Blind quality assessment of image and speech signals

Li, Qiaohong

2017

Li, Q. (2017). Blind quality assessment of image and speech signals. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/70758>

<https://doi.org/10.32657/10356/70758>

NANYANG TECHNOLOGICAL UNIVERSITY

Blind Quality Assessment of Image and Speech Signals

QIAOHONG LI

School of Computer Science and Engineering
A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

April, 2017

Abstract

Quality assessment of multimedia signals is of great interest to the researchers and practitioners in signal processing community. As most multimedia services and systems are provided for human consumption, it is of great importance to reproduce human judgement of perceived quality for objective quality assessment methods. Among all kinds of these methods, no-reference or blind methods that operate solely on the distorted signals are most desirable as the reference signals are not always available in many practical applications. However, blind quality assessment is a very challenging task due to the various distortion types and diverse content properties. In this thesis, I present a series of works on designing better blind models to automatically estimate perceptual quality of image and speech signals for modern multimedia systems.

The first work presented here deals with quality assessment on multiply-distorted images. We propose a novel structural feature as the gradient weighted histogram of local binary pattern calculated on the gradient map, which is effective to describe the complex degradation pattern introduced by multiple distortions.

In the second work we propose a general-purpose method to predict the visual quality of images degraded by various distortion types. By exploring the characteristics of the human visual system (HVS), two new perceptual features are extracted to represent the structural information and luminance changes in distorted images. We show that the complementary information provided by extracted statistical structural and luminance features plays an important role in image quality estimation. This work is later extended in the third work by two aspects: 1) we show that linear filter response can complement the widely used local contrast normalization response; 2) we fuse the luminance and structural information through a weighting scheme. These two works belong to perceptual feature based methods, accounting for the HVS properties in the feature design.

In the fourth work we explore the utilization of natural scene statistics (NSS) for general-purpose blind image quality assessment. We present a new model for natural images, by using multivariate Gaussian mixture model to approximate the joint distribution of log-contrast response. This is the first attempt to use joint NSS model for blind image quality assessment, and it has several advantages over related works.

The last work of this thesis presents a novel non-intrusive speech quality assessment method by adopting the bag-of-words model to speech feature extraction. It provides an effective way to producing global representation from local segments.

In all of these works we compare the proposed methods to the cutting edge of related works.

Acknowledgments

I would like to express my gratitude to my supervisor, Dr. Weisi Lin, for his great guidance and endorsement not only on my research work, but also on my daily life during my Ph.D. study. His professional attitude and expertise have greatly influenced my academic career. I have learned state-of-the-art works in signal processing domain. More importantly, I have learned how to find research problems and solve them under his professional supervision.

I express my appreciation to my co-supervisor Prof. Daniel Thalmann for his faith in my abilities to conduct independent research. I also acknowledge the funding received from Institute of Media Innovation (IMI) throughout my Ph.D. study.

I am also very grateful to my great team members, Dr. Yuming Fang, Dr. Jinjian Wu, Dr. Xinfeng Zhang, Dr. Lu Dong, Dr. Huan Yang, Dr. Yuan Yuan, Mr. Wentao Cheng, Mr. Yabin Zhang, Mr. Sharath Chandra Guntuku and other group members, for their friendship and insightful discussions.

My special thanks to the love and support from my parents and other family members. Especially to my husband and my little son, they are the sunshine of my life. Their trust in me is the driving force for me to move forward.

I would like to thank in advance the thesis examiners for accepting to be part of the committee, and for their comments and suggestions to improve this thesis.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	viii
List of Tables	xi
List of Abbreviations	xii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Objective and Scope	4
1.3 Thesis Contributions	4
1.4 Thesis Organization	7
2 Literature Review	9
2.1 Quality Assessment of Image Signals	10
2.1.1 Subjective quality assessment protocols	10
2.1.2 Objective IQA methods	11
2.2 Quality Assessment of Speech Signals	19
2.2.1 Subjective quality assessment protocols	19
2.2.2 Objective SQA Methods	20
3 Blind Quality Assessment for Multiply-Distorted Images in Gradient Domain	23
3.1 The Proposed Method	24
3.1.1 GLBP construction	24
3.1.2 Gradient weighted GLBP histogram	27
3.1.3 Regression model for quality prediction	28

