The Dissertation Committee for Anish Mittal
certifies that this is the approved version of the following dissertation:

Natural Scene Statistics-based Blind Visual Quality Assessment in the Spatial Domain

Committee:

Alan C. Bovik, Supervisor

Kristen Grauman

Gustavo de Veciana

Joydeep Ghosh

Wilson Geisler
Natural Scene Statistics-based Blind Visual Quality Assessment in the Spatial Domain

by

Anish Mittal, B.Tech., 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 2013
For all coaches, mentors, supporters and friends
Acknowledgments

I want to acknowledge and thank all the people whom I have met from different walks of life and got helped in various different ways. First in the list comes my parents Daleep and Sudha who brought me to this world. They nurtured me and imbued good moral and ethics in my personality. Not only did they provide unquestioned support throughout my life, but also provided me with a good education and encouragement to aim for doctorate degree from abroad even at the expense of their own discomfort. I also want to thank my siblings Esha, Parul, Ankit and Piyush for boosting me up during my emotional troughs.

I wish to express deep gratitude towards my mentors and coaches at UT - Alan for being a great advisor and teaching how to understand human psyche; Alice for pumping up positive energy and zeal in me through dance every week; Anush, Gautam, Michele and Rajiv for guiding, mentoring and providing motivation to read; Charlie and Steevee for giving basketball instructions and teaching team play; Don for giving swimming instructions; Guneet for being a patient sport instructor and a friend who I could always look towards to; Kristen for being a role model of a good teacher; Gustavo for being an ideal professor and Melanie for how to work with full dedication, passion and love. They all helped me grow as a more mature, responsible and complete
individual. I salute all of them.

I acknowledge all LIVE members for their support as friends and guidance by pointing out resources and helping me out when ever I felt lost. I thoroughly enjoyed all the moments I spent with LIVEies and made new friends for ever - Alan B., Ajay, Anush, Gautam, Joonsoo, Kalpana, Michele, Ming, Rajiv, Sina, Tony, Yang, Lark, Dinesh and Deepti.

Lastly, I want to thank all my friends from yore and late, who have companied me in the life journey getting introduced at different points in life. I would spare myself from listing them here and making errors on missing out some of them.
Natural Scene Statistics-based Blind Visual Quality Assessment in the Spatial Domain

Publication No. _________________

Anish Mittal, PhD.
The University of Texas at Austin, 2013

Supervisor: Alan C. Bovik

With the launch of networked handheld devices which can capture, store, compress, send and display a variety of audiovisual stimuli; high definition television (HDTV); streaming Internet protocol TV (IPTV) and websites such as Youtube, Facebook and Flickr etc., an enormous amount of visual data of visual data is making its way to consumers. Because of this, considerable time and resources are being expanded to ensure that the end user is presented with with a satisfactory quality of experience (QoE). While traditional QoE methods have focused on optimizing delivery networks with respect to throughput, buffer-lengths and capacity, perceptually optimized delivery of multimedia services is also fast gaining importance. This is especially timely given the explosive growth in (especially wireless) video traffic and expected shortfalls in bandwidth. These perceptual approaches attempt to deliver an
optimized QoE to the end-user by utilizing objective measures of visual quality. In this thesis, we shall cover a variety of such algorithms that predict overall QoE of an image or a video, depending on the amount of information available for the algorithm design.

Typically, quality assessment (QA) algorithms are classified on the basis of the amount of information that is available to the algorithm. This thesis will primarily focus on blind QA algorithms, where blind or no-reference (NR) QA refers to automatic quality assessment of an image/video using an algorithm which only utilizes the distorted image/video whose quality is being assessed. NR QA approaches are further classified on the basis of whether the algorithm had access to subjective/human opinion prior to deployment. Algorithms which use machine learning techniques along with human judgements of quality during the ‘training’ phase may be labelled ‘opinion aware’ algorithms. The first part of the thesis deals with such approaches.

While such opinion aware-NR algorithms demonstrate good correlation with human perception on controlled databases, it is impossible to anticipate all of the different distortions that may occur in a practical system and hence train on them. In such cases, it is of interest to design QA algorithms that are not limited in their performance by training data. Approaches which operate without the knowledge of human judgements during the training phase are labelled as ‘opinion unaware’ (OU) algorithms. We propose such an approach in the second part of the thesis.

Further, we propose new VQA algorithms in the last part of the disser-
tation to address the completely blind VQA problem. The proposed approach quantifies disturbances introduced due to distortions and thereby predicts the quality of distorted content even without any external knowledge about the pristine natural sources and hence zero shot models.
# Table of Contents

Acknowledgments

Abstract

List of Tables

List of Figures

Chapter 1. Introduction

Chapter 2. Literature Review

2.1 Blind Image Quality Assessment

2.2 Blind Video Quality Assessment

Chapter 3. Opinion Aware Image Quality Assessment

3.1 Blind Spatial Image Quality Assessment

3.1.1 Natural Scene Statistics in the Spatial Domain

3.1.2 Quality Evaluation

3.2 Performance Evaluation

3.2.1 Correlation with Human Opinions

3.2.2 Variation with window size

3.2.3 Statistical Significance and Hypothesis Testing

3.2.4 Classification Accuracy

3.2.5 Two-stage Performance

3.2.6 Database Independence

3.2.7 Computational Complexity

3.3 Application to Blind Image Denoising

3.4 Discussion and Conclusion
# Chapter 4. Opinion Unaware Image Quality Assessment

4.1 No Reference Opinion-Unaware Distortion-Unaware IQA Model
   4.1.1 Spatial Domain NSS
   4.1.2 Patch Selection
   4.1.3 Characterizing Image Patches
   4.1.4 Multivariate Gaussian Model
   4.1.5 NIQE Index

4.2 Performance Evaluation
   4.2.1 Correlation with Human Judgments of Visual Quality
   4.2.2 Number of Natural Images
   4.2.3 Performance variation with patch size
   4.2.4 Performance variation with sharpness percentile

4.3 Prediction Error Analysis
   4.3.1 Quality based error analysis
   4.3.2 Distortion-type based error analysis

4.4 Discussion and Conclusion

# Chapter 5. Completely Blind Video Quality Assessment

5.1 Natural Video Statistic Model
   5.1.1 Subband Decomposition and Divisive Normalization
   5.1.2 Characterization of Video Patches
   5.1.3 Inter Subband Correlations
      5.1.3.1 Local
      5.1.3.2 Global
   5.1.4 Proposed Index

5.2 Computational Complexity
   5.2.0.1 Frame differencing
   5.2.0.2 Filtering
   5.2.0.3 Divisive normalization
   5.2.0.4 Model based features
   5.2.0.5 Temporal feature differences
   5.2.0.6 Local and global temporal inter subband correlations
List of Tables

3.1 A summary of features extracted in order to classify and quantify distortions. ........................................... 27
3.2 Median spearman rank ordered correlation coefficient (SROCC) across 1000 train-test combinations on the LIVE IQA database. *Italics* indicate no-reference algorithms. ......................... 31
3.3 Median linear correlation coefficient across 1000 train-test combinations on the LIVE IQA database. *Italics* indicate no-reference algorithms. .................................................. 32
3.4 Median spearman rank ordered correlation coefficient (SROCC) across 1000 train-test combinations on the LIVE IQA database for different window sizes. *Italics* indicate no-reference algorithms. .............................. 32
3.5 Results of one sided t-test performed between SROCC values of various IQA algorithms. A value of ‘1’ indicates that the row algorithm is statically superior to the column algorithm; ‘-1’ indicates that the row is worse than the column; a value of ‘0’ gives indicates that the two algorithms are statistically indistinguishable. *Italics* indicate no-reference algorithms. .... 34
3.6 Median classification accuracy across 1000 train-test trials. . 35
3.7 Median spearman rank ordered correlation coefficient (SROCC) across 1000 train-test combinations on the LIVE IQA database. *Italics* indicate no-reference algorithms. ............................. 36
3.8 Spearman’s rank ordered correlation coefficient (SROCC) on the TID2008 database. *Italicized* algorithms are NR IQA algorithms, others are FR IQA algorithms. ................................. 37
3.9 Informal complexity analysis of BRISQUE. Tabulated values reflect the percentage of time devoted to each of the steps in BRISQUE. ......................................................... 38
3.10 Complexity analysis of BRISQUE: A comparison of the amount of time taken to compute various quality measures for a 512 × 768 image on a 1.8 Ghz single-core PC with 2 GB of RAM. . 39
4.1 Median spearman rank ordered correlation coefficient (SROCC) across 1000 train-test combinations on the LIVE IQA database. *Italics* indicate (OA/OU)-DA no-reference algorithms and **bold face** indicates the new OU-DU model algorithm. .................... 53
4.2 Median linear correlation coefficient across 1000 train-test combinations on the LIVE IQA database. *Italics* indicate (OA/OU)-DA no-reference algorithms and **bold face** indicates the new OU-DU model algorithm. ........................................... 54

5.1 SROCC of different VQA algorithms against DMOS on LIVE VQA database .................................................. 79

5.2 LCC of different VQA algorithms against DMOS on LIVE VQA database ....................................................... 80

5.3 Median SROCC of different blind VQA algorithms against DMOS on every possible combination of train/test set splits. 80% of content used for training on LIVE VQA database .............. 81

5.4 Median LCC of different blind VQA algorithms against DMOS on every possible combination of train/test set splits. 80% of content used for training on LIVE VQA database .............. 81
List of Figures

3.1 Underlying Gaussianity of natural images. (a) and (b) show examples of natural and artificial images from the TID database [104] respectively. (c) shows that normalized luminance coefficients follow a nearly Gaussian distribution for the natural image (a). This property does not hold true for the empirical distribution of the artificial image (b). ................................................. 12

3.2 Scatter plot between neighboring values of: Original luminance coefficients (Left Column) and MSCN coefficients (Right Column). Rows: horizontal, vertical, main diagonal and secondary diagonal neighbors. Notice a high correlation between surrounding pixels with a diagonal kind of structure in the plots shown in the left column. The normalization procedure greatly reduces these dependencies as is apparent in the plots shown in the right column. ................................................................. 18

3.3 Effect of the normalization procedure. (a) Original Image $I$, (b) Local mean field $\mu$, (c) $I - \mu$, (d) Local variance field $\sigma$ and (e) MSCN coefficients $(I - \mu)/\sigma$ ........................................ 19

3.4 Histogram of MSCN coefficients for a natural undistorted image and its various distorted versions. Distortions from the LIVE IQA database [125] – JPEG2000 (jp2k) and JPEG compression (jpeg), Additive white Gaussian noise (WN), Gaussian blur (blur), and a Rayleigh fast-fading channel simulation (ff). 20

3.5 (a) 2-D scatter plot between shape and scale parameters obtained by fitting GGD to the empirical distributions of MSCN coefficients of pristine images of Berkeley image segmentation database [72] and simulated distorted images where similar kinds of distortions as present in the LIVE IQA database [125] - JPEG 2000, JPEG, White Noise, Gaussian Blur, and Fast fading channel errors were introduced in each image at varying degrees of severity. (b) 3-D scatter plot between shape, left scale and right scale obtained by fitting AGGD to horizontal paired products using same set of images as (a). ............................... 21

3.6 Various paired products computed in order to quantify neighboring statistical relationships. Pairwise products are computed along four orientations – horizontal, vertical, main-diagonal and secondary-diagonal at a distance of 1 pixel. ......................... 23
3.7 Histograms of paired-products of MSCN coefficients of a natural undistorted image and various distorted versions of it. (a) Horizontal, (b) Vertical, (c) Main-diagonal, (d) Secondary-diagonal. Distortions from the LIVE IQA database [125] – JPEG2000 (jp2k) and JPEG compression (jpeg), Additive white Gaussian noise (WN), Gaussian blur (blur), and a Rayleigh fast-fading channel simulation (FF).

3.8 Correlation of features with human judgments of quality (DMOS) for different distortions

3.9 Mean SROCC and standard error bars for various algorithms across the 1000 train-test trials on LIVE IQA database

3.10 Mean confusion matrix for classifier across 1000 trials: which row (distortion) is confused as which column (distortion). Higher number indicates greater confusion

3.11 Accurate noise variance as input to the algorithm in [173] produces poorer quality denoised images: (a) Noisy Image ($\sigma = 0.0158$, MS-SSIM = 0.9063 ), (b) Denoised with $\sigma = 0.0158$ (MS-SSIM = 0.9176 ) and (c) Denoised with $\sigma = 0.0040$ (MS-SSIM = 0.9480)

3.12 Shows the mean quality and associated errors at each noise level across 2000 test images for our approach as well as the reference implementation of BM3D

4.1 The marked blocks in the images (a) and (b) depict instances of natural image patches selected using a local sharpness measure

4.2 Variation of performance with the number of natural images $K$. Error smear around each point indicate the standard deviation in performance across 100 iterations for $5 < K < 125$

4.3 Performance variation with the patch size. The best overall performance is achieved with patches of size $96 \times 96$. However, the performance remains fairly stable with changes in patch size

4.4 Performance variation with percentile sharpness, threshold based on which patches from an naturalistic image are selected

4.5 Figure plots the probability distribution of errors in DMOS prediction for a subset of images in LIVE IQA database [125] with quality range (a) All qualities (b) Excellent (c) Good (d) Fair (e) Poor and (f) Bad. Positive prediction error signifies that the quality of the distorted image is predicted to be better than its actual human judgment whereas negative error signifies that the quality of the distorted image is predicted to be worse than its actual human judgment
4.6 Figure plots the probability distribution of errors in DMOS prediction for a subset of images in LIVE IQA database [125] with quality range (a) All distortions (b) JPEG2000 (c) JPEG (d) White Noise (e) Gaussian Blur and (f) Fast fading. Positive prediction error signifies that the quality of the distorted image is predicted to be better than its actual human judgment whereas negative error signifies that the quality of the distorted image is predicted to be worse than its actual human judgment.

5.1 Block diagram of our proposed approach to completely blind video quality assessment.

5.2 Flow diagram depicting multiscale video decomposition followed by divisive normalization. The example uses consecutive frames from the sunflower sequence in the LIVE VQA database [117].

