may either be generated by the controlled introduction of distortions of varying degrees to high quality pristine images or may be captured by real-world imaging devices afflicted by complex mixtures of multiple distortions. Unlike the controlled subjective tests in laboratory conditions,

this paradigm of subjective testing is free from the constraints imposed by viewing conditions or display devices and helps us in gaining insight on the visual perception of the subjects under a much wide array of testing conditions.

Although conductive subjective experiments is a cumbersome process, it is imperative in providing the ground truth data required for the evaluation of the objective IQA algorithms. More details about the subjective testing frameworks and the major publicly available subjective databases have been mentioned in the next chapter.

1.1.2 Objective Quality Assessment

Depending on the availability of the reference image to judge the quality of the test image, the objective image quality assessment algorithms may be classified as:

- *Full-reference (FR) IQA algorithms*: This class of algorithms requires the reference image for the prediction of the quality of the test image. The simplest FR algorithm that has been used to judge image quality for multiple decades was the mean squared error metric, but it has been found to correlate very poorly with human perception. More sophisticated FR measures of signal fidelity have been proposed, but since the reference image is always not available against with the test image is compared, this severely limits the application areas of FR-IQA algorithms.

- *Reduced-reference (RR) IQA algorithms*: This class of algorithms requires the some statistical features extracted from the reference image to predict the quality of the test image. The systems employing RR-IQA models extract the features from the reference image at the sender side which are transmitted through an ancillary

channel to the receiver side. The receiver extracts the features from the test image and predicts the quality by comparing that with the features extracted from the reference image. The RR-IQA algorithms should aim at achieving a good balance between the bandwidth required to transmit the extracted features and the accuracy of the visual quality prediction [3].

- *No-reference (NR) IQA algorithms*: This class of objective IQA algorithm is the most challenging one because the reference image is not available to judge the quality of the test image. This method of IQA does away with older concepts of signal fidelity or fidelity [3]. Also, this paradigm of IQA is very useful for applications where there is no concept of a 'reference' image, such as those containing real world distortions arising from the multitude of image capturing devices used by people, or High Dynamic Range images created from fusing a stack of images shot at different exposures. The details about the different categories of NR-IQA algorithms have been outlined in Chapter 3.

1.2 Synthetic Scene Image Quality Assessment

In addition to videos captured with optical cameras, video traffic also often includes synthetic scenes, such as animated movies, cartoons and video games. The burgeoning popularity of multiplayer video games (on mobile platforms) is causing an exponential increase in synthetic video traffic[4]. The visual quality of synthetic scenes can be degraded both by the rendering process (e.g. video gaming on standalone devices) and by transmission over a wireless network (e.g. cloud gaming applications). Designing objective IQA algorithms to accurately predict the quality of the synthetic images distorted by

these artifacts is a challenging problem.

One approach in the performance of evaluation of visual quality in computer graphics is comparison of the results produced by an image processing algorithm with the de facto golden standard using a full-reference metric. This approach suffers from the disadvantage that with the development of new rendering techniques, the de facto standard itself might be replaced by an algorithm that produces better results, thereby resulting in a lack of standardization.

Many proposed no-reference algorithms are based on studying the overall statistical properties possessed by pristine images, which is inspired by natural images having certain statistical properties regardless of the image content, and are based on the assumption that distortions tend to deviate the Natural Scene Statistics (NSS) or Natural Video Statistics (NVS). However, these metrics for evaluating the quality of natural images have not been studied in the context of images generated using computer graphics. With the improvement of rendering technology, rendered images are becoming more and more photo-realistic, which has led me to hypothesize that NSS models can be applied in the domain of computer graphics with some modifications. Instead of conducting user studies, the NSS based no-reference algorithms can be used to quantify the perceptual quality of a rendered scene.

1.3 High Dynamic Range Image Quality Assessment

With the advent of 4K and HDTVs, the user expectations of clarity in video quality is bound to change in the coming years. Apart from increasing spatial and temporal resolution, there has been a lot of interest in high-dynamic range (HDR) videos,

for accurate representation of luminance variation in real scenes, from very bright sunlight to dark shadows. Unlike traditional Standard Dynamic Range (SDR) scenes with 8 bits/color/pixel, the range of the luminance levels in HDR scenes can range from 10,000 to 1[5]. In computer graphics, using HDR images result in more photo-realistic rendering with a rich level-of-detail. Recently, Amazon's Instant Video streaming service has started to stream HDR video content[6].

An image photographed at a single exposure may have overexposed and underexposed regions. The HDR image creation pipeline typically begins by a registered stack of images of the same scene at different exposures and fusing them to get the irradiance map, represented by 32 bit floating point images. However, displaying the irradiance map on ordinary displays meant for SDR images is not possible without tonemapping it to LDR format. Some applications skip the intermediate step of creating the floating point irradiance map and instead display the final SDR image by directly fusing the multi-exposure stack. HDR images created by commercial software like Adobe Photoshop or Photomatix are also followed by post-processing to increase the aesthetic appeal. These processes of tone-mapping, multi-exposure fusion or post-processing all give rise to annoying artifacts. Apart from the processing artifacts, the HDR images also suffer from compression artifacts for streaming applications.

Subjective and objective IQA for HDR images is a relatively new research topic. Compared to SDR images, a lower degree of subjectively evaluated HDR images is available. Most of the databases lack in variety of source content or the types of artifacts considered. Hence, conducting subjective studies is imperative to obtain ground truth data on which the HDR-specific objective IQA algorithms may be built on and evaluated.

1.4 Dissertation Summary

To summarize, I have contributed to the subjective and objective quality evaluation of both synthetic scenes and high dynamic range images.

1.4.1 Thesis Statement

In this dissertation, I defend the following statement:

Using scene statistics yields automated visual quality assessment algorithms for synthetic images and high dynamic range images that have high correlation with human visual quality evaluation.

1.4.2 Summary of Contributions

For synthetic image quality assessment, I have made the following contributions:

- *ESPL Synthetic Image Database*: I have designed the ESPL Synthetic Image Database, comprising of 25 high quality pristine images and 500 distorted images generated by controlled introduction of varying degrees of different types of processing, compression and transmission artifacts, such as interpolation, blur, additive noise, JPEG compression and Fast-Fading channels.

- *Laboratory Subjective Study of Synthetic Images*: I conducted subjective experiments for collecting data from 64 observers, and analyzed the data to reject the outlier subjects, calculated the differential mean opinion scores (DMOS) for each of the distorted images.

- *Performance evaluation of state-of-the-art IQA algorithms*: I compared the performance of more than 50 FR, RR and NR-IQA algorithms (originally designed for natural images) by correlating the scores obtained from the IQA algorithms with the synthetic image DMOS scores. For the FR-IQA algorithms I have identified the key distortion categories for which the natural images IQA algorithms show a lesser degree of correlation. I have shown that the NSS based NR-IQA algorithms can be used even for predicting the quality scores of distorted synthetic scenes.

For high dynamic range images, I have made the following contributions:

- *FR-IQA for Tonemapping Artifacts*: I improved the state-of-the-art FR-IQA algorithms for evaluating the quality of tonemapped images in comparison to the original HDR luminance map by incorporating models of natural scene statistics and visual saliency. In addition, the algorithm also showed a high degree of correlation on tonemapped images afflicted with JPEG compression artifacts.

- *ESPL-LIVE HDR Image Database*: I have designed the ESPL-LIVE HDR Image Database, comprising of more than 600 source HDR scenes, from which 1815 HDR images were created using different processing artifacts, such as tonemapping and multi-exposure fusion. In addition, I have also considered post-processing artifacts in HDR images.

- *Crowdsourced HDR Subjective Study*: I used the Amazon Mechanical Turk online crowdsourcing platform to garner ratings on the images of the ESPL-LIVE HDR

Image Database from a larger number of human subjects. To the best of our knowledge, presently this is the largest HDR image database in the world involving the largest number of source images and most number of human evaluations.

- *Scene Statistics based NR-IQA for HDR images*: I proposed a scene-statistics based NR-IQA algorithm in the gradient domain for evaluating HDR artifacts that outperforms the state-of-the-art NR-IQA algorithms on this class of distortion. For completeness, the algorithm has also been evaluated on SDR natural (LIVE Image Quality Database[7], LIVE Multiply Distorted Images[8]) and synthetic image databases (ESPL Synthetic Image Database).

1.5 Organization

The rest of the dissertation is organized as follows: Chapter 2 summarizes the design of the ESPL Synthetic Image Database and the steps taken to conduct the subjective test. It describes in detail the source content, different distortions that have been introduced and the methodologies employed for processing of the raw subjective scores. The analysis of the performance of the IQA algorithms on the ESPL Synthetic Image Database has been provided in Chapter 3. This includes the performance of the FR, RR and NR-IQA algorithms outlined with respect to measures of correlation with ground truth human subjective data and the run-time complexity of the different methods.

Chapter 4 gives a brief description of the process of creation of HDR content and explains the proposed FR-IQA algorithm for tonemapped HDR images. The crowdsourced subjective study of HDR images has been outlined in Chapter 5. This describes the

source images considered, the different types of processing artifacts, the subjective testing methodology and the performance of the state-of-the-art NR-IQA algorithms on HDR artifacts. Details about the proposed NR-IQA algorithm has been described in Chapter 6. Chapter 7 concludes the dissertation with a summary of the presented work and outlines possible avenues of future research.

Table 1.1: Table of acronyms

AMT	Amazon Mechanical Turk
BIQI	Blind Image Quality Index
BLIINDS-II	BLind Image Integrity Notator using DCT Statistics-II
BRISQUE	Blind/Referenceless Image Spatial QUality Evaluator
C-DIIVINE	Complex-DIIVINE
CORNIA	COdebook Representation for No-Reference Image Assessment
CPBD	Cumulative Probability of Blur Detection
CurveletQA	NR-IQA based on Curvelets
DESIQUE	Derivative Statistics-based QUality Evaluator
DIIVINE	Distortion Identification-based Image Verity and INtegrity Evaluation
ESPL	Embedded Signal Processing Laboratory
FF	Fast Fading
FISH	Fast Wavelet-Based Image Sharpness Estimation
FNVE	Fast Noise Variance Estimation
FR	Full-reference
FSIM	Feature Similarity Index
G-IQA-1	Gradient-Image Quality Assessment-1
G-IQA-2	Gradient-Image Quality Assessment-2
GM-LOG	Gradient Magnitude and Laplacian of Gaussian based NR-IQA
GMSD	Gradient Magnitude Similarity Deviation
GN	Gaussian Noise
GRNN	General Regression Neural Network IQA
GSM	Gradient Similarity Measure
HDR	High Dynamic Range
HDR-VDP-2	High Dynamic Range VDP
HIT	Human Intelligence Task
IFC	Information Fidelity Criterion
IQA	Image Quality Assessment
IW-SSIM	Information Content Weighted SSIM
JPEG	Joint Photographic Experts Group
JPEG-NR	NRIQA of JPEG compressed images
JNBM	Just-Noticeable Blur
LIVE	Laboratory for Image and Video Engineering
LPC-SI	Local Phase Coherence based sharpness index
MAD	Most Apparent Distortion
MEF	Multi-Exposure Fusion
MSVF	Metric based on Singular Value Decomposition
MS-SSIM	Multi-scale Structural Similarity Index
NIQR	Natural Image Quality Evaluator
NJQA	NRIQA of JPEG compressed images via Quality Relevance Map
NLWT	Noise-level Estimation using weak textured patches
NQM	Noise Quality Measure
NR	No-reference
PHA	Peak Signal-to-Noise ratio-Human Visual System-A
PHMA	Peak Signal-to-Noise ratio-Human Visual System(modified)-A
PHVS	Peak Signal-to-Noise ratio-Human Visual System
PHVSM	Peak Signal-to-Noise ratio-Human Visual System(modified)
PSNR	Peak Signal-to-Noise Ratio
QSSIM	Quarternion Structural Similarity Index
RFSIM	Riesz-transform based Feature Similarity Metric
RR	Reduced-reference
RRDNT	Reference based RRIQA with Divisive Normalization
RRED	Reduced-Reference Entropic Differences
RRIQA	Reduced-Reference Image Quality Assessment (Wavelet Domain)
S_3	Spectral and Spatial Measure of Local Perceived Sharpness
SR-SIM	Spectral Residual Based Similarity
SSIM	Structural Similarity Index
TM-IQA	Topic Model based IQA
UQI	Universal Quality Index
VIF	Visual Information Fidelity
VSI	Visual Saliency-Induced Index
VDP	Visual Difference Predictor
VSNR	Visual Signal-to-Noise ratio
WSNR	Weighted Signal-to-Noise ratio

Chapter 2

Subjective Quality Evaluation of Lightly Distorted Synthetic Images

2.1 Prior Work

[1]A subjective study with human observers is the most reliable way to gauge perceptual quality of images. Although a subjective study is difficult to design and time-consuming to conduct, the ground-truth data obtained from human observers is valuable for benchmarking objective IQA algorithms that aim to automate the process of visual quality assessment. The subjective experiments are also imperative for understanding the gap in performance between the state-of-the-art IQA algorithms and human perception.

To aid in the development of objective image quality assessment (IQA) algorithms, many natural image databases have been created that contain the subjective ratings of the images by human observers. Some of the largest natural image databases annotated by quality scores from humans are the LIVE Image Quality Database [7], the Tampere Image Database 2013 [10], the Categorical Image Quality Database [11] and EPFL JPEG XR codec [12]. Most of the commonly occurring distortions in these databases are the processing artifacts such as blur, additive noise, contrast changes, and chromatic distortions, compression artifacts resulting from JPEG or JPEG2000 standards, and transmission

[1]Contents of this capter has been published in [9]

artifacts resulting from sending the images over a Rayleigh fading channel.

In comparison, subjective quality evaluation data is not as available for synthetic images such as those commonly encountered in video games or animated movies. Recently Cadík[13] developed a synthetic image database of computer graphics generated imagery afflicted by distortions such as noise, aliasing, brightness changes, light leakage and tone mapping artifacts. Traditionally, compression artifacts, such as JPEG blocking artifacts were not studied for synthetic images, but with the advent of cloud gaming(such as the Nvidia Grid[14]), we do need to render the synthetic scenes on the server side (the clients may be "dumb" clients, having nothing more than a video playback facility), compress them, and send them over a wireless network, whereby, the rendered image might suffer due to packet loss, or low bit-rate connections.

In the development of the ESPL Synthetic Image Database, I considered a larger number images with a higher degree of source complexity and a broader class of distortions (transmission and compression artifacts) than the previous work by Cadík[13][15] so that the database better represents the types of images and artifacts encountered when watching animated movies and playing video games. These have not been considered in any previous subjective study to the best of our knowledge. With the advent of more powerful Graphics Processing Units, the degree of realism of graphical images[16] has vastly narrowed between natural scenes and high quality synthetic scenes. Compared to Cadík's database, our database spans a wider range of scene complexity, as outlined in Section 2.2.1.

2.2 Human Subjective Study

This section describes the source and distorted images considered, the methods employed in generating the synthetic scenes, the subjective testing framework and the methodology of analyzing the raw scores.

2.2.1 Source Images

A total of 25 synthetic images were chosen from video games and animated movies. These high quality color images from the Internet are 1920×1080 pixels in size. The video games that were considered included multiplayer role playing games (such as War of Warcraft), first person shooter games (such as Counter Strike), motorcycle and car racing games, and games with more realistic content (such as FIFA). Some of the animated movies, from which the images were collected, are, The Lion King, the Tinkerbell series, Avatar, Beauty and the Beast, Monster series, Ratatouille, the Cars series, etc. [2] We incorporated natural and non-photorealistic renderings of human figures and human-made objects, renderings of fantasy figures such as fairies and monsters, close-up shots, wide angle shots, images showing both high and low degrees of color saturation, and background textures without a foreground object. Fig. 2.1 shows the 25 reference images.

2.2.2 Source Complexity

The complexity of the source images gives an indication of the "richness" of the content in terms of edge distribution, local textures, contrast variation and colorfullness.

[2] All images are copyright of their rightful owners, and the authors do not claim ownership. No copyright infringement is intended. The database is to be used strictly for non-profit educational purposes.

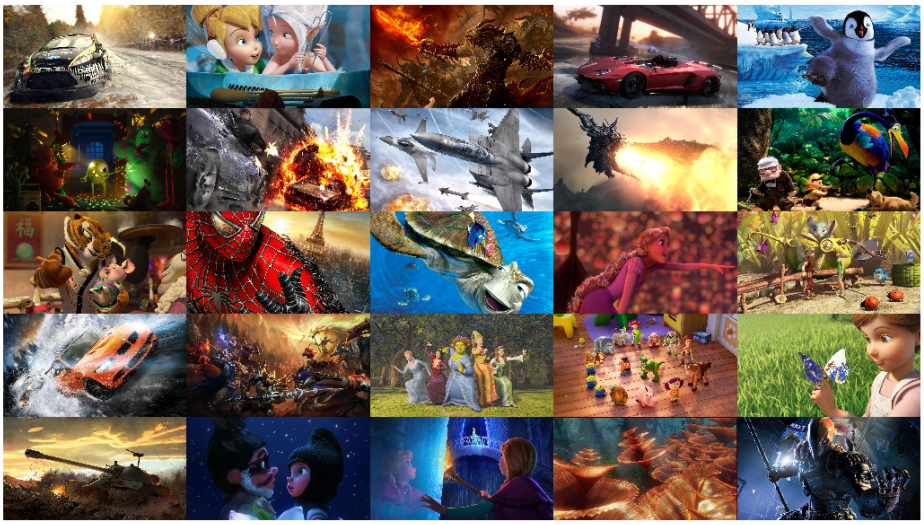


Figure 2.1: Sources images in the ESPL database [17]

A database is characterized by the variety of the images considered in order to better represent the real-world scenarios. The source complexity of the database has been analyzed using the following two quantitative metrics, as outlined in [18].

2.2.2.1 Spatial Information (SI)

This measure indicates the degree of presence of edges in an image. The luminance of the RGB image can be obtained by $Y = 0.299R + 0.587G + 0.114B$, which is filtered along the horizontal and vertical directions with the Sobel kernel to yield s_h and s_v respectively. The edge magnitude at every pixel is given by $s_r = \sqrt{{s_h}^2 + {s_v}^2}$. The final SI value of the image is obtained by the root mean square of the edge magnitudes at every pixel.

$$SI = \sqrt{L/1080}\sqrt{\sum {s_r}^2/P} \tag{2.1}$$

where P is the number of pixels in the filtered image, L is the vertical resolution. The factor $\sqrt{L/1080}$ has been included to make the computed SI somewhat scale/resolution agnostic.

2.2.2.2 Colorfullness (CF)

This measure indicates the variety and intensity of colors in the image. Let $rg = R - G$ and $yb = 0.5(R + G) - B$. Colorfulness is defined as:

$$CF = \sqrt{{\sigma_{rg}}^2 + {\sigma_{by}}^2} + 0.3\sqrt{{\mu_{rg}}^2 + {\mu_{by}}^2} \tag{2.2}$$

Fig 2.2 shows a scatter plot of spatial information vs. colorfulness computed for the images in the ESPL Synthetic Image Database and three other publicly available image quality assessment databases. (Cadík's[13], LIVE[7] and TID[10] databases). The scatter plots from the ESPL database, shown in Fig 2.2(a), show that spatial information and colorfulness span a similar range of scene complexity as the other natural image databases as shown in Fig 2.2(c) and Fig 2.2(d). In Fig 2.2(b), Cadík's Synthetic Image database shows a larger range but sparsely covers the range.

2.2.3 Distortion Simulations

Distortions in synthetic images differ from those in natural images. This is because the distortions in synthetic images arise from two sources: firstly, the image might have

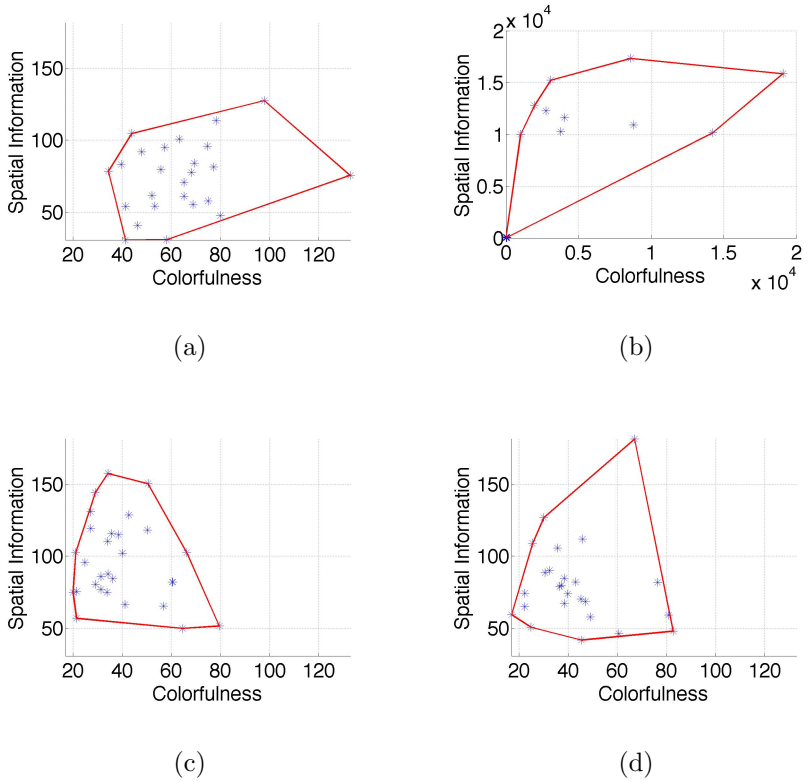


Figure 2.2: Spatial Information vs. Colorfulness scatter plots for the source images in the following databases (a) ESPL Synthetic, (b) Cadík's Synthetic Image[13], (c) LIVE[7],(d) TID 2013[10]. Red lines indicate the convex hull of the points in the scatter plot, which approximates the range of scene complexity.

artifacts from the rendering process, display and other processing steps, such as tone mapping and contrast amplification, and secondly, some distortions might be introduced due to encoding at a low bit-rate or transmission over a network, such as JPEG block artifacts and transmission noise. Other distortions may arise, such as unnaturalness of shading, which can be evaluated only given access to both the rendered 2D scene, and the information provided by the 3D depth buffer. This initial database does not contain these

other kinds of distortions. Since we did not have access to the proprietary 3D models and the lighting information that were used to render the scenes, we chose to introduce distortions on the rendered image themselves.

Three categories of processing artifacts are considered: interpolation (which arises frequently in texture maps, and causes jaggedness of crisp edges), blurring and additive Gaussian noise. With the advent of cloud gaming, where the rendered 2D game images are streamed from the server to 'dumb' clients (having only a video playback facility), we chose to study the effect of compression and transmission artifacts on computer graphics generated images (which had been previously considered only for natural scenes). Thus in the ESPL database, JPEG compression and Rayleigh fast-fading wireless channel artifacts are considered. For each artifact type, the intensity of distortion was varied to create four distorted versions of the same pristine image, so that they range from barely noticeable to a high degree of visual impairment. MATLAB was used for all the distortion categories. The following paragraphs briefly describe the types of distortions considered, and the methodology used in their generation.

1) Interpolation: The original images were downsampled using integer downsampling factors ranging from 3 to 6, which are upsampled back using nearest neighbor interpolation. This can be used to simulated the jagged edges caused by 'aliasing' in rendering. Since bilinear and trilinear interpolation eliminate jagged edges, to retain a higher degree of jaggedness and perceptual separation of these pictures, simple nearest neighbor (zeroeth order) interpolation was used.

2) Gaussian Blur: The RGB color channels were filtered using a circularly symmetric 2D Gaussian kernel with standard deviation ranging from 1.25 to 3.5 pixels. The

same kernel was employed on each of the color channels at every pixel location. Natural photographic images often suffer from severe blur as a consequence of lens defocus and/or motion of the camera. However in computer graphics, generating the degree of blur (motion blur or depth-of-field blur for aesthetic purposes) is generally controlled. For this reason, serious blur conditions (e.g. in the LIVE IQA database[7]) were avoided. Depth-of-field blur can be synthesized by placing sharper foreground objects on a uniformly blurred background. Hence, evaluation of images with global blur is an important component of judging the quality of these images. A future avenue of work would be to introduce localized types of blur. However as the first step, we chose to study global blur for synthetic images in order to get an idea of how humans evaluate the quality of blurred synthetic scenes. Later databases could be dedicated to capturing isolated blur events. This also can serve as a check when evaluating IQA algorithms originally intended for use on natural scene for when applying them to synthetic scenes because most existing natural image databases [7][10] contains uniformly blurred images.

3) Gaussian Noise: Zero mean white Gaussian noise was added to the RGB components of the images (same noise variance were used for all the color channels). The noise standard deviation ranged from 0.071 to 0.316 pixels, using the *imnoise* MATLAB function. Noise can occur in the generation of synthetic images using random sampling based rendering methods, such as Monte Carlo. In creating the current database, high levels of noise were not considered because synthetic images may be re-rendered in such cases. Since no assumption was made with regards to the distribution of the noise, we decided to simulate Gaussian noise distributed uniformly across the image. Future avenues of work could consider more specialized distortions, such as Perlin noise used in texture

synthesis.

4) JPEG compression: The MATLAB *imwrite* function compressed the reference images into JPEG format. The bits-per-pixel (bpp) ranged from 0.0445 to 0.1843. Higher bpp images were not considered, in order to better simulate playing a cloud video game under restricted bandwidth conditions. Blockiness in images arises from independent coding of spatially correlated adjacent blocks[19]. This can occur in both JPEG compression (using DCT basis functions) or H.264/HEVC (using integer transform basis functions). Since here we are studying still computer graphics images, JPEG was used. Further subjective studies involving computer graphics generated videos could probably model gameplay videos compressed by H.264/HEVC.

5) Simulated Fast Fading Channel: The reference images were compressed into JPEG2000 bitstreams (with wireless error resilience features enabled and 64 x 64 tiles) and then transmitted over a simulated Rayleigh-fading channel. The signal-to-noise ratio (SNR) was varied at the receiver from 14 to 17 dB to introduce different degrees of transmission errors. SNRs greater than 17 dB did not introduce perceptible distortions due to the error resilience feature of the JPEG2000 codec.

2.2.4 Testing Methodology

Since the number of images to be evaluated (525) was prohibitively high for a double stimulus setup, a single stimulus continuous evaluation testing procedure with hidden reference [20] was used.

Every image in the database was viewed by each subject, over three sessions of one hour each, with each session separated by roughly 24 hours. Each session was divided into

two sub-sessions of 25 minutes with a break of five minutes to reduce visual fatigue and eye strain. The 64 subjects who participated in the test were graduate and undergraduate students at The University of Texas at Austin (Fall 2014), with ages ranging from 18-30 years, mostly without prior experience participating in subjective tests or image quality assessment. The gender ratio of the subjects was roughly 1:1.

Before the start of the experiment, the procedure was explained to each subject and verbal confirmation of normal vision was obtained. Subjects viewed approximately 175 test images during each session which were randomly ordered using a random number generator, and randomized for each subject. In order to familiarize themselves with the testing setup, each testing session was preceded by a short training session comprising of around 10 images, which had different content but same type of distortions as the test images.

2.2.4.1 Subjective Testing Display

The user interface for the study was designed on two identical PCs in MATLAB, using the Psychology Toolbox[21]. Both PCs used identical NVIDIA Quadro NVS 285 GPUs and were interfaced to identical Dell 24 inch U2412M displays, which were roughly of the same age with identical display settings. The monitors had 16:10 aspect ratio, 1,000:1 static contrast ratio. Any additional digital processing of the monitor was turned off. It was found that the peak luminance of the monitors is $339 cd/m^2$, minimum black level is $0.04 cd/m^2$ and color gamut is 71% NTSC, 74.3% Adobe RGB, 95.8% sRGB. Each image was displayed on the screen for 12 s and the experiment was carried out under normal office illumination conditions. The ambient lighting was measured using

a 200,000 Lux Docooler Digital LCD Pocket Light Meter and was found to be 540lux. Subjects viewed the images from about 2 - 2.25 times the display height.

The screen resolution was set at 1920×1200 pixels, but the images were displayed at their normal resolution (1920×1080) without any distortion introduced by interpolation. The pixels per degree was found to be 43.63, assuming a viewing distance of 0.66m. The top and bottom portions of the display were mid gray. At the end of the image display duration, a continuous quality scale was displayed on the screen, where the default location of the slider was at the center of the scale. It was marked with five qualitative adjectives: "Bad", "Poor", "Fair", "Good", and "Excellent" placed at equal distances along the scale. After the subject entered a rating for the image, the location of the slider along the scale was converted into a numerical score lying between [0,100], after rounding to the nearest integer. The subject could take as much time as needed to decide the score, but there was no provision for changing the score once entered or viewing the image again. The next image was automatically displayed once the score was recorded.

2.2.5 Processing of Raw Subjective Scores

The raw subjective scores were analyzed using the ITU-R BT.500-13 recommendations [20]. Let s_{ijk} be the score assigned by subject i to image j in session $k = 1, 2, 3$, and $s_{i_{ref}jk}$ be the score assigned by the same subject to the corresponding reference image. The difference between the scores of the test image and the score of the corresponding reference image was calculated for each subject to take into account the preference of certain subjects for certain images.

$$d_{ijk} = s_{ijk} - s_{i_{ref}jk} \tag{2.3}$$

Since any reference image and its distorted version were shown in the same testing session, it is assumed that the quality scale used by the subject remained the same for any single session. The difference scores for the reference images were 0 and were not taken into consideration in subsequent processing steps. The difference scores per session was converted to the Z-scores per session:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{1}^{N_{ik}} d_{ijk} \tag{2.4}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik} - 1} \sum_{j=1}^{N_{ik}} (d_{ijk} - \mu_{ik})^2} \tag{2.5}$$

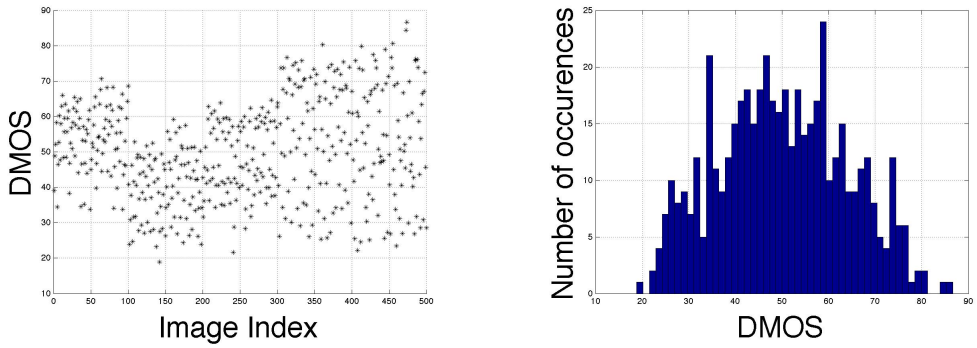
$$z_{ijk} = \frac{d_{ijk} - \mu ik}{\sigma_{ik}} \tag{2.6}$$

where N_{ik} is the number of test videos seen by subject $i$in session k. Thus, the Z-scores take into account any differences in subject preferences for reference images, use of the quality scale between subjects and differences in use of the quality scale by a subject between sessions.

A subject rejection procedure as outlined in ITU-R BT.500-13 recommendations [20], was used to discard scores from unreliable subjects. [3]First, it was determined whether

[3]The performance of the FR-IQA algorithms was evaluated by calculating the DMOS considering all subjects without removing the outliers. No noticeable difference was observed in the relative ranking of the different algorithms although the different algorithms showed small changes in the values of the correlation measures.

the scores assigned by the subject are normally distributed by computing the kurtosis of the scores. If the kurtosis falls between 2 and 4, the scores were assumed to be normally distributed. If the scores are normally distributed, a subject was rejected whenever more than 5% of the scores assigned by her falls outside the range of two standard deviations from the mean scores. If the distribution of the scores deviates from a normal distribution, a subject was rejected whenever more than 5% of the scores assigned by her falls outside the range of 4.47 standard deviations from the mean scores. Out of a total of 64 subjects, 12 were treated as outliers and the ratings obtained from the remaining 52 subjects were considered in the calculation of the final DMOS. The 5% criterion used in the subject rejection procedure translates to 26 images in the ESPL Synthetic Image Database.



(a) (b)

Figure 2.3: (a) Scatter Plot and (b) Histogram of DMOS scores for test images obtained from the study, the DMOS scores span a wide perceptual quality range.