3.1.4	Comparison with other NR features	29
3.2	Performance Evaluation	31
3.2.1	Experimental protocols	31
3.2.2	Implementation details	32
3.2.3	Performance comparison with FR-IQA methods	35
3.2.4	Performance comparison with BIQA methods	38
3.2.5	Run-time comparison	40
3.3	Summary	40
4	BIQA Using Statistical Structural and Luminance Features	42
4.1	The Proposed Method	43
4.1.1	Local contrast normalization	43
4.1.2	The structural histogram	44
4.1.3	The luminance histogram	47
4.1.4	Analysis of two feature sets	49
4.1.5	Regression model for quality prediction	52
4.2	Performance Evaluation	53
4.2.1	Database description	53
4.2.2	Implementation details	55
4.2.3	Overall performance on individual database	57
4.2.4	Performance on individual distortion type	61
4.2.5	Cross-database validation	63
4.2.6	Computational complexity	64
4.3	Summary	65
5	BIQA Based on Structural Degradation	67
5.1	The Proposed Method	68
5.1.1	Feature set I	68
5.1.2	Feature set II	71
5.1.3	Analysis of two feature sets	74
5.1.4	Regression model for quality prediction	76
5.2	Performance Evaluation	78

5.2.1	Database description	78
5.2.2	Implementation details	78
5.2.3	Overall performance on individual database	78
5.2.4	Performance on individual distortion type	81
5.2.5	Cross-database validation	82
5.2.6	Run-time comparison	83
5.3	Summary	83
6	BIQA Based on Joint Log-contrast Statistics	85
6.1	The Proposed Method	86
6.1.1	Joint directional log-contrast statistics	86
6.1.2	Joint NSS model	88
6.1.3	Quality-aware feature extraction	90
6.1.4	Partial least square regression	92
6.2	Performance Evaluation	93
6.2.1	Implementation details	93
6.2.2	Parameter choice	93
6.2.3	Overall performance on individual database	94
6.2.4	Performance on individual distortion type	96
6.2.5	Cross-database validation	100
6.3	Summary	100
7	Bag-of-Words Representation for Non-Intrusive SQA	101
7.1	The Proposed Method	101
7.1.1	Local PLP feature extraction	102
7.1.2	Codebook construction	103
7.1.3	BoW representation of speech signals	103
7.1.4	SVR mapping	104
7.2	Performance Evaluation	104
7.2.1	Experimental setup	104
7.2.2	Experimental results and analysis	105
7.3	Summary	106

8 Summary and Future Work	108
8.1 Summary	108
8.1.1 Distortion-specific BIQA for multiple distortions	108
8.1.2 General-purpose BIQA based on structural and luminance features . . .	109
8.1.3 General-purpose BIQA based on structural degradation	110
8.1.4 General-purpose BIQA based on joint NSS model	111
8.1.5 Non-intrusive SQA based on bag-of-words representation	112
8.2 Future Work	112
Publication List	115
References	117

List of Figures

1.1	The organization of this thesis.	7
2.1	The classification of existing objective IQA methods.	11
2.2	A typical image communication system which consists of several processes. Each process may introduce some characteristic artifacts.	17
3.1	GLBP maps under different multiple distortions. (a) Pristine image. (b) Image with GB+JPEG. (c) Image with GB+WN. (d) Image with GB+JPEG+WN. (e), (f), (g), (h) are the corresponding GLBP maps of images in the first row.	27
3.2	Scatter plots of t-SNE 2D embedding of feature space of representative BIQA models on MLIVE database. The distortion types are coded by shape and the quality scores are coded by color.	30
3.3	The SRCC performance of GWH-GLBP vs. the number of scales on two benchmark databases.	34
3.4	Box-plot of SRCC distributions of IQA models on MLIVE database.	36
3.5	Box-plot of SRCC distributions of IQA models on MDID2013 database.	37
4.1	Flowchart of the proposed NRSL metric.	43
4.2	Impact of distortions on structural histogram. (a) Average structural histograms on LIVE. (b) Average structural histograms on MLIVE.	47
4.3	Impact of distortions on luminance distribution. (a) Average luminance distri- butions on LIVE. (b) Average luminance distributions on MLIVE.	49
4.4	Validation the independence assumption by approximating the joint distribu- tion by the factorized marginal distribution. (a) The joint distribution $t(g_c, g_0 - g_c)$ which represents the concurrence of horizontally adjacent pixels. (b) The average error produced by factorization of $t(g_c, g_0 - g_c)$ into $t(g_c)$ and $t(g_0 - g_c)$	50