5.3 Feature extraction followed by local temporal inter subband correlation computation. The same example frames were used as in Fig. 5.2.

5.4 Distribution of local temporal inter subband correlations on undistorted and distorted videos in the LIVE VQA database [117] for (a) $\alpha$ and (b) $\beta$ features.

5.5 Bar graphs depicting global temporal inter subband correlations on undistorted and distorted natural videos in the LIVE VQA database [117] for (a) $\alpha$ and (b) $\beta$ features.

5.6 Method of estimating global temporal inter subband correlations. The computations were done using frames from the sunflower sequence in the LIVE VQA database [117].
Chapter 1

Introduction

The growing acceptance and use of mobile wireless media devices has become a major focus of managing the wireless network infrastructure that was originally designed for voice and low-level data, owing to an explosion of high-bitrate content in the form of high bandwidth images and videos. This poses an interesting challenge to redefine the quality of service in terms of quality of experience (QoE) at the end user - a difficult task - since QoE not only depends on the time varying nature of the channel due to shadowing, dynamic interference, mobility, and changing loads but also on a variety of other factors, including the content, amount of compression and the dynamic nature of the videos. For instance, the QoE might be deemed as poor for a high-motion action scene, whilst that of a low-motion scene may be acceptable under the same channel conditions. Modeling each of these factors individually is not only challenging but also impractical. An alternate, highly practical solution is that of abstracting these variations away, by attempting to measure the QoE from the perspective of the end user. This is achieved using sophisticated models of human visual perception and the statistical properties of the videos being viewed. In this thesis, we shall cover a variety of such algorithms that predict overall QoE of an image or a video, depending on the amount of
information available for the algorithm design.

The performance of any IQA model is best gauged by its correlation with human subjective judgements of quality, since the human is the ultimate receiver of the visual signal. Such human opinions of visual quality are generally obtained by conducting large-scale human studies, referred to as subjective quality assessment, where human observers rate a large number of distorted (and possibly reference) signals. When the individual opinions are averaged across the subjects, a mean opinion score (MOS) or differential mean opinion score (DMOS) is obtained for each of the visual signals in the study, where the MOS/DMOS is representative of the perceptual quality of the visual signal. The goal of an objective quality assessment (QA) algorithm is to predict quality scores for these signals such that the scores produced by the algorithm correlate well with human opinions of signal quality (MOS/DMOS). Practical application of QA algorithms requires that these algorithms compute perceptual quality efficiently. The Spearmans rank ordered correlation coefficient (SROCC) and Pearsons (linear) correlation coefficient (LCC) between the predicted score from the algorithm and DMOS are generally used to access QA performance.

Typically, QoE algorithms are classified on the basis of the amount of information that is available to the algorithm. This thesis will focus on blind QA algorithms. Blind or No-reference (NR) QA refers to automatic quality assessment of an image/video using an algorithm which only utilizes the distorted image/video whose quality is being assessed. NR QA approaches are
further classified on the basis of whether the algorithm had access to subjective/human opinion prior to deployment. Algorithms which use machine learning techniques along with human judgements of quality during the ‘training’ phase may be labelled ‘opinion aware’ algorithms. The first part of the thesis deals with such approaches.

While such opinion aware-NR algorithms demonstrate good correlation with human perception on controlled databases, it is impossible to anticipate all of the different distortions that may occur in a practical system and hence train on them. In such cases, it is of interest to design QA algorithms that are not limited in their performance by training data. Approaches which operate without the knowledge of human judgements during the training phase are labelled as ‘opinion unaware’ (OU) algorithms. Likewise, among algorithms derived from OU models, distorted images may or may not be available during IQA model creation or training. For example, in highly unconstrained environments, such as a photograph upload site, the \textit{a priori} nature of distortions may be very difficult to know. Thus a model may be formulated as ‘distortion aware’ (DA) by training on (and hence tuning to) specific distortions, or it may be ‘distortion unaware’ (DU), relying instead only on exposure to naturalistic source images or image models to guide the QA process. While this may seem as an extreme paucity of information to guide design, it is worth observing that very successful FR IQA models (such as the structural similarity index (SSIM) [155]) are DU.

Our contribution, which we describe in the second part of the thesis,
is the development of a NSS-based modeling framework for OU-DU NR IQA
design, resulting in a first of a kind NSS-driven blind OU-DU IQA model
which does not require exposure to distorted images a priori, nor any training
on human opinion scores. The new NR OU-DU IQA quality index performs
better than the popular FR peak signal-to-noise-ratio (PSNR) and structural
similarity (SSIM) index and delivers performance at par with top performing
NR OA-DA IQA approaches.

Further, we note that even such models, are also limited in that they
can only capture common baseline characteristics of a specific collection of
non-distorted content and do not universally represent content specific intrin-
sic characteristics. Also, the construction of such a database requires the
unbiased selection and maintainence of hundreds of natural undistorted con-
tent. This also raises the question of how much exemplar content is needed
to design an accurate natural video model, and how distinctive these need to
be relative to each other and to the world of videos. Finally, given the limita-
tions of image/video camera capture, distortions are inevitably introduced in
the capture process and hence the procurement of perfectly natural ‘pristine’
content is practically impossible.

We propose new VQA algorithms to address such issues and bridge
the gap to address the completely blind VQA problem in the last part of the
thesis. They quantify disturbances introduced due to distortions and thereby
predict the quality of distorted content even without any external knowledge
about the pristine natural sources and hence zero shot models.
Before we delve deep into the technical details of the proposed algorithms, we undertake a brief literature survey of existing approaches in the field.
Chapter 2

Literature Review

In this chapter, we undertake a brief overview regarding the previous work in the topics to be described in the rest of this dissertation. This chapter only summarizes relevant literature in the NR-QA domain and is by no means a comprehensive survey.

2.1 Blind Image Quality Assessment

Most existing blind IQA models proposed in the past assume that the image whose quality is being assessed is afflicted by a particular kind of distortion [7, 26, 35, 90, 113, 123, 139, 147]. These approaches extract distortion-specific features that relate to loss of visual quality, such as edge-strength at block-boundaries. However, a few general purpose approaches for NR IQA have been proposed recently.

Li devised a set of heuristic measures to characterize visual quality in terms of edge sharpness, random noise and structural noise [66] while Gabarda and Cristobal, modeled anisotropies in images using Renyi entropy [38]. The authors in [170] use gabor filter based local appearance descriptors to form a visual codebook, and learn DMOS score vector, associating each word with
a quality score. However, in the process of visual codebook formation, each feature vector associated with an image patch is labeled by DMOS assigned to the entire image. This is questionable as each image patch can present a different level of quality depending on the distortion process the image is afflicted with. In particular, local distortions such as packet loss might afflict only a few image patches. Also, the approach is computationally expensive limiting its applicability in real time applications.

Tang et al. [141] proposed an approach which learns an ensemble of regressors trained on three different groups of features - natural image statistics, distortion texture statistics and blur/noise statistics. Another approach [126] is based on a hybrid of curvelet, wavelet and cosine transforms. Although these approaches work on a variety of distortions, each set of features (in the first approach) and transforms (in the second) caters only to certain kinds of distortion processes. This limits the applicability of their framework to new distortions.

We have also developed previous NR QA models in the past, following our philosophy, first fully developed in [122], that NSS models provide powerful tools for probing human judgements of visual distortions. Our work on NSS based FR QA algorithms [122–124], more recent RR models [131] and very recent work on NSS based NR QA [84, 85, 109] have led us to the conclusion that visual features derived from NSS lead to particularly potent and simple QA models [154].

Our recently proposed NSS based NR IQA model, dubbed the Distor-
tion Identification-based Image INtegrity and Verity Evaluation (DIIVINE) index, deploys summary statistics derived from an NSS wavelet coefficient model, using a two stage framework for QA: distortion-identification followed by distortion-specific QA [85]. The DIIVINE index performs quite well on the LIVE IQA database [125], achieving statistical parity with the full-reference structural similarity (SSIM) index [155].

A complementary approach developed at the same time, named BLind Image Notator using DCT Statistics (BLIINDS-II index) is a pragmatic approach to NR IQA that operates in the DCT domain, where a small number of features are computed from an NSS model of block DCT coefficients [109]. Efficient NSS features are calculated and fed to a regression function that delivers accurate QA predictions. BLIINDS-II is a single-stage algorithm that also delivers highly competitive QA prediction power. Although BLIINDS-II index is multiscale, the small number of feature types (4) allow for efficient computation of visual quality and hence the index is attractive for practical applications.

While both DIIVINE and BLIINDS-II deliver top NR IQA performance (to date), each of them has certain limitations. The large number of features that DIIVINE computes implies that it may be difficult to compute in real time. Although BLIINDS-II is more efficient than DIIVINE, it requires non-linear sorting of block based NSS features, which slows it considerably.

We would like the reader to note that all the approaches discussed above are opinion aware. To the date of writing this dissertation, we are not
aware of contributions from other authors in the field of OU NR IQA.

2.2 Blind Video Quality Assessment

The topic of NR VQA has been extensively studied and surveyed [31, 55, 75, 89, 99, 102, 111, 166, 168, 169]. Here we conduct a brief review of progress in the area. Almost all prior NR VQA models have been ‘distortion specific’, meaning they are designed to predict the effect of a specific type of distortion on perceived quality. For example, Tan and Ghanbari [140], Vlahchos [148], Suthaharan [138] and Muijs and Kirenko [88] proposed methods to assess blocking severity in distorted videos.

Methods for assessing multiple coincident distortion types have also been contemplated. Caviedes and Oberti compute a set of blocking, blurring, and sharpness [23], Babu et. al. calculate a measure of blocking and packet-loss [5] and Farias and Mitra [34] measures blockiness, bluriness and noisiness. Massidda et. al. propose an HVS based NR metric for blur detection in 2.5G/3G systems which measures blockiness, bluriness and moving artifacts [75]. Dosselmann and Yang estimate quality by measuring three types of impairments - noise, blocking and bit-error based color impairments [31]. Yang et. al. proposed an NR VQA algorithm that measures spatial distortion between the block under consideration and its motion compensated block in the previous frame, where temporal distortion is computed as a function of the mean of the motion vectors [168]. Kawayoke and Horita propose a model for NR VQA comprised of a frame quality measure and correction, asymmetric
tracking and mean value filtering [55].

Yang et. al. measure dropping severity as a function of the number of frames dropped using timestamp information from the video stream [169]. Lu proposed a method to measure blur caused by video compression [68], Pastrana-Vidal and Gicquel proposed an algorithm to measure frame-drops [102], Yamada et. al. proposed an algorithm to measure error-concealment effectiveness [166] and Naccari et. al. proposed a model for channel induced distortion for H.264/AVC coded videos in [89]. Keimel et. al. proposed an NR VQA algorithm specifically for compressed HD videos [56] whereas Ong et. al. proposed a measure to monitor the quality of streamed videos via by modeling the jerkiness between frames[99]. Saad and Bovik recently proposed a spatio-temporal natural scene statistics (NSS) model in the DCT domain that predicts the perceptual severity of MPEG-2, H.264, and two types of packet loss [111].

All of the blind VQA algorithms require knowledge of (one of possibly multiple) distortion types, or of the artifacts introduced by them, or of human opinion scores on distorted images. There exists no blind VQA model to date which can predict video quality in the absence of such strong a priori information.

Having surveyed the relevant literature in NR IQA and VQA domain, we now proceed towards the crux of this dissertation where we talk about our contributions in a series of next three chapters and relationships between each of them.
Opinion Aware Image Quality Assessment

An IQA model is called ‘opinion-aware’ (OA) if it has been trained on a database(s) of human rated distorted images and associated subjective opinion scores. Our approach to ‘opinion-aware’ NR IQA is based on the principle that natural images\(^1\) possess certain *regular* statistical properties that are measurably modified by the presence of distortions. Figure 3.1(a) and (b) shows examples of natural and artificial images from the TID database [104] respectively. The normalized luminance coefficients (explained later) of the natural image closely follow Gaussian-like distribution, as shown in Fig. 3.1(c) while the same does not hold for the empirical distribution of the artificial image shown in Fig. 3.1(d).

Deviations from the regularity of natural statistics, when quantified appropriately, enable the design of algorithms capable of assessing the perceptual quality of an image without the need for any reference image. By quantifying natural image statistics and refraining from an explicit characterization of dis-

\(^1\)‘Natural’ images are not necessarily images of natural environments such as trees or skies. Any natural light image that is captured by an optical camera and is not subjected to artificial processing on a computer is regarded as a natural image. Of course, image sensors may capture natural radiation other than visible light, but the images formed may obey different NSS than those considered here.
Figure 3.1: Underlying Gaussianity of natural images. (a) and (b) show examples of natural and artificial images from the TID database [104] respectively. (c) shows that normalized luminance coefficients follow a nearly Gaussian distribution for the natural image (a). This property does not hold true for the empirical distribution of the artificial image (b).
tortions, our approach to quality assessment is not limited by the type of distortions that afflict the image. Such approaches to NR IQA are significant since most current approaches are distortion-specific [26, 35, 90, 113, 123, 139, 147], i.e., they are capable of performing blind IQA only if the distortion that afflicts the image is known beforehand, e.g., blur or noise or compression and so on (see below). Previously, we have proposed other NSS-based distortion-generic approaches to NR IQA that statistically model images in the wavelet domain [85] and in the DCT-domain [109]. Our contribution here is a new NR IQA model that is purely spatial; that relies on a spatial NSS model which does not require a mapping to a different co-ordinate domain (wavelet, DCT, etc.) and so is ‘transform-free’; that demonstrates better ability to predict human judgments of quality than other popular FR and NR IQA models; that is highly efficient; and that is useful for perceptually optimizing image processing algorithms such as denoising.

While the presence of a reference image or information regarding the reference simplifies the problem of quality assessment, practical applications of such algorithms are limited in real-world scenarios where reference information is generally unavailable at nodes where quality computation is undertaken. Further, it can be argued that FR and to a large-extent RR approaches are not quality measures in the true sense, since these approaches measure fidelity relative to a reference image. Moreover, the assumption of a pristine nature of any reference is questionable, since all images are ostensibly distorted [13].