The Z-scores are assumed to be distributed as a standard normal distribution, 99.9% of the scores in our study fell in the range of [-3,3]. The scores were rescaled to lie

in the range of [0,100] by using the linear mapping:

$$z_{ij}\prime = \frac{100(z_{ij}+3)}{6} \tag{2.7}$$

The DMOS score for each test image was calculated as the mean of the rescaled Z-scores from the $M = 52$ subjects remaining after outlier rejection.

$$DMOS_j = \frac{1}{M}\sum_{i=1}^{M} z_{ij}\prime \tag{2.8}$$

Compared to computing MOS by averaging the ratings obtained from the human subjects, DMOS calculated from the Z-scores removes the bias of the human subjects towards scene content and helps us focus only on the distortions.

The standard error in the DMOS scores was 0.6212 across distorted images. One major goal of an image database to be used for perceptual quality assessment is that the images should span over a wide range of visual quality. To illustrate this, the scatter plot and histogram of the DMOS scores of the test images are shown in Fig. 2.3. We see that the DMOS scores of the ESPL Synthetic Image Database spans the range from 18 to 87. Assuming that the Z-scores assigned by a subject comes from a standard normal distribution, 99% of the Z-scores should lie in the interval [-3,3], which translates to DMOS scores in the range of [0,100]. [18,87] on the DMOS scale corresponds to mean Z-scores in the range of [-1.92,2.22], which covers approximately 96% of the area of the standard normal distribution.

The high number of outliers resulted from the borderline reliability of some of the subjects. However we find that the scores obtained from the subjects after outlier rejection shows remarkable consistency. In order to evaluate the degree of consensus among the

subjects in judging quality, the subjects were divided into two groups, the DMOS scores for all the images were calculate using the ratings obtained from each group, and the rank correlation was measured between the two sets of DMOS scores thus obtained. The mean of the Pearson's linear correlation coefficient thus obtained was found to be 0.9813 over 50 such randomized splits. This shows a high level of agreement among the users in evaluating the quality of the images.

2.3 Conclusion

In this section, I have summarized the contributions in creating the ESPL Synthetic image database (comprising of 25 pristine images and 500 distorted images) and conducting the human subjective test to obtain the 'ground-truth' score for every image. The following chapter outlines the results obtained by evaluating how the state-of-the-art IQA algorithms, meant primarily for natural images, perform on synthetic images.

Chapter 3

Objective Quality Evaluation of Lightly Distorted Synthetic Images

3.1 Introduction

[1]The previous chapter dealt with subjective quality evaluation of synthetic scenes. This chapter is concerned with the objective quality evaluation of the different artifacts occurring in synthetic images. As explained in chapter 1, in order to automate perceptual quality evaluation, two broad categories of objective IQA algorithms have been developed: with-reference and blind or no-reference methods, based on the availability (or not) of a reference image. With-reference methods may have access to either the complete reference image or some statistical features extracted from it. The former defines full-reference (FR) IQA algorithms, while the latter defines reduced-reference (RR) IQA algorithms. The performance of several publicly available state-of-the-art FR-IQA algorithms has been evaluated on popular natural image databases [24][25]. Cadík *et al.* [13] evaluated the performance of six FR-IQA algorithms and demonstrated that they were sensitive to brightness and contrast changes, could not distinguish between plausible and implausible shading, and failed to localize distortions precisely.

When information about the reference image is not available, no-reference (NR)

[1]Contents of this capter has been published in [22] and [23]

IQA metrics are better suited. Many NR metrics rely on machine learning approaches using features expressive of statistical regularities possessed by pristine images, commonly called natural scene statistics (NSS) models [26][27]. NSS models for good quality natural images hold reliable well irrespective of image content and it is assumed that distortions tend to deviate from these statistical regularities. NR-IQA algorithms have not yet been studied in the context of images generated using computer graphics. Herzog *et al.* [15] proposes an NR-IQA metric for quantifying rendering distortions based on machine learning. The features were chosen heuristically, instead of being based on properties of pristine synthetic images.

I evaluate the performance of more than 50 state-of-the-art FR, RR and NR IQA algorithms on the synthetic scenes and compared them to the subjective test results. The performance of the algorithms was extensively tested using hypothesis testing and statistical significance analysis. It is hypothesized that with some modifications, NSS based NR-IQA metrics could be successfully applied to graphics images having sufficient degree of scene complexity. Here we take an important first step towards evaluating scene statistics based NR-IQA methods on synthetic scenes, expressed both in the spatial as well as various transform domains, and quantified how the presence of distortions change the scene statistics of synthetic images. Top performing NSS-based NR-IQA algorithms show a high degree of correlation with human perception on synthetic scenes, which is a promising development in regards to the successful automatic prediction of the perceptual quality of computer graphics generated imagery for which no 'ground truth' information is available.

3.2 Synthetic Scene Statistics

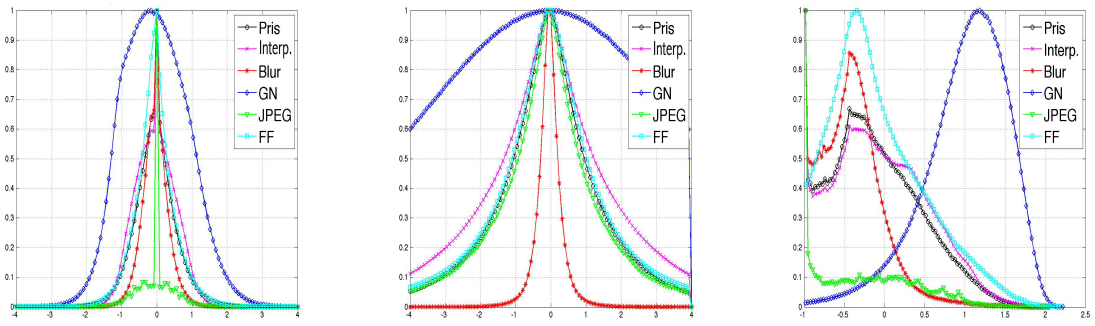
In this section we discuss scene statistics for synthetic scenes, and how the statistics of distorted images deviate from those of pristine images. In [22], we model the mean-subtracted-contrast-normalized (MSCN) coefficients [28] obtained from synthetic scenes using Generalized Gaussian and Symmetric α-stable distributions and found that as long as the image was devoid of distortions, irrespective of natural or synthetic content, the distribution of the MSCN coefficients still shows a Gaussian-like signature. This has led us to the hypothesis that, like natural scenes, scene-statistics based approaches can be used to evaluate the distortions present in synthetic images. Indeed, the presence of distortion reliably affect the statistics of synthetic images in the spatial as well as in bandpass transfer domains, as shown in Fig. 3.1.

3.3 Objective IQA Algorithms

More than 50 publicly available objective IQA algorithms were evaluated on the ESPL Synthetic Image Database. The full-reference, reduced-reference and no-reference IQA algorithms considered have been summarized in Table 3.1.

3.3.1 Full-Reference IQA Algorithms

For the sake of brevity, we group the IQA algorithms under consideration into categories, and then summarize each IQA algorithm in that category, as follows:



(a) (b) (c)

Figure 3.1: Histograms of (a) MSCN pixels, (b) Steerable Pyramid Wavelet Coefficients and (c) Curvelet Coefficients of pristine and distorted image patches obtained from the ESPL Synthetic Image Database. The figure shows how distortions change the statistics of pristine images. The legends Pris, Interp., Blur, GN, JPEG, FF refer to pristine images, images with interpolation distortion, blur distortion, additive white Gaussian noise, JPEG compression and simulated transmission over a Rayleigh fast-fading wireless channel, respectively.

3.3.1.1 Mean Square Error based algorithms

The Mean Square Error (MSE) between the reference and the test image is the simplest distortion measure between images. Peak Signal to Noise Ratio (PSNR) is a function of the MSE between the reference and the test image. For example, in [29] Shnayderman *et al.* propose a metric where the MSE between the singular values of the reference and test image block is computed.

3.3.1.2 Structural Similarity based algorithms

The Structural Similarity Index (SSIM)[30] and its multiscale version MS-SSIM[31] take into account luminance, contrast and structure changes. MS-SSIM allows for a wider variation in display resolution and distance of the viewer from the image plane, by computing the perceptual quality of the image over multiple scales. The Universal Quality Index (UQI)[33] is an older metric based on similar ideas. The Quarternion Structural Similarity Index (QSSIM)[32] represents the R,G, and B color channels using a quarternion.

3.3.1.3 Human Visual System model based algorithms

Different HVS properties such as the contrast sensitivity function (CSF), luminance masking, etc have been incorporated into number of algorithms like the Visual Difference Predictor (VDP)[34] and High Dynamic Range VDP (HDR-VDP-2) [35], which also incorporate viewing distance and display device characteristics. The Noise Quality Measure (NQM)[36], Weighted Signal-to-Noise ratio (WSNR) [37], PSNR-HVS[38], PSNR-HVSM[40], PSNR-HMA[39], PSNR-HA[39] are some other IQA algorithms based

on mean-shifting, CSF and between coefficient contrast masking of DCT basis functions.

3.3.1.4 Information Theoretic algorithms

Here, the test image is considered to be obtained by passing the reference image through a distortion channel and the resulting loss of visual information is hypothesized to be related to the capacity of this communication channel[78]. The Information Fidelity Criterion (IFC)[41] indicates the mutual information between the reference and the test image. The Visual Information Fidelity metric (VIF, VIFP)[42] is based on a natural scene statistics model and measures the Shannon information between the reference and the test images.

3.3.1.5 Feature Similarity based algorithms

These algorithms are based on extracting different types of low-level local features (that correlate closely with visual perception) from the reference and the test image, such as: (1) phase correlation, used in Feature Similarity Index (FSIM[44] (2) change in gradients, used in Gradient Similarity Measure (GSM)[46] and Gradient Magnitude Similarity Deviation (GMSD)[45] (3) Riesz transform based features in Riesz-transform based Feature Similarity Metric (RFSIM)[47] etc.

3.3.1.6 Visual Saliency based algorithms

Visual saliency (VS) aim to understand the areas of the image that will attract the attention of the viewers. Some algorithms which take into account saliency models to pool the localized quality scores are the Visual Saliency-Induced Index (VSI)[48] and

the Spectral Residual Based Similarity (SR-SIM)[49]. In the Information Weighted SSIM metric (IW-SSIM)[43], the local information content of the image is used as a saliency measure which weighs the local SSIM value.

3.3.1.7 Strategy based algorithms

These algorithms are based on different strategies employed by the HVS depending on whether the distortions are near-threshold or supra-threshold. The most prominent of these are the Most Apparent Distortion algorithm (MAD)[11] and the Visual Signal-to-Noise ratio (VSNR)[50].

3.3.2 Reduced-Reference IQA Algorithms

For RR-IQA algorithms, contrary to full-reference methods, partial information descriptive of the reference image may be made available to predict the quality of the test image. RR-IQA algorithms predict visual quality of the test image using as few features of the reference image as possible.

3.3.2.1 Natural Scene Statistics Feature based

Some examples in this category are [51], [52], and [53]. These algorithms are based on statistical features of the steerable pyramid representation of the images, coupled with divisive normalization.

3.3.2.2 Image Feature based

Many RR-IQA algorithms employ image features such as edge-maps[56]. In [54], the authors propose an RR-IQA algorithm based on the image gradient magnitude following a Weibull distribution. Sub-image similarity, coupled with edge-based features are used in [55].

3.3.3 No-Reference IQA Algorithms

This part of our work is aimed primarily at understanding the usefulness at scene-statistics oriented learning based NR-IQA algorithms of the latter category for quality evaluation of synthetic scenes, but for the sake of completeness one of the top performing publicly available algorithms of the former category has also been considered. The following section outlines the two classes of NR-IQA algorithms:

3.3.3.1 Artifact Based Methods

Some examples of NR-IQA algorithms for blur estimation are based on local phase coherence (LPCM[57]), pooling strategies (CPBD[58]), spectral and spatial domain features(S_3[60],FISH[61]). To detect blocking artifacts resulting from JPEG compression, Sheikh *et al.* in [62] proposed a no-reference blockiness measure using the power spectrum of the test image. In [63], the authors use a quality relevance map to determine whether the blocks are naturally uniform or have been made uniform by JPEG compression. For blind estimation of the noise level of the images, the authors of [64] estimate the noise level from image patches using principal component analysis after selecting weakly textured patches from the images. In [65], the noise level is estimated by a local 3×3 mask

which is insensitive to the Laplacian of the image.

3.3.3.2 Learning Based Methods

The NSS based NR-IQA use statistical features descriptive of good quality of undistorted images. Leading NR-IQA models are based on the premise that natural images occupy a small subspace of all possible two dimensional signals, and that distortions move them from this subspace.

• *Spatial Domain Features*: In [79], it was observed that the MSCN of natural images tend to follow a Gaussian-like distribution. The distribution of MSCN pixels and products of adjacent pairs of them have been employed in the Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE)[66] and the Natural Image Quality Evaluator (NIQE)[28]. The Derivative Statistics-based QUality Evaluator (DESIQUE)[67] supplements BRISQUE by using log-derivative distributions of MSCN pixels. Using the gradient magnitude (GM) map and the Laplacian of Gaussian (LOG) response, the NR-IQA metric (GM-LOG)[68], uses Gaussian partial derivative filters along the horizontal and vertical directions. Two gradient log-derivative statistics based NR-IQA algorithms, G-IQA-1 and G-IQA-2, proposed in the LAB color space has also been evaluated[80].

• *Transform Domain Features*: Neurons employed in early stages of the visual pathway capture information over multiple orientations and scales, motivating multiscale processing in many NR-IQAs: log-Gabor decomposition (DESIQUE[67]), steerable pyramid wavelets (DIIVINE[69], C-DIIVINE[70]), Daubechies 9/7 wavelets (BIQI[71]), DCT (BLIINDS-II[72]), phase congruency (GRNN[73]), curvelets (CurveletQA[74]), expected image entropy upon a set of predefined directions (Anisotropy[75]). By contrast, COde-

book Representation for No-Reference Image Assessment (CORNIA)[76] uses a supervised learning technique to learn a dictionary of different distortions from the raw image patches instead of using a fixed set of features. Mittal[77] applies a "topic model" to the visual words extracted from the pristine and distorted images.

3.4 Results

This section outlines the results of evaluating the performance of state-of-the-art IQA algorithms on the ESPL Synthetic Image Database. The performance metrics and the methods of statistical evaluation is also provided.

3.4.1 Correlation Measures

The performance of the objective IQA algorithms outlined in the previous section were evaluated using two correlation measures: the Spearman Rank Order Correlation Coefficient (SROCC) (for measuring prediction monotonicity) and the Pearson Linear Correlation Coefficient (PLCC) (for measuring prediction accuracy) after non-linear regression on the objective IQA scores using a five-parameter monotonic logistic function following the procedure outlined in [81].

3.4.2 Root Mean Square Error

The accuracy of the quality scored predicted by the IQA algorihtms have been quantified using Root Mean Square Error (RMSE) between the DMOS scores and the objective IQA scores (after non-linear regression).

3.4.3 Outlier Ratio

Prediction consistency of the objective IQA algorithms was evaluated by measuring the outlier ratio (OR) [82]. Let Q'_j be the objective IQA algorithm score obtained for image j on the ESPL Synthetic Image Database after the logistic fit. Let $Z'_j = \{z_{ij}\}, i = 1, 2, \ldots, M$ be the Z-scores obtained for image j for M observers and σ_j be the corresponding standard deviation. An image is defined as an outlier if $Q'_j - DMOS_j > 2\sigma_j$. The outlier ratio is given by the ratio of the number of outliers to the total number of images (expressed as %).

3.4.4 Statistical Significance and Hypothesis Testing

The correlation measures were used to measure the differences in performance of the different IQA algorithms considered. However, to understand whether these differences are statistically significant based on the number of sample points used, we used two variance-based F-tests: based on individual quality scores and on DMOS scores respectively, following similar procedures as in [81].

3.5 Discussion of IQA Algorithm Performance

This section outlines trends observed and conclusions drawn from the experimental results of the IQA algorithms in Section 4.6.

3.5.1 Discussion of results for FR-IQA algorithms

We evaluated performance of 27 state-of-art FR-IQA algorithms on the ESPL Synthetic Image Database, where the source code for the FR-IQA algorithms came from

[24] and [84]. The single-scale algorithms have been evaluated on images rescaled by a factor dependent upon the image dimension and viewing distance[83]. This part of our study aims at benchmarking performance of different categories of IQA algorithms over different distortion categories. We have isolated distortion categories on which the FR-IQA algorithms perform worse and gained insight on the factors that lead certain types of FR-IQA algorithms to perform better, such as using color information instead of only luminance, efficient pooling strategy and the role played by strategy.

In Table 3.2, PSNR (row 26) is outperformed by other objective IQA algorithms (except for SSIM on row 27 and MSVD on row 28), but it performs reasonably well for additive noise and fast-fading artifacts since it captures high-frequency distortions. The SSIM and MS-SSIM IQA algorithms, which perform exceedingly well on the LIVE database[7], shows a less impressive performance on our database, primarily due to the very low degree of correlation with human judgment on certain classes of distortions, such as interpolation, which has not been studied in any of the existing databases of natural images before. However, SSIM is a single-scale measure; hence, it is very important to find the precise scale that depends both on the image dimensions and the viewing distance. Based on the rule-of-thumb proposed in [83], if the SSIM index is computed on the downsampled images, much better degree of correlation is achieved with the human ground truth subjective data, as shown in row 12. If the scale is chosen appropriately, SSIM-D (in row 14) outperforms MS-SSIM (in row 23).

Almost all of the existing IQA algorithms fail to accurately predict the subjective ratings of the interpolation artifact. Only MAD[11] achieves reasonable performance, which advocates multiple strategies for determining the overall image quality, based on

whether the distortions are near-threshold or supra-threshold. Low down-sampling factors result in near-threshold artifacts, which might appear almost imperceptible, especially at normal viewing distances. Although both interpolation and JPEG compression lead to blocking artifacts, the algorithms which perform exceedingly well on JPEG compression distortion (such as FSIM[44]) show much-less impressive performance on interpolation artifacts. This is because the two types of blocking artifacts deviate the statistics of the pristine scenes in different ways (Fig. 3.1). We would like to study the effects of varying display sizes on error visibility for interpolated images, which could prove valuable for display designers of game consoles. Blurred images also led to a lower degree of correlation with human scores compared to other categories. In computer graphics, motion blur is added artificially in many video games in order to create more realistic aesthetically pleasing images. Hence, the presence of blur in an image may not always correspond to a lower subjective score. Thus our subjective test reveals a significant performance gap for certain distortion categories between synthetic and natural images on which future researchers can work.

Overall, some of the recently proposed FR-IQA algorithms, such as GMSD[45], FSIM[44], VSI[48], SR-SIM[49] and MAD[11] correlate rather well with human perception in terms of SROCC. GMSD uses the standard deviation of the gradient map as a pooling strategy. FSIM takes into account image gradient magnitude and phase congruency (a dimensionless measure of significance of local structure) and then uses it as a pooling strategy. VSI and SR-SIM use more sophisticated pooling strategies based on visual fixations. Hence, we see that, irrespective of whether the image is natural or synthetic, IQA algorithms that use more efficient pooling strategies by taking into account

the localized distortions perform better than other IQA algorithms, as corroborated by [24]. This shows that irrespective of the content of the scene being natural or synthetic, finding interesting regions of the image that grab the attention of the HVS will improve the performance of IQA algorithms. Some of the IQA algorithms which model different aspects of the human visual system (HVS), such as NQM, VSNR, PSNR-HVSM, perform worse than the top performing signal driven IQA algorithms. Significant progress has been made towards understanding the functioning of the HVS, but on synthetic scenes studying higher level cognitive factors might be useful in understanding user gaze based on image saliency and also how the HVS differently perceives synthetic scenes as compared to natural images.

Table 3.3 shows the RMSE, reduced $\tilde{\chi}^2$ statistic between scores predicted by the algorithms and the DMOS for various FR-IQA Algorithms (after logistic function fitting) and outlier ratio. The top performing algorithm GMSD also show zero outlier ratio, which shows that all the predicted scores lie within two times the standard deviation of the DMOS scores.

3.5.2 Discussion of results for RR-IQA algorithms

RR-IQA algorithms show a lower degree of correlation with human subjective scores as compared to state-of-the-art FR-IQA algorithms as shown in Table 3.4. Among the NSS based RR-IQA algorithms, RRED shows the best overall performance (which is also the best performing RR-IQA algorithm). RRED also shows the best performance for the interpolation distortion category, since it captures the differences in wavelet coefficient statistics between the images with interpolation artifacts and that of the pristine images.

The NSS based RR-IQA algorithms perform better than the other edge-map based RR-IQA algorithms primarily due to their poor performance on the interpolation artifact category. Also, as per Table 3.5, the RMSE and outlier ratios of the best performing RR-IQA algorithms is worse than that of the best-performing FR-algorithms.

3.5.3 Discussion of results for NR-IQA algorithms

In this section, we discuss the performance of the NR-IQA algorithms in predicting the type of distortion in the test image and also the quality score. Many NR-IQA algorithms operate in two steps: classification of the type of distortion present in the test image and using the features of the detected class to map the extracted image features to a quality score. Table 3.8 shows the classification accuracy of the features extracted for the learning based NR-IQA algorithms in identifying the different category of distortions. Algorithms like GM-LOG, C-DIIVINE, BRISQUE and DESIQUE show good performances in distortion identification. Gaussian Noise was easiest to detect among all the distortion categories by most of the learning based NR-IQA algorithms.

Table 6.9 compares the performances of 26 NR-IQA algorithms which comprise both learning based methods and artifact based methods in terms of SROCC and PLCC. For rows 1-9 (learning based methods), after the feature extraction step, a mapping is obtained from the feature space to the DMOS scores using a regression method, which provides a measure of the perceptual quality. We used a support vector machine regressor (SVR), specifically LibSVM [85] to implement ϵ-SVR with the radial basis function kernel. The training set had 80% of the reference images (and their corresponding distorted versions) and the test set had the remaining 20% of the reference images (and their

corresponding distorted versions). The process was repeated 100 times to eliminate any bias due to varying spatial content.

Tables 3.8 and 6.9 show that G-IQA-1 (Luminance), DESIQUE, BRISQUE, C-DIIVINE and GM-LOG features perform the best in classifying distortion and deducing the mapping between the feature space and DMOS scores.

Fig. 3.4 shows box plots of the distribution of SROCC values for each of the 1000 trials of random train-test splits enable us to study the robustness of performance of the algorithms with variations of the choice of the training set. DESIQUE, BRISQUE and C-DIIVINE shows smaller variation in the degree of correlation with human perception.

Compared to learning based models, NIQE and TMIQA use unsupervised learning models and are not trained on corpus of distorted images. As such, these models perform worse on synthetic images in spite of showing competitive performance on natural images. This might occur due to higher amount of variability in the distribution of the MSCN coefficients for synthetic images as compared to natural scenes[22]. The performance of artifact based NR-IQA algorithms have been outlined in rows 17-21 (blur), 22-23 (noise) and 24-25 (JPEG blocking). To the best of our knowledge, we did not find any artifact based NR-IQA algorithm meant only for images having interpolation or fast-fading artifacts. For blur, noise and JPEG blocking, the learning based NR-IQA algorithms perform better than artifact based NR-IQA algorithms.

Table 3.7 shows that the high outlier ratio for some of the algorithms result from the high outliers obtained for the JPEG and Fast-Fading distortion category.

Figures 3.2 and 3.3 show scatter plots between predicted scores and DMOS scores

on ESPL Synthetic Image Database for a selected few IQA algorithms.

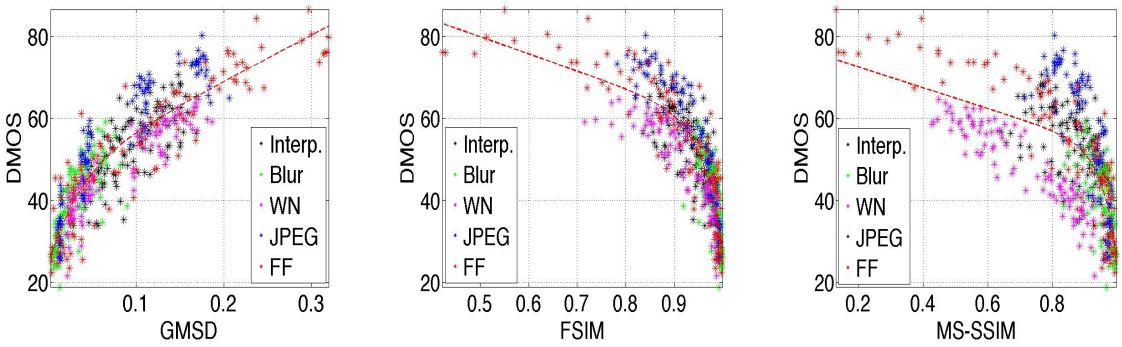
DMOS takes into account the preference of the subjects to certain source content by subtracting out the score provided by her to the source image. For this reason, the NR-IQA algorithms were also trained on the DMOS scores in the same way as [69][72][66]. On the other hand, the MOS scores do not take into account the score assigned by the user to the reference image. A comparison between the performance of the NR-IQA algorithms based on DMOS and MOS scores has also been provided in Tables 3.12 and 3.13. Comparison between Tables 6.9 and 6.9 shows the top-performing IQA algorithms show similar behavior irrespective of whether MOS or DMOS values are used for training them.

3.5.4 Determination of Statistical Significance

Results of statistical significance are summarized in Tables 6.5, 3.10 and 3.11. For this purpose, ten representative IQA algorithms were selected. For the learning based methods, the statistical significance tests were carried out for multiple training-test splits, using 60 test images each time, and similar results were obtained. The tables outline the results obtained for one such representative trial. For the F-Test based on quality scores provided by individual human observers, the variance of the residuals obtained from the null-model and the ten selected IQA algorithms, along with the number of samples considered in each category and the threshold F-ratio at 95% significance are shown in Table 3.10. None of the IQA algorithms tested was found to be statistically equivalent to the null-model corresponding to human judgment in any of the distortion categories. Similar conclusions were reached in [81]. GMSD shows the least variance of the residuals

2

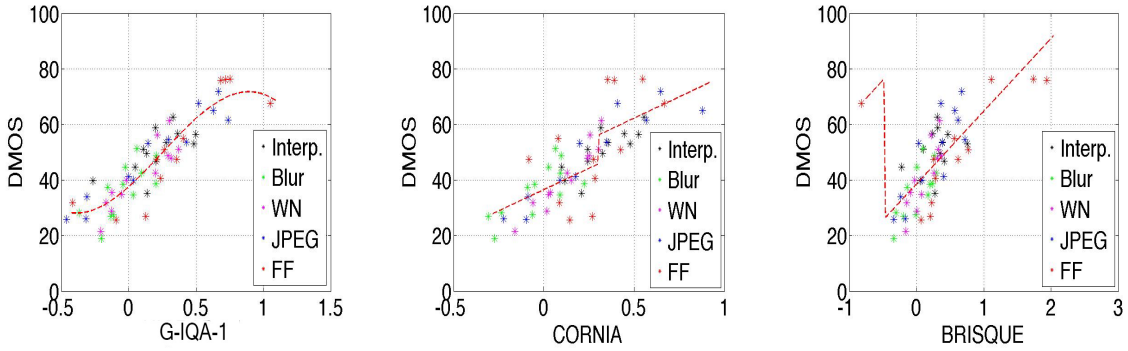
Figure 3.2: Predicted IQA scores vs. DMOS scatter plots for some selected full-reference and reduced-reference IQA algorithms. The red line indicates the logistic regression fit.



(a) (b) (c)

(d) (e)

Figure 3.3: Predicted IQA scores vs. DMOS scatter plots for some selected no-reference IQA algorithms. The red line indicates the logistic regression fit.



(a) (b) (c)

(d) (e)

for the overall database among the ten IQA algorithms.

For the F-test based on the DMOS scores, Table 3.11 outlines the variance of the residuals obtained from the ten selected IQA algorithms, along with the number of samples considered in each category, and the threshold F-ratio at 95% significance. For some of the cases, it was found that the assumption of Gaussianity of the residuals did not hold. However, we still believe that the F-test can be used in these cases due to the large number of samples.

To determine whether the IQA algorithms are significantly different from each other, the F-statistic, as in [7][81], was used to determine the statistical significance between the variances of the residuals after a non-linear logistic mapping between the two IQA algorithms, at the 95% confidence interval. Table 6.5 shows the results for ten selected IQA algorithms and all distortions. Overall, the FR-IQA algorithms are found to be statistically superior to the NR-IQA algorithms.

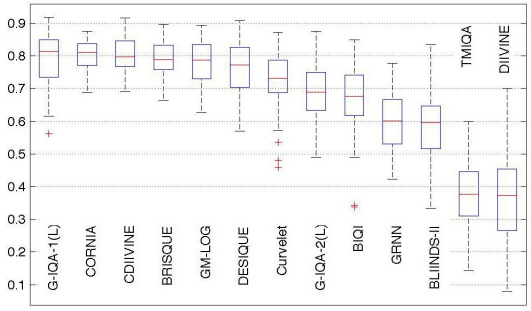


Figure 3.4: Box plot of SROCC of learning based NR-IQA algorithms on images in the ESPL Synthetic Image Database for 4:1 train-test splits over 100 trials. For each box, median is the central box, edges of the box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points, and the outliers are plotted individually.

3.5.5 Computational Complexity

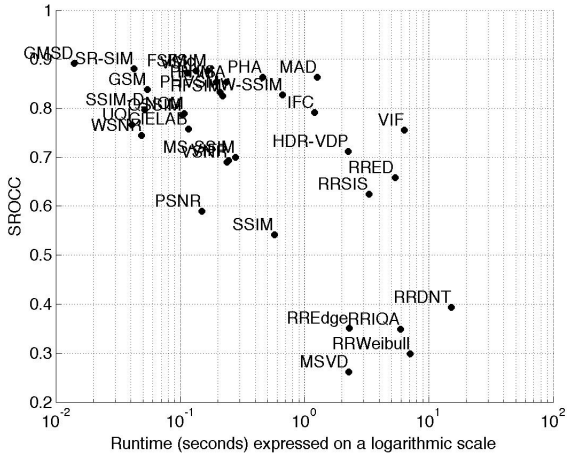


Figure 3.5: Scatter plot of SROCC of FR and RR-IQA algorithms with images in the ESPL Synthetic Image Database vs. runtime.

Fig. 3.5 shows the scatter plot of SROCC vs. execution time for the FR and RR IQA algorithms considered in this paper and Fig. 3.6 shows the similar plot for NR-IQA algorithms. All the IQA algorithms have been profiled using the original source codes provided publicly by the respective authors. FR-IQA metrics like SR-SIM and GMSD achieve a high degree of correlation with human perception and is computationally less intensive. As expected, the learning based NR-IQA algorithms (like BRISQUE, DESIQUE, C-DIIVINE, BLIINDS-II) achieve comparable performance results as the best performing FR-IQA algorithms, but they are computationally more intensive because they need to compute the image features and deploy them in a machine learning framework for quality prediction. RRED shows intermediate performance between FR-IQA and NR-IQA algorithms both in terms of correlation with human judgment and time complexity.

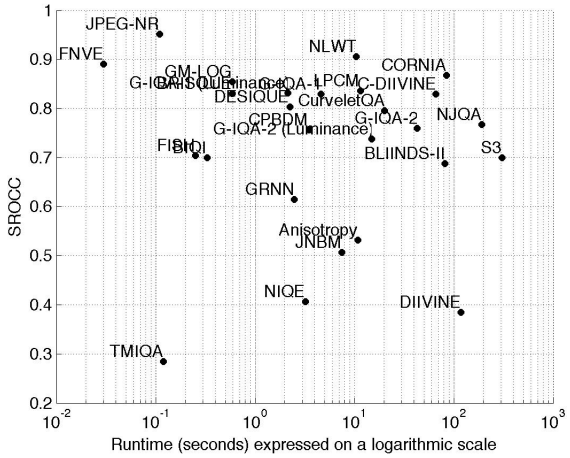


Figure 3.6: Scatter plot of SROCC of NR-IQA algorithms with images in the ESPL Synthetic Image Database vs. runtime.

3.6 Conclusion

We present the publicly available ESPL Synthetic Database comprising pristine source images and images containing five different types of distortions, annotated by 26,000 quality scores from 52 subjects. We evaluate the performance of more than 50 state-of-the-art IQA algorithms.

For FR-IQA algorithms, we observe the importance of saliency based spatial pooling strategies and strategies for evaluating the quality of the image, based on whether the artifacts are subthreshold or suprathreshold. GMSD offers the best trade-off between performance and run-time complexity. RR-IQA algorithms perform worse than FR-IQA and NR-IQA algorithms. RRED is the best performing RR-IQA algorithm. For NR-IQA, we see that the deviation in statistical regularity caused by distortions can be used to successfully evaluate the quality of synthetic images also. Scene statistics based algorithms

take a longer time to run. Algorithms such as GMSD, SR-SIM, GM-LOG, and DESIQUE show high correlation with human perception and reasonable runtime. We find that for synthetic images, interpolation distortion is the most challenging category for the IQA algorithms, but the scene statistics based NR-IQA algorithms shows a better performance for quantifying this artifact.