4.5	Image samples to show the complementary of structural and luminance information. (a) Image with DMOS = 68.99. (b) Image with DMOS = 7.23. (c) Image with DMOS = 75.74. (d), (e) and (f) are the structural and luminance histograms of images in the first row.	51
4.6	t-SNE 2D embedding of the NRSL feature space on LIVE database. (a) Distortion types are coded by color. (b) Quality scores are coded by color.	52
4.7	Parameter Choice. (a) The SRCC performance vs. the number of bins of luminance histogram. (b) The SRCC performance vs. constant α	55
5.1	Flowchart of the proposed BSD metric.	69
5.2	(a) Reference image. (b) Image with WN. (c) Image with GB. (d) Image with JPEG. (e) Image with JP2K.	71
5.3	(a) The gradient magnitude distributions of images with different distortion types. (b) The Weibull parameters of gradient magnitude distributions.	72
5.4	(a) The gradient map of Figure 5.2(a) . (b) The texture map of Figure 5.2(a).	72
5.5	The contribution of the weighting scheme. The four rows show the TLBP map, weighting map, TLBP histogram, and weighted TLBP histogram. (a*) Reference image. (b*) Image with WN. (c*) Image with GB. (d*) Image with JPEG. (e*) Image with JP2K.	75
5.6	(a) One image example. (b) The contrast enhanced version of (a) using linear mapping. (c) Local contrast normalization response of (a). (d) Local contrast normalization response of (b).	77
5.7	Box-plot of SRCC distributions of IQA models on LIVE database.	81
5.8	$SRCC_{wa}$ versus running time (log scale) of different methods.	84
6.1	Squared symmetric neighborhood for different (P, R)	88
6.2	Examples of image blocks with different types and levels of distortions. The distortion types for the four rows are: WN, GB, JPEG, JP2K. (Block size = 128×128 .)	89
6.3	The first principal components of joint log-contrast response of distorted images. (a) WN. (b) GB. (c) JPEG. (d) JP2K. MOS is the subjective quality score of each image.	89

6.4	The SRCC performance vs. the number of GMM components.	95
6.5	Box-plot of SRCC distributions of IQA models on four databases. (a) LIVE. (b) CSIQ. (c) TID2013. (d) CLIVE.	96
7.1	Flowchart of the proposed BoW metric for non-intrusive SQA.	102
7.2	Flowchart of the PLP feature extraction.	103
7.3	Scatter plot of the condition-averaged MOS versus condition-averaged objec- tive scores predicted by the BoW. Each point denotes one of the distortion conditions in the database.	106

List of Tables

3.1	The SRCC comparison for different gradient operators.	33
3.2	The SRCC comparison for different feature extraction methods.	35
3.3	Performance comparison with state-of-the-art FR-IQA metrics. The best three IQA metrics are highlighted in boldface.	37
3.4	Performance comparison with state-of-the-art BIQA metrics. The best three BIQA metrics are highlighted in boldface.	39
3.5	Results of two sample t-test and Wilcoxon rank-sum test performed between SRCC obtained by BIQA methods on two databases. 1(-1) indicates GWH-GLBP is statistically better (worse) than the method in the column. 0 indicates GWH-GLBP is statistically equivalent to the method in the column.	40
3.6	Run-time comparison of BIQA models.	41
4.1	SRCC comparison on LIVE database to validate the complementary of structural and luminance information.	56
4.2	SRCC comparison of BIQA models on seven benchmark databases.	59
4.3	PLCC comparison of BIQA models on seven benchmark databases.	60
4.4	RMSE comparison of BIQA models on seven benchmark databases.	60
4.5	Statistical significance test. 1(-1) indicates NRSL is statistically better (worse) than the method in the column. 0 indicates NRSL is statistically equivalent to the method in the column.	61
4.6	SRCC comparison of BIQA models on individual distortion type.	62
4.7	SRCC comparison on the whole CSIQ database.	63
4.8	SRCC comparison on the whole TID2013 database.	63
4.9	SRCC comparison on cross-database validation.	64