The performance of any IQA model is best gauged by its correlation
with human subjective judgements of quality, since the human is the ultimate receiver of the visual signal. Such human opinions of visual quality are generally obtained by conducting large-scale human studies, referred to as subjective quality assessment, where human observers rate a large number of distorted (and possibly reference) signals. When the individual opinions are averaged across the subjects, a mean opinion score (MOS) or differential mean opinion score (DMOS) is obtained for each of the visual signals in the study, where the MOS/DMOS is representative of the perceptual quality of the visual signal. The goal of an objective quality assessment (QA) algorithm is to predict quality scores for these signals such that the scores produced by the algorithm correlate well with human opinions of signal quality (MOS/DMOS). Practical application of QA algorithms requires that these algorithms compute perceptual quality efficiently.

The regularity of natural scene statistics (NSS) has been well established in the visual science literature, where regularity has been demonstrated in the spatial domain [108], and in the wavelet domain [134]. For example, it is well known that the power spectrum of natural images is a function of frequency and takes the form $1/f^\gamma$, where $\gamma$ is an exponent that varies over a small range across natural images.

The product of our research is the *Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)* which utilizes an NSS model framework of locally normalized luminance coefficients and quantifies ‘naturalness’ using the parameters of the model. BRISQUE introduces a new model of the statistics
of pair-wise products of neighboring (locally normalized) luminance values. The parameters of this model further quantify the naturalness of the image. Our claim is that characterizing locally normalized luminance coefficients in this way is sufficient not only to quantify naturalness, but also to quantify quality in the presence of distortion.

3.1 Blind Spatial Image Quality Assessment

Much recent work has focused on modeling the statistics of responses of natural images using multiscale transforms (e.g., Gabor filters, wavelets etc.) [134]. Given that neuronal responses in area V1 of visual cortex perform scale-space-orientation decompositions of visual data, transform domain models seem like natural approaches, particularly in view of the energy compaction (sparsity) and decorrelating properties of these transforms when combined with divisive normalization strategies [18, 154]. However, successful models of spatial luminance statistics have also received attention from vision researchers [108].

3.1.1 Natural Scene Statistics in the Spatial Domain

The spatial approach to NR IQA that we have developed can be summarized as follows. Given a (possibly distorted) image, first compute locally normalized luminances via local mean subtraction and divisive normalization [108]. Ruderman observed that applying a local non-linear operation to log-contrast luminances to remove local mean displacements from zero log-contrast
and to normalize the local variance of the log contrast has a decorrelating effect [108]. Such an operation may be applied to a given intensity image \( I(i, j) \) to produce:

\[
\hat{I}(i, j) = \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + C} + C
\]  

(3.1)

where, \( i \in 1, 2 \ldots M, \ j \in 1, 2 \ldots N \) are spatial indices, \( M, N \) are the image height and width respectively, \( C = 1 \) is a constant that prevents instabilities from occurring when the denominator tends to zero (eg., in the case of an image patch corresponding to the plain sky) and

\[
\mu(i, j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} I_{k,l}(i, j)
\]  

(3.2)

\[
\sigma(i, j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l}(I_{k,l}(i, j) - \mu(i, j))^2}
\]  

(3.3)

where \( w = \{w_{k,l}|k = -K, \ldots, K, l = -L, \ldots L\} \) is a 2D circularly-symmetric Gaussian weighting function sampled out to 3 standard deviations and rescaled to unit volume. In our implementation, \( K = L = 3 \). We show how performance varies with changes in the window size in the performance evaluation section.

Ruderman also observed that these normalized luminance values strongly tend towards a unit normal Gaussian characteristic [108] for natural images. Such an operation can be used to model the contrast-gain masking process in

16
early human vision [149], [18]. We utilize the pre-processing model (3.1) in our QA model development and refer to the transformed luminances $\hat{I}(i, j)$ as mean subtracted contrast normalized (MSCN) coefficients. As illustrated in the left column of Fig. 3.2, there is high correlation between surrounding pixels because image functions are generally piecewise smooth aside from sparse edge discontinuities. Hence we observe a diagonal kind of structure in the plots shown in the left column. The normalization procedure greatly reduces dependencies between neighboring coefficients as is apparent in the plots shown in the right column.

In order to help the reader visualize what the non-linear transformation $\hat{\circ} \circ$ does to an image, Figure 3.3 plots an image from the LIVE IQA database [125], its local mean field $\mu(i, j)$ and local variance field, $\sigma(i, j)$ and the MSCN field. The variance field highlights object boundaries and other local high contrast phenomenon. The MSCN field, while clearly not entirely decorrelated, exhibits a largely homogeneous appearance with a few low-energy residual object boundaries.

Our hypothesis is that the MSCN coefficients have characteristic statistical properties that are changed by the presence of distortion, and that quantifying these changes will make it possible to predict the type of distortion affecting an image as well as its perceptual quality. In order to visualize how the MSCN coefficient distributions vary as a function of distortion, Fig. 3.4 plots a histogram of MSCN coefficients for a natural undistorted image and for various distorted versions of it. Notice how the reference image exhibits
Figure 3.2: Scatter plot between neighboring values of: Original luminance coefficients (Left Column) and MSCN coefficients (Right Column). Rows: horizontal, vertical, main diagonal and secondary diagonal neighbors. Notice a high correlation between surrounding pixels with a diagonal kind of structure in the plots shown in the left column. The normalization procedure greatly reduces these dependencies as is apparent in the plots shown in the right column.
Figure 3.3: Effect of the normalization procedure. (a) Original Image $I$, (b) Local mean field $\mu$, (c) $I - \mu$, (d) Local variance field $\sigma$ and (e) MSCN coefficients $((I - \mu)/\sigma)$
a Gaussian-like appearance, as observed by Ruderman [108], while each distortion modifies the statistics in its own characteristic way. For example, blur creates a more Laplacian appearance, while white-noise distortion appears to reduce the weight of the tail of the histogram. We have found that a generalized Gaussian distribution (GGD) can be used to effectively capture a broader spectrum of distorted image statistics, which often exhibit changes in the tail behaviour (i.e. kurtosis) of the empirical coefficient distributions [119] where the GGD with zero mean is given by:

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

(3.4)
Figure 3.5: (a) 2-D scatter plot between shape and scale parameters obtained by fitting GGD to the empirical distributions of MSCN coefficients of pristine images of Berkeley image segmentation database [72] and simulated distorted images where similar kinds of distortions as present in the LIVE IQA database [125]- JPEG 2000, JPEG, White Noise, Gaussian Blur, and Fast fading channel errors were introduced in each image at varying degrees of severity. (b) 3-D scatter plot between shape, left scale and right scale obtained by fitting AGGD to horizontal paired products using same set of images as (a).

where

\[
\beta = \sigma \sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(3/\alpha)}} \tag{3.5}
\]

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

\[
\Gamma(a) = \int_0^{\infty} t^{a-1} e^{-t} dt \quad a > 0 \tag{3.6}
\]

The shape parameter \( \alpha \) controls the ‘shape’ of the distribution while \( \sigma^2 \) control the variance. We choose the zero mean distribution, since (generally) MSCN coefficient distributions are symmetric. The parameters of the GGD
\((\alpha, \sigma^2)\), are estimated using the moment-matching based approach proposed in [119].

We deploy this parametric model to fit the MSCN empirical distributions from distorted images as well as undistorted ones. For each image, we estimate 2 parameters \((\alpha, \sigma^2)\) from a GGD fit of the MSCN coefficients. These form the first set of features that will be used to capture image distortion. To show that pristine and distorted images are well separated in GGD parameter space, we took a set of pristine images from the Berkeley image segmentation database [72]. Similar kinds of distortions as present in the LIVE IQA database [125] - JPEG 2000, JPEG, white noise, Gaussian blur, and fast fading channel errors were introduced in each image at varying degrees of severity to form the distorted image set. As shown in Fig. 3.5(a), pristine and distorted images occupy different regions in this parameter space. White noise is very clearly separated from the pristine image set making it one of the easiest to gauge the quality of. JPEG2000 and fast fading have a high degree of overlap as fast fading images in LIVE database are actually multidistorted, first compressed into a bitstream using a JPEG2000 codec, then passed through a Rayleigh fast fading channel to simulate packet loss [125].

We also model the statistical relationships between neighboring pixels. While MSCN coefficients are definitely more homogenous for pristine images, the signs of adjacent coefficients also exhibit a regular structure, which gets disturbed in the presence of distortion. We model this structure using the empirical distributions of pairwise products of neighboring MSCN coefficients.
Figure 3.6: Various paired products computed in order to quantify neighboring statistical relationships. Pairwise products are computed along four orientations – horizontal, vertical, main-diagonal and secondary-diagonal at a distance of 1 pixel.

along four orientations – horizontal ($H$), vertical ($V$), main-diagonal ($D1$) and secondary-diagonal ($D2$), as illustrated in Fig. 3.6. Specifically,

$$H(i, j) = \hat{I}(i, j)\hat{I}(i, j + 1)$$  \hspace{1cm} (3.7)

$$V(i, j) = \hat{I}(i, j)\hat{I}(i + 1, j)$$ \hspace{1cm} (3.8)

$$D1(i, j) = \hat{I}(i, j)\hat{I}(i + 1, j + 1)$$ \hspace{1cm} (3.9)

$$D2(i, j) = \hat{I}(i, j)\hat{I}(i + 1, j - 1)$$ \hspace{1cm} (3.10)

for $i \in \{1, 2 \ldots M\}$ and $j \in \{1, 2 \ldots N\}$.

Under the Gaussian coefficient model, and assuming the MSCN coefficients are zero mean and unit variance, these products obey the following distribution in the absence of distortion [95]:

$$f(x, \rho) = \frac{\exp\left(\frac{|x|\rho}{1-\rho^2}\right)}{\pi \sqrt{1-\rho^2}} K_0 \left(\frac{|x|}{1-\rho^2}\right)$$ \hspace{1cm} (3.11)
Figure 3.7: Histograms of paired-products of MSCN coefficients of a natural undistorted image and various distorted versions of it. (a) Horizontal, (b) Vertical, (c) Main-diagonal, (d) Secondary-diagonal. Distortions from the LIVE IQA database [125] – JPEG2000 (jp2k) and JPEG compression (jpeg), Additive white Gaussian noise (WN), Gaussian blur (blur), and a Rayleigh fast-fading channel simulation (FF).
where \( f \) is an asymmetric probability density function, \( \rho \) denotes the correlation coefficient of adjacent coefficients, and \( K_0 \) is the modified bessel function of the second kind. While we have found that this density function is a good model of the empirical histograms of products of adjacent normalized coefficients, it has only a single parameter, and as such, does not provide a good fit to the empirical histograms of coefficient products (Fig. 3.2) from *distorted* images. Further, it is not finite at the origin. Hence, as a practical alternative, we adopt the very general asymmetric generalized Gaussian distribution (AGGD) model [59]. In order to visualize how paired products vary in the presence of distortion, in Fig. 3.7, we plot histograms of paired products along each of four orientations, for a reference image and for distorted versions of it.

The AGGD with zero mode is given by:

\[
f(x; \nu, \sigma_l^2, \sigma_r^2) = \begin{cases} 
\frac{\nu}{(\beta_l + \beta_r)\Gamma(\frac{1}{\nu})} \exp \left(-\left(\frac{x}{\sqrt{\beta_l}}\right)^\nu\right) & x < 0 \\
\frac{\nu}{(\beta_l + \beta_r)\Gamma(\frac{1}{\nu})} \exp \left(-\left(\frac{x}{\sqrt{\beta_r}}\right)^\nu\right) & x \geq 0 
\end{cases}
\]  

(3.12)

where

\[
\beta_l = \sigma_l \sqrt{\frac{\Gamma\left(\frac{1}{\nu}\right)}{\Gamma\left(\frac{3}{2}\right)}}
\]  

(3.13)

\[
\beta_r = \sigma_r \sqrt{\frac{\Gamma\left(\frac{1}{\nu}\right)}{\Gamma\left(\frac{3}{2}\right)}}
\]  

(3.14)

The shape parameter \( \nu \) controls the ‘shape’ of the distribution while \( \sigma_l^2 \) and \( \sigma_r^2 \) are scale parameters that control the spread on each side of the mode, respectively. The AGGD further generalizes the generalized Gaussian distribution (GGD) [119] and subsumes it by allowing for asymmetry in the distribution. The skew of the distribution is a function of the left and right scale parameters.
parameters. If $\sigma^2_l = \sigma^2_r$, then the AGGD reduces to the GGD. Although the AGGD is infrequently used, it has been deployed to model skewed heavy-tailed distributions of image texture [59]. The parameters of the AGGD $(\nu, \sigma^2_l, \sigma^2_r)$, are estimated using the moment-matching based approach proposed in [59]. Figure 3.5(b) shows the 3-D scatter plot between $(\nu, \sigma^2_l, \sigma^2_r)$ for horizontal paired products using the same set of images as used for showing separation in GGD parameter space. It can be visualized that different distortions occupy different parts of the space. Also, we expect images to have a better separation when modeled in the high dimensional space of parameters obtained by fitting AGGD distributions to paired products from different orientations and scales together. This figure also motives the use of (3.12) to better capture the finite empirical density function.

The parameters $(\eta, \nu, \sigma^2_l, \sigma^2_r)$ of the best AGGD fit are extracted where $\eta$ is given by:

$$
\eta = (\beta_r - \beta_l) \frac{\Gamma \left( \frac{2}{\nu} \right)}{\Gamma \left( \frac{1}{\nu} \right)}
$$  

(3.15)

Thus for each paired product, 16 parameters (4 parameters/orientation $\times$ 4 orientations) are computed, yielding the next set of features. Table 3.1 summarizes the features utilized.