The chapter concluded the objective quality evaluation of synthetic images of the ESPL Synthetic Image Database. The next chapter outlines FR-IQA algorithms for objective quality evaluation of high dynamic range images.

Table 3.1: List of Image Quality Assessment algorithms evaluated in this study.

Category of IQA	Method		Algorithm
Full Reference	Mean Square Error		Peak Signal-to-Noise Ratio
			Metric based on Singular Value Decomposition (MSVD)[29]
	Structural Similarity based		Structural Similarity Index (SSIM)[30]
			Multi-scale Structural Similarity Index (MS-SSIM)[31]
			Quarternion Structural Similarity Index (QSSIM)[32]
			Universal Quality Index (UQI)[33]
	Human Visual System model based		Visual Difference Predictor (VDP)[34]
			High Dynamic Range VDP (HDR-VDP-2) [35]
			Noise Quality Measure (NQM)[36]
			Weighted Signal-to-Noise ratio (WSNR)[37]
			Peak Signal-to-Noise ratio-Human Visual System (PHVS)[38]
			Peak Signal-to-Noise ratio-Human Visual System-A (PHA)[39]
			Peak Signal-to-Noise ratio-Human Visual System(modified) (PHVSM)[40]
			Peak Signal-to-Noise ratio-Human Visual System(modified)-A (PHMA)[39]
	Information Theory based		Information Fidelity Criterion (IFC)[41]
			Visual Information Fidelity (VIF)[42]
			Information Content Weighted SSIM (IW-SSIM)[43]
	Feature Similarity based		Feature Similarity Index (FSIM)[44]
			Gradient Magnitude Similarity Deviation (GMSD)[45]
			Gradient Similarity Measure (GSM)[46]
			Riesz-transform based Feature Similarity Metric (RFSIM)[47]
	Visual Saliency based		Visual Saliency-Induced Index (VSI)[48]
			Spectral Residual Based Similarity (SR-SIM)[49]
	Strategy based		Most Apparent Distortion algorithm (MAD)[11]
			Visual Signal-to-Noise ratio (VSNR)[50]
Reduced Reference	Natural Scene Statistics based		Reduced-Reference Image Quality Assessment (Wavelet Domain) (RRIQA)[51]
			RRIQA with Divisive Normalization (RRDNT)[52]
			Reduced-Reference Entropic Differences (RRED) [53]
	Image Feature based		RRIQA with Weibull Statistics[54]
			RRIQA with Sub-Image Similarity[55]
			RRIQA with Edge-Pattern map[56]
No-Reference	Artifact based	Blur	Local Phase Coherence based sharpness index LPC-SI[57]
			Metric based on Cumulative Probability of Blur Detection (CPBD)[58]
			Metric based on Just-Noticeable Blur (JNBM)[59]
			Spectral and Spatial Measure of Local Perceived Sharpness (S_3)[60]
			Fast Wavelet-Based Image Sharpness Estimation (FISH)[61]
		Blocking	NRIQA of JPEG compressed images (JPEG-NR)[62]
			NRIQA of JPEG compressed images via Quality Relevance Map (NJQA)[63]
		Noise	Noise-level Estimation using weak textured patches (NLWT)[64]
			Fast Noise Variance Estimation (FNVE)[65]
	Learning based	Spatial Domain	Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE)[66]
			Natural Image Quality Evaluator (NIQE)[28]
			Derivative Statistics-based QUality Evaluator (DESIQUE)[67]
			Gradient-Image Quality Assessment (G-IQA-1 and G-IQA-2)(Proposed)
			Gradient Magnitude and Laplacian of Gaussian based NR-IQA (GM-LOG)[68]
		Transform Domain	Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE)[69]
			Complex-DIIVINE (C-DIIVINE)[70]
			Blind Image Quality Index (BIQI)[71]
			BLind Image Integrity Notator using DCT Statistics-II (BLIINDS-II)[72]
			General Regression Neural Network IQA (GRNN)[73]
			NR-IQA based on Curvelets (CurveletQA)[74]
			NR-IQA based on Anisotropy (Anisotropy)[75]
			COdebook Representation for No-Reference Image Assessment (CORNIA)[76]
			Topic Model based IQA (TM-IQA)[77]

Table 3.2: Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between the algorithm scores and the DMOS for various FR-IQA Algorithms along with algorithm computation time (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU). PSNR is Peak Signal-to-Noise Ratio. The table has been sorted in the descending order of SROCC for the "Overall" category. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, computed by bootstrapping using 100 samples. Bold values indicate the best performing algorithm for that category. SSIM-D computes SSIM on images downsampled by a factor determined by image dimensions and viewing distance[83]

	IQA	Interp.		Blur		Additive Noise		JPEG Blocking		Fast Fading		Overall (Confidence Interval)		Time
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	(seconds)
1	GMSD	0.727	0.743	0.827	0.838	**0.923**	0.925	0.918	0.954	0.922	0.915	**0.892** (0.877,0.905)	**0.890** (0.871, 0.905)	0.014
2	SR-SIM	0.752	0.772	0.823	0.729	0.916	0.878	0.925	0.832	0.920	0.913	0.880 (0.853, 0.902)	0.873 (0.834, 0.891)	0.042
3	FSIMc	0.694	0.697	0.802	0.808	0.902	0.917	0.938	0.874	0.911	0.907	0.877 (0.855, 0.896)	0.874 (0.850, 0.891)	0.133
4	FSIM	0.692	0.697	0.801	0.809	0.902	0.917	**0.940**	**0.965**	0.907	0.902	0.876 (0.857, 0.898)	0.872 (0.854, 0.892)	0.165
5	VSI	0.692	0.663	0.811	0.814	0.914	0.883	0.880	0.844	0.923	0.917	0.872 (0.856, 0.897)	0.873 (0.855, 0.889)	0.114
6	MAD	**0.788**	**0.806**	0.813	0.815	0.909	0.915	0.933	0.950	**0.927**	0.917	0.863 (0.834, 0.880)	0.869 (0.846, 0.889)	1.257
7	PHA	0.716	0.717	0.781	0.772	0.842	0.883	0.898	0.927	0.905	0.900	0.863 (0.844, 0.884)	0.861 (0.840, 0.879)	0.458
8	PHMA	0.737	0.755	0.823	0.822	0.852	0.889	0.924	0.953	0.911	0.904	0.853 (0.822, 0.878)	0.859 (0.837, 0.881)	0.234
9	PHVS	0.717	0.718	0.778	0.771	0.876	0.885	0.896	0.926	0.903	0.897	0.853 (0.832, 0.874)	0.846 (0.822, 0.863)	0.195
10	GSM	0.676	0.630	0.780	0.655	0.919	**0.927**	0.903	0.881	0.921	0.678	0.839 (0.811, 0.866)	0.627 (0.584, 0.697)	0.054
11	PHVSM	0.736	0.748	**0.839**	**0.840**	0.854	0.874	0.925	0.954	0.905	0.902	0.833 (0.808, 0.857)	0.838 (0.813, 0.862)	0.207
12	IW-SSIM	0.761	0.793	0.823	0.836	0.902	0.921	0.933	0.959	0.925	**0.922**	0.827 (0.796, 0.849)	0.831 (0.790, 0.847)	0.663
13	RFSIM	0.706	0.717	0.763	0.766	0.906	0.912	0.907	0.930	0.891	0.886	0.825 (0.794, 0.846)	0.826 (0.796, 0.850)	0.218
14	SSIM-D	0.688	0.681	0.772	0.777	0.915	0.922	0.904	0.943	0.914	0.906	0.796 (0.758, 0.823)	0.801 (0.775, 0.833)	0.052
15	IFC	0.728	0.722	0.792	0.789	0.837	0.845	0.913	0.922	0.850	0.858	0.791 (0.757, 0.829)	0.786 (0.742, 0.814)	1.199
16	NQM	0.751	0.767	0.831	0.837	0.879	0.893	0.919	0.936	0.859	0.854	0.789 (0.760, 0.818)	0.796 (0.761, 0.822)	0.107
17	QSSIM	0.697	0.693	0.774	0.647	0.913	0.925	0.905	0.940	0.918	0.915	0.786 (0.758, 0.815)	0.793 (0.753, 0.812)	0.104
18	UQI	0.707	0.704	0.780	0.678	0.816	0.824	0.869	0.889	0.848	0.848	0.767 (0.718, 0.791)	0.776 (0.748, 0.818)	0.040
19	CIELAB	0.575	0.572	0.623	0.627	0.840	0.870	0.910	0.925	0.875	0.878	0.758 (0.716, 0.795)	0.772 (0.736, 0.812)	0.116
20	VIF	0.716	0.737	0.788	0.802	0.874	0.903	0.901	0.925	0.761	0.778	0.755 (0.710, 0.799)	0.748 (0.705, 0.782)	6.337
21	WSNR	0.627	0.638	0.773	0.777	0.821	0.825	0.886	0.911	0.839	0.845	0.744 (0.705, 0.780)	0.745 (0.700, 0.775)	0.048
22	HDR-VDP	0.662	0.699	0.766	0.795	0.854	0.861	0.791	0.790	0.856	0.863	0.712 (0.666, 0.753)	0.738 (0.698, 0.768)	2.245
23	MS-SSIM	0.623	0.635	0.646	0.650	0.908	0.924	0.871	0.891	0.903	0.900	0.699 (0.660, 0.742)	0.712 (0.678, 0.764)	0.276
24	VIFP	0.651	0.661	0.624	0.623	0.895	0.912	0.878	0.887	0.791	0.802	0.693 (0.655, 0.729)	0.695 (0.655, 0.730)	0.244
25	VSNR	0.607	0.619	0.611	0.600	0.848	0.889	0.756	0.771	0.884	0.882	0.690 (0.639, 0.734)	0.696 (0.652, 0.741)	0.237
26	PSNR	0.565	0.591	0.481	0.492	0.864	0.897	0.695	0.702	0.846	0.858	0.590 (0.529, 0.632)	0.603 (0.556, 0.645)	0.149
27	SSIM	0.463	0.476	0.440	0.455	0.909	0.927	0.633	0.653	0.797	0.815	0.542 (0.482, 0.590)	0.531 (0.481, 0.592)	0.570
28	MSVD	0.165	0.160	0.403	0.397	0.415	0.423	0.652	0.630	0.363	0.400	0.261 (0.176, 0.341)	0.253 (0.167, 0.321)	2.272

Table 3.3: Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for various FR-IQA Algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category. The bold values indicate the best performing algorithm for that category.

	IQA	Interp.			Blur			Additive Noise			JPEG Blocking			Fast Fading			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	GMSD	5.675	3.202	**0.0**	6.400	0.632	**0.0**	4.789	1.746	**0.0**	8.756	4.411	**0.0**	12.355	2.632	1.0	10.689	2.065	**0.0**
2	SR-SIM	6.935	1.230	**0.0**	7.065	2.159	1.0	4.641	1.048	**0.0**	7.463	0.801	2.0	12.549	3.105	14.0	10.808	3.539	7.6
3	FSIMc	7.308	2.886	**0.0**	6.856	0.768	**0.0**	5.301	1.376	**0.0**	8.093	1.461	**0.0**	8.166	2.626	6.0	9.182	3.043	4.2
4	FSIM	6.876	2.964	**0.0**	5.885	1.094	**0.0**	5.835	1.782	**0.0**	7.285	1.559	**0.0**	9.382	2.195	5.0	9.373	2.265	4.6
5	VSI	5.441	0.860	1.0	5.128	1.141	**0.0**	3.904	2.757	**0.0**	6.999	0.657	**0.0**	9.227	1.883	13.0	7.725	1.014	5.6
6	MAD	6.225	1.682	**0.0**	6.012	1.492	**0.0**	4.113	1.020	**0.0**	7.264	0.509	**0.0**	8.122	2.979	**0.0**	8.145	6.005	0.4
7	PHA	6.261	1.164	**0.0**	5.950	2.803	1.0	4.665	4.098	**0.0**	5.960	0.281	**0.0**	7.589	2.138	1.0	7.483	2.957	0.4
8	PHMA	5.981	2.862	**0.0**	**5.069**	1.439	**0.0**	5.016	3.620	**0.0**	4.756	1.733	**0.0**	7.481	2.077	**0.0**	8.111	3.507	1.0
9	PHVS	**5.298**	1.164	**0.0**	5.594	3.014	1.0	4.143	2.923	**0.0**	5.521	0.621	**0.0**	**6.918**	0.829	**0.0**	6.886	2.016	0.6
10	GSM	6.402	1.127	**0.0**	5.506	2.548	1.0	4.000	2.546	**0.0**	7.670	0.405	**0.0**	10.626	2.612	17.0	9.214	1.534	7.6
11	PHVSM	6.157	2.863	**0.0**	5.094	1.431	**0.0**	**3.701**	3.704	**0.0**	**4.740**	1.576	**0.0**	7.087	1.669	1.0	**6.335**	3.284	0.4
12	IW-SSIM	6.402	4.254	**0.0**	5.491	1.511	**0.0**	3.989	1.383	**0.0**	7.721	1.821	1.0	10.765	3.006	3.0	9.283	2.109	1.0
13	RFSIM	8.607	1.263	**0.0**	5.424	1.340	**0.0**	4.704	0.893	**0.0**	8.455	0.601	1.0	11.731	4.082	**0.0**	10.437	2.434	2.6
14	SSIM-D	7.213	2.718	**0.0**	8.213	1.025	**0.0**	4.462	1.403	**0.0**	11.477	1.486	3.0	8.847	1.946	4.0	11.171	4.429	7.6
15	IFC	6.344	1.422	2.0	6.866	1.314	**0.0**	6.206	0.638	**0.0**	9.522	0.608	**0.0**	9.612	1.632	3.0	8.818	1.729	6.8
16	NQM	7.409	1.118	**0.0**	7.021	1.040	**0.0**	5.375	3.101	**0.0**	6.946	1.146	**0.0**	8.859	1.064	**0.0**	10.415	1.934	2.4
17	QSSIM	8.813	3.107	**0.0**	8.578	3.258	**0.0**	9.142	0.694	**0.0**	12.267	0.665	**0.0**	16.062	2.783	1.0	13.426	5.622	5.4
18	UQI	6.697	1.550	**0.0**	7.307	1.928	1.0	4.150	0.318	**0.0**	7.883	1.379	7.0	10.196	1.177	3.0	10.017	3.893	2.6
19	CIELAB	6.447	0.234	**0.0**	5.872	1.052	**0.0**	5.362	4.063	**0.0**	9.675	0.590	**0.0**	8.779	0.651	1.0	9.357	2.711	3.0
20	VIF	7.038	1.417	**0.0**	7.497	1.560	**0.0**	5.123	4.230	**0.0**	10.647	1.648	**0.0**	8.230	2.038	7.0	10.015	6.305	4.2
21	WSNR	5.742	1.580	1.0	5.200	0.095	**0.0**	4.660	0.910	1.0	6.363	1.019	1.0	9.028	0.974	4.0	8.567	1.058	4.8
22	HDR-VDP	5.980	1.766	**0.0**	5.322	1.515	**0.0**	4.846	0.493	**0.0**	4.785	1.263	5.0	7.316	1.598	3.0	7.370	0.667	4.6
23	MS-SSIM	6.535	3.464	**0.0**	5.880	1.601	**0.0**	4.503	0.417	**0.0**	8.727	2.223	**0.0**	10.012	2.142	5.0	10.247	6.758	8.4
24	VIFP	6.093	2.373	**0.0**	5.693	1.058	1.0	4.448	3.016	**0.0**	6.001	2.211	1.0	10.858	1.500	9.0	9.209	2.751	5.4
25	VSNR	6.899	0.544	1.0	7.103	0.201	1.0	4.072	0.267	**0.0**	7.201	0.392	6.0	9.955	2.400	1.0	11.006	4.417	6.8
26	PSNR	6.681	1.753	**0.0**	6.822	2.059	1.0	5.591	6.533	**0.0**	9.249	1.316	8.0	13.111	2.197	1.0	12.697	1.682	9.2
27	SSIM	7.325	1.278	2.0	7.278	1.727	**0.0**	5.005	1.156	**0.0**	6.006	0.237	11.0	8.183	2.069	6.0	8.818	1.167	11.2
28	MSVD	6.260	1.880	2.0	5.934	0.603	1.0	4.697	2.038	2.0	5.936	2.877	17.0	7.554	0.880	27.0	7.168	3.128	16.8

Table 3.4: Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between the algorithm scores and the DMOS for various RR-IQA Algorithms along with algorithm computation time (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU). PSNR is Peak Signal-to-Noise Ratio. The table has been sorted in the descending order of SROCC for the "Overall" category. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, computed by bootstrapping using 100 samples. Bold values indicate the best performing algorithm for that category.

	IQA	Interp.		Blur		Additive Noise		JPEG Blocking		Fast Fading		Overall (Confidence Interval)		Time
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	(seconds)
1	RRED	**0.691**	**0.694**	**0.813**	**0.815**	0.908	0.923	0.878	0.892	0.798	0.802	**0.658** (0.593,0.702)	**0.666** (0.611, 0.706)	5.380
2	RRSIS	0.381	0.471	0.772	0.805	0.888	0.900	**0.938**	**0.955**	0.838	**0.853**	0.624 (0.537, 0.676)	0.635 (0.584, 0.686)	3.290
3	RRDNT	0.478	0.508	0.643	0.657	0.918	**0.928**	0.703	0.745	0.657	0.677	0.394 (0.311, 0.488)	0.406 (0.335, 0.487)	15.100
4	RREdge	0.424	0.489	0.578	0.589	0.842	0.871	0.747	0.809	0.690	0.707	0.351 (0.261, 0.420)	0.359 (0.297, 0.421)	2.290
5	RRIQA	0.206	0.243	0.613	0.628	0.822	0.840	0.621	0.686	0.669	0.738	0.349 (0.264, 0.429)	0.348 (0.268, 0.416)	5.920
6	RRWeibull	0.401	0.302	0.789	0.793	**0.918**	0.919	0.860	0.869	**0.844**	0.842	0.299 (0.203, 0.385)	0.400 (0.337, 0.463)	7.100

Table 3.5: Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for various RR-IQA Algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category. The bold values indicate the best performing algorithm for that category.

	IQA	Interp.			Blur			Additive Noise			JPEG Blocking			Fast Fading			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	RRED	**6.490**	3.579	**0.0**	**5.486**	2.816	**0.0**	4.061	0.611	**0.0**	7.173	0.670	**0.0**	9.885	1.553	6.8	**10.264**	6.322	**6.1**
2	RRSIS	7.887	1.069	**0.0**	5.818	1.742	**0.0**	5.206	2.342	**0.0**	**4.798**	0.883	**0.0**	**8.645**	2.373	3.4	10.621	2.606	7.7
3	RRDNT	7.823	0.762	**0.0**	7.057	0.559	**0.0**	**3.854**	1.045	**0.0**	11.184	2.179	8.0	12.578	1.201	12.5	12.566	1.826	14.3
4	RREdge	7.876	0.860	**0.0**	7.593	2.045	**0.0**	5.026	0.464	**0.0**	9.873	1.885	5.0	12.713	2.972	9.0	12.952	2.503	17.0
5	RRIQA	8.772	0.428	2.3	7.288	1.655	**0.0**	5.768	1.361	**0.0**	12.226	2.473	14.8	11.951	2.017	11.4	12.894	1.390	16.1
6	RRWeibull	8.544	0.200	**0.0**	6.049	0.869	**0.0**	4.350	4.098	**0.0**	11.321	1.330	10.2	8.933	0.967	**2.3**	12.650	4.955	15.5

Table 3.6: Median Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between algorithm scores and DMOS for various NR-IQA algorithms (described in Section 3.3.3) along with algorithm computation time needed (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU) across 100 train-test (4:1) combinations on the ESPL Synthetic Image Database (50 trials for CORNIA in row 2). Italicized entries are NR-IQA algorithms meant for particular distortion categories. Italicized algorithms indicate the values obtained when the mentioned NR-IQA algorithms were applied for distortion categories other than what they were originally intended for. For these algorithms, the correlation values quoted in the "Overall" category is same as the correlations in the distortion category for which the algorithm was originally meant for. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, obtained by considering the maximum and minimum values of the correlations obtained over a 100 trials. The table has been sorted in the descending order of SROCC for the "Overall" category. Bold values indicate the best performing algorithm for that category.

	IQA	Interp.		Blur		GN		JPEG		FF		Overall (Confidence Interval)		Time
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	(s)
1	G-IQA-1 (L)	0.605	0.646	0.612	0.640	0.858	0.904	0.901	0.927	**0.774**	**0.833**	**0.813** (0.562, 0.918)	**0.819**(0.626, 0.911)	2.134
2	CORNIA	**0.808**	**0.823**	**0.775**	**0.801**	0.793	0.821	0.898	0.918	0.706	0.763	0.810 (0.687, 0.875)	0.807(0.682, 0.880)	84.330
3	C-DIIVINE	0.702	0.760	0.730	0.769	0.847	0.896	0.841	0.879	0.738	0.802	0.798 (0.691, 0.916)	0.808(0.712, 0.912)	65.720
4	BRISQUE	0.631	0.643	0.720	0.782	0.840	0.902	0.898	0.935	0.717	0.740	0.789 (0.663, 0.897)	0.795(0.690, 0.895)	0.590
5	GM-LOG	0.680	0.711	0.653	0.694	0.853	0.906	0.912	0.944	0.701	0.746	0.787 (0.627, 0.893)	0.791(0.594, 0.892)	0.590
6	G-IQA-1	0.580	0.647	0.474	0.508	0.871	0.920	0.922	0.942	0.726	0.758	0.774 (0.552, 0.893)	0.786(0.569, 0.887)	4.641
7	DESIQUE	0.595	0.678	0.590	0.617	**0.886**	**0.922**	**0.934**	**0.955**	0.714	0.737	0.773 (0.570, 0.909)	0.781(0.588, 0.901)	2.250
8	G-IQA-2	0.510	0.584	0.565	0.576	0.857	0.906	0.865	0.879	0.728	0.762	0.743 (0.387, 0.888)	0.744(0.406, 0.877)	42.693
9	CurveletQA	0.658	0.695	0.695	0.753	0.880	0.916	0.854	0.880	0.553	0.595	0.731 (0.460, 0.872)	0.734(0.490, 0.863)	20.130
10	G-IQA-2 (L)	0.509	0.563	0.488	0.529	0.859	0.906	0.874	0.909	0.668	0.729	0.689 (0.489, 0.876)	0.714(0.538, 0.881)	14.893
11	BIQI	0.665	0.733	0.732	0.764	0.837	0.903	0.735	0.769	0.538	0.593	0.676 (0.338, 0.849)	0.676(0.414, 0.858)	0.330
12	GRNN	0.537	0.592	0.371	0.409	0.811	0.896	0.738	0.790	0.408	0.551	0.602 (0.422, 0.777)	0.643(0.422, 0.802)	2.480
13	BLIINDS-II	0.388	0.444	0.499	0.556	0.794	0.839	0.680	0.754	0.548	0.608	0.596 (0.333, 0.834)	0.622(0.382, 0.835)	81.790
14	Anisotropy	0.364	0.354	0.357	0.400	0.835	0.871	0.385	0.449	0.392	0.439	0.470 (0.379, 0.513)	0.431(0.391, 0.483)	10.780
15	NIQE	0.428	0.496	0.425	0.528	0.740	0.511	0.732	0.834	0.606	0.623	0.377 (0.144, 0.600)	0.395(0.181, 0.601)	3.240
16	DIIVINE	0.421	0.523	0.441	0.490	0.484	0.537	0.444	0.489	0.439	0.513	0.372 (0.080, 0.700)	0.404(0.121, 0.705)	118.040
17	TMIQA	0.367	0.376	0.437	0.353	0.741	0.681	0.159	0.227	0.411	0.469	0.220 (0.097, 0.300)	0.311(0.223, 0.387)	0.120
18	*LPCM*	*0.415*	*0.444*	**0.836**	**0.847**	*0.623*	*0.621*	*0.211*	*0.231*	*0.108*	*0.237*	**0.836**(0.791, 0.890)	**0.847**(0.792, 0.885)	11.570
19	*CPBDM*	*0.676*	*0.720*	0.757	0.766	*0.746*	*0.815*	*0.765*	*0.749*	*0.347*	*0.405*	0.757 (0.678, 0.808)	0.766(0.669, 0.830)	3.500
20	*FISH*	*0.222*	*0.305*	0.705	0.716	*0.823*	*0.870*	*0.196*	*0.252*	*0.432*	*0.472*	0.705 (0.548, 0.787)	0.716(0.631, 0.793)	0.250
21	S_3	*0.409*	*0.449*	0.700	0.756	*0.747*	*0.786*	*0.151*	*0.189*	*0.402*	*0.450*	0.700 (0.554, 0.792)	0.756(0.692, 0.818)	308.150
22	*JNBM*	*0.598*	*0.635*	0.506	0.528	*0.756*	*0.816*	*0.536*	*0.512*	*0.448*	*0.455*	0.506 (0.327, 0.627)	0.528(0.336, 0.676)	7.520
23	*NLWT*	*0.324*	*0.334*	*0.024*	*0.141*	**0.872**	**0.888**	*0.000*	*0.187*	*0.559*	*0.589*	**0.872**(0.821, 0.905)	**0.888**(0.847, 0.928)	10.410
24	*FNVE*	*0.320*	*0.332*	*0.463*	*0.553*	0.863	0.887	*0.517*	*0.543*	*0.461*	*0.459*	0.863 (0.817, 0.894)	0.887(0.838, 0.915)	0.030
25	*JPEG-NR*	*0.540*	*0.570*	*0.593*	*0.650*	*0.748*	*0.865*	**0.928**	**0.954**	*0.464*	*0.607*	**0.928**(0.878, 0.952)	**0.954**(0.940, 0.969)	0.110
26	*NJQA*	*0.373*	*0.406*	*0.333*	*0.367*	*0.878*	*0.808*	0.743	0.819	*0.420*	*0.437*	0.743 (0.649, 0.854)	0.819(0.732, 0.869)	192.590

Table 3.7: Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for various NR-IQA Algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category. The bold values indicate the best performing algorithm for that category.

	IQA	Interp.			Blur			Additive Noise			JPEG Blocking			Fast Fading			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	G-IQA-1 (L)	6.981	0.006	**0.000**	8.326	0.665	**0.000**	4.690	3.608	**0.000**	14.908	0.172	20.000	13.615	0.202	17.500	9.209	4.099	3.000
2	CORNIA	**0.112**	0.057	**0.000**	**0.131**	0.719	**0.000**	**0.136**	0.051	**0.000**	**0.151**	0.662	**0.000**	**0.262**	0.445	**0.000**	**0.190**	8.032	**0.000**
3	C-DIIVINE	5.897	0.084	**0.000**	8.290	0.306	**0.000**	5.067	0.131	**0.000**	14.858	0.429	20.000	14.269	0.019	20.000	9.631	4.586	4.000
4	BRISQUE	6.747	0.007	**0.000**	6.804	0.017	**0.000**	5.087	1.105	**0.000**	15.202	0.005	25.000	14.214	0.017	20.000	9.231	2.427	4.000
5	GM-LOG	6.182	0.000	**0.000**	7.858	0.166	**0.000**	4.856	2.231	**0.000**	14.953	0.006	20.000	14.846	0.009	20.000	9.579	1.390	5.000
6	G-IQA-1	6.803	0.124	**0.000**	8.256	0.076	**0.000**	4.720	0.689	**0.000**	14.985	0.012	25.000	13.532	0.030	10.000	9.933	10.419	6.000
7	DESIQUE	6.799	0.107	**0.000**	7.993	0.025	**0.000**	4.527	3.408	**0.000**	15.207	0.010	25.000	14.205	0.462	20.000	9.799	1.119	5.000
8	G-IQA-2	7.287	0.201	**0.000**	8.207	0.009	**0.000**	4.956	0.401	**0.000**	15.200	0.003	25.000	13.386	1.144	15.000	10.870	2.906	8.500
9	CurveletQA	6.535	0.215	**0.000**	7.136	0.069	**0.000**	4.735	0.466	**0.000**	15.152	0.004	25.000	15.279	0.434	25.000	11.272	6.938	9.000
10	G-IQA-2 (L)	7.480	0.155	**0.000**	8.250	0.280	**0.000**	4.912	5.519	**0.000**	15.204	0.002	25.000	14.095	0.923	20.000	10.836	14.526	8.000
11	BIQI	6.177	0.520	**0.000**	8.216	0.970	**0.000**	4.915	0.002	**0.000**	14.838	0.143	20.000	14.514	0.893	25.000	10.741	3.509	9.000
12	GRNN	6.725	0.296	**0.000**	8.318	1.415	**0.000**	5.089	0.778	**0.000**	15.065	0.004	25.000	15.193	0.772	20.000	11.336	4.263	9.500
13	BLIINDS-II	7.546	0.884	**0.000**	7.884	0.686	**0.000**	5.826	0.000	**0.000**	15.312	0.002	25.000	14.689	0.009	20.000	11.060	6.710	9.000
14	Anisotropy	8.496	0.406	**0.000**	9.113	0.934	1.000	2.956	2.626	1.000	9.561	1.618	16.000	14.354	1.308	27.000	10.846	3.328	12.800
15	NIQE	7.683	0.030	**0.000**	8.095	0.234	**0.000**	8.582	0.346	**0.000**	10.994	0.002	5.000	12.394	1.493	10.000	12.490	2.538	14.000
16	DIIVINE	7.682	0.000	**0.000**	8.133	0.028	**0.000**	8.172	0.126	**0.000**	14.874	0.004	20.000	14.724	0.172	25.000	12.632	5.402	14.000
17	TMIQA	14.342	1.373	1.000	10.219	0.478	2.000	5.275	2.338	**0.000**	6.478	4.082	27.000	10.586	1.102	22.000	13.245	2.466	15.200
18	*LPCM*	-	-	-	**4.968**	1.019	**0.000**	-	-	-	-	-	-	-	-	-	**4.968**	1.019	**0.000**
19	*CPBDM*	-	-	-	6.485	0.440	**0.000**	-	-	-	-	-	-	-	-	-	6.485	0.440	**0.000**
20	*FISH*	-	-	-	6.603	0.324	**0.000**	-	-	-	-	-	-	-	-	-	6.603	0.324	**0.000**
21	S_3	-	-	-	6.339	0.162	**0.000**	-	-	-	-	-	-	-	-	-	6.339	0.162	**0.000**
22	*JNBM*	-	-	-	7.952	0.360	1.000	-	-	-	-	-	-	-	-	-	7.952	0.360	1.000
23	*NLWT*	-	-	-	-	-	-	4.611	3.620	**0.000**	-	-	-	-	-	-	**4.611**	3.620	**0.000**
24	*FNVE*	-	-	-	-	-	-	4.626	6.129	**0.000**	-	-	-	-	-	-	4.626	6.129	**0.000**
25	*JPEG-NR*	-	-	-	-	-	-	-	-	-	6.949	1.088	**0.000**	-	-	-	6.949	1.088	**0.000**
26	*NJQA*	-	-	-	-	-	-	-	-	-	9.279	1.453	8.000	-	-	-	9.279	1.453	8.000

Table 3.8: Mean classification accuracy (in percentage) for various NR-IQA algorithms (described in Section 3.3.3) across 100 train-test (4:1) combinations on the ESPL Synthetic Image Database.

IQA	Alias	Blur	GN	JPEG	FF	All
GM-LOG	**99.8**	**96.2**	**100.0**	**96.8**	92.5	**97.1**
C-DIIVINE	91.7	95.3	**100.0**	95.5	**93.3**	95.2
BRISQUE	90.3	95.6	**100.0**	92.8	87.2	93.2
DESIQUE	90.7	87.3	**100.0**	89.1	85.3	90.5
BIQI	89.3	87.9	94.0	92.4	83.0	89.3
G-IQA-1	78.5	83.4	**100.0**	90.2	87.7	88.0
BLIINDS-II	86.2	84.6	**100.0**	81.1	81.8	86.7
CurveletQA	87.0	87.0	**100.0**	81.2	69.5	84.9
DIIVINE	21.8	74.7	80.8	45.1	51.7	54.8

Table 3.9: Results of the F-test performed on the residuals between model predictions and DMOS scores.. Each cell in the table is a codeword consisting of 5 symbols that correspond to "Interpolation"', "Blur", "Gaussian Noise", "JPEG Blocking", "Fast Fading" and "Overall" distortions. "1"("0") indicates that the performance of the row IQA is superior(inferior) to that of the column IQA. - indicates that the statistical performance of the row IQA is equivalent to that of the column IQA. The matrix is symmetric.