4.10	Computational complexity and run-time comparison of BIQA models.	65
5.1	The effect of the weighting scheme.	75
5.2	Benchmark databases for IQA performance validation.	78
5.3	Performance comparison of twelve IQA models on four benchmark databases. The best two BIQA models are shown in boldface.	80
5.4	Weighted-average performance on four benchmark databases.	80
5.5	SRCC comparison of ten BIQA models on individual distortion type. The best two models are shown in boldface.	82
5.6	SRCC comparison on cross-database validation when models are trained on LIVE.	83
6.1	Overall performance comparison of BIQA methods on four benchmark databases. The best two methods are shown in boldface.	97
6.2	SRCC comparison of BIQA methods on individual distortion type for LIVE and CSIQ. The best two methods are shown in boldface.	98
6.3	SRCC comparison of BIQA methods on individual distortion type for TID2013. The best two methods are shown in boldface.	99
6.4	SRCC comparison on cross-database validation when models are trained on LIVE.	100
7.1	The ITU-T P.Supplement-23 database.	104
7.2	PLCC comparison on the seven databases of ITU-T P.Supplement-23.	107
7.3	RMSE comparison on the seven databases of ITU-T P.Supplement-23.	107

List of Abbreviations

ADM	Additive and Detailed Loss based Measure
AGGD	Asymmetric GGD
BIQA	Blind Image Quality Assessment
BIQES	Blind Image Quality Evaluator based on Scales
BIQI	Blind Image Quality Index
BIQI	Blind Image Quality Index
BLIINDS	Blind Image Integrity Notator using DCT Statistics
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
BSD	BIQA based on Structural degradation
CBIQ	Codebook based Image Quality
CORNIA	Codebook Representation for No-Reference Image Assessment
CSF	Contrast Sensitivity Function
CSIQ	Categorical Subject Image Quality
DCT	Discrete Cosine Transform
DIIVINE	Distortion Identification-based Image Verity and Integrity Evaluation
DMOS	Differential Mean Opinion Score
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DWT	Discrete Wavelet Transform
FR	Full-Reference
FSIM	Feature-Similarity Index
FSIMc	Feature-Similarity Index for Color Images
FT	Fourier Transform
GB	Gaussian blur
GGD	Generalized Gaussian Distribution
GMM	Gaussian mixture model

GSM	Gaussian Scale Mixture
GSSIM	Gradient based Structural Similarity
HVS	Human Visual System
IFC	Information Fidelity Criterion
IGM	Internal Generative Mechanism
IQA	Image Quality Assessment
ITU	International Telecommunication Union
IVC	IRCCyN/IVC Image Quality Database
IWSSIM	Information-Content Weighted Structural Similarity Index
JNB	Just Noticeable Blur
JP2K	JPEG2000 compression
JPEG	JPEG compression
KRCC	Kendalls Rank-Order Correlation Coefficient
LIVE	Laboratory for Image and Video Engineering
LoG	Laplacian of Gaussian
MAD	Most Apparent Distortion
MAE	Mean Absolute Error
MIT	Multimedia Information and Communication Technology
MKL	Multiple Kernel Learning
MOS	Mean Opinion Score
MSCN	Mean Subtracted Contrast Normalized
MSSSIM	Multiscale Structural Similarity Index
MVG	Multivariate Gaussian Model
NIQE	Natural Image Quality Evaluator
NR	No-Reference
NRQI	No-Reference Quality Index
NRSL	NRIQA based on Structural and Luminance features
NSS	Natural Scene Statistics
PC	Phase Congruency
PFT	Phase Fourier Transform
PGN	Additive Pink Gaussian Noise
PLCC	Pearsons Linear Correlation Coefficient
PSNR	Peak-Signal-to-Noise Ratio
QA	Quality Assessment
QAC	Quality Aware Clustering