Images are naturally multiscale, and distortions affect image structure across scales. Further, as research in quality assessment has demonstrated, incorporating multiscale information when assessing quality produces QA al-
<table>
<thead>
<tr>
<th>Feature ID</th>
<th>Feature Description</th>
<th>Computation Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1 - f_2$</td>
<td>Shape and variance</td>
<td>Fit GGD [119] to MSCN coefficients</td>
</tr>
<tr>
<td>$f_3 - f_6$</td>
<td>Shape, mean, left variance, right variance</td>
<td>Fit AGGD [59] to $H$ pairwise products</td>
</tr>
<tr>
<td>$f_7 - f_{10}$</td>
<td>Shape, mean, left variance, right variance</td>
<td>Fit AGGD [59] to $V$ pairwise products</td>
</tr>
<tr>
<td>$f_{11} - f_{14}$</td>
<td>Shape, mean, left variance, right variance</td>
<td>Fit AGGD [59] to $D1$ pairwise products</td>
</tr>
<tr>
<td>$f_{15} - f_{18}$</td>
<td>Shape, mean, left variance, right variance</td>
<td>Fit AGGD [59] to $D2$ pairwise products</td>
</tr>
</tbody>
</table>

algorithms that perform better in terms of correlation with human perception [109, 159]. Hence, we extract all features listed in Table 3.1 at two scales - the original image scale, and at a reduced resolution (low pass filtered and downsampled by a factor of 2). We observed that increasing the number of scales beyond 2 did not contribute to performance much. Thus, a total of 36 features – 18 at each scale, are used to identify distortions and to perform distortion-specific quality assessment. In Fig. 3.8, we plot the Spearman’s rank ordered correlation coefficient (SROCC) between each of these features and human DMOS from the LIVE IQA database, for each of the distortions in the database – JPEG and JPEG2000 compression, additive white Gaussian noise, Gaussian blur and a Rayleigh fast fading channel distortion, to ascertain how well the features correlate with human judgments of quality. Note that no training is undertaken here, the plot is simply to illustrate that each feature captures quality information and to show that images are affected differently by different distortions.
Figure 3.8: Correlation of features with human judgments of quality (DMOS) for different distortions
3.1.2 Quality Evaluation

A mapping is learned from feature space to quality scores using a regression module, yielding a measure of image quality. The framework is generic enough to allow for the use of any regressor. In our implementation, a support vector machine (SVM) regressor (SVR) [114] is used. SVR has previously been applied to image quality assessment problems [85, 91, 92]. For example, a learning driven feature pooling approach using SVR was proposed in [92]. Wavelet-domain NSS and singular value decomposition features have been used to map quality to human ratings via SVR in [85] and [91] respectively. SVR is generally noted for being able to handle high dimensional data [16]. We utilize the LIBSVM package [25] to implement the SVR with a radial basis function (RBF) kernel.

3.2 Performance Evaluation

3.2.1 Correlation with Human Opinions

We used the LIVE IQA database [125] to test the performance of BRISQUE, which consists of 29 reference images with 779 distorted images spanning five different distortion categories – JPEG2000 (JP2K) and JPEG compression, additive white Gaussian noise (WN), Gaussian blur (Blur), and a Rayleigh fast-fading channel simulation (FF). Each of the distorted images has an associated difference mean opinion score (DMOS) which represents the subjective quality of the image.

Since the BRISQUE approach requires a training procedure to cali-
brate the regressor module, we divide the LIVE database into two randomly chosen subsets – 80% training and 20% testing – such that no overlap between train and test content occurs. We do this to ensure that the reported results do not depend on features extracted from known spatial content, which can artificially improve performance. Further, we repeat this random train-test procedure 1000 times and report the median of the performance across these 1000 iterations, in order to eliminate performance bias.

The Spearman’s rank ordered correlation coefficient (SROCC) and Pearson’s (linear) correlation coefficient (LCC) between the predicted score from the algorithm and DMOS were used to access QA performance. Before computing LCC, the algorithm scores were passed through a logistic non-linearity as described in [125]. A value close to 1 for SROCC and LCC indicate good performance in terms of correlation with human opinion. These performance indices are tabulated in Tables 3.2 and 3.3 respectively.

We also tabulated the performance of three full-reference indices: peak-signal-to-noise ratio (PSNR), structural similarity index (SSIM) [155] and multi-scale structural similarity index (MS-SSIM) [159]. Although PSNR is a poor measure of perceptual quality, it is often used to benchmark for QA algorithms [42, 153]. The SSIM and MS-SSIM indices are popular owing to their performance and simplicity. We also include the performance of the previously

---

Further, note that due to randomness of the 1000 trials, there may be a slight discrepancy between results reported here and elsewhere, however, these differences in correlations are not statistically significant, and are simply an artifact of the random train-test sampling.
Table 3.2: Median spearman rank ordered correlation coefficient (SROCC) across 1000 train-test combinations on the LIVE IQA database. *Italics* indicate no-reference algorithms.

<table>
<thead>
<tr>
<th></th>
<th>JP2K</th>
<th>JPEG</th>
<th>WN</th>
<th>Blur</th>
<th>FF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.8646</td>
<td>0.8831</td>
<td>0.9410</td>
<td>0.7515</td>
<td>0.8736</td>
<td>0.8636</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.9389</td>
<td>0.9466</td>
<td>0.9635</td>
<td>0.9046</td>
<td>0.9393</td>
<td>0.9129</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.9627</td>
<td>0.9785</td>
<td>0.9773</td>
<td>0.9542</td>
<td>0.9386</td>
<td>0.9535</td>
</tr>
<tr>
<td>CBIQ</td>
<td>0.8935</td>
<td>0.9418</td>
<td>0.9582</td>
<td>0.9324</td>
<td>0.8727</td>
<td>0.8954</td>
</tr>
<tr>
<td>LBIQ</td>
<td>0.9040</td>
<td>0.9291</td>
<td>0.9702</td>
<td>0.8983</td>
<td>0.8222</td>
<td>0.9063</td>
</tr>
<tr>
<td>BLIINDS-II</td>
<td>0.9323</td>
<td>0.9331</td>
<td>0.9463</td>
<td>0.8912</td>
<td>0.8519</td>
<td>0.9124</td>
</tr>
<tr>
<td>DIIVINE</td>
<td>0.9123</td>
<td>0.9208</td>
<td>0.9818</td>
<td>0.9373</td>
<td>0.8694</td>
<td>0.9250</td>
</tr>
<tr>
<td>Pointwise</td>
<td>0.7957</td>
<td>0.8593</td>
<td>0.9608</td>
<td>0.8759</td>
<td>0.7773</td>
<td>0.8297</td>
</tr>
<tr>
<td>Pairwise</td>
<td>0.9007</td>
<td>0.9510</td>
<td>0.9773</td>
<td>0.9502</td>
<td>0.8741</td>
<td>0.9302</td>
</tr>
<tr>
<td>BRISQUE</td>
<td>0.9139</td>
<td>0.9647</td>
<td>0.9786</td>
<td>0.9511</td>
<td>0.8768</td>
<td>0.9395</td>
</tr>
</tbody>
</table>

summarized general purpose no-reference algorithms - CBIQ [170], LBIQ [141], BLIINDS-II [109] and DIIVINE index [85]. We requested quality scores from authors for CBIQ [170] and LBIQ [141]. We also reported the correlations obtained by modeling empirical distributions of MSCN coefficients (pointwise) alone and pairwise products alone to compare their relative importance.

### 3.2.2 Variation with window size

As observed from the Table 3.4, the performance of BRISQUE remains relatively stable with respect to variation in the window size used to compute the local mean and variances. However, the performance starts to decrease when it becomes fairly large as the computations become non-local.
Table 3.3: Median linear correlation coefficient across 1000 train-test combinations on the LIVE IQA database. Italics indicate no-reference algorithms.

<table>
<thead>
<tr>
<th></th>
<th>JP2K</th>
<th>JPEG</th>
<th>WN</th>
<th>Blur</th>
<th>FF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.8762</td>
<td>0.9029</td>
<td>0.9173</td>
<td>0.7801</td>
<td>0.8795</td>
<td>0.8592</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.9405</td>
<td>0.9462</td>
<td>0.9824</td>
<td>0.9004</td>
<td>0.9514</td>
<td>0.9066</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.9746</td>
<td>0.9793</td>
<td>0.9883</td>
<td>0.9645</td>
<td>0.9488</td>
<td>0.9511</td>
</tr>
<tr>
<td>CBIQ</td>
<td>0.8898</td>
<td>0.9434</td>
<td>0.9533</td>
<td>0.9338</td>
<td>0.8951</td>
<td>0.8955</td>
</tr>
<tr>
<td>LBIQ</td>
<td>0.9103</td>
<td>0.9345</td>
<td>0.9761</td>
<td>0.9104</td>
<td>0.8382</td>
<td>0.9087</td>
</tr>
<tr>
<td>BLIINDS-II</td>
<td>0.9386</td>
<td>0.9426</td>
<td>0.9635</td>
<td>0.8994</td>
<td>0.8790</td>
<td>0.9164</td>
</tr>
<tr>
<td>DIVINE</td>
<td>0.9233</td>
<td>0.9347</td>
<td>0.9867</td>
<td>0.9370</td>
<td>0.8916</td>
<td>0.9270</td>
</tr>
<tr>
<td>Pointwise</td>
<td>0.7947</td>
<td>0.8447</td>
<td>0.9711</td>
<td>0.9087</td>
<td>0.8151</td>
<td>0.8258</td>
</tr>
<tr>
<td>Pairwise</td>
<td>0.8968</td>
<td>0.9511</td>
<td>0.9830</td>
<td>0.9438</td>
<td>0.8952</td>
<td>0.9309</td>
</tr>
<tr>
<td>BRISQUE</td>
<td>0.9229</td>
<td>0.9734</td>
<td>0.9851</td>
<td>0.9506</td>
<td>0.9030</td>
<td>0.9424</td>
</tr>
</tbody>
</table>

Table 3.4: Median spearman rank ordered correlation coefficient (SROCC) across 1000 train-test combinations on the LIVE IQA database for different window sizes. Italics indicate no-reference algorithms.

<table>
<thead>
<tr>
<th>K,L</th>
<th>JPEG2000</th>
<th>JPEG</th>
<th>White noise</th>
<th>Gaussian Blur</th>
<th>Fast fading</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.9120</td>
<td>0.9581</td>
<td>0.9764</td>
<td>0.9535</td>
<td>0.8839</td>
<td>0.9388</td>
</tr>
<tr>
<td>5</td>
<td>0.9083</td>
<td>0.9510</td>
<td>0.9742</td>
<td>0.9497</td>
<td>0.8790</td>
<td>0.9360</td>
</tr>
<tr>
<td>6</td>
<td>0.9043</td>
<td>0.9483</td>
<td>0.9706</td>
<td>0.9417</td>
<td>0.8725</td>
<td>0.9309</td>
</tr>
<tr>
<td>7</td>
<td>0.9040</td>
<td>0.9482</td>
<td>0.9700</td>
<td>0.9407</td>
<td>0.8720</td>
<td>0.9305</td>
</tr>
<tr>
<td>8</td>
<td>0.8950</td>
<td>0.9405</td>
<td>0.9631</td>
<td>0.9321</td>
<td>0.8633</td>
<td>0.9208</td>
</tr>
</tbody>
</table>
Figure 3.9: Mean SROCC and standard error bars for various algorithms across the 1000 train-test trials on LIVE IQA database.

3.2.3 Statistical Significance and Hypothesis Testing

Figure 3.9 plots the mean SROCC across the 1000 trials and the standard deviations of performance across these 1000 trials for each of the algorithms considered here.

Although there exist differences in the median correlations between the different algorithms (see Table 3.2), these differences may not be statistically relevant. Hence, to evaluate the statistical significance of performance of each of the algorithms considered, we performed hypothesis testing based on the $t$-test \cite{127} on the SROCC values obtained from the 1000 train-test trials, and we tabulated the results in Table 3.5. The null hypothesis is that the mean correlation for the (row) algorithm is equal to mean correlation for the (column) algorithm with a confidence of 95\%. The alternate hypothesis is that the mean correlation of row is greater than or lesser than the mean correlation.
Table 3.5: Results of one sided t-test performed between SROCC values of various IQA algorithms. A value of ‘1’ indicates that the row algorithm is statically superior to the column algorithm; ‘-1’ indicates that the row is worse than the column; a value of ‘0’ gives indicates that the two algorithms are statistically indistinguishable. *italics* indicate no-reference algorithms.

<table>
<thead>
<tr>
<th></th>
<th>PSNR</th>
<th>SSIM</th>
<th>MSSSIM</th>
<th>CBIQ</th>
<th>LBIQ</th>
<th>BLINDS-II</th>
<th>DIVINE</th>
<th>BRISQUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>SSIM</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>MSSSIM</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CBIQ</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>LBIQ</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>BLINDS-II</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>DIVINE</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>BRISQUE</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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

From Table 3.5 we conclude that BRISQUE is highly competitive with all no reference algorithms tested and statistically better than the full reference algorithms PSNR and SSIM. Given that these measures require additional information in the form of the reference image, this is by no means a small achievement. This result suggests that to the extent distortions can be trained on, one can replace full reference algorithms such as SSIM with the proposed BRISQUE without any loss of performance. We note that BRISQUE remains slightly inferior to the FR MS-SSIM, indicating that there may still be some room for improvement in performance.
3.2.4 Classification Accuracy

In order to demonstrate that BRISQUE features can also be used for explicit distortion-identification [83], we report the median classification accuracy of the classifier for each of the distortions in the LIVE database, as well as across all distortions in Table 3.6.

Further, in order to visualize which distortions are ‘confused’ the most, Fig. 3.10 plots the confusion matrix for each of the distortions, where the sum of each row in the confusion matrix is 1 and actual values represent the mean confusion percentage across the 1000 train-test trials. We see from Fig. 3.10 that FF and JP2K are most confused with each other which is not surprising, since FF distortion is a combination of JP2K followed by packet-loss errors. JP2K and JPEG are also confused sometimes. WN and Blur are generally not confused with other distortions.

3.2.5 Two-stage Performance

We also investigated the possibility of replacing the one stage framework, where features are directly mapped to quality, with a two-stage framework, similar to that proposed in [83]. In this approach, the same set of features are used to identify the distortion afflicting the image as are then used for distortion-specific QA. Such a two-stage approach was used with recent
success for NSS-based blind IQA [85]. In Table 3.7, we tabulate the median SROCC value across 1000 trials for the two-stage realization of BRISQUE. We also list the performances of BRISQUE for comparison purposes. The slight dip in the performance can be attributed to imperfect distortion identification in the first stage of the two-stage framework.