	GMSD	FSIM	MS-SSIM	PSNR	RRED	G-IQA-1	CORNIA	BRISQUE	DESIQUE	DIIVINE
GMSD	- - - - - -	- - - - 1 1	- - - - - 1	- - - 1 1 1	- - - 1 1 1	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1
FSIM	- - - - 0 0	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1
MS-SSIM	- - - - - 0	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1	- - 1 1 1 1
PSNR	- - - 0 0 0	- - - - - -	- - - - - -	- - - - - -	- - 0 - - -	- - 1 - 1 -	- - 1 - 1 -	- - 1 - 1 -	- - 1 - 1 -	- - 1 - 1 -
RRED	- - - 0 0 0	- - - - - -	- - - - - -	- - 1 - - -	- - - - - -	- - 1 - 1 1	- - 1 - 1 1	- - 1 - 1 1	- - 1 - 1 1	- - 1 - 1 1
G-IQA-1	- - 0 0 0 0	- - 0 0 0 0	- - 0 0 0 0	- - 0 - 0 -	- - 0 - 0 0	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
CORNIA	- - 0 0 0 0	- - 0 0 0 0	- - 0 0 0 0	- - 0 - 0 -	- - 0 - 0 0	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
BRISQUE	- - 0 0 0 0	- - 0 0 0 0	- - 0 0 0 0	- - 0 - 0 -	- - 0 - 0 0	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
DESIQUE	- - 0 0 0 0	- - 0 0 0 0	- - 0 0 0 0	- - 0 - 0 -	- - 0 - 0 0	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -
DIIVINE	- - 0 0 0 0	- - 0 0 0 0	- - 0 0 0 0	- - 0 - 0 -	- - 0 - 0 0	- - - - - -	- - - - - -	- - - - - -	- - - - - -	- - - - - -

Table 3.10: Variance of the residuals between individual subjective scores and IQA algorithm predictions. Boldfaces indicate the lowest variance of the model residual for that distortion category. Residuals were found to be normally distributed for all the cases.

IQA	Interp.	Blur	GN	JPEG	FF	All
Samples	624	624	624	624	624	3120
F-ratio	1.14	1.14	1.14	1.14	1.14	1.06
GMSD	134.61	168.49	118.82	**143.03**	**119.41**	**151.12**
FSIM	144.37	176.06	115.92	173.38	168.64	196.02
MS-SSIM	**132.91**	166.71	111.32	184.97	129.48	199.03
PSNR	161.85	183.06	126.82	251.87	152.79	236.90
RRED	139.04	**146.70**	**107.66**	214.15	171.08	192.69
G-IQA-1	164.04	195.01	224.36	354.90	422.75	304.58
CORNIA	162.34	194.80	223.70	346.17	426.44	305.56
BRISQUE	163.71	194.91	224.80	347.69	423.59	306.30
DESIQUE	163.20	195.69	224.72	348.85	423.89	305.32
DIIVINE	163.90	196.38	228.08	354.78	426.64	311.40
Null Model	105.23	110.70	101.30	122.82	112.08	110.28

Table 3.11: Variance of the residuals between DMOS values and IQA algorithm predictions. Boldfaces indicate the lowest variance of the model residual for that distortion category. Residuals were found to be normally distributed 80% of the cases.

IQA	Interp.	Blur	GN	JPEG	FF	All
Damples	12	12	12	12	12	60
F-ratio	2.82	2.82	2.82	2.82	2.82	1.54
GMSD	32.01	62.95	19.08	**22.01**	**7.98**	**41.52**
FSIM	42.63	71.18	15.92	55.07	61.60	87.16
MS-SSIM	**30.15**	61.01	10.91	67.69	18.95	90.22
PSNR	61.67	78.81	27.79	140.56	44.34	128.73
RRED	36.83	**39.21**	**6.93**	99.48	64.25	83.77
G-IQA-1	64.06	91.82	134.03	252.77	338.37	197.52
CORNIA	62.21	91.60	133.31	243.27	342.38	198.52
BRISQUE	63.70	91.72	134.51	244.92	339.27	199.27
DESIQUE	63.14	92.57	134.42	246.18	339.61	198.28
DIIVINE	63.90	93.32	138.08	252.64	342.61	204.46

Table 3.12: Median Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between algorithm scores and MOS for various NR-IQA algorithms along with algorithm computation time needed (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU) across 100 train-test (4:1) combinations on the ESPL Synthetic Image Database (50 trials for CORNIA in row 2). Italicized entries are NR-IQA algorithms meant for particular distortion categories. Italicized algorithms indicate the values obtained when the mentioned NR-IQA algorithms were applied for distortion categories other than what they were originally intended for. For these algorithms, the correlation values quoted in the "Overall" category is same as the correlations in the distortion category for which the algorithm was originally meant for. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, obtained by considering the maximum and minimum values of the correlations obtained over a 100 trials. The table has been sorted in the descending order of SROCC for the "Overall" category. Bold values indicate the best performing algorithm for that category.

	IQA	Interp.		Blur		GN		JPEG		FF		Overall (Confidence Interval)		Time
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	(s)
1	CORNIA	**0.892**	**0.902**	**0.877**	**0.889**	0.912	0.947	0.928	0.951	0.771	0.817	**0.868** (0.826, 0.931)	**0.867**(0.817, 0.919)	84.330
2	GM-LOG	0.817	0.832	0.702	0.756	0.929	0.969	0.931	0.953	0.770	0.807	0.855 (0.730, 0.916)	0.853(0.715, 0.914)	0.590
3	G-IQA-1 (L)	0.691	0.777	0.638	0.679	0.948	0.975	0.938	0.956	**0.814**	**0.846**	0.831 (0.554, 0.951)	0.844(0.603, 0.947)	2.134
4	BRISQUE	0.659	0.681	0.797	0.814	0.914	0.962	0.925	0.945	0.738	0.793	0.830 (0.691, 0.944)	0.833(0.692, 0.938)	0.590
5	C-DIIVINE	0.747	0.800	0.734	0.779	0.932	0.969	0.893	0.912	0.762	0.805	0.830 (0.697, 0.925)	0.838(0.701, 0.936)	65.720
6	G-IQA-1	0.635	0.721	0.571	0.612	0.938	0.970	0.944	0.957	0.786	0.837	0.829 (0.664, 0.906)	0.831(0.702, 0.914)	4.641
7	DESIQUE	0.707	0.744	0.660	0.696	**0.952**	**0.976**	**0.945**	**0.962**	0.735	0.769	0.803 (0.472, 0.935)	0.816(0.473, 0.933)	2.250
8	CurveletQA	0.759	0.784	0.704	0.753	0.905	0.960	0.911	0.927	0.602	0.663	0.795 (0.606, 0.877)	0.794(0.620, 0.877)	20.130
9	G-IQA-2	0.571	0.627	0.565	0.600	0.928	0.965	0.902	0.913	0.726	0.760	0.760 (0.580, 0.907)	0.767(0.586, 0.911)	42.693
10	G-IQA-2 (L)	0.483	0.536	0.563	0.586	0.926	0.963	0.911	0.926	0.738	0.805	0.737 (0.396, 0.902)	0.757(0.441, 0.897)	14.893
11	BIQI	0.703	0.804	0.819	0.841	0.880	0.935	0.739	0.765	0.504	0.576	0.700 (0.539, 0.823)	0.692(0.531, 0.808)	0.330
12	BLIINDS-II	0.553	0.552	0.580	0.611	0.862	0.920	0.802	0.862	0.683	0.740	0.688 (0.483, 0.872)	0.701(0.479, 0.860)	81.790
13	GRNN	0.468	0.457	0.246	0.336	0.823	0.940	0.745	0.800	0.489	0.594	0.615 (0.461, 0.765)	0.633(0.485, 0.790)	2.480
14	Anisotropy	0.392	0.433	0.363	0.360	0.893	0.921	0.469	0.505	0.476	0.506	0.532 (0.468, 0.585)	0.481(0.409, 0.514)	10.780
15	NIQE	0.347	0.446	0.453	0.492	0.773	0.522	0.741	0.848	0.644	0.661	0.406 (0.193, 0.629)	0.443(0.253, 0.647)	3.240
16	DIIVINE	0.435	0.476	0.462	0.518	0.526	0.567	0.503	0.540	0.459	0.541	0.385 (0.127, 0.723)	0.460(0.203, 0.718)	118.040
17	TMIQA	0.308	0.307	0.492	0.518	0.772	0.815	0.180	0.138	0.389	0.459	0.285 (0.181, 0.383)	0.330(0.228, 0.409)	0.120
18	*LPCM*	*0.415*	*0.444*	**0.836**	**0.847**	*0.623*	*0.621*	*0.211*	*0.231*	*0.108*	*0.237*	**0.836**(0.791, 0.890)	**0.847**(0.792, 0.885)	11.570
19	*CPBDM*	*0.676*	*0.720*	0.757	0.766	*0.746*	*0.815*	*0.765*	*0.749*	*0.347*	*0.405*	0.757 (0.678, 0.808)	0.766(0.669, 0.830)	3.500
20	*FISH*	*0.222*	*0.305*	0.705	0.716	*0.823*	*0.870*	*0.196*	*0.252*	*0.432*	*0.472*	0.705 (0.548, 0.787)	0.716(0.631, 0.793)	0.250
21	S_3	*0.409*	*0.449*	0.700	0.756	*0.747*	*0.786*	*0.151*	*0.189*	*0.402*	*0.450*	0.700 (0.554, 0.792)	0.756(0.692, 0.818)	308.150
22	*JNBM*	*0.598*	*0.635*	0.506	0.528	*0.756*	*0.816*	*0.536*	*0.512*	*0.448*	*0.455*	0.506 (0.327, 0.627)	0.528(0.336, 0.676)	7.520
23	*NLWT*	*0.361*	*0.350*	*0.000*	*0.056*	**0.905**	**0.943**	*0.000*	*0.206*	*0.622*	*0.638*	**0.905** (0.875, 0.927)	**0.943**(0.923, 0.956)	10.410
24	*FNVE*	*0.304*	*0.328*	*0.497*	*0.528*	0.891	0.939	*0.511*	*0.526*	*0.368*	*0.472*	*0.891* (0.869, 0.918)	0.939(0.907, 0.953)	0.030
25	*JPEG-NR*	*0.611*	*0.509*	*0.605*	*0.613*	*0.775*	*0.915*	**0.951**	**0.966**	*0.497*	*0.625*	**0.951**(0.937, 0.966)	**0.966**(0.952, 0.975)	0.110
26	*NJQA*	*0.448*	*0.415*	*0.307*	*0.284*	*0.872*	*0.932*	0.768	0.837	*0.479*	*0.511*	0.768 (0.667, 0.849)	0.837(0.748, 0.876)	192.590

Table 3.13: Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the MOS for various NR-IQA Algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category. The bold values indicate the best performing algorithm for that category.

	IQA	Interp.			Blur			Additive Noise			JPEG Blocking			Fast Fading			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	CORNIA	**0.099**	0.039	**0.000**	0.113	0.620	**0.000**	**0.097**	0.059	**0.000**	**0.144**	0.415	**0.000**	**0.266**	0.059	**0.000**	**0.190**	5.979	**0.000**
2	GM-LOG	4.664	2.040	**0.000**	5.973	0.695	**0.000**	2.553	0.067	**0.000**	5.106	3.931	**0.000**	12.013	0.066	15.000	8.426	14.715	7.000
3	G-IQA-1 (L)	5.476	0.081	**0.000**	6.540	0.216	**0.000**	2.330	0.066	**0.000**	5.128	0.111	**0.000**	10.231	1.120	12.500	12.344	3.825	22.000
4	BRISQUE	6.170	0.005	5.000	5.185	1.565	**0.000**	3.139	0.104	**0.000**	5.608	3.309	**0.000**	11.717	2.501	15.000	8.325	23.099	7.500
5	C-DIIVINE	5.147	0.531	**0.000**	5.845	0.300	**0.000**	2.842	0.033	**0.000**	6.840	0.364	**0.000**	11.420	0.150	20.000	12.257	3.478	22.000
6	G-IQA-1	5.789	0.115	**0.000**	6.988	0.083	**0.000**	2.359	0.075	**0.000**	4.961	0.301	**0.000**	10.248	0.111	10.000	12.377	2.543	22.000
7	DESIQUE	5.584	0.158	**0.000**	6.415	0.856	**0.000**	2.184	0.044	**0.000**	4.760	0.040	**0.000**	12.557	1.810	20.000	12.163	2.659	21.000
8	CurveletQA	5.174	1.097	**0.000**	6.053	0.309	**0.000**	3.162	0.004	**0.000**	6.797	0.001	5.000	13.932	0.191	30.000	10.206	3.488	14.000
9	G-IQA-2	6.378	0.168	5.000	7.038	0.593	5.000	2.621	0.095	**0.000**	7.176	0.559	5.000	11.929	1.861	15.000	10.345	3.044	14.000
10	G-IQA-2 (L)	6.736	0.001	5.000	7.106	0.018	5.000	3.150	0.524	**0.000**	6.651	0.346	5.000	12.164	1.531	20.000	12.332	2.555	21.500
11	BIQI	5.272	0.183	**0.000**	4.885	3.948	**0.000**	3.917	0.238	**0.000**	10.977	1.539	15.000	13.896	0.103	25.000	10.423	2.344	17.000
12	BLIINDS-II	6.480	0.258	5.000	7.081	0.359	5.000	4.550	1.289	**0.000**	9.042	0.295	10.000	12.336	1.879	20.000	11.238	3.506	17.500
13	GRNN	7.065	1.777	5.000	8.078	0.411	5.000	4.416	0.004	**0.000**	10.472	1.110	15.000	14.277	0.195	25.000	11.538	3.025	19.500
14	Anisotropy	6.023	1.886	8.000	**0.000**	2.168	5.000	7.664	0.316	**0.000**	8.567	3.153	33.000	11.341	2.090	33.000	11.882	3.302	23.400
15	NIQE	7.079	0.005	5.000	7.909	0.113	5.000	7.974	0.100	5.000	8.565	0.716	15.000	12.550	0.256	20.000	12.105	12.093	22.000
16	DIIVINE	7.004	0.512	5.000	7.676	0.941	5.000	7.734	0.152	5.000	13.124	0.286	30.000	14.026	0.138	25.000	12.569	4.401	24.000
17	TMIQA	7.718	2.035	8.000	7.338	0.322	5.000	9.033	1.359	**0.000**	11.744	6.088	35.000	11.287	0.705	27.000	12.338	1.578	24.400
18	*LPCM*	-	-	-	**4.968**	1.019	**0.000**	-	-	-	-	-	-	-	-	-	**4.968**	1.019	**0.000**
19	*CPBDM*	-	-	-	6.485	0.440	**0.000**	-	-	-	-	-	-	-	-	-	6.485	0.440	**0.000**
20	*FISH*	-	-	-	6.603	0.324	**0.000**	-	-	-	-	-	-	-	-	-	6.603	0.324	**0.000**
21	S_3	-	-	-	6.339	0.162	**0.000**	-	-	-	-	-	-	-	-	-	6.339	0.162	**0.000**
22	*JNBM*	-	-	-	7.952	0.360	1.000	-	-	-	-	-	-	-	-	-	7.952	0.360	1.000
23	*NLWT*	-	-	-	-	-	-	3.181	1.989	**0.000**	-	-	-	-	-	-	**3.181**	1.989	**0.000**
24	*FNVE*	-	-	-	-	-	-	3.337	1.793	**0.000**	-	-	-	-	-	-	3.337	1.793	**0.000**
25	*JPEG-NR*	-	-	-	-	-	-	-	-	-	6.815	1.421	5.000	-	-	-	6.815	1.421	5.000
26	*NJQA*	-	-	-	-	-	-	-	-	-	9.937	1.232	13.000	-	-	-	9.937	1.232	13.000

Chapter 4

Objective evaluation of tone-mapping artifacts in HDR images

4.1 Introduction

[1]The range of radiance values encountered in the real world far exceeds the range that can be captured by a photographic sensor. To deal with this limitation, recent years have seen a huge growth in the popularity of High Dynamic Range (HDR) images due to their ability to accurately represent the wide range of variation of illumination in real scenes. Unlike traditional Standard Dynamic Range (SDR) scenes with 8 bits/color/pixel, the range of the luminance levels in HDR scenes can range from 10,000 to 1[5]. Apart from natural scenes, HDR rendering also finds its use in computer graphics where the lighting calculations are performed over a wider dynamic range. This results in a better contrast variation leading to a higher degree of detail preservation.

For photographs taken under challenging lighting conditions, an image taken at a single exposure may contain overexposed and underexposed regions. Hence, a widely used approach for generating well-exposed SDR images is to begin with a stack of registered images taken at different exposures[87] (typically taken in the Automatic Exposure Bracketing mode of digital SLR cameras to intentionally underexpose and overexpose the

[1]Contents of this capter has been published in [86]

scene) and performing the following three post-processing steps:

- Estimating the radiometric response function of the camera from the registered images.
- Estimating a radiance map by merging pixels from different exposures to get the HDR image.
- Tone-mapping the HDR image to an SDR image to visualize the images on standard displays meant for HDR images. The resulting HDR is more visually appealing and informative than any single-exposure image.

However, a different class of algorithms is used in many consumer electronic devices in order to generate a sufficiently detailed well-exposed image by bypassing the intermediate step of constructing an HDR radiance map[88]. These multi-exposure fusion (MEF) algorithms take as input a stack of registered images taken at different exposures and outputs an image in which the details are clearly visible both in the underexposed and overexposed regions.

Different tone-mapping operators (TMO) or MEF algorithms may result in different SDR visualization, so a natural question is how to gauge the quality of the images obtained. In addition, quality evaluation of compressed tone-mapped images is an emerging problem that involves the joint optimization of tone-mapping and compression parameters. This chapter focuses on FR-IQA algorithms for comparing the tone-mapped SDR image with the original HDR image.

In a previous work, the Tone-Mapped image Quality Index (TMQI)[5] compares the original HDR image with the rendered SDR image. TMQI quantifies distortions locally and pools them by uniform averaging, in addition to measuring naturalness of the SDR image. For SDR images, perceptual pooling strategies have improved correlation of image quality assessment (IQA) algorithms with subjective scores. In this chapter we outline different perceptual pooling strategies for the TMQI IQA algorithm, propose a NSS based model for quantifying image naturalness and test the proposed methods on JPEG compressed tone-mapped images and tone-mapped images for SDR displays using human subjective scores.

This chapter is organized as follows: Section 4.2 summarizes the tone-mapping algorithms used to generate the SDR visualizations (the MEF algorithms will be described in the next section), section 4.3 outlines the previous FR-IQA algorithms proposed for evaluating tone-mapping artifacts, section 4.4 describes the TMQI IQA algorithm, the proposed contributions have been outlined in 4.5, the experimental results have beem mentioned in section4.6. Section 4.7 concludes the chapter.

4.2 HDR Image Creation

The process of generating well-exposed SDR scenes involves estimating the scene radiance map, followed by tone-mapping it to the displayable gamut of the SDR displays.

4.2.1 Creating scene radiance map

Some of the earliest algorithms for estimating the radiance map of a natural scene in the HDR format was proposed in [89][90][91] by using photographs taken with conventional

digital cameras. From the multiple photographs of the same scene taken with different degrees of exposures, the algorithm first recovers the camera response function (up to a factor of scale) and uses that to fuse multiply exposed images into a single HDR radiance map whose pixel values are proportional to the true radiance values of the scene.

It is presumed that the scene is static and the series of images photographed by deliberately changing the exposure is taken in quick succession so that the lighting changes in the actual scene can be safely ignored. In [89] the digitized images are assumed to be taken with the same camera position with different known exposure durations t_j. E_i represents the irradiance values for each pixel (assumed to be constant), Z_{ij} represents the pixel values, where i is the spatial index and j indexes over the exposure times t_j. The sensor reciprocity equation is expressed as:

$$Z_{ij} = f(E_i t_j) \tag{4.1}$$

Since f is assumed to be monotonic, it is invertible. Hence 4.1 can be expressed as:

$$f^{-1}(Z_{ij}) = E_i t_j \tag{4.2}$$

Taking the natural logarithm on both sides we get,

$$\ln f^{-1}(Z_{ij}) = \ln(E_i) + \ln(t_j) \tag{4.3}$$

Let $g = \ln f^{-1}$. Then we have the set of equations:

$$g(Z_{ij}) = \ln(E_i) + \ln(t_j) \tag{4.4}$$

where i ranges over the pixels and j indexes the exposure durations. The unknown exposures E_i and the response values $g_k = g(k)$ (discretized according to the 256 pixel values commonly observed in eight-bit images) is solved via least squares.

In computer graphics, HDR rendering is becoming increasingly popular now-a-days and is supported by OpenGL, the de-facto standard for rendering 2D and 3D graphics. Instead of clamping the color values in the frame-buffer between the values of 0.0 and 1.0 after each fragment shader run, HDR OpenGL rendering allows the colors to be saved in true floating point values outside the default range of 0.0 and 1.0. This leads to the preservation of more detail as well as provides the designer with the ability to configure the lighting of the scene better with more realistic lighting parameters [92].

4.2.2 Tone-Mapping algorithms

Once the radiance map is obtained, either from the multiply exposed images in case of photographs taken with optical images or the floating point rendered color values in computer graphics, it is tonemapped to a lower gamut (8 bit/color/pixel) of the SDR display. These algorithms try to replicate the local-adaptation behavior of the human visual system. The human eye deals with the vast range of real-world illuminations by changing their sensitivity to be responsive at different illumination levels in a highly localized fashion, enabling us to see the details both in the bright and dark regions [93]. The tone-mapping algorithms compute either a spatially varying transfer function or shrinks the image gradients to fit into the available dynamic range [87]. Depending on the type of processing involved, the tone-mapping algorithms may be classified into the following classes:

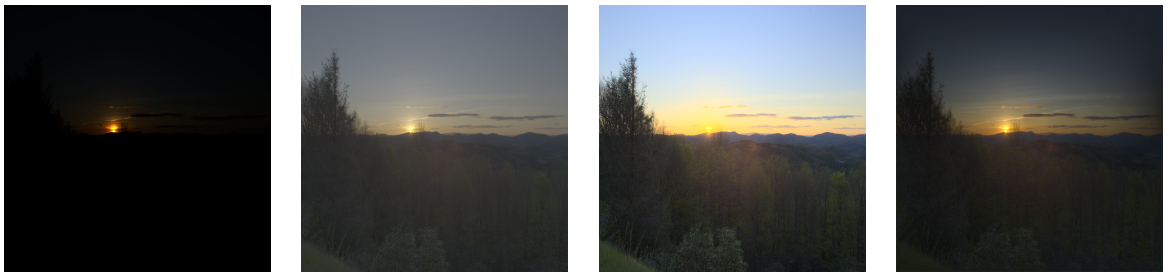
- *Global methods*: A global transfer curve maybe used to map an HDR image to the displayable gamut [94]. Each color channel may be processed separately or the input image may be split up into luminance and chrominance channels and the global gamma curve maybe applied only on the luminance channel.

- *Local Adaptation methods*: The global approach is found to be less successful for images having a wide range of exposures. The local adaptation methods work on the principle of dividing each pixel by the average luminance of the region around the pixel and replicates the "dodging-and-burning" technique employed by photographes [95]. This corresponds to subtracting out the low-pass filtered version from the original image in the logarithm domain. However, linear filtering does not preserve the edges, hence the tone-mapped image might end with unnatural halos. Instead, different edge-preserving filters may be used, such as: bilateral filtering[96], weighted least square filtering [97], or guided filtering [98].

- *Gradient Domain methods*: These class of algorithms compress the gradient of the log-luminance image by a spatially varying attenuation factor and takes into account gradients at different scales. The modified gradient field is re-integrated by solving a first order variational problem [99].

4.3 Visual Quality Evaluation

Subjective testing is important in evaluating the visual quality of images produced by different algorithms. A faster and less expensive alternative is objective quality evaluation. Recently, full-reference IQA (FR-IQA) algorithms[5][100][101] were proposed for



(a) (b) (c) (d)

Figure 4.1: (a) Input HDR image, linearly, mapped to [0-255] (b) Gamma compressed (c) Processed with Reinhard's "dodging-and-burning" method [95] (d) Processed with gradient domain method [99]

evaluating tone-mapped SDR images in comparison to the reference HDR image. In [5], Yeganeh *et al.* carried out a subjective study with various tone-mapped SDR images and proposed the tone-mapped image quality index (TMQI) based on the structural similarity metric in order to ensure that the details in the original HDR image are represented faithfully in the SDR version. It is combined with a naturalness measure based on scene statistics in order to ensure that the rendered image looks realistic.

Tone-mappped FR-IQA metrics employ average pooling that weights all pixels equally. Using different pooling strategies for combining local quality scores to yield the final quality index of the processed SDR image is well-researched[24][43][49]. Using perceptual pooling methods for quality evaluation of HDR images is less well studied. In [102], the authors have proposed Saliency weighted Tone-Mapped Quality Index (STMQI) that employ an Attention based on Information Maximization[103] model to find the salient regions of the image.

Petit *et al.*[104] propose a variation of the SDR image saliency measure by Itti *et al.*[105] to make it suitable for HDR images. In this work, we use this method with TMQI for pooling the local quality scores. Although this is found to show good correlation with the ground-truth eye-tracking data obtained from human subjects, computing the saliency map using Gaussian Dyadic Pyramid and Gabor filters is computationaly expensive. To reduce complexity, we also propose simple local information content based pooling strategies that improve the performance of the TMQI algorithm.

We also investigate a natural scene satistics model based on mean-subtracted-contrast-normalized (MSCN) pixels that has been widely used for blind quality prediction of natural SDR images [66][28]. This does not need any previous training on the corpus of natural images, unlike the model in [5] that fits a Gaussian and a Beta probability distribution to the histograms of the means and standard deviations of these images.

Quality evaluation of compressed tone-mapped images is an emerging problem that involves the joint optimization of tone-mapping and compression parameters. The proposed FR-IQA algorithms have not been evaluated until now for this application. In this chapter, we evaluate the performance of the proposed algorithm on multiple artifacts arising from tone-mapping and JPEG compression of HDR images.

4.4 Tone Mapped Quality Index

The TMQI algorithm is based on the combination of two image quality indicators: 1) a multi-scale image fidelity metric based on a modified structural similarity (SSIM)[30] index and 2) a measure based on natural scene statistics (NSS): the mean and standard deviation of pixel intensities. Since the dynamic range of the HDR images is much higher

than that of standard SDR displays, the TMOs cannot preserve all the details of the HDR versions; however it must ensure that the SDR image is structurally similar to the HDR version. The SSIM-inspired component takes into account this aspect of signal fidelity. In addition, the SDR image must also ensure that it looks natural because the human visual system is trained on NSS that appear irrespective of image content. The TMQI algorithm takes into account only the pixel luminances. Both the HDR and SDR images are converted from RGB color space to the Yxy color space and the algorithm is applied only to the Y component.

4.4.1 Structural Fidelity

The SSIM index (and its multi-scale version MS-SSIM) measures changes in luminance, structure and contrast between the images. Tone mapping operators change local intensity and contrast[5], so TMQI redefines the structural fidelity term as:

$$S_{local}(x,y) = \frac{2\sigma_x{}'\sigma_y{}' + C_1}{\sigma_x{}'^2 + \sigma_y{}'^2 + C_1} . \frac{\sigma_{xy} + C_2}{\sigma_x\sigma_y + C_2} \tag{4.5}$$

where x and y are image patches in the HDR and the corresponding tone-mapped SDR image, σ_x, σ_y and σ_{xy} are the local standard deviations and cross-correlations between them, $\sigma_x{}'$ and $\sigma_y{}'$ are the nonlinearly mapped versions of σ_x and σ_y in (4.6). The algorithm penalizes only those cases where the signal strength is significant in one of the image patches, but insignificant in the other. To distinguish between significant and insignificant signal strength, the local standard deviation in mapped nonlinearly through a psychometric function (related to the visual sensitivity of contrast) which takes the form

of a cumulative normal distribution, given by

$$\sigma' = \frac{1}{\sqrt{2\pi}\theta_\sigma} \int_{-\infty}^{\sigma} \exp\left[-\frac{(x-\tau_\sigma)^2}{2{\theta_\sigma}^2}\right] \mathrm{d}x \tag{4.6}$$

where σ' is the mapped version of σ, τ_σ is the modulation threshold, and θ_σ is the standard deviation of the normal distribution. It is bounded between 0 and 1. τ_σ is proportional to the inverse of the visual contrast sensitivity[106]. At each scale, the scaled version of map is pooled by averaging to output a single score:

$$S_l = \frac{1}{N_l} \sum_{i=1}^{N_l} S_{local}\left(x_i, y_i\right) \tag{4.7}$$

where x_i and y_i are the i-th patches in the HDR and SDR images respectively and N_l is the number of patches in the l-th scale. The overall structural similarity metric is obtained by multiplying the structural similarity scores from the various scales:

$$S = \prod_{l=1}^{L} {S_l}^{w_l} \tag{4.8}$$

where L is the total number of scales and w_l is the weight assigned to the l-th scale.

4.4.2 Image Naturalness

Apart from maintaining structural fidelity, the tone-mapped SDR images should also satisfy some criterion of natural fidelity. In[5], the authors have used naturalness measures based on brightness and contrast of the tone-mapped images. The histograms of the means and standard deviations of natural images have been found to fit a Gaussian and Beta probability distribution respectively. The naturalness measure is the product of

these two distributions since natural scene statistics of brightness and contrast are largely independent quantities. The final Tone Mapped image Quality Index (TMQI) is given by:

$$Q = aS^{\gamma} + (1-a)N^{\delta} \tag{4.9}$$

where $0 \leq a \leq 1$ adjusts the relative importance of the structual measure (S) and the naturalness measure (N), and γ and δ control their sensitivities.

4.5 Proposed IQA algorithm

This section outlines the modifications made to the TMQI algorithm to take into account perceptual pooling strategies and a NSS model based on the distribution of the MSCN pixels to quantify image naturalness.

4.5.1 Visual saliency measure

In[105], the authors build a master "saliency map" using features like color, intensity and orientations at different scales. Instead of just using intensity differences for HDR images,[104] uses intensity contrast between the scales and normalizes the orientation features also over the intensity channel. We use this method of saliency detection to improve performance of the TMQI algorithm. In addition, we explore the role of the local contrast of the tone-mapped images in pooling the quality scores since the quality of tone-mapped images on SDR displays depends on the degree of detail-preservation. Measures of edge density and local contrast tend to be greater at the points of fixation

than at other locations [107][108]. Regions of higher contrast in the tone-mapped SDR image should be given higher weight in pooling the local structural fidelity score at every scale.

The local contrast of the test image (the tone-mapped SDR image or the JPEG compressed tone-mapped image) is measured with two simple methods. (1) σ-map of the image obtained by (4.12), and (2) local entropy of the image at every pixel location (using a rectangular window). Since entropy is a measure of uncertainty of the random variables, it can be used to capture the local contrast also. For example, if a tone-mapping operator leads to over-exposed uniformly bright regions (such as the sky), these regions are expected to show higher entropy than a region having aesthetically rendered foliage.

4.5.2 Natural Scene Statistics

For this work, we model the scene statistics of tone mapped images in the spatial domain, MSCN pixels and the σ-field of the image. The pixels of the image are preprocessed by mean subtraction and divisive normalization. Let $M \times N$ be the dimension of the image I, and $I(i,j)$ be the pixel value in the (i,j)-th spatial location, $i \in \{1,2,..,M\}$, $j \in \{1,2,...,N\}$. MSCN pixels are generated by

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + 1} \tag{4.10}$$

where the local mean $\mu(i,j)$ and standard deviation $\sigma(i,j)$ are defined as:

$$\mu(i,j) = \sum_{k=-K}^{k=K} \sum_{l=-L}^{l=L} w_{k,l} I(i+k, j+l) \tag{4.11}$$

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{k=K} \sum_{l=-L}^{l=L} w_{k,l}[I(i+k,j+l) - \mu(i,j)]^2} \tag{4.12}$$

$w = \{w_{k,l} | k = -K, .., K, l = -L, ..., L\}$ is a symmetric local convolution window centered at the (i,j)-th pixel. K and L determine the size of local patch considered in the calculation of the mean and standard deviation. In [66], the authors considered 7×7 image patches, and a circularly symmetric 2D Gaussian kernel; however, experiments show that the distribution of the MSCN patches are not very sensitive to the window size, or the convolution kernel.