RMS	Root Mean Squared
RMSE	Root Mean Squared Error
RR	Reduced-Reference
SQA	Speech Quality Assessment
SRCC	Spearman's Rank-Order Correlation Coefficient
SRNSS	Sparse Representation of Natural Scene Statistics
SSIM	Structural Similarity Index
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regressor
TID2008	Tampere Image Quality Database 2008
TID2013	Tampere Image Quality Database 2013
VIF	Visual Information Fidelity
VQEG	Video Quality Experts Group
VSNR	Visual-Signal-to-Noise Ratio
WN	Additive White Gaussian Noise

Chapter 1

Introduction

1.1 Background and Motivation

Recent years have witnessed an explosion of personal and commercial multimedia applications, such as digital television home entertainment systems, mobile multimedia applications, video tele-conferencing, and immersive multimedia applications [1]. The fast development of information and communication technology has brought forward the requirement for high quality of experience (QoE) of end-users [1, 2, 3]. Multimedia signals have been processed from the time of acquisition to the time of human consumption [1]. For example, the multimedia signals need to be compressed to reduce the storage requirement. They also need to be transmitted from one user to another through data links for delivering media to human consumers. And these signals need to be reproduced on TV screen or smart phone for human beings to listen and view [1]. As most multimedia systems and devices are designed for human consumption, the perceptual quality of received signals has become an important factor to the success of these systems [1, 2, 3].

Traditionally, multimedia signals are evaluated by human subjects, which is referred as subjective quality assessment [1]. In order to guarantee the accuracy and reliability of subjective ratings, dozens of human subjects are recruited and trained to evaluate the perceived quality of multimedia signals following strict assessment protocols [1, 3]. Despite the fact that subjective assessment is the most straightforward and reliable way for quality evaluation, there are many

limitations for this method. It's quite expensive and time-consuming, and thus companies and research institutes cannot afford to conduct such experiments for most applications. Also, there are many practical applications that require real-time monitoring and quality prediction, where subjective assessment would become useless for these scenarios [1, 2, 3]. Therefore, subjective assessment methods are generally used to benchmark automatic objective quality assessment models [4].

Objective quality assessment methods use computational models to automatically predict the perceived multimedia quality consistent with human subjective ratings [3]. They are much desired in practical applications, as most multimedia signal processing services and devices are devoted to maintaining or improving the perceived quality for human consumption. Refs [1, 3] have summarized the most widely practical applications where objective methods can be deployed.

First, they can be deployed in quality control systems for quality monitoring. For example, during image or speech acquisition, quality metrics can be deployed to monitor the capture process. During multimedia streaming applications such as voice over internet (VoIP) or high-definition television (HDTV) services, quality metrics can be used for real-time monitoring of streaming signals.

Second, they can be used to benchmark signal processing algorithms and systems. For example, there are many speech and image enhancement algorithms proposed to improve the quality of multimedia signals, and quality metrics can be used to determine the algorithm with the best performance.

Third, they can be embedded into signal processing systems to optimize the algorithms. For example, during image or speech denoising, the quality metric can be used the guidance for optimization, which can provide high-quality output signals.

The wide applications of quality assessment have attracted much research efforts to design objective methods that are consistent with human judgement. During the last decades, various

computational models have been proposed for image and speech quality assessment, and some of them have achieved quite promising performance. However, much room is still available for performance improvement and generalization to wide applications.

Most existing quality assessment methods are full-reference methods, where the quality score is calculated by comparing the distorted signal with the original reference one. In practice, the clean reference signals are not available in many applications, such as transmission, enhancement and reproduction. Thus, the design of no-reference/blind quality assessment methods are more valuable. Although many works have been conducted to develop no-reference methods, their performance is still inferior to that of full-reference methods, due to the diverse contents and distortion types. Thus, more research efforts are required to develop effective no-reference methods.