### 3.2.6 Database Independence

Having evaluated BRISQUE on the LIVE IQA database, we now demonstrate that the performance of BRISQUE is not bound by the database on
Table 3.8: Spearman’s rank ordered correlation coefficient (SROCC) on the TID2008 database. *Italicized* algorithms are NR IQA algorithms, others are FR IQA algorithms.

<table>
<thead>
<tr>
<th></th>
<th>JP2K</th>
<th>JPEG</th>
<th>WN</th>
<th>Gblur</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.825</td>
<td>0.876</td>
<td>0.918</td>
<td>0.934</td>
<td>0.870</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.963</td>
<td>0.935</td>
<td>0.817</td>
<td>0.960</td>
<td>0.902</td>
</tr>
<tr>
<td>BRISQUE</td>
<td>0.832</td>
<td>0.924</td>
<td>0.829</td>
<td>0.881</td>
<td>0.896</td>
</tr>
</tbody>
</table>

which it is tested. To show this, we trained BRISQUE on the entire LIVE IQA database and then applied BRISQUE to the TID2008 database [104].

The TID database consists of 25 reference images and 1700 distorted images over 17 distortion categories [104]. Since there are only 24 natural images, and our algorithm is based on the statistics of natural images, we test our approach only on these 24 images. Further, although there exist 17 distortion categories, we tested BRISQUE only on these distortions that it is trained for: JPEG, JPEG2000 compression (JP2K), additive white noise (WN) and Gaussian Blur (blur) – FF distortion does not exist in the TID database.

The results of applying BRISQUE on TID are tabulated in Table 3.8, where we also list the performance of PSNR and SSIM for comparison purposes. It should be clear that BRISQUE performs well in terms of correlation with human perception of quality and that the performance does not depend on the database.
Table 3.9: Informal complexity analysis of BRISQUE. Tabulated values reflect the percentage of time devoted to each of the steps in BRISQUE.

<table>
<thead>
<tr>
<th>Step</th>
<th>Percentage of Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSCN</td>
<td>50.9</td>
</tr>
<tr>
<td>GGD</td>
<td>8.6</td>
</tr>
<tr>
<td>Pairwise Products and AGGD</td>
<td>40.6</td>
</tr>
</tbody>
</table>

3.2.7 Computational Complexity

Our description of BRISQUE focused on the relationship of the statistical features to natural scene statistics and the effect that distortions have on such statistics. However, given the small number of features that are extracted (18 per scale) and the fact that parameter estimation needs to be performed only 5 times for an entire image, in comparison to parameter estimation for each block as in BLIINDS-II [109], the reader will appreciate the fact that BRISQUE is extremely efficient. Having demonstrated that BRISQUE performs well in terms of correlation with human perception, we also now show that BRISQUE has low complexity. In Table 3.9 we list the relative percentage of time each of the stages of BRISQUE uses as a percentage of the time taken to compute the quality of an image (once trained).

We also compare the overall computational complexity of BRISQUE with the FR PSNR and the NR BLIINDS-II and DIIVINE, and in Table 3.10, we list the time taken (in seconds) to compute each quality measure on an image of resolution 512 × 768 on a 1.8 Ghz single-core PC with 2 GB of RAM. We use unoptimized MATLAB code for all of these algorithms in order to ensure a fair comparison. We also list the efficiency as a fraction of the time taken to
Table 3.10: Complexity analysis of BRISQUE: A comparison of the amount of time taken to compute various quality measures for a 512 $\times$ 768 image on a 1.8 Ghz single-core PC with 2 GB of RAM.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.05</td>
</tr>
<tr>
<td>DIIVINE</td>
<td>149</td>
</tr>
<tr>
<td>BLIINDS-II</td>
<td>70</td>
</tr>
<tr>
<td>BRISQUE</td>
<td>1</td>
</tr>
</tbody>
</table>

compute PSNR, to allow for a machine-independent comparison across algorithms. As Table 3.10 demonstrates, BRISQUE is quite efficient, outperforming the DIIVINE index and the BLIINDS-II index by a large amount. This suggests that the spatial-domain BRISQUE an ideal candidate for real-time blind assessment of visual quality.

3.3 Application to Blind Image Denoising

The computational efficiency and excellent quality prediction performance makes BRISQUE an attractive option for practical applications. One such application could be using a quality measure to augment the performance of image repair algorithms. In this section, we describe one such approach, where the BRISQUE features are used to transform a non-blind image denoising algorithm into a blind image denoising algorithm.

Blind image denoising algorithms seek to reduce the amount of noise present in corrupted images, without any additional information such as the noise variance. Although image denoising is a well studied problem in image
Figure 3.11: Accurate noise variance as input to the algorithm in [173] produces poorer quality denoised images: (a) Noisy Image ($\sigma = 0.0158$, MS-SSIM = 0.9063), (b) Denoised with $\sigma = 0.0158$ (MS-SSIM = 0.9176) and (c) Denoised with $\sigma = 0.0040$ (MS-SSIM = 0.9480)
processing [14, 28, 33, 70, 105], blind image denoising remains relatively under-explored [28, 173]. The proposed algorithms typically address parameter estimation in an *ad-hoc* fashion without regard to natural scene statistics. Here, we demonstrate a systematic perception-based parameter estimation approach that results in better denoising performance. We augment a state-of-the-art image denoising algorithm by using BRISQUE feature-based parameter prediction to improve performance.

The work closest in concept to this approach is the one proposed in [173] where image content measures were used to predict the noise variance in the image, which was then used for image denoising; however the approach is computationally intensive and the measure of content in the image may not be the ideal measure to predict noise variance. In [173], the noisy image is denoised multiple times and quality is estimated using their proposed no-reference content evaluation algorithm. Amongst the large set of denoised images produced, the image with the best content-quality is selected as the denoised image. As an alternative, we propose a learning based framework where noise parameters are estimated using natural scene statistics based on BRISQUE features.

The denoising algorithm that we use is the one proposed in [28], which requires as input the noise variance in the image. However, our experiments suggest that when the algorithm is fed the true accurate noise-variance, the performance of the denoiser is sub-par. The performance of the algorithm drastically improves, if a (systematically) different parameter selected based
on perceptual quality is fed as input to the algorithm. In order to demonstrate this, in Fig. 3.11, we plot an image denoised using the true noise variance and that arrived at using the noise variance from our approach (described below). Notice that our approach produces better visual quality, and better objective quality, as gauged by the multi-scale structural similarity index (MS-SSIM) [159].

We design our training framework to account for this discrepancy and to ensure that the denoised image attains the highest visual quality. Our approach proceeds as follows. Given a large set of noisy images afflicted with different levels of noise, we denoise each image using the denoising algorithm – BM3D [28] – by providing as input images distorted with various values of noise variance. The denoised images so obtained are judged for their quality using MS-SSIM and the noise parameter corresponding to the image with the maximum denoised quality is set as the input to the algorithm. These noise variances are then used in a training phase, where BRISQUE features are mapped on to the noise-prediction parameter, using SVM regression as before [25]. Once trained, the automatic parameter prediction approach is capable of predicting the level of input noise to BM3D, so that the output denoised image has the highest visual quality. We note that our training approach resembles that of [173].

Given a new (unseen) test noisy image, the BRISQUE augmented BM3D approach predicts the accurate input to BM3D and denoises the image with (as we shall soon see) much higher visual quality than the baseline. No-
tice that BRISQUE augmentation is not limited to the BM3D algorithm; and any non-blind algorithm could be improved by using BRISQUE natural scene features to produce a blind image denoiser.

To show the effectiveness of our algorithm and to demonstrate its robustness across a large variety of images and distortion levels, we created a noisy image dataset from the 300 images present in the Berkeley image segmentation database [72]. We introduced 10 different levels of Gaussian noise to each image yielding a total of 3000 noisy images. The noise variance ranged from 0.001 to 0.5, uniformly sampled on a logarithmic scale. 1000 images were then used for training and 2000 for testing thereby ensuring no content overlap between the two sets. The regression model described above was trained on 1000 training images and then used to predict the input parameter on the test images.

Once denoised images are obtained, we compare their quality (using MS-SSIM) using our approach as well for the default implementation of the BM3D algorithm and in Fig. 3.12, we plot the mean quality and the associated standard errors at each noise level across the 2000 test images for both these approaches. It is clear that BRISQUE augmented BM3D produces much higher quality images than the baseline BM3D. We also analyzed whether the differences observed in the quality of the denoised images between our approach and the reference BM3D implementation are statistically significant using the $t$-test [46]. Our analysis indicates that for all noise variances simulated in the present data, our approach is statistically superior to the reference
Figure 3.12: Shows the mean quality and associated errors at each noise level across 2000 test images for our approach as well as the reference implementation of BM3D

BM3D implementation in terms of perceived visual quality at the 95% confidence level, excepting when the noise variance is a tiny 0.0316 - where the two approaches become statistically indistinguishable.

3.4 Discussion and Conclusion

We proposed a natural scene statistic based distortion-generic blind/no-reference (NR) quality assessment algorithm – the Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) – which operates in the spatial domain. No distortion specific features such as ringing, blur or blocking were modeled in the algorithm. The algorithm only quantifies the ‘naturalness’ (or lack thereof) in the image due to presence of distortions.
We detailed the algorithm and the statistical features extracted, and demonstrated how each of these features correlate with human perception. We then undertook a thorough evaluation of the BRISQUE index in terms of correlation with human perception and demonstrated that BRISQUE is statistically better than FR PSNR and SSIM as well as highly competitive to all NR algorithms compared with. We demonstrated that BRISQUE performance is independent of database content and BRISQUE features may be used for distortion-identification as well. Further, we also showed that BRISQUE is computationally efficient and that its efficiency is superior to other distortion-generic approaches to NR IQA, thus making BRISQUE an attractive option for practical applications like image denoising. We demonstrated this application by augmenting non-blind image denoising algorithms using the BRISQUE features to produce blind image denoising algorithms.

In the next chapter, we will discuss the limitations of ‘opinion aware’ approaches and propose a new approach using which we can get away with the information about distorted images and human judgments on them.
Chapter 4

Opinion Unaware Image Quality Assessment

‘Opinion-aware’ IQA models described in the previous chapter are necessarily limited, since they can only assess quality degradations arising from the distortion types that they have been trained on. Given the impracticality of obtaining collections of distorted images with co-registered human scores, models that do not require training on databases of human judgments of distorted images, and hence are ‘opinion unaware’ (OU), are of great interest. One such effort was made in this direction by the authors of [79]. However, their model requires knowledge of the expected image distortions.

Likewise, among algorithms derived from OU models, distorted images may or may not be available during IQA model creation or training. For example, in highly unconstrained environments, such as a photograph upload site, the \textit{a priori} nature of distortions may be very difficult to know. Thus a model may be formulated as ‘distortion aware’ (DA) by training on (and hence tuning to) specific distortions, or it may be ‘distortion unaware’ (DU), relying instead only on exposure to naturalistic source images or image models to guide the QA process. While this may seem as an extreme paucity of information to guide design, it is worth observing that very successful FR IQA models (such
as the structural similarity index (SSIM) [155]) are DU.

Our contribution in this direction is the development of a NSS-based modeling framework for OU-DU NR IQA design, resulting in a first of a kind NSS-driven blind OU-DU IQA model which does not require exposure to distorted images \textit{a priori}, nor any training on human opinion scores. The new NR OU-DU IQA quality index performs better than the popular FR peak signal-to-noise-ratio (PSNR) and structural similarity (SSIM) index and delivers performance at par with top performing NR OA-DA IQA approaches.

4.1 No Reference Opinion-Unaware Distortion-Unaware IQA Model

Our new NR OU-DU IQA model is based on constructing a collection of ‘quality aware’ features and fitting them to a multivariate Gaussian (MVG) model. The quality aware features are derived from a simple but highly regular natural scene statistic (NSS) model. The quality of a given test image is then expressed as the distance between a multivariate Gaussian (MVG) fit of the NSS features extracted from the test image, and a MVG model of the quality aware features extracted from the corpus of natural images.

4.1.1 Spatial Domain NSS

Our ‘completely blind’ IQA model is founded on spatial NSS features extracted from local image patches that effectively capture the essential low-order statistics of natural images.
The coefficients (3.1) have been observed to reliably follow a Gaussian distribution when computed from natural images that have suffered little or no apparent distortion [108]. This ideal model, however, is violated when the images do not derive from a natural source (e.g., computer graphics) or when natural images are subjected to unnatural distortions. The degree of modification can be indicative of perceptual distortion severity.

The NSS features used in the NIQE index are similar to those used in a prior OA-DA IQA BRISQUE model [78]. However, NIQE only uses the NSS features from a corpus of natural images while BRISQUE is trained on features obtained from both natural and distorted images and also on human judgments of the quality of these images. Therefore, BRISQUE is limited to the types of distortions it has been tuned to. By comparison, the NIQE Index is not tied to any specific distortion type, yet, as will be shown, delivers nearly comparable predictive power on the same distortions the BRISQUE index has
been trained on, with a similar low complexity.

4.1.2 Patch Selection

Once the image coefficients (3.1) are computed, the image is partitioned into $P \times P$ patches. Specific NSS features are then computed from the coefficients of each patch. However, only a subset of the patches are used for the following reason.

Every image is subject to some kind of limiting distortion [13]. For instance, there is a loss of resolution due to defocus blur in parts of most images due to the limited depth of field (DOF) of any single-lens camera. Since humans appear to more heavily weight their judgments of image quality from the sharp image regions [45], more salient quality measurements can be made from sharp patches. Setting aside the question of the aesthetic appeal of having some parts of an image sharper than others, any defocus blur represents a potential loss of visual information.

We use a simple device to preferentially select from amongst a collection of natural patches those that are richest in information and less likely to have been subjected to a limiting distortion. This subset of patches is then used to construct a model of the statistics of natural image patches.