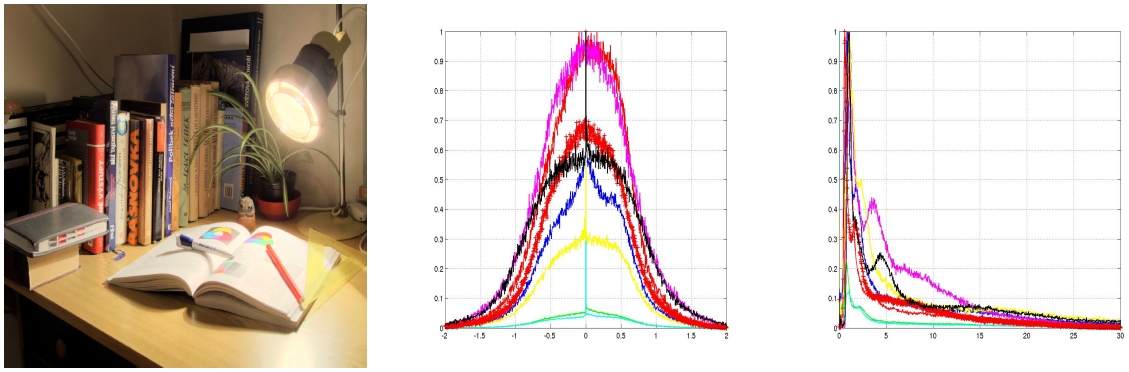
The variance normalized image ($\hat{I}$) tends to be more uniform than the original image, and almost looks like a noise pattern, except at object boundaries. Also, their histograms seem to show a Gaussian like distribution. The standard deviation image σ looks more like the original image, highlighting object boundaries and attenuating textures. The MSCN pixels have been modeled using an Asymmetric Generalized Gaussian Distribution and used in image quality assessment[66][28].

As a measure of the image naturalness, we consider the scale parameter of the distribution of the MSCN pixels (β) and standard deviation of the σ-field, obtained from (4.12). Let ϕ be the variance of the σ field. The modified TMQI index is given by:

$$Q = aS^{\gamma} + \frac{1}{2}(1-a)\beta^{\delta_1} + \frac{1}{2}(1-a)\phi^{\delta_2} \tag{4.13}$$

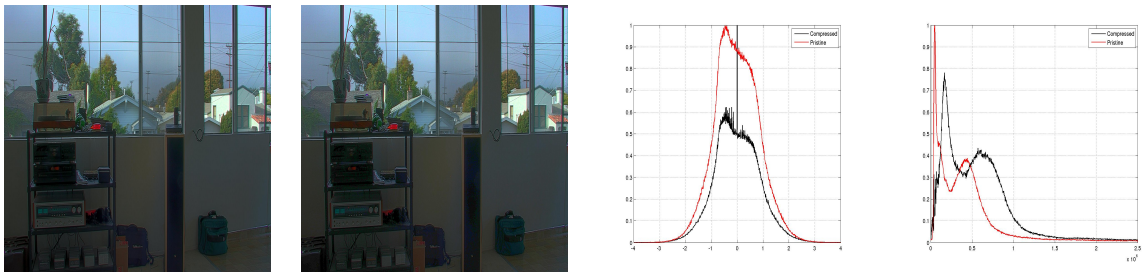
4.6 Experimental Results

This section outlines the performance of the proposed algorithms on two HDR datasets. The first one ("TMQI Database")[5] contains 15 reference natural HDR images



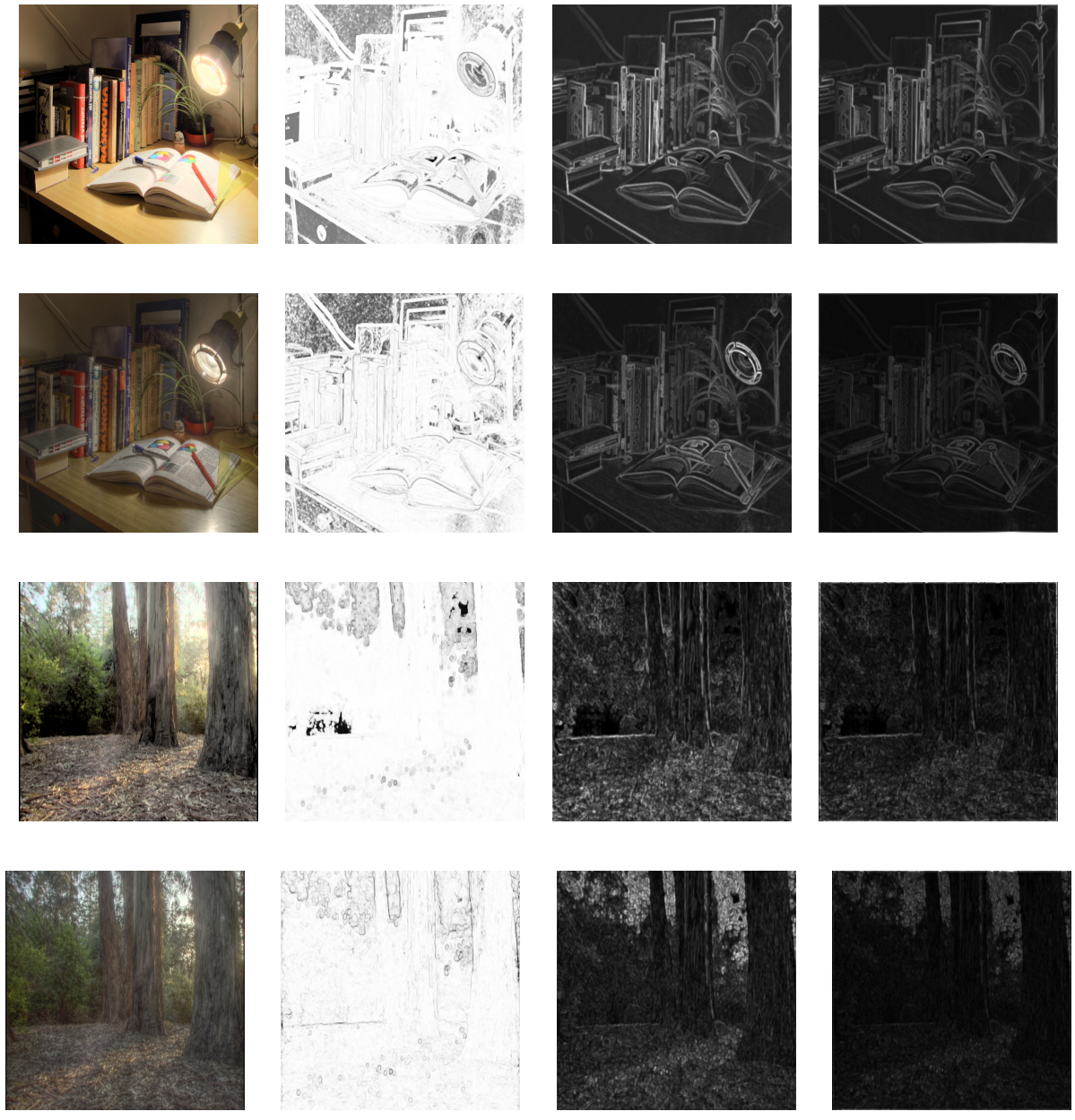
(a) (b) (c)

Figure 4.2: (a) An image from the TMQI database[5] and the corresponding histograms of (b) MSCN pixels and (c) σ-field of the tone-mapped SDR images. The figures show how different tone-mapping operators result in different distribution of the MSCN pixels and the σ-field, which can be quantified into the naturalness measure of FR-IQA algorithms.



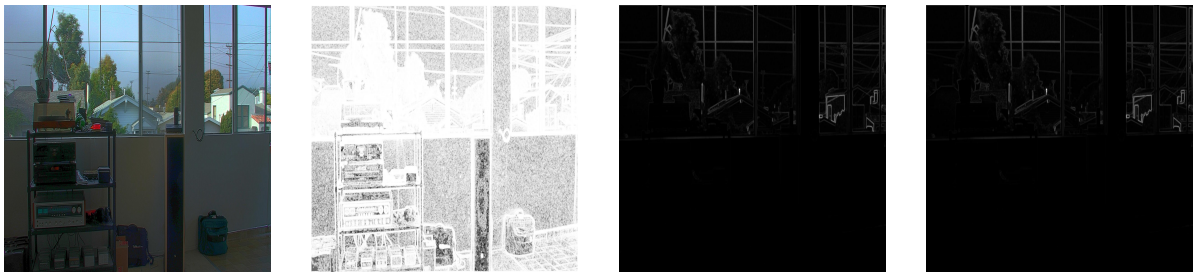
(a) (b) (c) (d)

Figure 4.3: (a) An reference image from the HDR-JPEG database[109], (b) A compressed image and the corresponding histograms of (c) MSCN pixels and (d) σ-field of both the images. The figures show how JPEG compression, coupled with tone-mapping operators result in different distribution of the MSCN pixels and the σ-field, which can be quantified into the naturalness measure of FR-IQA algorithms.



(d) (s) (m) (e)

Figure 4.4: Results from the TMQI database: Tone-mapped LDR image (d), the corresponding local structural similarity map (s), the fidelity maps obtained by the product of the structural similarity map and the σ-map (m) and that obtained by the product of the structural similarity map and the local entropy map (e). Brighter gray level means higher similarity. Results are shown only for the coarsest scale.



(d) (s) (m) (e)

Figure 4.5: Results from the HDR-JPEG database: Compressed HDR image (d), the corresponding structural similarity map(s), the fidelity maps obtained by the product of the structural similarity map and the σ-map (m) and that obtained by the product of the structural similarity map and the local entropy map (e). Brighter gray level means higher similarity. Results are shown only for the coarsest scale.

and 8 tone-mapped SDR images for each of them, generated using different algorithms. The SDR images were ranked according to the quality from 1 (best) to 8 (worst) by 20 subjects. The second one is a tone mapping based HDR compression dataset ("HDR-JPEG Database")[109] comprising of 10 different still images and 14 distorted versions obtained by JPEG compression of the original one with 7 different bitrates and using two different optimization criteria (Mean Squared Error and the Structural Similarity Index Metric[30]). This database contains both natural and synthetic scenes. The Visual Information Fidelity metric[42] has also been included for comparison because it is a wavelet domain FR-IQA method that correlates well with human perception for SDR images.

The performance of TMQI, FSITM, TMQI-II, and STMQI FR-IQA algorithms have been evaluated using the MATLAB source codes provided by the authors. TMQI-

Table 4.1: Spearman Rank-order Correlation Coefficient (SROCC), Pearson's Linear Correlation Coefficient (PLCC) and Kendall's Correlation Coefficient (KCC) between the algorithm scores for various IQA algorithms and the DMOS scores for TMQI database[5] along with runtime (in seconds). The table has been sorted in the descending order of SROCC. Bold values indicate the best performing algorithm. Red indicates the proposed methods.

IQA	SROCC	PLCC	KCC	Runtime
TMQI-NSS-σ	**0.8810**	**0.9439**	**0.7857**	0.3212
TMQI-NSS-Entropy	**0.8810**	0.9438	0.7143	1.2759
SHDR-TMQI	**0.8810**	0.9346	0.7143	0.8010
FSITM-TMQI [101]	0.8571	0.9230	**0.7857**	0.9428
STMQI [102]	0.8503	0.9382	0.7638	1.5385
TMQI-II [100]	0.8333	0.8790	0.7143	**0.2002**
FSITM [101]	0.8333	0.8948	0.7143	0.4741
TMQI [5]	0.8095	0.9082	0.6429	0.5206
VIF [42]	0.3810	0.6136	0.2857	1.3935

NSS-σ uses the TMQI index in conjunction with the MSCN based natural scene statistics model and the σ-map as the local pooling strategy. SHDR-TMQI and TMQI-NSS-Entropy employ a similar scheme but use the saliency detection method proposed in [104] and local entropy respectively for pooling the structural fidelity score. These pooling based IQA algorithms employing the MSCN based naturalness measure outperform the state-of-the-art FR-IQA algorithms both for tone-mapping artifacts (Table 4.1) as well as for multiply distorted HDR images having both tone-mapping and JPEG compression artifacts (Table 4.2).

For the different variations of the TMQI algorithm, the relative weightage of the structural similarity term with respect to the naturalness term has been kept constant ($a = 0.8012$). Five levels have been considered for all the IQA algorithms except for

Table 4.2: Spearman Rank-order Correlation Coefficient (SROCC), Pearson's Linear Correlation Coefficient (PLCC) and Kendall's Correlation Coefficient (KCC) between the algorithm scores for various IQA algorithms and the DMOS scores for HDR-JPEG database[109] along with runtime (in seconds). The table has been sorted in the descending order of SROCC. Bold values indicate the best performing algorithm. Red indicates the proposed methods. For the HDR-VDP-2 algorithm (in row 6), in absence of any information on the display size at [109], the correlation values have been computed assuming that a 24 inch HDR monitor was used for the subjective experiments.

FR-IQA algorithms	SROCC	PLCC	KCC	Runtime
SHDR-TMQI	**0.8510**	0.8533	0.6700	3.0003
TMQI-NSS-σ	0.8485	0.8520	0.6659	1.6470
TMQI-NSS-Entropy	0.8454	**0.8645**	**0.6719**	6.7424
VIF [42]	0.8004	0.8178	0.6143	9.5000
TMQI [5]	0.7947	0.8057	0.6127	3.4394
HDR-VDP-2[35]	0.6389	0.6479	0.4737	19.5031
FSITM-TMQI [101]	0.6300	0.6584	0.4762	8.3486
TMQI-II [100]	0.5096	0.5137	0.3642	**1.3424**
FSITM [101]	0.4720	0.5167	0.3422	5.2617
STMQI [102]	0.3464	0.3244	0.2449	11.9965

SHDR-TMQI, where two levels have been considered in order to ensure that the size of the image do not fall below 128×128; an implementation restriction imposed by the authors of Itti's saliency measure[105]. The source code of our proposed algorithm can be downloaded from[110].

Spearman Rank-order Correlation Coefficient (SROCC), Pearson's Linear Correlation Coefficient (PLCC) and Kendall's Correlation Coefficient (KCC) have been used to evaluate the performance of FR-IQA algorithms. Execution time (in seconds) for each algorithm (on a Linux desktop having 12 GB RAM, Intel Xeon CPU, 3.33 GHz clock) has also been evaluated. Results for the TMQI and HDR-JPEG Databases are summarized in Tables 4.1 and 4.2 respectively.

4.7 Conclusion

In this chapter we show that simple perceptual pooling techniques that take into account the local contrast improve the performance of the TMQI algorithm for full-reference quality evaluation and propose a different NSS model to better qualify the image naturalness. We show that in addition to tone-mapping artifacts, the proposed methods show good correlation with human observers for JPEG compressed tone-mapped images. However, the currently available HDR databases are limited in the number of images considered or the number of subjects participating in the subjective study. In the next chapter we describe a massive online crowdsourced subjective study for judging the quality of HDR images providing a much larger corpus of source images, processing artifacts and the number of subjects involved.

Chapter 5

Crowdsourced evaluation of HDR images

5.1 Introduction

Most available HDR IQA databases suffer from the limitation of having a relatively small number of images and small number of human subjects participating in the experiments, typically conducted in a stringently controlled visual environment. In addition, most of these databases either ask the subjects to rank multiple versions of the same HDR scene created using different processing algorithms or implement a two-alternative forced choice method of subjective evaluation. These approaches severely restrict the number of the source images that can be considered, the type of processing artifacts and the number of subjects participating in the experiments.

The previous chapter outlined a FR assessment method for tonemapping artifacts in HDR images. In applications in which the reference 32-bit irradiance map is not available for comparison, no-reference IQA is the way to go. In the present legacy databases, since the HDR images are annotated with a rank relative to other images instead of an absolute raw quality score, they are unsuitable for blind IQA for the perceptual quantification of HDR artifacts as has been done for SDR images. In order to address these limitations, I propose the following contributions:

- Designing the ESPL-LIVE HDR Image Database, comprising of 1,811 HDR pro-

cessed images created from 605 high quality source HDR scenes. The images have been obtained by eleven HDR processing algorithms involving tonemapping, multi-exposure fusion algorithms. In addition I have also considered post-processing artifacts in HDR image creation, typically found in commercial HDR softwares.

- Conducting subjective experiments for collecting data from thousands of observers over a online crowdsourcing platform, and analyzing the data to reject the outlier subjects, and calculate the mean opinion scores (MOS) for each image.

5.2 Related Work

Some of the HDR IQA databases have addressed two typical HDR processing methods: tonemapping and multi-exposure fusion. In [5], Yeganeh *et al.* carried out a subjective study with 15 reference natural HDR images and 8 tone-mapped SDR images for each of them, generated using different algorithms. The SDR images were ranked according to the quality from 1 (best) to 8 (worst) by 20 subjects. Ma *et al* conducted a subjective experiment with 17 reference HDR images and 8 images created using different multi-exposure fusion algorithm for each of them. 25 subjects participated in the study.

HDR compression artifacts have been subjectively evaluated in [109] and [111]. [109] comprises of 10 different still images (both natural and synthetic) and 14 distorted versions obtained by JPEG compression of the original one with 7 different bitrates and using two different optimization criteria (Mean Squared Error and the Structural Similarity Index Metric[30]). In [111], Hanhart *et al.* conducted a subjective experiment with 240 images obtained by tonemapping 20 HDR images using a display adaptive tone-mapping

algorithm and compressed using different profiles of the JPEG XT [112] compression algorithm. In [113], the authors have considered 192 images created from 6 source HDR images inflicted with four types of disotortions (namely, JPEG/JPEG2K compression, white noise injection and Gaussian blurring) at 8 different levels with 25 naive participants.

For SDR images, one of the earliest crowdsourced subjective experiements [114] garnered ratings from 40 subjects with 116 JPEG compressed images. In [115], the authors developed the LIVE In the Wild Image Quality Challenge Database comprising of 1,162 images containing real world distortions involving more than 8,100 unique subjects. For HDR images, crowdsourcing has been used before in [116] for evaluating privacy. To the best of our knowledge, crowdsourcing has not been used before for the purpose of subjective quality evaluation of HDR images at this large scale.

5.3 ESPL-LIVE HDR Database

This section describes the types of source images considered, the method of capturing them and the HDR processing algorithms used to generate different versions of the source images in the ESPL-LIVE HDR Database.

5.3.1 Source Content

The images considered in this database comprised of real-world HDR scenes of nature, lakes, snow, forests, cities, man-made structures, historical architectures etc. The database consists of images shot both during the day and the night and includes both indoor and outdoor scenes. Figure 5.1 shows some of the sample images of our database.

Figure 5.2(a) and (b) show the distribution of the source scenes under various imaging conditions. The high dynamic range images used in this database has been obtained by combining photographs of the same scene shot with multiple exposures using a modern digital SLR camera. The auto-bracketing feature of modern SLR cameras allow the photo of the scene captured at a number of exposures with one depression of the shutter release.

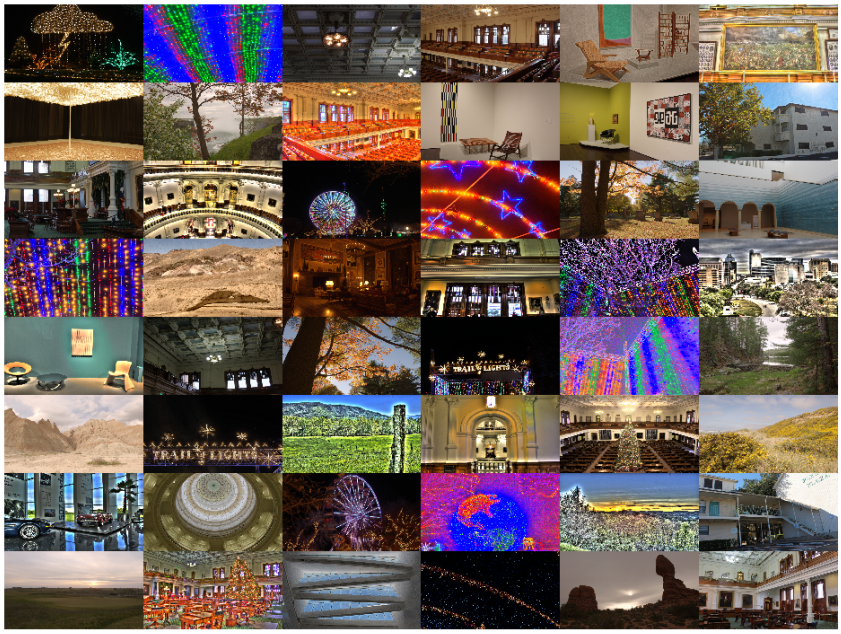
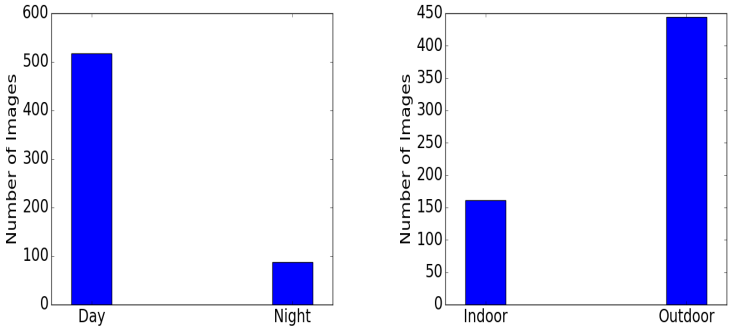


Figure 5.1: Sample images from the ESPL-LIVE HDR Image Quality Database. The images include pictures taken during day and night under different illuminating conditions. Both indoor and outdoor photos have been considered, along with scenes containing natural or man-made objects.

106 images have been obtained from the HDR Photographic Survey [117]. For



(a) (b)

Figure 5.2: Bar chart showing the number of source images taken (a) during day and night, (b) indoors and outdoors

these images, a Nikon D2x was used with a selection of lenses. Most of the images were obtained with a Nikon 17-55mm f/2.8 ED-IF AF-S DX Zoom-Nikkor lens. The D2x is a professional digital SLR with a 12.4 Megapixel CMOS sensor. The auto-bracketing function allowed for nine exposures to be made with one stop increments in exposure time at a fixed aperture. This, combined with the high speed of 5 frames/second , allowed nine-exposure HDR sequences covering a nine-stop exposure range to be made in less than two seconds with sufficient light, a feature that is helpful for subjects that might tend to move. The images have a resolution of 4288×2848.

The rest of the images were captured using a Canon Rebel T5 and Nikon D5300 digital SLR camera, with an 18 Megapixel CMOS sensor. An 18-55mm standard zoom lens was used. The auto-bracketing function allowed three exposures to be captured for every scene. The exact range of the exposures considered varied from scene to scene depending on the subject and the available lighting conditions. For low lighting conditions,

a tripod was used to prevent inadvertent camera shakes. These images have a resolution of 5184×3456. All images were saved in raw electronic format (NEF for Nikon and CR2 for Canon cameras).

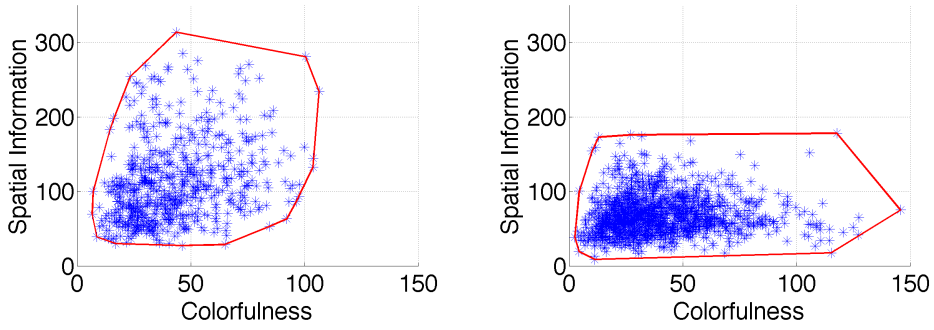
Lastly, in order to minimize the degree of ghosting artifacts arising from moving objects in the scene, care has been taken to ensure that no high motion objects appear in the scene, especially in the foreground.

5.3.2 Source Complexity

The source complexity of the image database has been evaluated using two metrics: *spatial information*, that gives an indication of the richness of the edge distribution in the image, and *colorfullness*, that quantifies color saturation. Details of these measures may be found in Chapter 2 (Section 2.2.2). However, since for HDR images the scenes are captured at multiple exposures, the scene complexity is determined from the medium exposure image. Figure 5.3 shows the scatter plot between the spatial information and colorfulness of the source scenes.

5.3.3 HDR Processing Algorithms

Unlike the legacy subjective image quality assessment databases that come with clearly marked distortion categories (such as "Blur", "JPEG Compression", and "Color Saturation"), it is hard to come up with such classification schemes for HDR images. Depending on the scene and the type of processing algorithm considered, the image might be inflicted by a complex interplay of luminacne, structual or chromatic artifacts that are hard to categorize. Furthermore, many of the commercial HDR processing programs post-



(a) (b)

Figure 5.3: Spatial Information vs. Colorfulness scatter plots for the source images in the following databases (a) ESPL-LIVE HDR, (b) LIVE In the Wild Challenge[115]. Red lines indicate the convex hull of the points in the scatter plot, which approximates the range of scene complexity.

process the images by changing the local contrast and color saturation, thereby leading to a wide perceptual gamut of images.

Prior to fusion of the exposure stack, the bracketed photos need to be registered because of small misalignments due to camera movement between bracketed shots. In addition, even if the camera is held fixed (as with a tripod), the scene may have moving objects, and since the merging process assumes that the pixels in the bracketed stack are aligned perfectly, the moving objects results in ghosting or blurring artifacts depending on whether the amount of motion is high or low respectively[118]. If the trailing 'ghost' of the moving objects are not removed, the observers may be annoyed by the artifacts. Hence, in this section we outline the different HDR algorithms used for creating the images instead of outlining 'distortion' categories. Figure 5.4 shows the distribution of the different algorithms considered in our database.

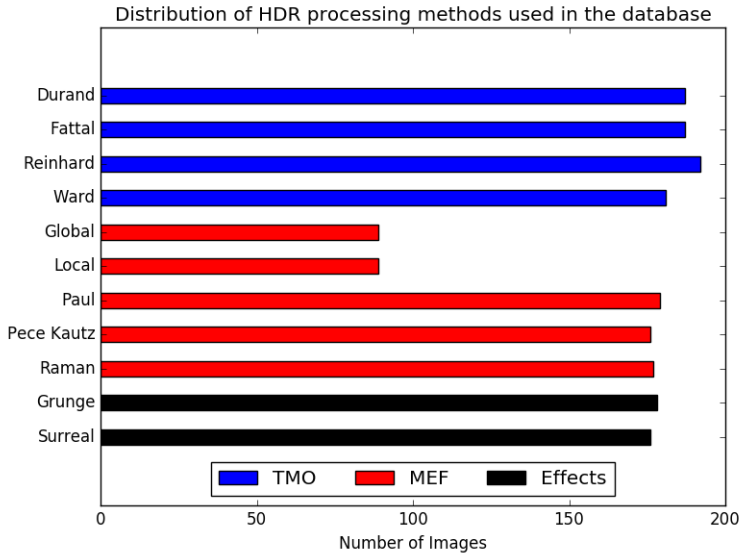


Figure 5.4: Bar chart showing the number of images in the database created by the different HDR algorithms. 'TMO', 'MEF', 'Effects' stand for Tone-Mapping Operators, Multi-Exposure Fusion Algorithms and Post Processing respectively.

Most of the algorithms have been obtained from the HDR Toolbox [119], implemented in MATLAB. The remaining source codes has been provided by the authors.The final images displayed to the subjects had a resolution of 960×540 for landscape orientation and 304×540 for portrait images (downsampled from the original resolution using 'imresize' functionality in MATLAB with bicubic interpolation method). This was done to ensure that the images fit comfortably within the monitors of smaller display size and the subjects do not face any issues with delay in loading the images over lower bandwidth internet connections.

5.3.4 Images generated by Tone Mapping Operators (TMO)

Chapter 4 outlines the process of generating well-exposed SDR scenes by estimating the scene radiance map, followed by tone-mapping it to the displayable gamut of the SDR displays. For every scene, the raw exposure stack was registered and combined into a 32-bit floating point irradiance map (in OpenEXR format) using Photomatix software with minimal processing. Apart from capturing photographs of the same scene at multiple exposures, some OpenEXR images were also obtained from [120]. The tonemapped images were created by using four representative TMOs proposed by Ward[94], Fattal[99], Durand[96] and Reinhard[95]. The resulting image was downsampled to a resolution of 960×540 for landscape orientation and 304×540 for portrait images.

5.3.5 Images generated by Multi-Exposure Fusion (MEF)

The bracketed stack of images, after being downsampled to the display resolution was first registered using a SIFT based image alignment method[119] and the aligned

images were cropped so that every pixel is visible in all the image of the stack, devoid of "black border" artifact. The exposure images are then blended using a MEF algorithm, which can broadly be expressed as [88]:

$$Y(i) = \sum_{k=1}^{K} W_k(i) X_k(i) \tag{5.1}$$

where K is the number of bracketed images, Y is the fused image, and $X_k(i)$ and $W_k(i)$ indicates the luminance or color either in the spatial modain or some coefficient in a transform domain, and the weight at the i-th pixel in the k-th exposure image respectively. W_k represents the relative weightage given to the spatial locations and the different exposure levels based on the perceptual information content. Different MEF algorithms differ in the ways of computing the weights, but has the end goal of maintaining the details both in the underexposed and overexposed regions. These methods bypass the intermediate step of creating the HDR irradiance map and instead creates an SDR image that can displayed directly on standard displays.

The algorithms that have been used to create the multi-exposure fused images are: local and global energy weighting methods, Raman's method based on bilateral filtering[121], the method by Pece *et. al.* that also deghosts along with multi-exposure fusion[122] and Paul *et. al*'s method based on blending the luminance component in the gradient domain.

5.3.6 Post Processed Images

Many HDR images created by professional and amateur photographers are post-processed in order to convey different 'feels' about the scene, that can drastically alter the final look of the image. Hence, for our database, we have also included these post-processed HDR images for subjective evaluation; an issue that was not addressed in any of the existing databases. For our implementation, we first created the irradiance map using Photomatix and tonemapped it using their default tone-mapping algorithm, followed by post-processing using two commonly used effects, namely "Surreal' and "'Grunge" by using different parameters of color saturation, color temperature and detail contrast preservation.

5.4 Subjective Study Setup

Crowdsourced subjective image quality assessment studies provides a wider range of challenges to us compared to a traditional subjective study in a laboratory study primarily due to the lack of control over the precise experimental setup. However in order to demonstrate the effective of such a setup, we also conducted a small-scale laboratory subjective test with a small subset of the HDR images that was used as the control group or 'gold standard' in the large scale crowdsourced subjective study. This section describes the set up of the laboratory and online subjective test, methods used to check the consistency of ratings and the techniques used to analyze the raw scores. In addition, we also show the dependency of the subjective scores on various demographical factors such as age and gender and viewing conditions like distance from the display screen and type of display used.

5.4.1 Laboratory Subjective Evaluation

Fifteen graduate students comprising of five women and ten men roughly of the age group of 20-30 years participated in the laboratory subjective study conducted at the Department of Electrical and Computer Engineering at the University of Texas at Austin in Spring 2016. Most of the subjects did not have prior experience of participating in a visual subjective test. A single stimulus continuous evaluation testing procedure with hidden reference [20] was used. The subjects viewed a total of 38 images of a range of qualities processed by a variety of HDR algorithms. The actual testing session comprised of 27 images and it was preceded by a short training phase, where the subject was shown 11 images. The training phase was provided in order to make the subject familiar with the experimental setup and hence, the scores entered by the subject during this phase were not considered. On an average, each subject took roughly 15 minutes to complete the task.

The user interface for the study was designed on a PC on MATLAB, using the Psychology Toolbox[21] with NVIDIA Quadro NVS 285 GPUs and were interfaced with Dell 24-inch U2412M display. Each image was displayed on the screen for 12 seconds and the experiment was carried out under normal office illumination conditions. The subjects viewed the images from about 2 - 2.25 times of the display height.

The screen resolution was set at 1920×1200 pixels, but the images were displayed at their normal resolution (1920×1080) without any distortion introduced by interpolation. The top and bottom portions of the display were gray. At the end of the image display duration, a continuous quality scale was displayed on the screen, where the default location of the slider was at the center of the scale. It was marked with five qualitative adjectives:

"Bad", "Poor", "Fair", "Good", and "Excellent" placed at equal distances along the scale. After the subject entered a rating for the image, the location of the slider along the scale was converted into a numerical score lying between [0,100], after rounding to the nearest integer. The subject could take as much time as needed to decide the score, but there was no provision for changing the score once entered or to view the image again. The next image was automatically displayed once the score was recorded.

3 of the 15 subjects were found to be outliers and the mean opinion score (MOS) for every image was calculated using the scores of the remaining 12 subjects. In order to take into account the variability among the subjects in using the quality scale, the raw subjective scores were converted to Z-scores before calculating the MOS for every image. A method similar to section 2.2.5 was followed for rejecting the outlier subjects and processing the raw scores. Based on the MOS scores, five images were chosen as the gold standard to approximately span the entire quality scale.