The variance field (3.3) has been largely ignored in the past in NSS based image analysis, but it is a rich source of structural image information that can be used to quantify local image sharpness. Letting the $P \times P$ sized patches be indexed $b = 1, 2, \ldots, B$, a direct approach is to compute the average
local deviation field of each patch indexed $b$:

$$
\delta(b) = \sum\sum_{(i,j) \in \text{patch}_b} \sigma(i, j) \quad (4.1)
$$

where $\delta$ denotes local activity/sharpness.

Once the sharpness of each patch is found, those having a suprathreshold sharpness $\delta > T$ are selected. The threshold $T$ is picked to be a fraction $p$ of the peak patch sharpness over the image. In our experiments, we used the nominal value $p = 0.75$. Examples of this kind of patch selection are shown in Fig. 4.1. We have observed only small variations in performance when $p$ is varied in the range $[0.6, 0.9]$.

4.1.3 Characterizing Image Patches

Given a collection of natural image patches selected as above, their statistics are characterized by ‘quality aware’ NSS features computed from each selected patch [78]. Prior studies of NSS based image quality have shown that the generalized Gaussian distribution effectively captures the behavior of the coefficients (3.1) of natural and distorted versions of them [82].

4.1.4 Multivariate Gaussian Model

A simple model of the NSS features computed from natural image patches can be obtained by fitting them with an MVG density, providing
a rich representation of them:

\[
f_X(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \nu)^T \Sigma^{-1}(x - \nu)\right) \tag{4.2}
\]

where \((x_1, \ldots, x_k)\) are the computed NSS features, and \(\nu\) and \(\Sigma\) denote the mean and covariance matrix of the MVG model, which are estimated using a standard maximum likelihood estimation procedure [10]. We selected a varied set of 125 natural images with sizes ranging from \(480 \times 320\) to \(1280 \times 720\) to obtain the multivariate Gaussian model. Images were selected from copyright free Flickr data and from the Berkeley image segmentation database [72] making sure that no overlap occurs with the test image content. The images may be viewed at http://live.ece.utexas.edu/research/quality/pristinedata.zip.

### 4.1.5 NIQE Index

The new OU-DU IQA index, called NIQE, is applied by computing the 36 identical NSS features from patches of the same size \(P \times P\) from the image to be quality analyzed, fitting them with the MVG model (4.2), then comparing its MVG fit to the natural MVG model. The sharpness criterion (4.1) is not applied to these patches because loss of sharpness in distorted images is indicative of distortion and neglecting them would lead to incorrect evaluation of the distortion severity. The patch size was set to \(96 \times 96\) in our implementation. However, we observed stable performance across patch sizes ranging from \(32 \times 32\) to \(160 \times 160\).

Finally, the quality of the distorted image is expressed as the distance between the quality aware NSS feature model and the MVG fit to the features
extracted from the distorted image:

\[ D(\nu_1, \nu_2, \Sigma_1, \Sigma_2) = \sqrt{\left( (\nu_1 - \nu_2)^T \left( \frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (\nu_1 - \nu_2) \right)} \]  

(4.3)

where \( \nu_1, \nu_2 \) and \( \Sigma_1, \Sigma_2 \) are the mean vectors and covariance matrices of the natural MVG model and the distorted image’s MVG model.

### 4.2 Performance Evaluation

#### 4.2.1 Correlation with Human Judgments of Visual Quality

To test the performance of the NIQE index, we used the LIVE IQA database [125] of 29 reference images and 779 distorted images spanning five different distortion categories – JPEG and JPEG2000 (JP2K) compression, additive white Gaussian noise (WN), Gaussian blur (blur) and a Rayleigh fast fading channel distortion (FF). A difference mean opinion score (DMOS) associated with each image represents its subjective quality.

Since all of the OA IQA approaches that we compare NIQE to require a training procedure to calibrate the regressor module, we divided the LIVE database randomly into chosen subsets for training and testing. Although our blind approach and the FR approaches do not require this procedure, to ensure a fair comparison across methods, the correlations of predicted scores with human judgments of visual quality are only reported on the test set. The dataset was divided into 80% training and 20% testing – taking care that no overlap occurs between train and test content. This train-test procedure was
Table 4.1: Median spearman rank ordered correlation coefficient (SROCC) across 1000 train-test combinations on the LIVE IQA database. *Italics* indicate (OA/OU)-DA no-reference algorithms and **bold face** indicates the new OU-DU model algorithm.

<table>
<thead>
<tr>
<th></th>
<th>JP2K</th>
<th>JPEG</th>
<th>WN</th>
<th>Blur</th>
<th>FF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.8646</td>
<td>0.8831</td>
<td>0.9410</td>
<td>0.7515</td>
<td>0.8736</td>
<td>0.8636</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.9389</td>
<td>0.9466</td>
<td>0.9635</td>
<td>0.9046</td>
<td>0.9393</td>
<td>0.9129</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.9627</td>
<td>0.9785</td>
<td>0.9773</td>
<td>0.9542</td>
<td>0.9386</td>
<td>0.9535</td>
</tr>
<tr>
<td>CBIQ</td>
<td>0.8935</td>
<td>0.9418</td>
<td>0.9582</td>
<td>0.9324</td>
<td>0.8727</td>
<td>0.8954</td>
</tr>
<tr>
<td>LBIQ</td>
<td>0.9040</td>
<td>0.9291</td>
<td>0.9702</td>
<td>0.8983</td>
<td>0.8222</td>
<td>0.9063</td>
</tr>
<tr>
<td>BLIINDS-II</td>
<td>0.9323</td>
<td>0.9331</td>
<td>0.9463</td>
<td>0.8912</td>
<td>0.8519</td>
<td>0.9124</td>
</tr>
<tr>
<td>DIVINE</td>
<td>0.9123</td>
<td>0.9208</td>
<td>0.9818</td>
<td>0.9373</td>
<td>0.8694</td>
<td>0.9250</td>
</tr>
<tr>
<td>BRISQUE</td>
<td>0.9139</td>
<td>0.9647</td>
<td>0.9786</td>
<td>0.9511</td>
<td>0.8768</td>
<td>0.9395</td>
</tr>
<tr>
<td>TMIQ</td>
<td>0.8412</td>
<td>0.8734</td>
<td>0.8445</td>
<td>0.8712</td>
<td>0.7656</td>
<td>0.8010</td>
</tr>
<tr>
<td>NIQE</td>
<td><strong>0.9172</strong></td>
<td><strong>0.9382</strong></td>
<td><strong>0.9662</strong></td>
<td><strong>0.9341</strong></td>
<td><strong>0.8594</strong></td>
<td><strong>0.9135</strong></td>
</tr>
</tbody>
</table>

repeated 1000 times to ensure that there was no bias due to the spatial content used for training. We report the median performance across all iterations.

We use Spearman’s rank ordered correlation coefficient (SROCC), and Pearson’s (linear) correlation coefficient (LCC) to test the model. The NIQE scores are passed through a logistic non-linearity [125] before computing LCC for mapping to DMOS space. We compared NIQE with three FR indices: PSNR, SSIM [155] and multiscale SSIM (MS-SSIM) [159], five general purpose OA-DA algorithms - CBIQ [170], LBIQ [141], BLIINDS-II [110], DIVINE [85], BRISQUE [78] and the DA-OU approach TMIQ [79].

As can be seen from Tables 4.1 and 4.2, NIQE performs better than the FR PSNR and SSIM and competes well with all of the top performing OA-DA NR IQA algorithms. This is a fairly remarkable demonstration of the relationship between quantified image naturalness and perceptual image quality.
Figure 4.2: Variation of performance with the number of natural images $K$. Error smear around each point indicate the standard deviation in performance across 100 iterations for $5 < K < 125$.

Table 4.2: Median linear correlation coefficient across 1000 train-test combinations on the LIVE IQA database. *Italic* indicates (OA/OU)-DA no-reference algorithms and *bold face* indicates the new OU-DU model algorithm.

<table>
<thead>
<tr>
<th>Method</th>
<th>JP2K</th>
<th>JPEG</th>
<th>WN</th>
<th>Blur</th>
<th>FF</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.8762</td>
<td>0.9029</td>
<td>0.9173</td>
<td>0.7801</td>
<td>0.8795</td>
<td>0.8592</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.9405</td>
<td>0.9462</td>
<td>0.9824</td>
<td>0.9004</td>
<td>0.9514</td>
<td>0.9066</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.9746</td>
<td>0.9793</td>
<td>0.9883</td>
<td>0.9645</td>
<td>0.9488</td>
<td>0.9511</td>
</tr>
<tr>
<td>CBIQ</td>
<td>0.8898</td>
<td>0.9434</td>
<td>0.9333</td>
<td>0.9338</td>
<td>0.8951</td>
<td>0.8955</td>
</tr>
<tr>
<td>LBIQ</td>
<td>0.9103</td>
<td>0.9345</td>
<td>0.9761</td>
<td>0.9104</td>
<td>0.8382</td>
<td>0.9087</td>
</tr>
<tr>
<td>BLINDS-II</td>
<td>0.9386</td>
<td>0.9426</td>
<td>0.9635</td>
<td>0.8994</td>
<td>0.8790</td>
<td>0.9164</td>
</tr>
<tr>
<td>DIIVINE</td>
<td>0.9233</td>
<td>0.9347</td>
<td>0.9867</td>
<td>0.9370</td>
<td>0.8916</td>
<td>0.9370</td>
</tr>
<tr>
<td>BRISQUE</td>
<td>0.9229</td>
<td>0.9734</td>
<td>0.9851</td>
<td>0.9506</td>
<td>0.9030</td>
<td>0.9424</td>
</tr>
<tr>
<td>TMIQ</td>
<td>0.8730</td>
<td>0.8941</td>
<td>0.8816</td>
<td>0.8530</td>
<td>0.8234</td>
<td>0.7856</td>
</tr>
<tr>
<td>NIQE</td>
<td>0.9370</td>
<td>0.9564</td>
<td>0.9773</td>
<td>0.9525</td>
<td>0.9128</td>
<td>0.9147</td>
</tr>
</tbody>
</table>
4.2.2 Number of Natural Images

We addressed the question: ‘How many natural images are needed to obtain a stable model that can correctly predict image quality?’ Such an analysis provides an idea of the quality prediction power of the NSS features and how well they generalize with respect to image content.

To undertake this evaluation, we varied the number of natural images $K$ from which patches are selected and used for model fitting. Figure 4.2 shows the performance against the number of images. An error band is drawn around each point to indicate the standard deviation in performance across 100 iterations of different sample sets of $K$ images. It may be observed that a stable natural model can be obtained using a small set of images.

4.2.3 Performance variation with patch size

We evaluate how changing the patch size, affects the performance of Naturalness Index in terms of correlation with human judgments of visual quality. One would hypothesize that large patches would have limited capability to capture the local behaviour, but conversely feature estimates are nosier if the patches are too small. Figure 4.3 shows the variation in performance with patch size. The best performance was achieved for patches of intermediate patch size $96 \times 96$. The variation, however, was minimal with stable performance achieved over a wide range of patch sizes.
4.2.4 Performance variation with sharpness percentile

To make the natural image model, patches with sharpness greater than a threshold percentile were picked. Such kind of model has relevance from human perception viewpoint of visual quality. As noted by [45], humans also judge image quality with greater emphasis on the sharp image patches.

We evaluate how changing this threshold percentile affects the performance of Naturalness Index in terms of correlation with human judgments of visual quality. Figure 4.4 shows the performance variation with percentile sharpness threshold determining which patches from an image are selected. One would expect that blurry patches would tend to be selected at low percentiles, making the naturalistic model noisy, and hence less sensitive towards predicting the quality. Yet the performance only modestly varies for a wide
Figure 4.4: Performance variation with percentile sharpness, threshold based on which patches from an naturalistic image are selected.

range of percentile values, dipping only when the percentile falls towards 0.1.

4.3 Prediction Error Analysis

To be able to use an IQA algorithm in practical applications, its limitations must be understood. Thus we analyse the performance of the Naturalness Index in terms of the probability of error in DMOS prediction on distortion-by-distortion and image quality basis. This kind of study can show error trends where an algorithm tends to over predict or under predict human judgments of visual quality. We also computed error probability distributions of DMOS prediction for the MS-SSIM and PSNR algorithms for comparison purposes.

To map Naturalness Index scores onto DMOS space, the predicted scores were passed through a logistic non-linearity [125]. The prediction er-
ror is computed as the difference between the mapped algorithm scores and DMOS scores.

4.3.1 Quality based error analysis

To analyse the variation of the prediction error with image quality, we divided the LIVE IQA database [125] into five subsets of human judgments: excellent, good, fair, poor and bad. Figure 4.5 plots the probability distribution of errors in DMOS prediction, where a positive prediction error signifies that the quality of the distorted image was predicted to be better than the true human judgment, whereas a negative error signifies that the quality of the distorted image is predicted to be worse than the human judgment. Based on LCC analysis, the zero error probability follows the order: MS-SSIM, Naturalness Index and PSNR.

Interestingly, there is a trend in the error distribution for intermediate quality images for both PSNR and the Naturalness Index. PSNR tends to over penalize the errors between reference and distorted image and hence images are predicted to be of worse quality than their actual human judgments. The Naturalness Index and MS-SSIM, on the other hand, tend to make negative prediction errors. However, all three algorithms tend to under predict DMOS scores for bad quality images and to over predict DMOS scores for the best quality images.
Figure 4.5: Figure plots the probability distribution of errors in DMOS prediction for a subset of images in LIVE IQA database [125] with quality range (a) All qualities (b) Excellent (c) Good (d) Fair (e) Poor and (f) Bad. Positive prediction error signifies that the quality of the distorted image is predicted to be better than its actual human judgment whereas negative error signifies that the quality of the distorted image is predicted to be worse than its actual human judgment.
4.3.2 Distortion-type based error analysis

This section answers the question of what kind of errors are made by different algorithms depending on the kind of distortion the image is afflicted with. To study this variation, we analysed images affected by the different distortion types in the LIVE IQA database [125] : JPEG2000, JPEG, white noise, Gaussian blur and fast fading. Figure 4.6 plots the error probability distribution in DMOS prediction where, if the quality of the distorted image is predicted to be better than its actual human judgment, the prediction error is positive and vice versa.

It can be observed from the figure that MS-SSIM and Naturalness Index make negative and positive prediction errors for JPEG2000 and JPEG compressed images respectively. PSNR, on the other hand makes positive prediction errors for JPEG and white noise.

4.4 Discussion and Conclusion

We have created a first of a kind blind IQA model that assesses image quality without knowledge of anticipated distortions or human opinions of them. The quality of the distorted image is expressed as a simple distance metric between the model statistics and those of the distorted image. The new model outperforms FR IQA models and competes with top performing NR IQA trained on human judgments of known distorted images. Such a model has great potential to be applied in unconstrained environments.
Figure 4.6: Figure plots the probability distribution of errors in DMOS prediction for a subset of images in LIVE IQA database [125] with quality range (a) All distortions (b) JPEG2000 (c) JPEG (d) White Noise (e) Gaussian Blur and (f) Fast fading. Positive prediction error signifies that the quality of the distorted image is predicted to be better than its actual human judgment whereas negative error signifies that the quality of the distorted image is predicted to be worse than its actual human judgment.
Having described ‘opinion unaware’ algorithms, we will talk about their limitations in the next chapter and see how they can be addressed using intrinsic statistics. The proposed approach will deal with the more difficult problem of video quality assessment.
Chapter 5

Completely Blind Video Quality Assessment

The image quality assessment approach proposed in the previous chapter was able to predict human judgments of visual quality without training on human judgments of visual quality or using distortion specific knowledge [81]. The approach we take here is even more spare in the use of underlying assumptions or information, and does not even require pristine natural videos to serve as model ground truth. No training of any kind is used. While this may seem as an extreme paucity of information, the use of perceptually relevant quantities yields results that are very promising. Indeed, the resulting algorithm predicts human judgments of video quality better than the ubiquitous full reference PSNR on the LIVE VQA database [117].

This new NR VQA approach is derived based on intrinsic statistical regularities that are observed in natural videos. Deviations from these regularities alter their visual impression. Quantifying measurements of regularity (or lack thereof) under a natural video statistic model makes it possible to develop a ‘quality analyzer’ which can predict the visual quality of the distorted video without any external knowledge and hence is zero shot.

The approach does not require any distortion knowledge, such as exem-
plar training videos containing anticipated distortions or human opinions of them. This is a significant advantage given that the creation of VQA databases containing distorted videos with co-registered human opinion scores is much more involved than is the creation of IQA databases [30] and [117].

In the past we have proposed the use of exemplar natural picture content as ground truth relative to which statistical regularity may be determined [81]. Such a model, however, may be limited in that it can only capture common baseline characteristics of a specific collection of non-distorted content, and is thereby not able to universally represent video specific intrinsic characteristics. Also, the construction of such a database requires the unbiased selection and maintainence of hundreds of natural undistorted videos. This also raises the question of how many exemplar videos are needed to design an accurate natural video model, and how distinctive these need to be relative to each other and to the world of videos. Finally, given the limitations of image/video camera capture, distortions are inevitably introduced in the capture process and hence the procurement of perfectly natural ‘pristine’ videos is practically impossible.

The topic of NR VQA has been extensively studied and surveyed [31, 55, 89, 99, 111, 166, 169]. All of the blind VQA algorithms require knowledge of (one of possibly multiple) distortion types, or of the artifacts introduced by them, or of human opinion scores on distorted images. There exists no blind VQA model to date which can predict video quality in the absence of such strong a priori information.
In this article, we explain the underlying ‘quality aware’ natural video statistics model in the space-time domain and describe perceptually relevant temporal features that are used to model inter subband correlations over both local and global time spans. The overall model is the basis on an algorithm for predicting video quality that is shown to correlate well with human judgments of visual quality. We also compare it’s performance to state-of-the-art FR and NR VQA approaches. Before we describe the model in detail, we review relevant prior work in the area of VQA.

5.1 Natural Video Statistic Model

We create a ‘zero shot’ NR VQA model by making measurements of perceptually relevant temporal video statistics. Natural videos contain regular structures and have generally piece-wise smooth luminances in space-time separated by sparse spatio-temporal edge discontinuities. This strong property of natural videos has been exploited in a variety of applications [53, 111, 118, 133, 174]. It induces self similarity over space and time which, for example, has been exploited for resolution enhancement [118, 174], action
Figure 5.2: Flow diagram depicting multiscale video decomposition followed by divisive normalization. The example uses consecutive frames from the sunflower sequence in the LIVE VQA database [117].
Figure 5.3: Feature extraction followed by local temporal inter subband correlation computation. The same example frames were used as in Fig. 5.2.

Self similarity statistics computed using differences between consecutive frames have been used to capture distortion-induced anomalous behavior and to conduct visual quality inference [111, 133]. Deriving inspiration from these examples, we model temporal self similarities using frame differences between consecutive frames.

5.1.1 Subband Decomposition and Divisive Normalization

Our ‘zero shot’ quality analyzer uses insights from the VQA model proposed in [133] which showed that the spatial bandpass filter coefficients of frame differences ($\Delta F_t$) capture temporal statistical regularities arising from
structures such as moving edges [133]. The filter coefficients were obtained using multiscale wavelet transforms. However, inspired by the success of a recently proposed fast spatial domain IQA model [81], we instead analyse the statistics of locally windowed frame differences $\Delta F^{(t)}$ of videos in the space-time domain.

Define the frame difference $\Delta F^t$ between consecutive frames $F^{t+1}$ and $F^t$ of size $M \times N$ as follows:

$$\Delta F^t = F^{t+1} - F^t \forall t \in \{1, 2, 3 \ldots T\} \quad (5.1)$$

where $T$ is the total number of frames.

Assuming circularly-symmetric Gaussian weighting functions sampled out to 3 standard deviations ($K_1 = K_2 = 3$) and rescaled to unit volume, we perform the scale space decomposition of the video as follows:

$$\Delta F^t_L(i, j) = \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} w_{k_1,k_2} \Delta F^t(i + k_1, j + k_2)$$
$$\Delta F^t_H(i, j) = \Delta F^t - \Delta F^t_L(i, j),$$

where $i \in \{1, 2 \ldots M\}$, $j \in \{1, 2 \ldots N\}$ are spatial indices, $M$ and $N$ are the image dimensions, and $w = \{w_{k_1,k_2} | k_1 = -K_1, \ldots, K_1, k_2 = -K_2, \ldots K_2\}$.

To obtain a finer scale-space representation, iterate the procedure.
\[
\Delta F^t_{(L,L)}(i,j) = \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} w_{k_1,k_2} \Delta F^t_L(i + k_1, j + k_2),
\]
\[
\Delta F^t_{(L,H)}(i,j) = \Delta F^t_L - \Delta F^t_{(L,L)}(i,j),
\]
\[
\Delta F^t_{(H,L)}(i,j) = \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} w_{k_1,k_2} \Delta F^t_H(i + k_1, j + k_2) \quad \text{and}
\]
\[
\Delta F^t_{(H,H)}(i,j) = \Delta F^t_H - \Delta F^t_{(H,L)}(i,j).
\]

Further iterations can be done, however, we have found that the second scale is enough to attain good quality prediction performance (atleast on the large database tested on). We note in passing that if larger-format videos or video databases are being analyzed; then finer scale representations using more scales may be advisable. Once these signal representations are obtained, a divisive contrast normalization nonlinearity is applied, to account for the contrast-gain masking process that occurs in early human vision [47] and [108]. Hence let:

\[
\overline{\Delta F}^t_{(A,B)}(i,j) = \frac{\Delta F^t_{(A,B)}}{\sigma^t_A(i,j) + 1}
\]

\[
\forall \ (A, B) \in \{ (L, L) , (L, H) , (H, L) , (H, H) \}
\]

where

\[
\sigma^t_A(i,j) = \sqrt{\sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} w_{k_1,k_2} [\Delta F^t_A(i + k_1, j + k_2) - \mu^t_A(i,j)]^2}
\]

69
estimates the local contrast and

$$\mu^t_A(i, j) = \sum_{k_1 = -K_1}^{K_1} \sum_{k_2 = -K_2}^{K_2} w_{k_1, k_2} \Delta F^t_A(i + k_1, j + k_2).$$

### 5.1.2 Characterization of Video Patches

Once the normalized coefficients (5.2) for each subband are computed, each coefficient map is partitioned into $P \times P$ patches. Features are extracted using the coefficients of each patch as shown in Fig. 5.3. The coefficients of each sub-band are modeled as obeying a generalized Gaussian distribution, which effectively captures the behavior of the coefficients of natural and distorted videos.

The zero mean generalized Gaussian distribution (GGD) is given by:

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

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

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

The parameters of the GGD $(\alpha, \beta)$ can be reliably estimated using the moment-matching based approach proposed in [119]. The parameters we use as features, $v = \langle \alpha, \beta \rangle$, are computed for every patch in each of the 4 sub-bands, for all patches $p \in \{1, 2 \ldots N_t\}$ where $N_t$ is the total number of patches in frame $t$, over all frequency subbands $\langle A, B \rangle \in \{(L, L), (L, H), (H, L), (H, H)\}$ and different frames $t \in \{1, 2 \ldots T\}$. 

70
5.1.3 Inter Subband Correlations

Change in local video statistics that are otherwise regular across consecutive frames for a given patch can be indicative of some kind of temporal inconsistency arising from distortion. To capture this behaviour, we compute feature differences $\Delta \nu_i(p, \langle A, B \rangle, t)$ at every patch location, as given by (5.5) and shown visually in Fig. 5.3:

$$\Delta \nu_i(p, \langle A, B \rangle, t) = \nu_i(p, \langle A, B \rangle, t) - \nu_i(p, \langle A, B \rangle, t)$$  \hspace{1cm} (5.5)

$\forall i \in 1, 2$

In the following, we analyse both local and global temporal intersubband correlations towards modeling local and global inconsistencies introduced due to the presence of video distortions.

5.1.3.1 Local

Local temporal subband correlations capture the inter frequency interactions between the low and high pass bands across all patches at a given frame instant. They are computed by estimating the correlations of observed feature differences across the low pass bands $\langle A, B \rangle \in \{\langle L, L \rangle, \langle L, H \rangle\}$ and the high pass bands $\langle A, B \rangle \in \{\langle H, L \rangle, \langle H, H \rangle\}$ at each frame instant $t$ as given by (5.6) and (5.7), respectively. For every $i \in \{1, 2\}$, define:

$$\nu_i^l(t) \triangleq \{\Delta \nu_i(1, \langle L, L \rangle, t), \Delta \nu_i(1, \langle L, H \rangle, t), \Delta v_i(2, \langle L, L \rangle, t),$$
$$\Delta v_i(2, \langle L, H \rangle, t), \ldots, \Delta \nu_i(N_t, \langle L, L \rangle, t), \Delta v_i(N_t, \langle L, H \rangle, t)\}$$  \hspace{1cm} (5.6)
Local $\alpha$ based temporal subband correlation

Probability

Pristine
Distorted

Local $\beta$ based temporal subband correlation

Probability

Pristine
Distorted

Figure 5.4: Distribution of local temporal inter subband correlations on undistorted and distorted videos in the LIVE VQA database [117] for (a) $\alpha$ and (b) $\beta$ features.

\[
\nu_i^2(t) \triangleq \{ \Delta v_i(1, \langle H, L \rangle, t), \Delta v_i(1, \langle H, H \rangle, t), \Delta v_i(2, \langle H, L \rangle, t), \\
\Delta v_i(2, \langle H, H \rangle, t), \ldots \Delta v_i(N_t, \langle H, L \rangle, t), \Delta v_i(N_t, \langle H, H \rangle, t) \} \tag{5.7}
\]

The correlation coefficient $\phi^f_i(\nu^1_i(t), \nu^2_i(t))$ between $\nu^1_i(t)$ and $\nu^2_i(t)$ that is estimated is:

\[
\phi^f_i(\nu^1_i(t), \nu^2_i(t)) = \frac{M[(\nu^1_i(t) - \mu_{\nu^1_i(t)})(\nu^2_i(t) - \mu_{\nu^2_i(t)})]}{\sigma_{\nu^1_i(t)}\sigma_{\nu^2_i(t)}} \tag{5.8}
\]

where $M[]$ is the sample mean, and $\mu_{\nu^1_i(t)}$, $\mu_{\nu^2_i(t)}$, and $\sigma_{\nu^1_i(t)}$, $\sigma_{\nu^2_i(t)}$ denote the means and standard deviations of $\nu^1_i(t)$ and $\nu^2_i(t)$ respectively.

Feature differences across different frequency subbands are generally decorrelated since different frequency subbands capture different spatio-temporal
details in undistorted videos. The correlation is further reduced by divisive normalization. However, the procedure operates imperfectly in a distorted video, thereby disturbing the ‘decorrelation regularity’. As shown in Fig. 5.4, correlation values across subbands are low for both features $\alpha$ and $\beta$ when videos are not visibly distorted. However, when distortions are present, causing the different frequency bands to be less decorrelated causes the correlation measure $\phi_t^i(\nu_1^i(t), \nu_2^i(t))$ to rise. Deviations from statistical regularity, when quantified appropriately, make it possible to capture the presence and severity of distortions present in the video.

A feature specific score for the video sequence is given by the weighted mean of quality scores obtained over all frames:

$$\lambda_i = \frac{\sum_t \omega(t)\phi_t^i(\nu_1^i(t), \nu_2^i(t))}{\sum_t \omega(t)} \quad (5.9)$$
where $\omega(t)$ is the temporal weight. We discuss different weighting schemes in the experiments section. The overall score is then computed as the product:

$$\Lambda = \prod_{i \in \{1,2\}} \lambda_i \quad (5.10)$$

### 5.1.3.2 Global

Inter frequency interactions between the low and high pass bands over all patches and across all frames capture global temporal distortion behaviour in videos. Similar to local temporal inter-frequency interactions, the statistics of global temporal correlations are lower for undistorted pristine videos as
compared to the distorted versions of them.