5.4.2 Challenges of crowdsourcing

In the recent days, there has been a growing popularity of using crowdsourcing platforms such as Amazon Mechanical Turk (AMT)[123], Microworkers[124], Crowdflower[125] etc in psychology research for effective large-scale collection of data from a diverse and distributed population from all over the world over the web. The registered 'requesters' advertise their tasks to registered 'workers' on the different platforms and the workers may choose to provide their inputs for data-collection in lieu of some monetary compensation. The following salient features should be kept in mind while designing a crowdsourced subjective experiment:

- While the reach of these online platforms to a large number of potential subjects does ensure that the requesters may collect a large number of image ratings in a much shorter time compared to the standard laboratory experiments, the requesters have limited control on the experimental setup such as the display devices used, distance from the display, and the illumination conditions of the viewing environment. Since these factors may have a compelling effect on the image ratings provided by the users, some information on these factors were accumulated from the users at the end of the viewing session by asking them to complete a short survey. In addition, we gathered information from them on their familiarity with HDR photography, the devices used to capture HDR content and the softwares used to process HDR images. Further details are outlined in the next section.

- The time spent by a subject on doing a subjective study via a crowdsourcing platform differs from a laboratory experiment. In the latter setup, the goal is make the subject evaluate each and every image in the dataset, hence the study may last for a couple of hours which is broken down into multiple sessions in order to avoid subject fatigue. Such setup was used to gather the subjective evaluations for the ESPL Synthetic Image Database (outlined in chapter 2), where each of the 64 participants viewed all 525 images in the database over three sessions, each lasting an hour. However, in a crowdsourced settings, since it is difficult to make workers participate in time-consuming tasks[126], the online tasks need to be segmented into smaller chunks. Hence, all the images in the database will not be viewed and evaluated by every participating worker.

5.4.3 Instructions, Training, and Testing

Although there have been image quality subjective studies before to judge image aesthetics, for this particular study, the subjects were instructed to focus on the image quality than image aesthetics. Care was taken to provide a wide array of images having different degrees of aesthetic appeal. On AMT, requesters present the tasks as Human Intelligence Tasks (HITs). The workers are shown the instructions page explaining the details of the study along with the monetary reimbursement offered. If the worker is interested to participate, she has to click the "Accept HIT" button to begin the actual task. At the end of the task, the worker has to submit her results to the requester by clicking on the "Submit Results" button.

5.4.3.1 Interface used

Figure 5.5 shows the screenshot of the instructions page shown to the workers. Apart from the instructions, the workers were also shown some representative images in the database along with a screenshot of the interface to be used to rate the images. Once the worker accepted the HIT, she was presented with a rating interface as shown in Figure 5.6 containing the image to be evaluated and a slider below it. A single stimulus continuous quality evaluation[20] method was used in the experiment. The subjects entered the ratings by dragging the horizontal slider bar that was divided into five segments and labeled as bad, poor, fair, good, and excellent to aid the subject in entering his judgment. Once she decided on the rating, she pressed the "Next Image" button, upon which the position of the slider was converted to a quality score between [1-100] and the next image was presented. Thus unlike the laboratory experiments where the subjects are

shown each image for a fixed amount of time, on the crowdsourced platform, the subjects can view each image for as long as they want.

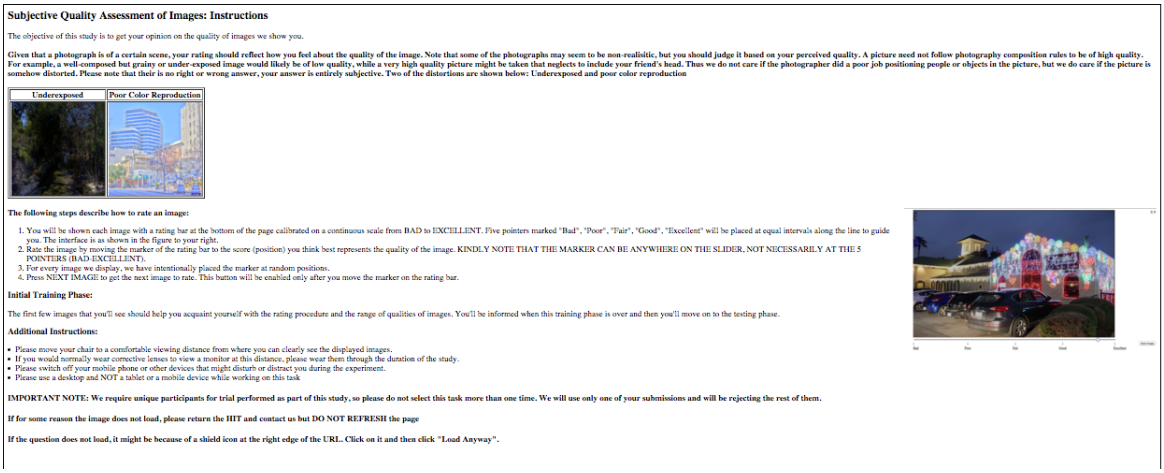


Figure 5.5: Instruction Screen for Amazon Mechanical Task HIT shown to the subjects for collecting the ratings.

Figure 5.6: Rating Screen for Amazon Mechanical Task HIT shown to the subjects for collecting the ratings.

5.4.3.2 Training and Testing Phase

Following similar procedure as the laboratory experiment, before the testing phase, each participant was shown a set of 11 training images in order to make them familiar

with the user interface, approximate range of image qualities, and the type of processing artifacts that they might encounter during the actual testing phase. The training set of images was the same for all participants.

The actual testing phase considered of 49 images selected randomly from the corpus of 1,811 images of our database, that were presented in a random order to each subject. The testing phase was followed by a short survey. On an average the subjects took 9 minutes to complete the task of evaluating a total of 60 images and they were reimbursed with 45 cents for their participation.

5.4.4 Subject Reliability and Rejection Strategies

Although AMT allows us to gather subjective evaluations from a large number of subjects in a relatively short period of time, stringent subject rejection strategies need to be implemented in order to ensure high quality reliable ratings. Following are the subject rejection methods considered in this work.

- *Intrinsic metric*: Only those workers on AMT having confidence values greater than 0.75 were allowed to participate in this study. This number, lying between 0 and 1 is based on their responses across all tasks they have completed on AMT. Although this number does not take into account the performance of the subject only on the visual tasks, a higher confidence number indicates a more reliable subject. Also, if the same worker picked up the study multiple times, it might bias the ratings. Hence, since we wanted unique participants for this study, if the same worker selected the task again, she was not allowed proceed beyond the instructions page.

- *Using corrective lens*: If any worker wore corrective lens in their day-to-day life, they were instructed to wear them during the entire duration of the study. At the end of the task, they are asked two questions on whether they normally wore corrective lens and whether they were wearing them during the task. If a certain worker, who was supposed to be wearing lens, reported that she was not using them during the study, her scores were rejected.

- *Repeated images*: Among the 49 test images, 5 of them were randomly chosen and presented twice to each subject during the testing phase. If the difference between the two scores provided by the worker to the same image exceeded a certain threshold for at least 3 of the 5 repeated images, the scores from the worker was rejected. During the initial phase of the study, the average standard deviation of the scores obtained from 400 workers was found to be 17 (rounding up to the nearest integer). 1.5 times the average standard deviation was considered as the threshold for rejecting subjects. This method helped to eliminate inattentive subjects who were providing arbitrary scores to the images.

- *Gold standard images*: 5 of the remaining 44 images were chosen from the laboratory subjective study. These images, referred to as "gold standard" was used to provide a control. The median value of the Pearson's linear correlation coefficient (PLCC) between the scores provided by each subject to these five images in the crowdsourced study and the corresponding MOS calculated from the laboratory subjective test was found to be **0.9465**[1] and the median root-mean-square-error

[1]All the correlation values between the IQA algorithm scores and/or human ground truth values are computed after using a non-linear logistic regression as outlined in [7].

between the subject scores and the ground truth MOS values was 5.4710. The high degree of agreement between the ground truth data obtained from the laboratory settings and that obtained from the online platform shows the high degree of reliability of the scores obtained by crowdsourcing.

5.4.5 Subject-Consistency Analysis

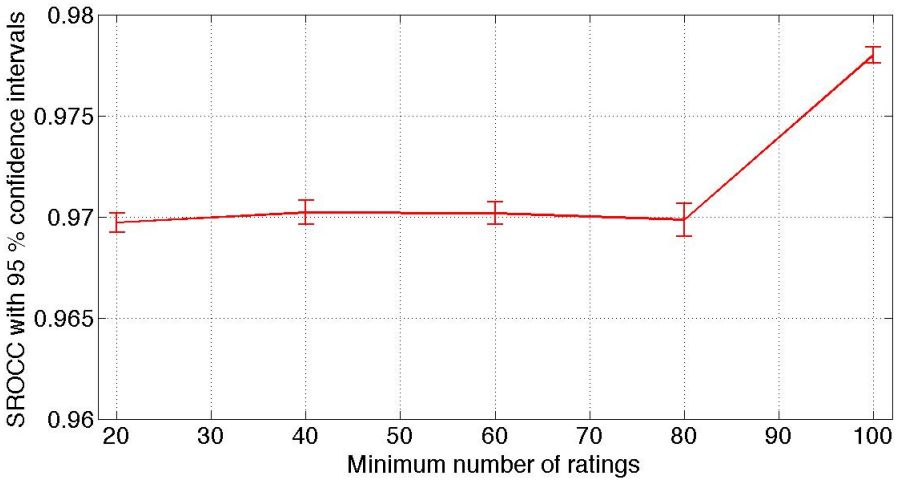


Figure 5.7: Variation of inter-subject consistency with number of ratings. The horizontal axis indicates the minimum number of ratings per image for the subject of the images used to check subject consistency. The vertical axis shows mean Spearman's Rank Ordered Correlation Coefficient between the MOS values of two halves of ratings randomized over 10 splits, along with the 95 % confidence intervals. As the number of ratings per image increases so does the inter-subject consistency.

The consistency of the scores obtained from the subjects has also been measured using the following methods:

- *Inter-Subject consistency*: For every image, the ratings were divided into two disjoint equal sized subsets and the MOS values were computed using each of them. The

procedure were repeated over 10 random splits and the mean Spearman's Rank Ordered Correlation Coefficient (SROCC) between the MOS between the two sets was found to be **0.9677**. Figure 5.7 shoes the inter-subject consistency variation with the increasing number of ratings per image.

- *Intra-Subject consistency*: Pearson's linear correlation coefficient was measured between the individual opinion scores and the MOS values for the gold standard images. The median PLCC of **0.8743** was obtained over all the subjects.

The high values of these metrics indicates good consistency between the scores obtained from the subjects for each image.

5.5 Analysis of subjective scores

We have gathered 327,720 ratings, the images of the database have being evaluated by 5,462 unique participants. 388 subjects were eliminated following the rejection criterion based on their performance on the "gold standard" images and/or for not following the instruction of wearing corrective lenses when they were supposed to. On an average every image has been evaluated by 110 observers. Figure 5.8 shows the histogram of the number of ratings per image received till now.

The MOS has been computed by averaging the Z-scores as outlined in Chapter 2. The range of the MOS values spans [16.941 - 68.502]. Figure 5.9 shows the scatter plot and histogram of the MOS scores for every image obtained from the Z-scores. The average standard deviation of the subjective scores obtained on every image was found to be 21.131.

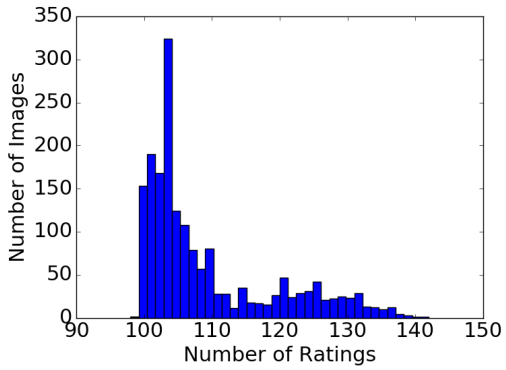
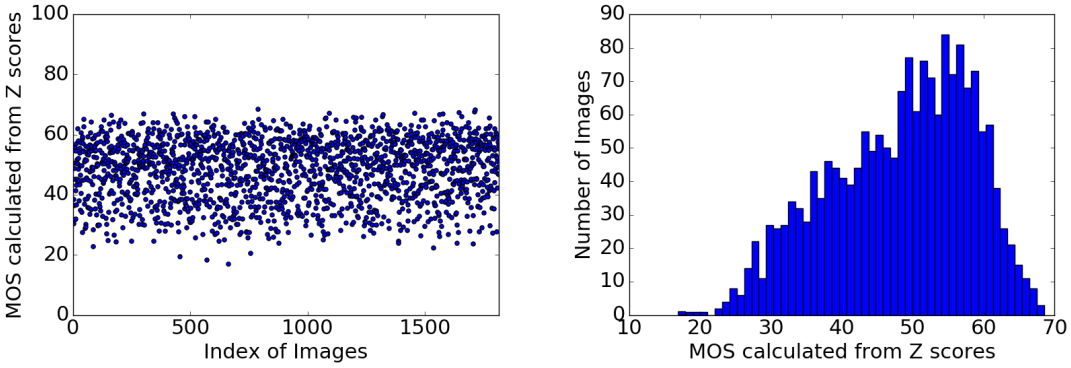


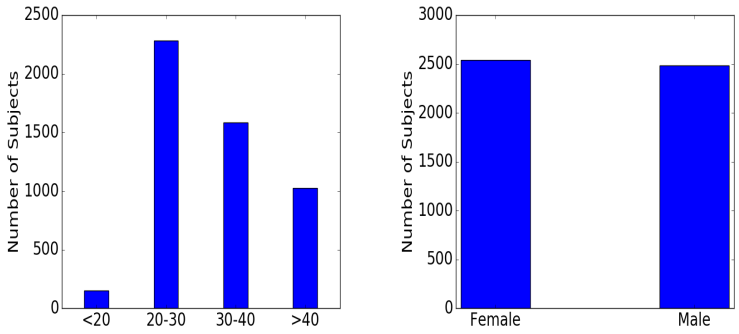
Figure 5.8: Distribution of number of ratings per image.



(a) (b)

Figure 5.9: (a) Scatter plot and (b) Histogram of MOS obtained from the human subjects. The range of the MOS values spans [16.941 - 68.502]

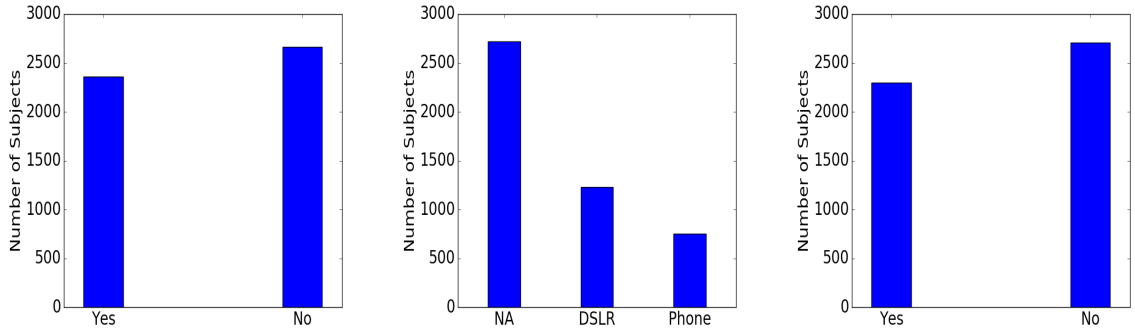
In this subjective study, we also gathered demographic information about the subjects, such as age and gender, as outlined in Figure 5.10. Since familiarity of the subjects with HDR photography might affect the quality scores provided by them, the subjects were also requested to provide information on that. Figure 5.11 shows the awareness of the subjects about HDR photography, the type of optical devices used by them to capture HDR content (if they indeed knew about HDR) and their familiarity with image processing softwares like Adobe Photoshop or Photomatix. The last factor was included in the survey because some of the images were created by adding special post-processing effects after HDR fusion.



(a) (b)

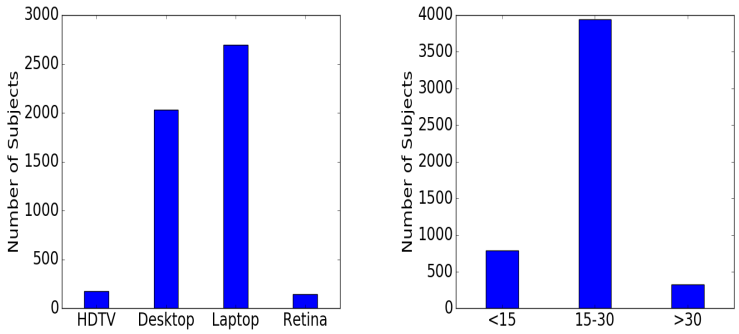
Figure 5.10: Demographics of a sample of subjects (a) age (b) gender

The subjects were instructed to work on the HIT only from personal computers instead of smartphones or tablets. The type of display devices used and the distance from the screen can affect visual quality of the image. The subjects also provided information on these aspects. Figure 5.12 shows the different displays used by the subjects and their estimated distance from the screen while completing the HIT.



(a) (b) (c)

Figure 5.11: HDR awareness of the subjects (a) Number of subjects aware of HDR images (b) The devices used to capture HDR content mostly. The bar titled 'NA' shows the subjects who are not familiar with HDR (c) Number of subjects familiar with softwares like Photoshop or Photomatix



(a) (b)

Figure 5.12: Display used by subjects (a) display devices used by the subjects (b) approximate distance of the subject to the display used

(a) MOS = 62.43 ± 2.043

(b) MOS = 52.90 ± 2.170

(c) MOS = 42.33±2.77

(d) MOS = 40.23±2.42

(e) MOS = 31.07±2.82

Figure 5.13: Sample images from HDR database used to illustrate the effect of increasing the number of participants on the calculated MOS. The caption of each image gives the MOS values and the associated 95% confidence intervals.

5.5.1 Variation of subjective scores with different factors

In section outlines the observations on how the perceptual quality of the subjects are affected by different parameters such as age, gender, display device used for participating in the subjective study, distance from the display and their familiarity with HDR image processing. Figure 5.13 shows some representative images upon which the effect of the above mentioned factors on the subjective scores was studied.

5.5.1.1 Age

Subjects who used a laptop during the study and were sitting about 15 - 30 inches away from the screen were considered to isolate the effect of age on perceived quality of the

images while keeping other factors constant. These display settings were selected because most of the subjects participated in the experiment using their laptops and reported to be sitting at about 15 - 30 inches away from the screen, thereby providing us with sufficient number of samples to study the effect of age on perceived quality. The individual ratings on the images shown in Fig. 5.13 were grouped according to three age categories: '20-30', '30-40' and '>40' and the MOS was computed for each group, as shown in Figure 5.14. For these images, no overall conclusion can be drawn, but subjects from the '20-30' group were found to assign lower scores to some of the images as compared to the other age groups.

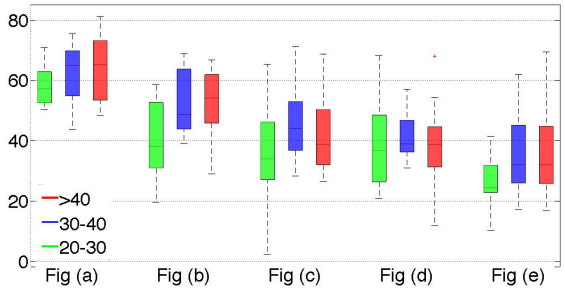


Figure 5.14: Individual Z-scores obtained from subjects of different ages who rated the images shown in Fig 5.13. For each vertical column, median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points.

5.5.1.2 Gender

Subjects between 2030 years of age, who used a laptop during the study and were sitting about 15 - 30 inches away from the screen were considered to isolate the effect of gender on perceived quality of the images while keeping other factors constant. These

display settings were selected because most of the subjects participated in the experiment using their laptops and reported to be sitting at about 15 - 30 inches away from the screen, thereby providing us with sufficient number of samples to study the effect of gender on perceived quality. The individual ratings on the images shown in Fig. 5.13 were grouped according to their gender and the MOS was computed for each group, as shown in Fig. 5.15. For these images, no overall conclusion can be drawn, but female subjects were found to assign lower scores to some of the images as compared to the male participants.

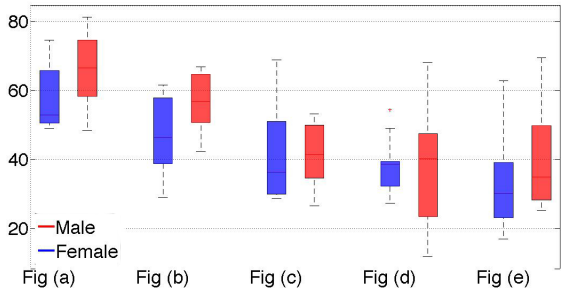


Figure 5.15: Individual Z-scores obtained from subjects of different genders who rated the images shown in Fig 5.13. For each vertical column, median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points.

5.5.1.3 HDR Awareness

One of the questions asked of the subjects was whether they were familiar with HDR images. Fig. 5.11 shows the distribution of the answer of the subjects to various HDR related questions. The individual ratings on the images shown in Figure 5.13 were grouped according to whether the users were familiar with HDR imaging and the MOS

was computed for each group, as shown in Figure 5.16. It was found that the subjects evaluated the perceptual quality of the images in a similar manner, irrespective of whether they were familiar with HDR imaging or not.

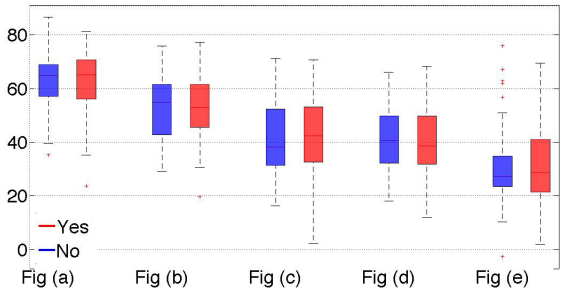


Figure 5.16: Individual Z-scores obtained from subjects familiar with or not familiar with HDR imaging who rated the images shown in Fig 5.13. For each vertical column, median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points.

5.5.1.4 Display device used

The subjects were asked to report the type of display device they used to participate in this study. The individual ratings on the images shown in Fig. 5.13 were grouped according to whether the users were using a desktop or a laptop computer and the MOS was computed for each group, as shown in Fig. 5.17. It was found that the subjects evaluated the perceptual quality of the images in a similar manner for these two types of displays.

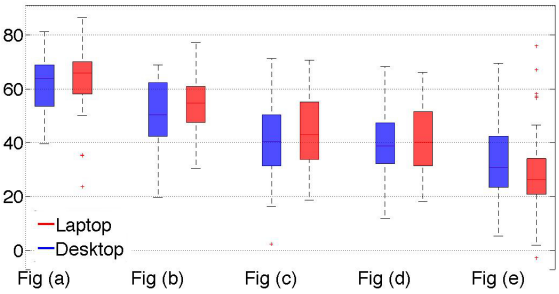


Figure 5.17: Individual Z-scores obtained from subjects using different display devices who rated the images shown in Fig 5.13. For each vertical column, median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points.

5.5.1.5 Distance from display

The subjects were asked to report how far they were sitting from the display while participating in this study. The individual ratings on the images shown in Figure 5.13 were grouped according to three distances: '<15', '15-30' and '>30' inches from the display and the MOS was computed for each group, as shown in Figure 5.18. It was found that the subjects evaluated the perceptual quality of the images in a similar manner for these different distances.

5.5.2 Variation of subjective scores with number of sujects

In order to study the effect of including more subjects to the final computed MOS scores for different images, we randomly selected five images of varying qualities from the database and plotted the MOS values for each against the number of subjective evaluations considered. Figure 5.19 shows that the computed MOS values are more or less constant to the number of subjects viewing the images but the standard error decreases upon

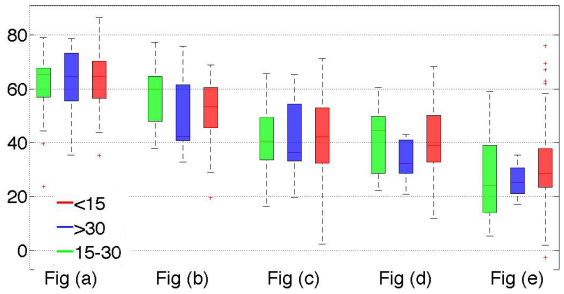


Figure 5.18: Individual Z-scores obtained from subjects viewing the images at different distances who rated the images shown in Fig 5.13. For each vertical column, median is the center of the central box, while the upper and lower edges of each box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points.

considering more and more subjects.

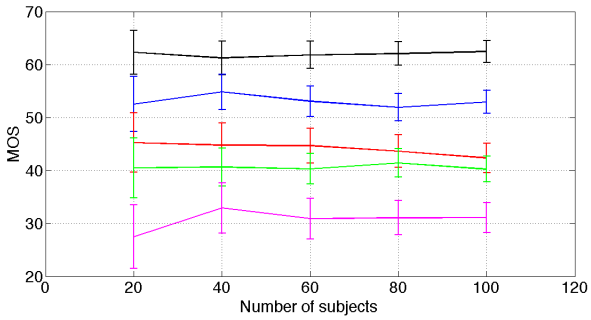


Figure 5.19: MOS plotted against the number of workers who viewed and rated the images shown in Fig 5.13 along with the 95 % confidence intervals.

5.6 Experiments Conducted

The performance of the leading NR-IQA algorithms were tested on this database to see the usefulness and limitations of the current algorithms to evaluate blindly the different HDR processing artifacts. The algorithms G-IQA-1 and G-IQA-2 are two proposed

gradient scene-statistics based NR-IQA algorithms in the LAB color space that correlates better with the human ground truth subjective score for HDR scenes, compared to the existing state-of-the-art NR-IQA algorithms. G-IQA-1 (L) and G-IQA-2 (L) indicate versions of the proposed algorithms using only the luminance channel (L). Details of these methods are outlined in the next chapter.

Most of the algorithms are based on training a machine learning based model with perceptually relevant features extracted from the images in different domains. We randomly split the data into disjoint training and testing sets at 4:1 ratio and the split was randomized over 100 iterations. Care was taken to ensure that the same source scene does not appear in the training and the testing sets in order to prevent artificial inflation of the results. The Spearman's rank ordered correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC) values between the predicted and the ground truth quality scores for every iteration and the median value of the correlations were reported. We discovered that there is significant room for improvement in using the present NR-IQA metrics to predict HDR artifacts. The results are summarized in Table 5.1.

5.7 Conclusion

In this chapter we have outlined the different sources of the HDR images, the algorithms used to process them and the crowdsourced subjective study framework to have the images evaluated by thousands of observers over the internet. We also showed the shortcomings of the present NR-IQA algorithms in judging the perceptual quality of HDR images and proposed a spatial-domain NR-IQA algorithm that shows correlates better with human perception. The goal of this subjective study is to gather ratings from

Table 5.1: Median Spearman"s Rank Ordered Correlaton Coefficient (SROCC) and Pearson"s Linear Correlation Coefficient (PLCC) between the algorithm scores for various IQA algorithms and the MOS scores for ESPL-LIVE HDR database. The table has been sorted in the descending order of SROCC of the 'Overall category'. Red indicates the proposed methods. The bold values indicate the best performing algorithm.

	IQA	Tone Mapping		Multi-Exposure Fusion		Post Processing		Overall	
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
1	G-IQA-1	0.728	0.764	**0.711**	**0.705**	**0.616**	**0.643**	**0.719** (0.671, 0.766)	**0.718**(0.652, 0.776)
2	G-IQA-2	**0.752**	**0.777**	0.706	0.690	0.529	0.552	0.711 (0.639, 0.792)	0.704(0.645, 0.788)
3	G-IQA-2 (L)	0.703	0.737	0.662	0.661	0.465	0.515	0.662 (0.575, 0.730)	0.663(0.571, 0.730)
4	G-IQA-1 (L)	0.672	0.702	0.634	0.637	0.551	0.582	0.661 (0.595, 0.732)	0.658(0.590, 0.738)
5	DESIQUE	0.542	0.553	0.572	0.584	0.529	0.563	0.570 (0.481, 0.657)	0.568(0.467, 0.650)
6	GM-LOG	0.549	0.562	0.545	0.541	0.578	0.599	0.556 (0.448, 0.638)	0.557(0.465, 0.639)
7	CurveletQA	0.584	0.623	0.517	0.535	0.481	0.506	0.547 (0.458, 0.610)	0.560(0.447, 0.631)
8	DIIVINE	0.523	0.530	0.453	0.472	0.392	0.447	0.482 (0.326, 0.578)	0.484(0.331, 0.583)
9	BLIINDS-II	0.412	0.442	0.446	0.459	0.486	0.510	0.444 (0.310, 0.519)	0.454(0.326, 0.545)
10	C-DIIVINE	0.453	0.453	0.423	0.460	0.432	0.470	0.434 (0.265, 0.551)	0.444(0.277, 0.538)
11	BRISQUE	0.340	0.370	0.494	0.516	0.468	0.483	0.418 (0.300, 0.500)	0.444(0.313, 0.528)

more than 5,000 unique observers. To the best of our knowledge, this is the world's largest and most comprehensive study of HDR image quality ever conducted. This can be used to construct better performing NR-IQA algorithms for HDR images. The next chapter outlines the details of the proposed NR-IQA algorithms. For the sake of completeness, the performance of these algorithms have been evaluated on the legacy natural and synthetic SDR images.

Chapter 6

Image Quality Evaluation Algorithm based on Natural Scene Statistics

6.1 Introduction

The previous chapter compares several NSS-based NR-IQA algorithms in terms of their correlation with human subjective evaluations of HDR image artifacts. For HDR image artifacts, this study shows that there is significant room for improvement in NSS-based NR-IQA algorithms. This chapter proposes two NSS-based NR-IQA algorithms that use the gradient domain and that show high degrees of correlation with human subjective scores for HDR artifacts. For completeness, the correlation performance of the proposed algorithms and other NSS-based NR-IQA algorithms has also been evaluated on natural and synthetic SDR images.

Among the existing algorithms, DErivative Statistics-based QUality Evaluator (DESIQUE) [67] has been found to show very good correlation with human ground truth scores for natural and synthetic images. This algorithm uses log-derivative statistics and combines features in both spatial and the frequency domain because certain class of distortions affect the scene statistics in complimentary transform domains. In this work, we propose a spatial domain NR-IQA algorithm using log-derivative statistics on the mean-subtracted contrast normalized (MSCN) transformed pixels[79]. In addition, we

also extract features from the σ-field of the image. Statistical features are also computed on the gradient field. Lastly, we take into account the chromatic artifacts by computing these statistics in perceptually relevant color spaces. The LAB color space has been used. The process is repeated over two scales.

6.2 Proposed algorithm

This section summarizes the proposed algorithm. This is based on the assumption that the log-derivative statistics of the pixels and the pixels gradient magnitudes change with the different types of HDR processing methods and this deviation may be used to predict the quality scores.

6.2.1 Computing Log-Derivatives

The log-derivative statistics of the images are based on the difference between a particular pixel and its neighbors after converting the pixels to the logarithm domain[67]. Let $M \times N$ be the dimension of the image I, and $I(i,j)$ be the pixel value in the (i,j)-th spatial location, $i \in \{1,2,..,M\}$, $j \in \{1,2,...,N\}$. The logarithm image is given by:

$$J(i,j) = \log[I(i,j) + C] \tag{6.1}$$

where C is a small constant added to avoid numerical instabilities. Considering the different neighboring directions, the following log-derivatives are defined:

$$D1 : \nabla_x J(i,j) = J(i,j+1) - J(i,j) \tag{6.2}$$

$$D2 : \nabla_y J(i,j) = J(i+1,j) - J(i,j) \tag{6.3}$$

$$D3 : \nabla_{xy} J(i,j) = J(i+1,j+1) - J(i,j) \tag{6.4}$$

$$D4 : \nabla_{yx} J(i,j) = J(i+1,j-1) - J(i,j) \tag{6.5}$$

$$D5 : \nabla_x \nabla_y J(i,j) = J(i-1,j) + J(i+1,j) - J(i,j-1) - J(i,j+1) \tag{6.6}$$

$$D6 : \nabla_{cx} \nabla_{cy} J(i,j)_1 = J(i,j) + J(i+1,j+1) - J(i,j+1) - J(i+1,j) \tag{6.7}$$

$$D7 : \nabla_{cx} \nabla_{cy} J(i,j)_2 = J(i-1,j-1) + J(i+1,j+1) - J(i-1,j+1) - J(i+1,j-1) \tag{6.8}$$

6.2.2 Spatial Domain Scene Statistics

For this work, we model the scene statistics of the images in the spatial domain, MSCN pixels and the σ-field of the image. The pixels of the image are preprocessed by mean subtraction and divisive normalization. MSCN pixels are generated by:

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + 1} \tag{6.9}$$

where the local mean $\mu(i,j)$ and standard deviation $\sigma(i,j)$ are defined as:

$$\mu(i,j) = \sum_{k=-K}^{k=K} \sum_{l=-L}^{l=L} w_{k,l} I(i+k, j+l) \tag{6.10}$$

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{k=K} \sum_{l=-L}^{l=L} w_{k,l} [I(i+k, j+l) - \mu(i,j)]^2} \tag{6.11}$$

$w = \{w_{k,l} | k = -K, .., K, l = -L, ..., L\}$ is a symmetric local convolution window centered at the (i,j)-th pixel. K and L determine the size of local patch considered in the calculation of the mean and standard deviation. In [66], the authors considered 7×7 image

patches, and a circularly symmetric 2D Gaussian kernel; however, experiments show that the distribution of the MSCN patches are not very sensitive to the window size, or the convolution kernel.