A feature-specific score is again found for each feature by correlating observed feature differences across the low pass bands $\langle A, B \rangle \in \{\langle L, L \rangle, \langle L, H \rangle\}$ and high pass bands $\langle A, B \rangle \in \{\langle H, L \rangle, \langle H, H \rangle\}$ over all frames as given by (5.11) and (5.12) respectively. For every $i \in \{1, 2\}$, define:

$$\xi_i^1 \overset{\Delta}{=} \{\nu_i^1(1), \nu_i^1(2) \ldots \nu_i^1(T)\} \quad (5.11)$$

and

$$\xi_i^2 \overset{\Delta}{=} \{\nu_i^2(1), \nu_i^2(2) \ldots \nu_i^2(T)\} \quad (5.12)$$

The correlation coefficient $\theta(\xi_i^1, \xi_i^2)$ between $\xi_i^1$ and $\xi_i^2$ that is estimated is:

$$\theta_i(\xi_i^1, \xi_i^2) = \frac{M[\langle \xi_i^1 - \mu_{\xi_i^1}, \xi_i^2 - \mu_{\xi_i^2} \rangle]}{\sigma_{\xi_i^1} \sigma_{\xi_i^2}}. \quad (5.13)$$

The overall score is computed as the product of the feature-specific scores predicted for all features.

$$\Theta = \prod_{i \in \{1, 2\}} \theta_i(\xi_i^1, \xi_i^2) \quad (5.14)$$

The measurement of global temporal correlations makes it possible to broadly predict the overall temporal quality experience perceived by a human viewer.

5.1.4 Proposed Index

The proposed index is the product

$$\Omega \overset{\Delta}{=} \Lambda \Theta \quad (5.15)$$
inter subband correlations measured over a local and global time spans. The proposed index captures inter subband inconsistencies over multiple spatial and temporal scales.

5.2 Computational Complexity

We analyse the computational complexity of each step of the proposed algorithm as shown in the block diagram shown in Figure 5.1.

5.2.0.1 Frame differencing

Computing the frame difference for each pixel takes constant time $O(1)$. Hence the complexity of computing frame differences over the entire video has linear time complexity $O(MNT)$.

5.2.0.2 Filtering

The scale space decomposition of every frame difference instant in the video is done using Gaussian weighting functions computed using the `imfilter` function of MATLAB and difference filters to compute the low and high pass components respectively. Low pass filtering uses fast FFT algorithms for performing the convolution operation, the complexity of which is $O(MN \log_2(NM))$, and for the entire video $O(MN \log_2(NM)T)$. Difference filter used to compute high pass signal has complexity $O(MN)$ and the entire video $O(MNT)$. 
5.2.0.3 Divisive normalization

It involves two steps. Computation of the normalizer 5.3 uses square and square root functions with linear time complexity $O(MN \log_2(NM)T)$ while the low pass Gaussian filtering operation has complexity $O(TM N \log_2(NM))$. Division of the filter responses by the normalizer is a constant time operation for every filter coefficient and hence has linear complexity $O(MNT)$ on the whole video.

5.2.0.4 Model based features

Moment based estimation method of the parameters of the GGD distribution is explained in detail in [120].

For a frame size $M \times N$ and patch size $P \times P$, the number of pixels $N_B$ in each block is $\frac{MN}{P^2}$. The first step involves computing of second order moments needed to find the parameter $\beta$ and first order moments for computing $\alpha$, the time complexity of both of which is linear in the number of coefficients present in the block, i.e. $O(N_B)$. Since the number of blocks in the video is $P \times P \times T$, the time complexity is of the order $O(MNT)$.

The second step involves the numerical estimation of the parameter $\alpha$, determined using a lookup table procedure which involves sweeping the interval $[0, I]$ in steps of $\epsilon$. Based on observed distributions of normalized filter coefficients, $\alpha < 5$, hence $I$ was chosen to be 5. The time complexity of the algorithm is of the order $O \left( \log \left( \frac{1}{\epsilon} \right) \right)$ for each block where we chose $\epsilon$ to be 0.01, and $O \left( P^2 \log \left( \frac{1}{\epsilon} \right) T \right)$ for the entire video.
5.2.0.5 Temporal feature differences

Feature difference computation between a pair of features takes constant $O(1)$ time. Since feature differences are computed only for block pairs between consecutive frame differences, the number of blocks are $P \times P \times T$. Hence the time complexity is $O(P^2T)$ for the entire video.

5.2.0.6 Local and global temporal inter subband correlations

Computation of local and global temporal correlations has linear time complexity in terms of the number of feature difference points. The number of points involved in computation of local temporal correlations for every frame is $P \times P$ and hence the computational complexity $O(P^2)$. Because the number of frames for which local temporal correlation is computed is $T$, the complexity is $O(P^2T)$. Similarly, the number of feature difference points involved in the computation of global temporal correlations is $P \times P \times T$ for the entire video, hence the computational complexity is $O(P^2T)$. We do not discuss the time complexity of obtaining frame weights $\omega(t)$ as it depends on the scheme used.

5.3 Dataset and Experiments

We evaluated the results of the proposed zero shot index on the LIVE Video Quality Assessment Database [117]. The LIVE Video Quality Assessment Database contains 10 reference videos and 150 distorted videos with a span of 4 distortion categories, including compression artifacts due to MPEG and H.264, errors induced by transmission over IP networks and errors in-
Table 5.1: SROCC of different VQA algorithms against DMOS on LIVE VQA database

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Wireless</th>
<th>IP</th>
<th>H.264</th>
<th>MPEG</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.6574</td>
<td>0.4167</td>
<td>0.4585</td>
<td>0.3862</td>
<td>0.5398</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.7289</td>
<td>0.6534</td>
<td>0.7313</td>
<td>0.6684</td>
<td>0.7364</td>
</tr>
<tr>
<td>MOVIE</td>
<td>0.8114</td>
<td>0.7192</td>
<td>0.7797</td>
<td>0.8170</td>
<td>0.8055</td>
</tr>
<tr>
<td>STRRED</td>
<td>0.7758</td>
<td>0.7709</td>
<td>0.7886</td>
<td>0.7716</td>
<td>0.8007</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.4640</td>
<td>0.6503</td>
<td>0.7443</td>
<td>0.5142</td>
<td>0.6086</td>
</tr>
</tbody>
</table>

introduced due to transmission over wireless networks. Six videos contain 250 frames at 25 fps, 1 video contains 217 frames at 25 fps and 3 videos contain 500 frames at 50 fps.

The patch size $P$ was set to 64 for the results presented in this work. The performance of any VQA model can be best evaluated by its correlation with human subjective judgments of quality, since the human is the ultimate receiver of the visual signal. We use Spearman’s rank ordered correlation coefficient (SROCC), and Pearson’s (linear) correlation coefficient (LCC) to test the model. The scores were passed through a logistic non-linearity [125] to map to DMOS before computing LCC.

Scores from all frames are given equal weight - $\{\omega(t) = \frac{1}{T} \forall t \in T\}$. Substituting the values of $\omega(t)$ in to (5.9) yields:

$$\lambda_i = \frac{\sum_t \phi_t^i(\nu^1_t(t), \nu^2_t(t))}{T} \quad (5.16)$$

We compared the proposed index with three FR indices: PSNR, MS-SSIM [159] and MOVIE [116], one RR index: STRRED [133] and one learning based blind index: V-BLIINDS [111]. We report the performance of FR and
Table 5.2: LCC of different VQA algorithms against DMOS on LIVE VQA database

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Wireless</th>
<th>IP</th>
<th>H.264</th>
<th>MPEG</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.6695</td>
<td>0.4689</td>
<td>0.5330</td>
<td>0.3986</td>
<td>0.5604</td>
</tr>
<tr>
<td>MS-SSIM</td>
<td>0.7157</td>
<td>0.7267</td>
<td>0.7020</td>
<td>0.6640</td>
<td>0.7379</td>
</tr>
<tr>
<td>MOVIE</td>
<td>0.8371</td>
<td>0.7383</td>
<td>0.7920</td>
<td>0.8252</td>
<td>0.8217</td>
</tr>
<tr>
<td>STRRED</td>
<td>0.7954</td>
<td>0.8114</td>
<td>0.7883</td>
<td>0.7988</td>
<td>0.8084</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.5414</td>
<td>0.6653</td>
<td>0.7503</td>
<td>0.6567</td>
<td>0.6271</td>
</tr>
</tbody>
</table>

RR indices on the entire database in Tables 5.1 and 5.2. As is evident from the results, the proposed index is able to correlate better with human judgments of visual quality than the full reference measure PSNR, which is remarkable given that the proposed index does not incorporate or depend on any kind of external information. We note that the proposed index remains inferior to MS-SSIM and MOVIE, which indicates that there may be still some room for improvement.

The only high performance general-purpose blind VQA algorithm available to compare the proposed index with is the learning based V-BLIINDS model [111], which requires a training procedure to calibrate the regressor module. Hence we divided the LIVE VQA database randomly into chosen subsets for training and testing. The dataset was divided into 80% training and 20% testing - taking care that no overlap occurs between train and test content. This train-test procedure was done on every possible combination of train/test set splits to ensure that there was no bias due to the spatial content used for training. We report the median performance across all iterations in Tables 5.3 and 5.4. As observed from Table 5.4, the proposed index per-
forms at par with V-BLIINDS when the videos in examination are afflicted with the same distortion. However, V-BLIINDS has better performance when videos afflicted with different distortions are examined together for their visual quality, which is not surprising given that it has the advantage of distortion specific calibration during training.

### 5.4 Discussion and Conclusion

We proposed a natural video statistics based zero shot NR quality assessment algorithm - Natural Video Integrity Evaluation (NVIE) - which operates in the spatial domain. It does not model any distortion specific information but only models the naturalness (or lack thereof) in the video. We described how the combination of inter subband correlations extracted at different temporal scales can be used to predict human judgments of visual quality.

Table 5.3: Median SROCC of different blind VQA algorithms against DMOS on every possible combination of train/test set splits. 80% of content used for training on LIVE VQA database

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Wireless</th>
<th>IP</th>
<th>H.264</th>
<th>MPEG</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.3810</td>
<td>0.7143</td>
<td>0.8810</td>
<td>0.6347</td>
<td>0.6294</td>
</tr>
<tr>
<td>V-BLIINSD</td>
<td>0.6190</td>
<td>0.6571</td>
<td>0.8571</td>
<td>0.8810</td>
<td>0.7373</td>
</tr>
</tbody>
</table>

Table 5.4: Median LCC of different blind VQA algorithms against DMOS on every possible combination of train/test set splits. 80% of content used for training on LIVE VQA database

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Wireless</th>
<th>IP</th>
<th>H.264</th>
<th>MPEG</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>0.7314</td>
<td>0.9393</td>
<td>0.9397</td>
<td>0.9348</td>
<td>0.6929</td>
</tr>
<tr>
<td>V-BLIINSD</td>
<td>0.7336</td>
<td>0.6893</td>
<td>0.8287</td>
<td>0.8569</td>
<td>0.7518</td>
</tr>
</tbody>
</table>
quality. Further, we also analyzed the time complexity of NVIE for every step in the algorithm. The algorithm is highly parallelizable as computations on the image blocks can be performed simultaneously. We also undertook a thorough evaluation of NVIE in terms of correlation with human judgments and demonstrated that NVIE correlates better than the FR PSNR metric. Future work would involve faster implementation of the algorithm for real time video monitoring applications. Also, we envision that the proposed algorithm can be used to help solve the resource allocation problem, by modeling video traffic with the intent of optimizing end-user perceptual experience.
Chapter 6

Conclusion and Future Work

In this dissertation, we detailed efficient yet high-performing solutions to complicated problems in image and video quality assessment which use different amounts of input information. A photo quality helper app for android phones has also been designed recently as a part of senior design project by an undergraduate team, which uses the principles discussed in this thesis.

One problem of great interest where proposed algorithms might prove helpful, is to find ways to automatically evaluate and control the perceptual quality of the visual content as a function of these multiple parameters. This problem finds great meaning today when consumers are drowning in digital visual content and finding ways to review and control of the quality of digital photographs is becoming quite challenging. At the same time, camera manufacturers continue to provide improvements in photographic quality and resolution. The raw captured images pass through multiple post processing steps in the camera pipeline, each requiring parameter tuning.

Another interesting application lies in real-time monitoring of content quality for utilization in real traffic networks. Such intelligent QA agents would successfully monitor the quality of incoming data through learning the
statistics of new distortions and perceptual principles and natural content in new environments. Such autonomous quality agents could help in multiple applications like next-generation video networking, surveillance, re-requests for videos, reallocation of resources, identification of faults and other corrective and control tasks.

Another exciting direction of investigation lies in the interaction between the assessment of visual quality and visual tasks. We believe that it is an important step to separate measurements of quality impairment (from capture, compression, processing, transmission) from scene-dependent factors, so their effects on detection, recognition, or other task can be identified and mitigated. This is particularly true in high-distortion environments, such as the increasingly crowded wireless/mobile environment. A principled, ground-up approach is needed whereby the effects of blindly measured video quality degradations on visual tasks can be established. This is of particular importance in forthcoming wireless vision applications, where intelligent, robust blind algorithms are needed, where severe distortions occur, and where human tracking is becoming increasingly important in security applications.


[77] F. Meng, X. Jiang, H. Sun, and S. Yang. Objective perceptual video quality measurement using a foveation-based reduced reference algo-


[120] K. Sharifi and A. Leon-Garcia. Estimation of shape parameter for
generalized gaussian distributions in subband decompositions of video.

[121] H. R. Sheikh and A. C. Bovik. A visual information fidelity approach to


[124] H. R. Sheikh, A. C. Bovik, and G. De Veciana. An information fi-
delity criterion for image quality assessment using natural scene statis-

[125] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of
recent full reference image quality assessment algorithms. *IEEE Trans

quality assessment of noisy, blurry, JPEG2000, and JPEG images. *IEEE


Vita

Anish Mittal graduated with the Bachelor of Technology from Indian Institute of Technology, Roorkee in 2009 and with Masters in Science from The University of Texas at Austin in 2011 respectively. He joined the Laboratory of Image and Video Processing in 2009 and worked as a graduate research assistant under Professor Alan Conrad Bovik during the span of four years of his stay at UT Austin. His research involves the use of human perception and machine learning concepts in solving the engineering problems of image and video processing.

Email address: mittal.anish@gmail.com

This thesis was typed by the author.