A zero mean Generalized Gaussian Distribution (GGD) can be used to model the MSCN coefficients $\hat{I}(i,j)$:

$$f(x;\alpha,\gamma^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} exp\left[-\left(\frac{|x|}{\beta}\right)^2\right] \tag{6.12}$$

where $\beta = \gamma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}}$ and $\Gamma(x) = \int_0^\infty t^{(x-1)}e^{-t}, x > 0$ is the gamma function. The parameters α and γ are known as shape and scale parameters respectively and are used as features that capture the deviation in the image statistics for the different HDR processing artifacts.

Log-derivatives of the adjacent MSCN coefficients are also modeled by a GGD. The shape(α) and scale(γ) parameters of the GGD fitted to the seven types of log-derivatives have also been used as features in the spatial domain.

We also extract two quantities from the σ-field: mean(Φ_σ) and square inverse of coefficient of variation(Ψ_σ). The quantities are defined as:

$$\Phi_\sigma = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\sigma(i,j) \tag{6.13}$$

$$\Sigma_\sigma(i,j) = \sqrt{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[\sigma(i+k,j+l) - \Phi_\sigma(i,j)]^2} \tag{6.14}$$

$$\Psi_\sigma = \left(\frac{\Phi_\sigma}{\Sigma_\sigma}\right)^2 \tag{6.15}$$

Table 6.1 shows a summary of the spatial domain features extracted in every scale and in each color channel.

Table 6.1: Spatial domain features considered for the proposed algorithm.

Feature ID	Feature Description
$f_1 - f_2$	Shape and Scale parameters of the GGD fitted to the MSCN coefficients as outlined in 6.9
$f_3 - f_{16}$	Shape and Scale parameters of the GGD fitted to the log-derivative of the seven types of neighbors as outlined in 6.2.1
$f_{17} - f_{18}$	Two parameters extracted from the σ-field as outlined in 6.11

6.2.3 Gradient Domain Scene Statistics

The gradient field of the image gives important information about the distribution of edges and variations in local contrast. The magnitude of the gradient gives information about the amount of local change in luminance and the orientation tells us the direction in which the change is the most rapid. Many of the HDR processing algorithms, such as tone-mapping or multi-exposure fusion algorithms are found to modify the local gradients of the multi-exposure stacks, and that results in changing the contrast of the resultant fused image both locally and globally. This has led us to believe that extracting the statistical features in the gradient domain may lead to better NR-IQA models. In this algorithm, the gradient information is incorporated in two different ways:

6.2.3.1 Using Gradient Magnitude

The gradient magnitude has been used in FR-IQA metrics[45]. The local gradient is computed by convolving the image with linear filters along the horizontal and vertical

directions. The Sobel operator is one such commonly used 3×3 gradient filter. The horizontal (H_x) and vertical (H_y) components of the Sobel operator are given by:

$$H_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{6.16}$$

$$H_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{6.17}$$

The gradient magnitude,M of the image $I(i,j)$ at the (i,j)-th spatial location is given by:

$$M(i,j) = \sqrt{(I * H_x)^2(i,j) + (I * H_y)^2(i,j)} \tag{6.18}$$

where $*$ denotes the convolution operator. The features, as summarized in Table6.1 are also extracted from the gradient magnitude field. The resultant algorithm that combines the spatial domain features of section 6.2.2 with these gradient magnitude features is referred to as Gradient-Image Quality Assessment-1 (**G-IQA-1**).

6.2.4 Using Gradient Structure Tensor

The gradient structure tensor[127] is an important operator that summarizes the predominant gradient directions over a local neighborhood. The 2D structure tensor is given by:

$$J = \begin{bmatrix} f(G_x) & f(G_x.G_y) \\ f(G_x.G_y) & f(G_y) \end{bmatrix} \tag{6.19}$$

where

$$f(V) = \sum_{l,k} w[i,j]V(i-l, j-k)^2 \tag{6.20}$$

$G_x(i,j)$ and $G_y(i,j)$ are the horizontal and vertical components of the gradient vector at pixel (i,j) respectively and w is a window of dimension of PXP over which the localized structure tensor is computed. The quantities $G_x(i,j)$ and $G_y(i,j)$ are computed by convolving with difference-of-Gaussians. The relative discrepancy of between the two eigenvalues indicates the degree of anisotropy of the local gradient. The coherence measure is defined by:

$$C = \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right)^2 \tag{6.21}$$

where λ_1 and λ_2 are the two eigenvalues of the gradient structure tensor. The coherene measure is computed over PXP non-overlapping blocks of the image and the mean, standard deviation, skewness and kurtosis values are considered as the features. The resultant algorithm that combines the spatial domain features of section 6.2.2 with the gradient structure tensor features is referred to as Gradient-Image Quality Assessment-2 (**G-IQA-2**).

Table 6.2 shows the correlation of each type of feature with the MOS on the ESPL-LIVE HDR database. The low correlations between each individual features and the MOS indicates the need to combine complementary feature in order to predict the quality scores of image inflicted with a wide range of artifacts. Figure 6.1 shows three images of the same scene (obtained from the ESPL-LIVE HDR database) tone-mapped using three different versions. Each of the tonemapping operators give rise to distinctly different images. Figures 6.2 and 6.3 show the corresponding changes in features in different domains.

Table 6.2: Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between each feature and DMOS across 50 train-test (4:1) combinations on the ESPL-LIVE HDR Database over a single image scale and considering the L-component. Low correlations between each individual feature and DMOS show that the features complement each other

Domain	Feature Description	SROCC	PLCC
Spatial	Shape and Scale parameters of the GGD fitted to the MSCN coefficients (6.9) $[f_1 - f_2]$	0.238	0.266
Spatial	Shape and Scale parameters of the GGD fitted to the log-derivative of the seven types of neighbors (Section 6.2.1) $[f_3 - f_{16}]$	**0.439**	0.436
Spatial	Two parameters extracted from the σ-field (6.11) $[f_{17} - f_{18}]$	0.369	0.358
Gradient	Shape and Scale parameters of the GGD fitted to the MSCN coefficients of gradient magnitude field (Section 6.2.3.1) $[f_{19} - f_{20}]$	0.250	0.277
Gradient	Shape and Scale parameters of the GGD fitted to the log-derivative of the seven types of neighbors of gradient magnitude field (Section 6.2.3.1) $[f_{21} - f_{34}]$	0.386	0.384
Gradient	Two parameters extracted from the σ-field of gradient magnitude field (Section 6.2.3.1) $[f_{35} - f_{36}]$	0.388	0.392
Gradient	Mean, standard deviation, skewness, and kurtosis of gradient structure tensor (Section 6.2.4) $[f_{37} - f_{38}]$	0.420	**0.466**

(a) MOS = 40.47 (b) MOS = 49.23 (c) MOS = 52.80

Figure 6.1: Image of the same scene tone-mapped using three different versions. (a) Method 1 (Durand TMO [96]) (b) Method 2 (Fattal TMO [99]) (c) Method 3 (Reinhard TMO [95]) The caption of each image shows the MOS.

6.3 Results

This section outlines the results of evaluating the performance of state-of-the-art NR-IQA algorithms on the ESPL-LIVE HDR Database. The performance of the proposed algorithms have been evaluated by measuring correlation with subjective scores and the results have also been analyzed to determine statistical significance.

Once the features were extracted, a mapping is obtained from the feature space to the DMOS scores using a regression method, which provides a measure of the perceptual quality. We used a support vector machine regressor (SVR), specifically LibSVM [85] to implement ϵ-SVR with the radial basis function kernel, γ is by default the inverse of the number of features.

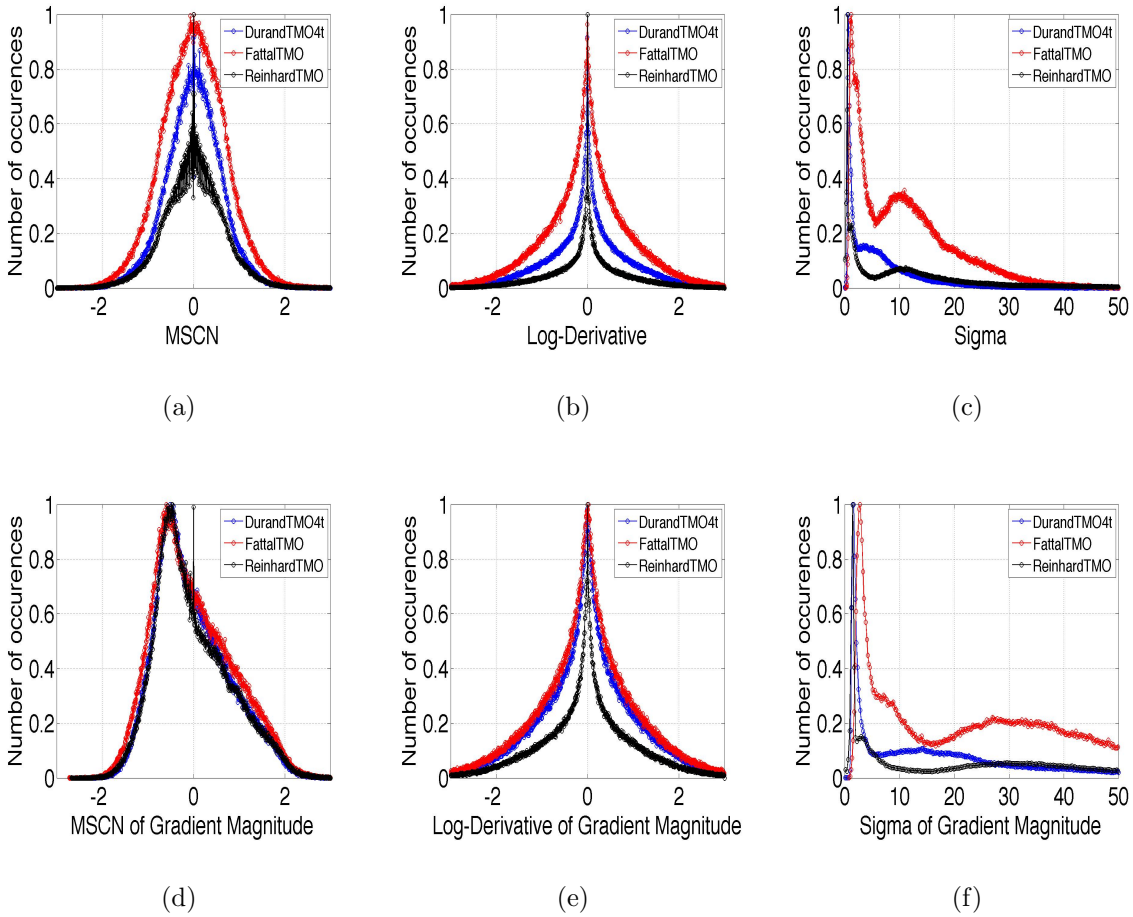


Figure 6.2: Histograms of (a) MSCN pixels, (b) Log-derviatives of the MSCN pixels (c) σ-field of the pixels (d) MSCN coefficients of the gradient magnitude field (e) Log-derviatives of the MSCN coefficients of the gradient magnitude field (f) σ-field of the gradient magnitude field. The legends "Method 1", "Method 2", and "Method 3" represents processing by Durand TMO [96], Fattal TMO [99], and Reinhard TMO [95] respectively as show in Fig 6.1.

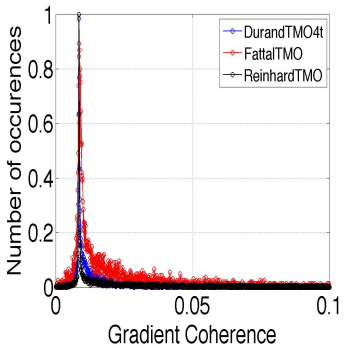


Figure 6.3: Histograms of coherence of the gradient structure tensor. The legends "Method 1", "Method 2", and "Method 3" represents processing by Durand TMO [96], Fattal TMO [99], and Reinhard TMO [95] respectively as show in Fig 6.1.

6.3.1 Experiments on ESPL-LIVE HDR Database

The proposed algorithms, G-IQA-1 and G-IQA-2 have been evaluated on the ESPL-LIVE HDR database. G-IQA-1 (L) and G-IQA-2 (L) indicate versions of the proposed algorithms using only the luminance channel (L). We randomly split the data into disjoint training and testing sets at 4:1 ratio and the split was randomized over 100 trials. Care was taken to ensure that the same source scene does not appear in the training and the testing sets in order to prevent artificial inflation of the results. The Spearmans rank ordered correlation coefficient (SROCC) and Pearsons linear correlation coefficient (PLCC) values between the predicted and the ground truth quality scores for every iteration and the median value of the correlations were reported. We discovered that there is significant room for improvement in using the present NR-IQA metrics to predict HDR artifacts. The results are summarized in table 6.3.

Table 6.4 shows the root-mean-squared-error (RMSE), reduced $\tilde{\chi}^2$ statistic between scores predicted by the algorithms and the MOS for various algorithms (after logistic func-

Table 6.3: Median Spearman's Rank Ordered Correlaton Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between the algorithm scores for various IQA algorithms and the MOS scores for ESPL-LIVE HDR database. The table has been sorted in the descending order of SROCC of the 'Overall category'. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, computed by bootstrapping using 100 samples. Red indicates the proposed methods. The bold values indicate the best performing algorithm.

	IQA	Tone Mapping		Multi-Exposure Fusion		Post Processing		Overall	
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
1	G-IQA-1	0.728	0.764	**0.711**	**0.705**	**0.616**	**0.643**	**0.719** (0.671, 0.766)	**0.718**(0.652, 0.776)
2	G-IQA-2	**0.752**	**0.777**	0.706	0.690	0.529	0.552	0.711 (0.639, 0.792)	0.704(0.645, 0.788)
3	G-IQA-2 (L)	0.703	0.737	0.662	0.661	0.465	0.515	0.662 (0.575, 0.730)	0.663(0.571, 0.730)
4	G-IQA-1 (L)	0.672	0.702	0.634	0.637	0.551	0.582	0.661 (0.595, 0.732)	0.658(0.590, 0.738)
5	DESIQUE	0.542	0.553	0.572	0.584	0.529	0.563	0.570 (0.481, 0.657)	0.568(0.467, 0.650)
6	GM-LOG	0.549	0.562	0.545	0.541	0.578	0.599	0.556 (0.448, 0.638)	0.557(0.465, 0.639)
7	CurveletQA	0.584	0.623	0.517	0.535	0.481	0.506	0.547 (0.458, 0.610)	0.560(0.447, 0.631)
8	DIIVINE	0.523	0.530	0.453	0.472	0.392	0.447	0.482 (0.326, 0.578)	0.484(0.331, 0.583)
9	BLIINDS-II	0.412	0.442	0.446	0.459	0.486	0.510	0.444 (0.310, 0.519)	0.454(0.326, 0.545)
10	C-DIIVINE	0.453	0.453	0.423	0.460	0.432	0.470	0.434 (0.265, 0.551)	0.444(0.277, 0.538)
11	BRISQUE	0.340	0.370	0.494	0.516	0.468	0.483	0.418 (0.300, 0.500)	0.444(0.313, 0.528)

tion fitting) and outlier ratio (expressed in percentage). The top performing algorithms also show lower values of RMSE and outlier ratio.

Fig. 6.4 shows box plots of the distribution of SROCC values for each of the 100 trials of random train-test splits on the ESPL-LIVE HDR Image Database. This enable us to study the robustness of performance of the algorithms with variations of the choice of the training set. The proposed method shows smaller variation in the degree of correlation with human subjective evaluation.

To analyze the degree of variation of SROCC between the scores predicted by the algorithm and the DMOS, the percentage of train/test splits was varied from 90% of the content used for training and the remaining 10% used for testing to 10% of the content used for training to 90% used for testing. The knee of the curve occurs roughly at 60:40 train:test splits. This shows that the results are not affected by overfitting or underfitting

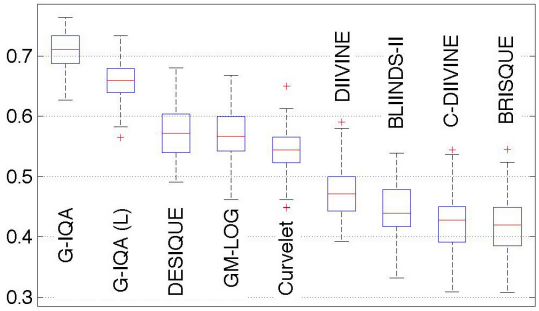


Figure 6.4: Box plot of SROCC of learning based NR-IQA algorithms on images in the ESPL-LIVE HDR Image Database for 4:1 train-test splits over 100 trials. For each box, median is the central box, edges of the box represent the 25th and 75th percentiles, the whiskers span the most extreme non-outlier data points, and the outliers are plotted individually.

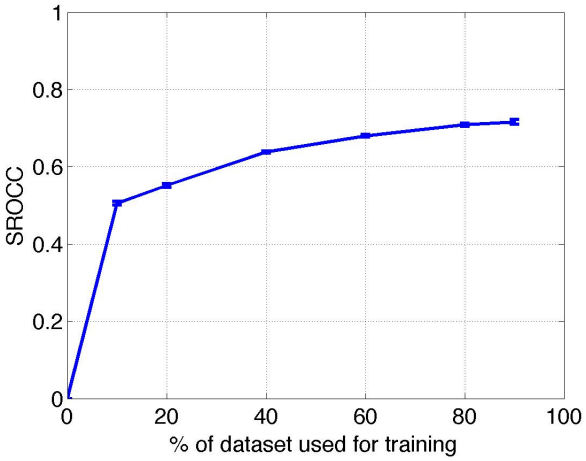


Figure 6.5: Mean SROCC between predicted and subjective DMOS scores for G-IQA-1 (and the associated 95% confidence intervals) as a function of the percentage of the content used for training on images in the ESPL-LIVE HDR Image Database over 50 trials.

Table 6.4: Root-mean-square error (RMSE), reduced $\tilde{\chi}^2$ statistic between the algorithm scores and the DMOS for various NR-IQA Algorithms (after logistic function fitting) and outlier ratio (expressed in percentage) for each distortion category for ESPL-LIVE HDR database. Red indicates the proposed methods. The bold values indicate the best performing algorithm for that category.

	IQA	Tone Mapping			Multi-Exposure Fusion			Post Processing			Overall		
		RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR	RMSE	$\tilde{\chi}^2$	OR
1	G-IQA-1	6.711	9.908	**0.000**	**6.884**	21.155	**0.000**	**6.884**	2.376	**0.000**	**7.033**	13.918	**0.275**
2	G-IQA-2	**6.643**	3.576	**0.000**	6.988	5.983	**0.000**	7.457	6.660	**0.000**	7.231	16.495	0.277
3	G-IQA-2 (L)	7.070	5.327	**0.000**	7.178	13.882	**0.000**	7.742	3.227	**0.000**	7.607	13.879	0.551
4	G-IQA-1 (L)	7.434	8.624	0.662	7.484	5.263	**0.000**	7.308	3.131	**0.000**	7.628	12.558	0.552
5	DESIQUE	8.577	12.079	0.683	7.862	11.588	0.687	7.402	1.851	**0.000**	8.296	19.614	0.829
6	GM-LOG	8.632	5.002	1.170	8.028	15.027	0.702	7.420	0.851	**0.000**	8.357	20.659	0.829
7	CurveletQA	8.177	17.408	0.694	8.054	10.754	0.714	7.922	2.892	**0.000**	8.511	15.253	0.829
8	DIIVINE	8.805	10.025	0.791	8.371	5.663	0.667	7.979	2.659	**0.000**	8.821	12.115	0.829
9	BLIINDS-II	9.330	7.565	0.697	8.517	19.979	0.752	7.818	1.976	**0.000**	8.975	21.948	0.828
10	C-DIIVINE	9.167	15.338	1.356	8.485	8.374	0.671	7.852	1.428	**0.000**	8.983	12.305	0.966
11	BRISQUE	9.535	16.712	1.356	8.227	5.681	0.685	7.894	7.146	**0.000**	9.049	17.259	0.831

to the training data. Figure 6.5 shows the results.

Figure shows the scatter plot between predicted scores and MOS scores on ESPL-LIVE HDR Database for a selected few NR-IQA algorithms.

6.3.2 Determination of Statistical Significance

For this purpose, nine representative NR-IQA algorithms were selected. The statistical significance tests were carried out for multiple training-test splits, using different 4:1 train-test splits of the database each time, and similar results were obtained. The table outlines the results obtained for one such representative trial. To determine whether the IQA algorithms are significantly different from each other, the F-statistic, as in [7][81], was used to determine the statistical significance between the variances of the residuals after a non-linear logistic mapping between the two IQA algorithms, at the 95% confidence interval. Table 6.5 shows the results for ten selected IQA algorithms and all

distortions. Overall, the proposed algorithms are found to be statistically superior to the other NR-IQA algorithms.

Table 6.5: Results of the F-test performed on the residuals between model predictions and MOS scores on ESPL-LIVE HDR database. Each cell in the table is a codeword consisting of 4 symbols that correspond to "Tone Mapping Operators"', "Multi-Exposure Fusion", "Post Processing", and "Overall" distortions. "1"("0") indicates that the performance of the row IQA is superior(inferior) to that of the column IQA. - indicates that the statistical performance of the row IQA is equivalent to that of the column IQA. The matrix is symmetric. Red indicates the proposed methods.

	G-IQA-1	G-IQA-2	DESIQUE	BRISQUE	GM-LOG	C-DIIVINE	DIIVINE	BLIINDS-II	CurveletQA
G-IQA-1	- - - -	- - - -	1 - - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1
G-IQA-2	- - - -	- - - -	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	1 1 - 1	- 1 - 1
DESIQUE	0 - - 0	0 0 - 0	- - - -	1 - - 1	- - - -	- - - -	- - - -	- - - -	- - - -
BRISQUE	0 0 - 0	0 0 - 0	0 - - 0	- - - -	- - - -	- - - -	- - - -	- - - -	0 - - -
GM-LOG	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -
C-DIIVINE	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -
DIIVINE	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	0 - - -
BLIINDS-II	0 0 - 0	0 0 - 0	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -	- - - -
CurveletQA	0 0 - 0	- 0 - 0	- - - -	1 - - -	- - - -	- - - -	1 - - -	- - - -	- - - -

6.3.3 Experiments on other databases

In addition to the ESPL-HDR database, for the sake of completeness, the performance of the proposed algorithm has also been tested on the legacy LIVE database[7], LIVE Multiply Distorted Database[8] and on the ESPL Synthetic Image Database[9]. Table 6.6 and Table 6.7 show the performance of the proposed algorithm on the LIVE database[7] and LIVE Multiply Distorted Image Database[8] respectively. Similar technique of splitting the data into disjoint training and testing sets at 4:1 ratio, randomized over 100 trials, was followed. The high degrees of correlation with the subjective data shows that the proposed methods can also capture the processing, compression and transmission artifacts arising in SDR images.

Table 6.6: Median Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between algorithm scores and DMOS for various NR-IQA algorithms across 100 train-test (4:1) combinations on the LIVE Database of natural images. Bold values indicate the best performing algorithm for that category. Performances of some FR-IQA algorithms (shown in italics) have been included for comparison. Red indicates the proposed methods. Italics indicate FR-IQA algorithms.

	IQA	JP2K		JPEG		Gaussian Noise		Blur		Fast Fading		Overall	
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
1	GM-LOG	0.882	0.904	0.878	**0.917**	0.978	0.988	0.915	0.925	**0.899**	**0.917**	**0.914** (0.860, 0.941)	**0.917**(0.857, 0.942)
2	G-IQA-1	0.905	**0.914**	**0.883**	0.915	**0.983**	**0.990**	0.917	0.925	0.836	0.860	0.906 (0.788, 0.952)	0.907(0.786, 0.952)
3	G-IQA-2	0.904	0.910	0.867	0.902	0.982	0.990	0.920	0.930	0.841	0.863	0.904 (0.810, 0.943)	0.903(0.819, 0.943)
4	BRISQUE	0.878	0.888	0.852	0.889	0.962	0.975	**0.941**	**0.942**	0.863	0.887	0.902 (0.798, 0.950)	0.900(0.786, 0.949)
5	C-DIIVINE	0.872	0.882	0.839	0.876	0.965	0.974	0.915	0.915	0.891	0.915	0.898 (0.817, 0.944)	0.905(0.816, 0.945)
6	BLIINDS-II	**0.907**	0.912	0.846	0.884	0.939	0.960	0.906	0.918	0.884	0.902	0.897 (0.775, 0.938)	0.900(0.746, 0.946)
7	DESIQUE	0.875	0.893	0.824	0.869	0.975	0.985	0.908	0.925	0.829	0.865	0.878 (0.805, 0.944)	0.884(0.797, 0.938)
8	G-IQA-1 (L)	0.848	0.853	0.839	0.870	0.955	0.960	0.865	0.891	0.788	0.836	0.866 (0.721, 0.934)	0.861(0.710, 0.930)
9	CurveletQA	0.816	0.824	0.827	0.836	0.969	0.979	0.896	0.900	0.826	0.866	0.863 (0.694, 0.916)	0.859(0.493, 0.911)
10	G-IQA-2 (L)	0.822	0.843	0.818	0.855	0.941	0.956	0.897	0.907	0.729	0.737	0.837 (0.600, 0.920)	0.840(0.621, 0.911)
11	DIIVINE	0.824	0.828	0.759	0.798	0.937	0.950	0.854	0.888	0.759	0.792	0.827 (0.451, 0.924)	0.829(0.452, 0.919)
12	GRNN	0.816	0.822	0.765	0.748	0.916	0.939	0.877	0.896	0.816	0.861	0.776 (0.652, 0.833)	0.784(0.688, 0.854)
13	BIQI	0.668	0.689	0.580	0.612	0.776	0.782	0.744	0.783	0.567	0.578	0.634 (0.173, 0.811)	0.642(0.194, 0.815)
14	*MS-SSIM*	0.963	0.975	0.979	0.979	0.977	0.988	0.954	0.965	0.939	0.949	0.954	0.951
15	*SSIM*	0.939	0.941	0.947	0.946	0.964	0.982	0.905	0.900	**0.939**	**0.951**	0.913	0.907
16	*PSNR*	0.865	0.876	0.883	0.903	0.941	0.917	0.752	0.780	0.8736	0.880	0.864	0.859

In order to show the database independence of the proposed method, it was trained on the LIVE database of natural images and the performance was evaluated on the TID 2013 database[10]. Among the distortions present in the TID2008 database: JPEG2000, JPEG, Gaussian noise, and blur were chosen. Table 6.8 shows the results obtained for the different types of artifacts. In addition, for the sake of comparison, the results obtained from some well-known NR-IQA and FR-IQA algorithms have also been included. The degree of correlation drops in contrast to the results obtained when the methods are trained and tested on disjoint sets of the same database, but still we see a sufficiently high degree of match with the human subjective scores. The results are also visually illustrated in figure 6.7.

The proposed algorithms have also been evaluated on the ESPL Synthetic Image Database in Chapter 2. Table 6.9 summarizes the results, along with the runtime.

Table 6.7: Spearman Rank-order Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between the algorithm scores for various IQA algorithms and the MOS scores across 100 train-test (4:1) combinations on the LIVE Multiply Distorted Image database (100 iterations considered for G-IQA-1 and G-IQA-2). Bold values indicate the best performing algorithm for that category. Performances of some FR-IQA algorithms (shown in italics) have been included for comparison. Red indicates the proposed methods.

NR-IQA algorithms	SROCC	PLCC
G-IQA-1	0.**9523**	**0.9589**
G-IQA-2	0.9541	0.9577
DESIQUE	0.9403	0.9511
BRISQUE	0.9111	0.9349
PSNR	0.6954	0.7637
MS-SSIM	0.8454	0.8825
VIF	0.8874	0.9083
IFC	0.8888	0.9137
NQM	0.9020	0.9160
VSNR	0.7844	0.8326
WSNR	0.7768	0.8408

6.4 Conclusion

In this chapter, I describe the different spatial domain features extracted in the proposed scene-statistics based NR-IQA algorithms. I also conduct a series of experiments on different IQA databases to evaluate its performance. The proposed methods show high degree of correlations for HDR artifacts and also performs well on the legacy natural and synthetic SDR image databases. The next chapter summarizes the dissertation and proposes avenues of future work.

Table 6.8: Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between algorithm scores and DMOS for various NR-IQA algorithms (mentioned in Table 3.1) across 100 train-test (4:1) combinations on a subset of the TID 2013 Database after training the algorithms on the LIVE database. Bold values indicate the best performing algorithm for that category. Performances of some FR-IQA algorithms (shown in italics) have been included for comparison. '-' indicates that the original paper did not report these values. Red indicates the proposed methods.

IQA	JPEG		JP2K		GN		Blur		Overall	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
G-IQA-1	0.8644	**0.9133**	0.9094	**0.9257**	0.8741	**0.8841**	**0.9268**	**0.9220**	0.8393	**0.8828**
BLIINDS-II	**0.8889**	-	**0.9147**	-	0.6956	-	0.8572	-	**0.8542**	-
BRISQUE	0.8355	0.8670	0.8704	0.8896	0.6955	0.6993	0.8159	0.8049	0.7789	0.8191
DESIQUE	0.7622	0.8310	0.8116	0.8053	0.7521	0.7699	0.7137	0.7565	0.7271	0.7494
GMSD	0.9507	0.9736	0.9657	0.9788	0.9462	0.9126	0.9113	0.8924	0.9508	0.9488
MS-SSIM	0.9172	0.9781	0.9486	0.9776	0.8641	0.9541	0.9619	0.9481	0.9135	0.9495

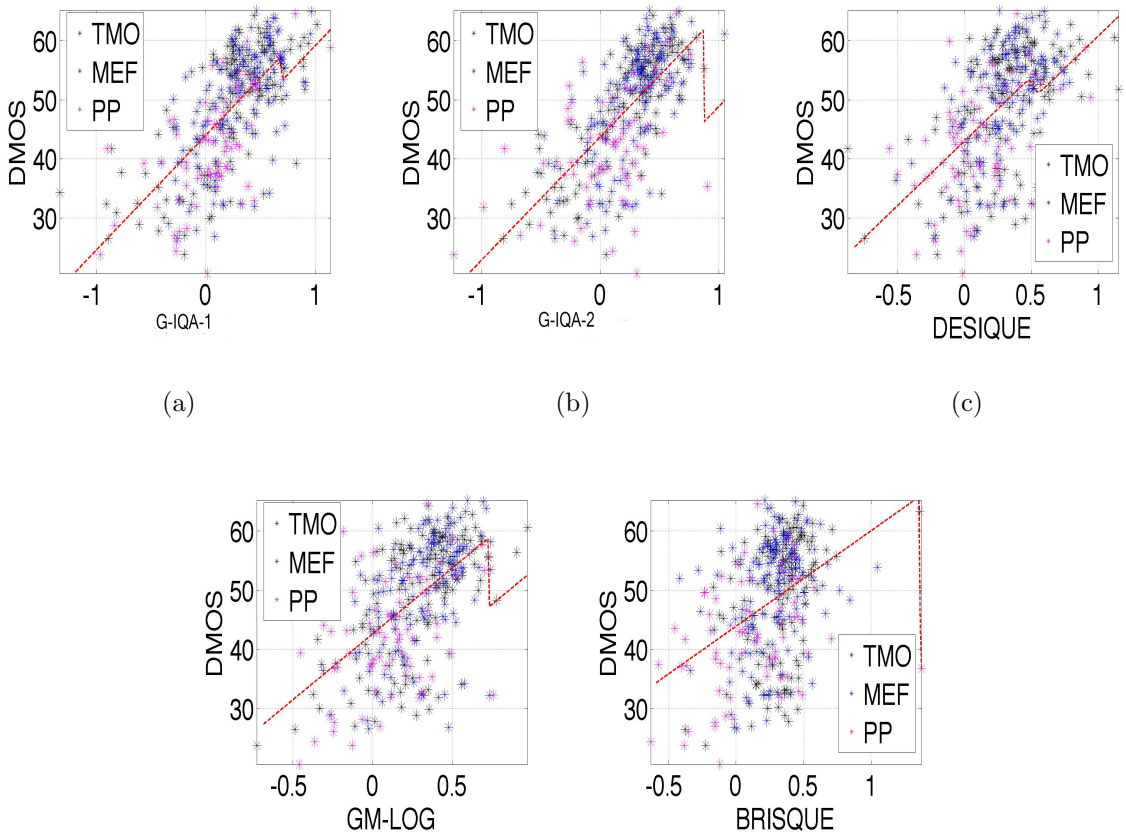


Figure 6.6: Scatter plot between predicted scores and MOS scores on ESPL-LIVE HDR Database for a selected few NR-IQA algorithms. The red line indicates the logistic regression fit. The abbreviations "TMO", "MEF", and "PP" indicate Tone-Mapping, Multi-Exposure Fusion and Post-Processing algorithms respectively.

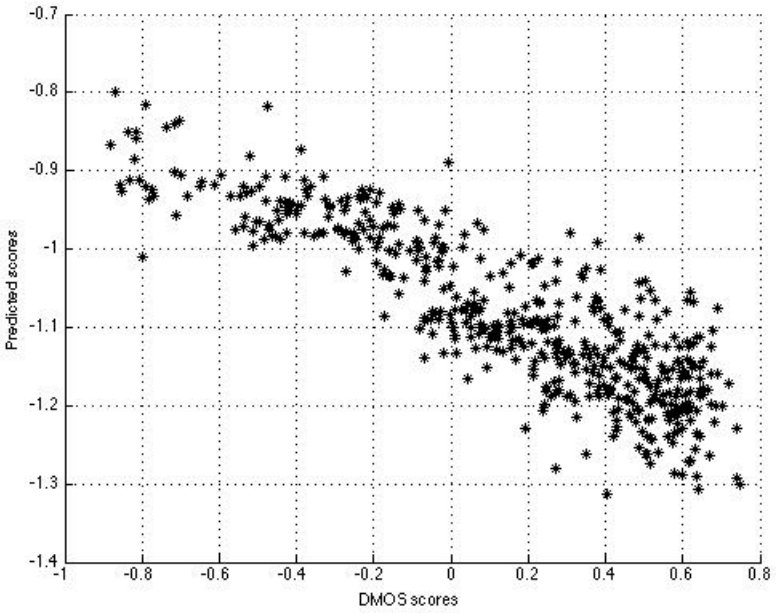


Figure 6.7: Scatter plot between predicted scores of G-IQA-1 Versus subjective MOS on TID2013 database[10] when the proposed algorithm is trained on LIVE database[7].

Table 6.9: Median Spearman's Rank Ordered Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC) between algorithm scores and DMOS for various NR-IQA algorithms (described in Section 3.3.3) along with algorithm computation time needed (on a Macintosh laptop having 8 GB RAM, 2.9 GHz clock, Intel Core i7 CPU) across 100 train-test (4:1) combinations on the ESPL Synthetic Image Database (50 trials for CORNIA in row 2). Italicized entries are NR-IQA algorithms meant for particular distortion categories. Italicized algorithms indicate the values obtained when the mentioned NR-IQA algorithms were applied for distortion categories other than what they were originally intended for. For these algorithms, the correlation values quoted in the "Overall" category is same as the correlations in the distortion category for which the algorithm was originally meant for. The numbers within parentheses in the "Overall" category show the confidence intervals on correlation values, obtained by considering the maximum and minimum values of the correlations obtained over a 100 trials. The table has been sorted in the descending order of SROCC for the "Overall" category. Red indicates the proposed methods. Bold values indicate the best performing algorithm for that category.

	IQA	Interp.		Blur		GN		JPEG		FF		Overall (Confidence Interval)		Time
		SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	(s)
1	G-IQA-1 (L)	0.605	0.646	0.612	0.640	0.858	0.904	0.901	0.927	**0.774**	**0.833**	**0.813** (0.562, 0.918)	**0.819**(0.626, 0.911)	2.134
2	CORNIA	**0.808**	**0.823**	**0.775**	**0.801**	0.793	0.821	0.898	0.918	0.706	0.763	0.810 (0.687, 0.875)	0.807(0.682, 0.880)	84.330
3	C-DIIVINE	0.702	0.760	0.730	0.769	0.847	0.896	0.841	0.879	0.738	0.802	0.798 (0.691, 0.916)	0.808(0.712, 0.912)	65.720
4	BRISQUE	0.631	0.643	0.720	0.782	0.840	0.902	0.898	0.935	0.717	0.740	0.789 (0.663, 0.897)	0.795(0.690, 0.895)	0.590
5	GM-LOG	0.680	0.711	0.653	0.694	0.853	0.906	0.912	0.944	0.701	0.746	0.787 (0.627, 0.893)	0.791(0.594, 0.892)	0.590
6	G-IQA-1	0.580	0.647	0.474	0.508	0.871	0.920	0.922	0.942	0.726	0.758	0.774 (0.552, 0.893)	0.786(0.569, 0.887)	4.641
7	DESIQUE	0.595	0.678	0.590	0.617	**0.886**	**0.922**	**0.934**	**0.955**	0.714	0.737	0.773 (0.570, 0.909)	0.781(0.588, 0.901)	2.250
8	G-IQA-2	0.510	0.584	0.565	0.576	0.857	0.906	0.865	0.879	0.728	0.762	0.743 (0.387, 0.888)	0.744(0.406, 0.877)	42.693
9	CurveletQA	0.658	0.695	0.695	0.753	0.880	0.916	0.854	0.880	0.553	0.595	0.731 (0.460, 0.872)	0.734(0.490, 0.863)	20.130
10	G-IQA-2 (L)	0.509	0.563	0.488	0.529	0.859	0.906	0.874	0.909	0.668	0.729	0.689 (0.489, 0.876)	0.714(0.538, 0.881)	14.893
11	BIQI	0.665	0.733	0.732	0.764	0.837	0.903	0.735	0.769	0.538	0.593	0.676 (0.338, 0.849)	0.676(0.414, 0.858)	0.330
12	GRNN	0.537	0.592	0.371	0.409	0.811	0.896	0.738	0.790	0.408	0.551	0.602 (0.422, 0.777)	0.643(0.422, 0.802)	2.480
13	BLIINDS-II	0.388	0.444	0.499	0.556	0.794	0.839	0.680	0.754	0.548	0.608	0.596 (0.333, 0.834)	0.622(0.382, 0.835)	81.790
14	Anisotropy	0.364	0.354	0.357	0.400	0.835	0.871	0.385	0.449	0.392	0.439	0.470 (0.379, 0.513)	0.431(0.391, 0.483)	10.780
15	NIQE	0.428	0.496	0.425	0.528	0.740	0.511	0.732	0.834	0.606	0.623	0.377 (0.144, 0.600)	0.395(0.181, 0.601)	3.240
16	DIIVINE	0.421	0.523	0.441	0.490	0.484	0.537	0.444	0.489	0.439	0.513	0.372 (0.080, 0.700)	0.404(0.121, 0.705)	118.040
17	TMIQA	0.367	0.376	0.437	0.353	0.741	0.681	0.159	0.227	0.411	0.469	0.220 (0.097, 0.300)	0.311(0.223, 0.387)	0.120
18	*LPCM*	*0.415*	*0.444*	**0.836**	**0.847**	*0.623*	*0.621*	*0.211*	*0.231*	*0.108*	*0.237*	**0.836**(0.791, 0.890)	**0.847**(0.792, 0.885)	11.570
19	*CPBDM*	*0.676*	*0.720*	0.757	0.766	*0.746*	*0.815*	*0.765*	*0.749*	*0.347*	*0.405*	0.757 (0.678, 0.808)	0.766(0.669, 0.830)	3.500
20	*FISH*	*0.222*	*0.305*	0.705	0.716	*0.823*	*0.870*	*0.196*	*0.252*	*0.432*	*0.472*	0.705 (0.548, 0.787)	0.716(0.631, 0.793)	0.250
21	S_3	*0.409*	*0.449*	0.700	0.756	*0.747*	*0.786*	*0.151*	*0.189*	*0.402*	*0.450*	0.700 (0.554, 0.792)	0.756(0.692, 0.818)	308.150
22	*JNBM*	*0.598*	*0.635*	0.506	0.528	*0.756*	*0.816*	*0.536*	*0.512*	*0.448*	*0.455*	0.506 (0.327, 0.627)	0.528(0.336, 0.676)	7.520
23	*NLWT*	*0.324*	*0.334*	*0.024*	*0.141*	**0.872**	**0.888**	*0.000*	*0.187*	*0.559*	*0.589*	**0.872**(0.821, 0.905)	**0.888**(0.847, 0.928)	10.410
24	*FNVE*	*0.320*	*0.332*	*0.463*	*0.553*	0.863	0.887	*0.517*	*0.543*	*0.461*	*0.459*	0.863 (0.817, 0.894)	0.887(0.838, 0.915)	0.030
25	*JPEG-NR*	*0.540*	*0.570*	*0.593*	*0.650*	*0.748*	*0.865*	**0.928**	**0.954**	*0.464*	*0.607*	**0.928**(0.878, 0.952)	**0.954**(0.940, 0.969)	0.110
26	*NJQA*	*0.373*	*0.406*	*0.333*	*0.367*	*0.878*	*0.808*	0.743	0.819	*0.420*	*0.437*	0.743 (0.649, 0.854)	0.819(0.732, 0.869)	192.590

Chapter 7

Conclusion and Future Work

While more and more visual data is being generated in the world, either from capturing natural scenes using optical cameras or rendering computer generated imagery, visual quality evaluation is an interesting and relevant problem to explore. In my dissertation, I have contributed in subjective and objective quality evaluation of synthetic scenes and high dynamic range images. I propose the following thesis statement:

Using scene statistics yields automated visual quality assessment algorithms for synthetic images and high dynamic range images that have high correlation with human visual quality evaluation.

In the following section, I discuss how my contributions in each chapter contribute toward defending this thesis statement. Section 7.1 discusses my contributions in each chapter toward defending the thesis statement. Section 7.2 presents future work to build on and extend the dissertation results.

7.1 Summary

In this section, I would like to summarize how my contribution in every chapter helps to defend my thesis statement. The premise of Natural Scene Statistics (NSS) is based on the fact that irrespective of content, natural images possess very unique

statistical properties and the presence of distortions make the image deviate from these statistics. The NSS based NR-IQA algorithms exploit this change in the image statistics to predict the visual quality of images. However, these methods have not been applied to distorted synthetic and HDR images. In addition, there has been a dearth of images annotated with human scores in these domains. Collecting ground truth subjective scores is very important both for synthetic and HDR scenes because the NR-IQA algorithms need to be trained on the human ratings.

In chapter 2, I outline the details of the SPL Synthetic Image Database, comprising of 25 high quality pristine images and 500 distorted images generated by controlled introduction of varying degrees of different types of processing, compression and transmission artifacts, such as interpolation, blur, additive noise, JPEG compression and Fast-Fading channels. I also explain the methods of processing the raw scores and rejecting outliers. In chapter 3, I compare the performance of more than 50 FR, RR and NR-IQA algorithms (originally designed for natural images) by correlating the scores obtained from the IQA algorithms with the synthetic image DMOS scores . For the FR-IQA algorithms I have identified the key distortion categories for which the natural images IQA algorithms show a lesser degree of correlation. I have shown that the NSS based NR-IQA algorithms can be used even for predicting the quality scores of distorted synthetic scenes.

For HDR images, in chapter 4, I improved the state-of-the-art FR-IQA algorithms for evaluating the quality of tonemapped images in comparison to the original HDR luminance map by incorporating models of natural scene statistics and visual saliency. The scene statistics model is based on mean-subtracted-contrast-normalized coefficients and the standard deviation field. In addition, the algorithm also showed a high degree of

correlation on tonemapped images afflicted with JPEG compression artifacts.

In chapter 5, I conducted a large scale online crowdsourced subjective test on a corpus of 1815 HDR images created using different processing artifacts in order to garner ratings from a larger number of human subjects. To the best of our knowledge, presently this is the largest HDR image database in the world involving the largest number of source images and most number of human evaluations. In chapter 6, I proposed a scene-statistics based NR-IQA algorithm in the gradient domain for evaluating HDR artifacts that outperforms the state-of-the-art NR-IQA algorithms on this class of distortion. For completeness, the algorithm has also been evaluated on SDR natural (LIVE Image Quality Database[7], LIVE Multiply Distorted Images[8]) and synthetic image databases (ESPL Synthetic Image Database).

All of these contributions aim at emphasizing the usefulness of scene statistics models for objective quality evaluation of synthetic images and HDR images.

7.2 Future Work

In this section I outline several interesting research directions in image quality assessment to which the researchers in image processing may potentially contribute to.

7.2.1 IQA for a larger number of graphics artifacts

In this dissertation, I have conducted subjective tests on high quality computer graphics generated images after the controlled addition of different types of distortions at varying levels. A follow-up subjective study can be done on images rendered directly by using a graphics rendering pipeline. Some of these aspects that may be considered are:

7.2.1.1 Dynamic resolution rendering

Instead of rendering every frame at the same display resolution, based on the scene content, the rendering resolution may be changed, leading to significant savings in computation time and power. Subjective tests need to be considered in order to gain some insight on how the rendering resolution may be varied locally depending upon the scene complexity.

7.2.1.2 Number and/or types of lights used

The number and types of light sources used can drastically affect the rendered scnene. However, the present algorithms depend on feedback from human subjects in order to optimally place the light sources. Designing objective metrics that give some idea of local and global contrast and correlates well with human perception need to be designed to make benchmarking of algorithms easier for this class of problems.

7.2.1.3 Motion Blur

In this disseration, we found that for some of the images, the observers evaluated the slightly blurred image higher than the corresponding reference and the present NR-IQA blur evaluation methods, that do well on natural scenes, shows a less impressive performance in predicting blur in synthetic scenes. We conjecture that, since in many computer graphics applications, blur is introduced intentionally in order to increase aesthetic quality (such as "soft shadowing" or motion blur), users found them to be less annoying compared to natural scenes. A follow up subjective test on motioned blurred computer graphics images would be a potential avenue of future research.

7.2.2 Using IQA algorithms for different applications

This dissertation shows the usefulness of the scene-statistics approaches in quantifying different types of visual distortions. Following are the potential areas of applications of these algorithms:

7.2.2.1 Cloud gaming

Unlike the other computer graphics databases, artifacts arising from compressing graphics images and sending them over a Rayleigh fading channel have been included in this work. The IQA metrics studied can be used in this context for quantifying the visual distortions arising from sending the rendered video game frames from the server to the dumb clients.

7.2.2.2 IQA for hybrid scenes

This dissertation shows that the presence of distortions deviate the scene-statistics of synthetic images in the same way as natural scenes. This leads us to the interesting problem of visual quality evaluation of hybrid natural and synthetic scenes that occur frequently in many augmented reality applications. The scene-statistics based approaches can lead to a synergistic IQA algorithm that would be useful for evaluating the artifacts in hybrid images having both natural and synthetic components.

7.2.3 NR-IQA algorithms of HDR images

This dissertation describes a large-scale crowdsourced study for HDR artifacts and proposes a NR-IQA algorithm based on scene-statistics using support vector regression

technique that does well on these artifacts. However, there is plenty of scope to improve upon these algorithms to come up with methods that correlate better with human perception. In future, researchers may look at improved features or advanced machine learning algorithms in order to exploit fully the subjective ratings obtained in this database.

7.2.4 Aesthetic quality assessment of HDR images

Many of the HDR post processing artifacts are added in order to improve the aesthetic quality of the images, and similar post-processing methods may result in drastically different levels of aesthetic pleasure based on the scene content. The present class of algorithms does not take into account this aspect of the HDR images. Future research endeavors may look at incorporating content specific aesthetics for evaluating the quality of HDR images.

Index

Bibliography

[1] I. Cisco Systems. (2015) Cisco visual networking index: Global mobile data traffic forecast update, 20142019, white paper. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf

[2] L. K. Choi, Y. Liao, and A. C. Bovik, "Video QoE metrics for the compute continuum," *IEEE Commun. Soc. Multimed. Tech. Comm (MMTC) E-lett.*, vol. 8, no. 5, pp. 26–29, 2013.

[3] M. A. Saad, "Blind image and video quality assessment using natural scene and motion models," May 2013, https://repositories.lib.utexas.edu/bitstream/handle/2152/21955/SAAD-DISSERTATION-2013.pdf.

[4] Entertainment Software Association, "Essential facts about the computer and video game industry," 2014. [Online]. Available: http://www.theesa.com/wp-content/uploads/2014/10/ESA_EF_2014.pdf

[5] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 657–667, Feb 2013.

[6] D. Katzmaier. (2015) Amazon beats netflix to deliver hdr video. [Online]. Available: http://www.cnet.com/news/amazon-beats-netflix-to-deliver-hdr-video/

[7] 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, Nov 2006.

[8] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on*, Nov 2012, pp. 1693–1697.

[9] D. Kundu and B. L. Evans, "Full-reference visual quality assessment for synthetic images: A subjective study," in *Proc. International Conference on Image Processing*, September 2015, http://users.ece.utexas.edu/\%7Ebevans/papers/2015/imagequality/index.html.

[10] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, no. 0, pp. 57 – 77, 2015.

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

[12] F. De Simone, L. Goldmann, V. Baroncini, and T. Ebrahimi, "Subjective evaluation of JPEG XR image compression," in *Proc. SPIE*, vol. 7443, 2009.

[13] M. Čadík, R. Herzog, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts," *ACM Trans. Graphics*, vol. 31, no. 6, pp. 1–10, Nov. 2012.

[14] (2015) The NVidia GRID game-streaming service lets you play popular PC games on NVidia Shield devices from the cloud. [Online]. Available: http://shield.nvidia.com/grid-game-streaming

[15] R. Herzog, M. Čadík, T. O. Aydin, K. I. Kim, K. Myszkowski, and H.-P. Seidel, "NoRM: No-Reference image quality metric for realistic image synthesis." *Computer Graphics Forum*, vol. 31, no. 2, pp. 545–554, 2012.

[16] S. Lyu and H. Farid, "How realistic is photorealistic?" *IEEE Signal Process. Lett.*, vol. 53, no. 2, pp. 845–850, Feb 2005.

[17] D. Kundu and B. L. Evans, "ESPL Synthetic Image Database Release 2," January 2015, http://signal.ece.utexas.edu/\%7Ebevans/synthetic/.

[18] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct 2012.

[19] A. Unterweger, "Compression artifacts in modern video coding and state-of-the-art means of compensation," *Multimedia Networking and Coding*, p. 28, 2012.

[20] ITU-R BT.500-13 methodology for the subjective assessment of the quality of television pictures. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf

[21] M. Kleiner, D. Brainard, D. Pelli, C. Broussard, T. Wolf, and D. Niehorster, "The Psychology Toolbox," http://psychtoolbox.org/.

[22] D. Kundu and B. L. Evans, "Spatial domain synthetic scene statistics," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Nov 2014.

[23] ——, "No-reference synthetic image quality assessment using scene statistics," in *Proc. Asilomar Conf. Signals, Systems and Computers., accepted.*, November 2015, http://users.ece.utexas.edu/\%7Ebevans/papers/2015/imagequalitynoref/index.html.

[24] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. IEEE International Conference on Image Processing*, Sept 2012, pp. 1477–1480.

[25] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: A survey," *Computing Research Repository*, vol. abs/1406.7799, 2014. [Online]. Available: http://arxiv.org/abs/1406.7799

[26] W. S. Geisler, "Visual Perception and the Statistical Properties of Natural Scenes," *Annual Review Psych.*, vol. 59, no. 1, pp. 167–192, 2008.

[27] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, pp. 1193–1216, 2001.

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

[29] A. Shnayderman, A. Gusev, and A. M. Eskicioglu, "A multidimensional image quality measure using singular value decomposition," in *Proc. SPIE*, vol. 5294, 2004, pp. 82–92.

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

[31] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals, Systems and Computers*, vol. 2, Nov 2003, pp. 1398–1402 Vol.2.

[32] A. Kolaman and O. Yadid-Pecht, "Quaternion structural similarity: A new quality index for color images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1526–1536, April 2012.

[33] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, March 2002.

[34] S. Daly, "Digital images and human vision," A. B. Watson, Ed. MIT Press, 1993, ch. The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity, pp. 179–206.

[35] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," in *SIGGRAPH*. ACM, 2011, pp. 1–14.

[36] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, pp. 636–650, 2000.

[37] T. Mitsa and K. L. Varkur, "Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms," in *Proc. International Conference on Acoustics Speech and Signal Processing*, vol. 5, April 1993, pp. 301–304 vol.5.

[38] K. Egiazarian, J. Astola, V. Lukin, F. Battisti, and M. Carli, "New Full-Reference Quality Metrics based on HVS," *Proc. Int. Work. Video Process. and Quality Metrics*, 2006.

[39] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, "Modified image visual quality metrics for contrast change and mean shift accounting," in *Proc. Int. Conf. Experience of Designing and Appl. of CAD Systems in Microelectronics*, Feb 2011, pp. 305–311.

[40] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On Between-Coefficient Contrast Masking of DCT Basis Functions," *Proc. Int. Work. Video Process. and Quality Metrics*, 2007.

[41] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec 2005.

[42] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb 2006.

[43] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[44] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug 2011.

[45] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb 2014.

[46] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, April 2012.

[47] L. Zhang, D. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *Proc. International Conference on Image Processing*, Sept 2010, pp. 321–324.

[48] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct 2014.

[49] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. International Conference on Image Processing*, Sept 2012, pp. 1473–1476.

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

[51] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," pp. 149–159, 2005. [Online]. Available: http://dx.doi.org/10.1117/12.597306

[52] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, April 2009.

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

[54] W. Xue and X. Mou, "Reduced reference image quality assessment based on Weibull statistics," in *Proc. IEEE International Conference on Quality of Multimedia Experience*, June 2010, pp. 1–6.

[55] X. Mou, W. Xue, and L. Zhang, "Reduced reference image quality assessment via sub-image similarity based redundancy measurement," in *Proc. SPIE*, vol. 8291, 2012, pp. 82 911S–82 911S–7.

[56] M. Zhang, W. Xue, and X. Mou, "Reduced reference image quality assessment based on statistics of edge," in *Proc. SPIE*, vol. 7876, 2011, pp. 787 611–787 611–7.

[57] R. Hassen, Z. Wang, and M. Salam, "Image sharpness assessment based on local phase coherence," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2798–2810, July 2013.

[58] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *Proc. IEEE International Conference on Quality of Multimedia Experience*, July 2009, pp. 87–91.

[59] 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, April 2009.

[60] C. Vu, T. Phan, and D. M. Chandler, "S3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Trans. Image Process.*, vol. 21, no. 3, September 2011.

[61] P. Vu and D. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 423–426, July 2012.

[62] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. International Conference on Image Processing*, vol. 1, 2002, pp. I–477–I–480 vol.1.

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

[64] X. Liu, M. Tanaka, and M. Okutomi, "Single-image noise level estimation for blind denoising," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5226–5237, Dec 2013.

[65] J. Immerkr, "Fast noise variance estimation," *Computer Vision and Image Understanding*, vol. 64, no. 2, pp. 300 – 302, 1996.

[66] 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, Dec 2012.

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

[68] 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 Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov 2014.

[69] 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, Dec 2011.

[70] 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," *Signal Processing: Image Communication*, vol. 29, no. 7, pp. 725 – 747, 2014.

[71] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 7, no. 5, May 2010.

[72] M. A. 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.

[73] C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 793–799, May 2011.

[74] L. Liu, H. Dong, H. Huang, and A. C. Bovik, "No-reference image quality assessment in curvelet domain," *Signal Processing: Image Communication*, vol. 29, no. 4, April 2014.

[75] S. Gabarda and G. Cristbal, "Blind image quality assessment through anisotropy," *J. Opt. Soc. Am.*, vol. 24, no. 12, December 2007.

[76] Peng Ye, Jayant Kumar, Le Kang, and David Doermann, "Unsupervised Feature Learning Framework for No-reference Image Quality Assessment," in *Proc. CVPR*, June 2012, pp. 1098–1105.

[77] A. Mittal, G. S. Muralidhar, J. Ghosh, and A. C. Bovik, "Blind image quality assessment without human training using latent quality factors," *IEEE Signal Process. Lett.*, vol. 19, no. 2, pp. 75–78, Feb 2012.

[78] K. Seshadrinathan and A. Bovik, "Unifying analysis of full reference image quality assessment," in *Proc. International Conference on Image Processing*, Oct 2008, pp. 1200–1203.

[79] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," in *Proc. NIPS*, 1993, pp. 551–558.

[80] D. Kundu, "*Subjective and Objective Quality Evaluation of Synthetic and High Dynamic Range Images*," Ph.D. dissertation, Dept. of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712, May 2016, http://users.ece.utexas.edu/%7Ebevans/students/phd/debarati_kundu/.

[81] 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.

[82] "Final report from the video quality experts group on the validation of objective models of video quality assessment," ftp://vqeg.its.bldrdoc.gov/Documents/Meetings/Hillsboro_VQEG_Mar_03/VQEGIIDraftReportv2a.pdf, 2003.

[83] H. R. S. Zhou Wang, Alan C. Bovik and E. P. Simoncelli, "The ssim index for image quality assessment," Feb 2003, https://ece.uwaterloo.ca/~z70wang/research/ssim/.

[84] M. Gaubatz, "Metrix mux visual quality assessment package," http://foulard.ece.cornell.edu/gaubatz/metrix_mux/.

[85] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, 2011, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[86] D. Kundu and B. L. Evans, "Visual attention guided quality assessment of tone-mapped images using scene statistics," in *Proc. International Conference on Image*

Processing, September 2016, http://users.ece.utexas.edu/~bevans/papers/2016/imagequality/index.html.

[87] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st ed. New York, NY, USA: Springer-Verlag New York, Inc., 2010.

[88] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, Nov 2015.

[89] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 369–378. [Online]. Available: http://dx.doi.org/10.1145/258734.258884

[90] Mann, Picard, S. Mann, and R. W. Picard, "On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures," in *Proceedings of IS&T*, 1995, pp. 442–448.

[91] S. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, 2000, pp. 472–479 vol.1.

[92] J. de Vries, "Learn OpenGL," http://learnopengl.com/#!Advanced-Lighting/HDR.

[93] P. Ledda and et al., "A local model of eye adaptation for high dynamic range images," in *Proc. of ACM Afrigraph 04*. ACM Press, 2004, pp. 151–160.

[94] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 3, no. 4, pp. 291–306, Oct. 1997. [Online]. Available: http://dx.doi.org/10.1109/2945.646233

[95] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002. [Online]. Available: http://doi.acm.org/10.1145/566654.566575

[96] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," in *Proc. ACM SIGGRAPH*, 2002, pp. 257–266. [Online]. Available: http://doi.acm.org/10.1145/566570.566574

[97] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," in *ACM SIGGRAPH 2008 Papers*, ser. SIGGRAPH '08. New York, NY, USA: ACM, 2008, pp. 67:1–67:10. [Online]. Available: http://doi.acm.org/10.1145/1399504.1360666

[98] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[99] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 249–256, Jul. 2002. [Online]. Available: http://doi.acm.org/10.1145/566654.566573

[100] K. Ma, H. Yeganeh, K. Zeng, and Z. Wang, "High dynamic range image compression by optimizing tone mapped image quality index," *IEEE Trans. Image Process.*,

vol. 24, no. 10, pp. 3086–3097, Oct 2015.

[101] H. Ziaei Nafchi, A. Shahkolaei, R. Farrahi Moghaddam, and M. Cheriet, "FSITM: A feature similarity index for tone-mapped images," *IEEE Signal Process. Lett*, vol. 22, no. 8, pp. 1026–1029, Aug 2015.

[102] H. R. Nasrinpour and N. D. Bruce, "Saliency weighted quality assessment of tone-mapped images," in *Proc. IEEE Int. Conf. on Image Process.*, Sept 2015, pp. 4947–4951.

[103] N. D. Bruce and J. Tsotsos, "Attention based on information maximization," *Journal of Vision*, vol. 7, no. 950, June 2007.

[104] J. Petit, R. Brémond, and J.-P. Tarel, "Saliency maps of high dynamic range images," in *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, ser. APGV '09. New York, NY, USA: ACM, 2009, pp. 134–134.

[105] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998. [Online]. Available: http://dx.doi.org/10.1109/34.730558

[106] P. G. J. Baren, *Contrast sensitivity of the human eye an its effects on image quality.* Bellingham, Washington: SPIE Press, 1999.

[107] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[108] P. Reinagel and A. M. Zador, "Natural scene statistics at the center of gaze," *Network: Computation in Neural Systems*, vol. 10, pp. 1–10, 1999.

[109] M. Narwaria, M. Perreira Da Silva, P. Le Callet, and R. Ppion, "Tone mapping-based high-dynamic-range image compression: study of optimization criterion and perceptual quality," *Optical Engineering*, vol. 52, no. 10, pp. pp. 102 008–1 – 102 008–15, Oct 2013.

[110] D. Kundu and B. L. Evans, "Full-reference high dynamic range image quality assessment, software release 1.0," January 2016, http://signal.ece.utexas.edu/%7Ebevans/HDRImaging/.

[111] P. Hanhart, M. V. B. M. Pereira, A. M. G. Pinheiro, and T. Ebrahimi, "Benchmarking of objective quality metrics for hdr image quality assessment," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, pp. 1–18, 2015. [Online]. Available: http://dx.doi.org/10.1186/s13640-015-0091-4

[112] T. Richter, "On the standardization of the JPEG XT image compression," in *Picture Coding Symposium (PCS), 2013*, Dec 2013, pp. 37–40.

[113] M. Liu, G. Zhai, S. Tan, Z. Zhang, K. Gu, and X. Yang, "HDR2014 - a high dynamic range image quality database," in *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on*, July 2014, pp. 1–6.

[114] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, Sept 2011, pp. 3097–3100.

[115] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016. [Online]. Available: http://dx.doi.org/10.1109/TIP.2015.2500021

[116] P. Korshunov, H. Nemoto, A. Skodras, and T. Ebrahimi, "Crowdsourcing-based evaluation of privacy in hdr images," pp. 913 802–913 802–11, 2014. [Online]. Available: http://dx.doi.org/10.1117/12.2054541

[117] M. D. Fairchild, "The HDR Photographic Survey," *15th Color Imaging Conference*, pp. 233–238, 2007.

[118] J. Hu, O. Gallo, K. Pulli, and X. Sun, "Hdr deghosting: How to deal with saturation?" in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 1163–1170.

[119] F. Banterle, "Hdr toolbox for matlab," https://github.com/banterle/HDR_Toolbox.

[120] R. Mantiuk, "HDR image gallery," http://pfstools.sourceforge.net/hdr_gallery.html.

[121] S. Raman and S. Chaudhuri, "Bilateral Filter Based Compositing for Variable Exposure Photography," in *Eurographics - Short Papers*, P. Alliez and M. Magnor, Eds. The Eurographics Association, 2009.

[122] F. Pece and J. Kautz, "Bitmap movement detection: Hdr for dynamic scenes," *Journal of Virtual Reality and Broadcasting*, vol. 10, no. 2, 2013. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:0009-6-36506

[123] "Amazon mechanical turk," https://www.mturk.com.

[124] "Microworkers," https://microworkers.com/.

[125] "Crowdflower," https://crowdflower.com/.

[126] T. Schulze, S. Seedorf, D. Geiger, N. Kaufmann, and M. Schader, "Exploring task properties in crowdsourcing - an empirical study on mechanical turk." in *ECIS*, V. K. Tuunainen, M. Rossi, and J. Nandhakumar, Eds., 2011. [Online]. Available: http://dblp.uni-trier.de/db/conf/ecis/ecis2011.html#SchulzeSGKS11

[127] J. Bigun, "G.h.: Optimal orientation detection of linear symmetry," in *In: Proceedings of the IEEE First International Conference on Computer Vision, London, Great Britain*, 1987, pp. 433–438.

Vita

Debarati Kundu was born in Kolkata, India on October 12, 1988. She received the Bachelor of Engineering degree in Electronics and Telecommunications Engineering from Jadavpur University, India in 2010. She joined the Department of Electrical and Computer Engineering at the University of Texas at Austin in Fall 2010 and obtained the Master of Science degree in Spring 2012. She started the Ph.D. program at the University of Texas at Austin in Fall 2012 under the supervision of Dr. Brian L. Evans. She joined the Embedded Signal Processing Laboratory in Fall 2011.

Her research interests include image and video quality assessment, computer graphics, computer vision, machine learning and prototyping of real-time systems. She is also an Indian classical music and dance enthusiast. She spends her spare time by painting, reading, and maintaining her own blog.

Permanent address: 56/1, Purba Sinthee Bye Lane
Dumdum, Kolkata - 700030, Kolkata, India

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

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