Currently, most image quality assessment (IQA) studies focus on predicting the visual quality of images distorted by a single distortion. In practice, images available to end users usually undergo several stages, and each stage might introduce certain specific distortion. It is very likely that images are simultaneously distorted by multiple distortions. Compared with images with a single type of distortion, it is more difficult for visual quality assessment of multiply-distorted images. Thus, it is much desired to design IQA methods for this application.

Additionally, most IQA methods are developed to estimate the quality of images degraded by common distortion types, such as JPEG and JPEG 2000 compression, Gaussian white noise, and Gaussian blurring. These distortions are generated by adding graded distortions to the clean reference images. However, images captured and shared by people are usually degraded by various authentic image distortions, which can not be well reproduced by the highly controlled simulation. The performance of most existing IQA methods deteriorate when tested on images with realistic distortions. Thus, more research efforts are required to handle the realistic distortions due to their wide usage.

Research on non-intrusive speech quality assessment (SQA) aims to accurately and automatically estimate the perceptual quality of speech signals without any prior information about

the reference signals. Compared to the research works on IQA, research works on SQA is limited. Generally, speech features are extracted on a frame basis, and the quality score is the scalar rated on the whole speech utterance. How to relate the local feature to the global quality is a fundamental problem, with very limited research efforts in the research community. In existing studies, to construct the global features which represent the characteristics of the whole speech utterance, statistical properties such as mean, variance, skewness and kurtosis across all frames are calculated and combined together as a whole feature vector. However, much information is lost by summarizing the frame-level features using several kinds of statistic moments. Thus, a more elegant way to construct global features from local ones is needed for SQA.

1.2 Objective and Scope

The objective of this thesis is to develop new objective quality assessment models for image and speech signals. In particular, we explore the following problems in this thesis:

- to design effective perceptual features for blind IQA (BIQA).
- to design novel natural scene statistics (NSS) model for BIQA.
- to evaluate images degraded by multiple distortions.
- to evaluate images degraded by realistic distortions.
- to form global features from local features for SQA.

1.3 Thesis Contributions

In this thesis, new computational models of image and speech quality assessment are proposed in the following aspects: distortion-specific BIQA for multiple distortions, general-purpose

BIQA based on structural and luminance features, general-purpose BIQA based on first- and second-order structures, general-purpose BIQA based on novel NSS model, non-intrusive SQA based on bag-of-words (BoW) framework. The main contributions of this thesis are outlined as follows:

- Propose a novel distortion-specific BIQA method for multiple distortions.

In practice, images available to consumers usually undergo several stages of processing including acquisition, compression, transmission and presentation, and each stage may introduce certain type of distortion. It is common that images are simultaneously distorted by multiple types of distortions. Most existing IQA methods are designed to estimate perceived quality of images corrupted by a single image processing stage. We propose a BIQA method to predict the visual quality of multiply-distorted images in gradient domain. The proposed metric achieves better performance than existing full-reference and no-reference methods.

- Propose a novel general-purpose BIQA method based on the human visual system (HVS) characteristics.

The main contribution of this study lies in three folds: first, by exploring the characteristics of the HVS, two new perceptual features are extracted to represent the structural information and luminance changes in distorted images, respectively. Second, we propose a novel perceptual BIQA metric for visual quality prediction. The proposed metric can consistently handle both synthetically-distorted and naturally-distorted images for quality evaluation. Furthermore, the proposed method has the advantages of high prediction accuracy, high generalization capability, low computational complexity and low feature dimension.

- Propose to use first- and second-order structures in general-purpose BIQA.

Research in visual science has evidenced that the human visual cortex has separate mechanisms to process first- and second-order patterns. Therefore, we take into account both the first- and second-order image structures and validate their effectiveness in the IQA tasks. Specifically, the Prewitt linear filters are used to extract first-order image structures and the local contrast normalization is employed to extract second-order image structures. An effective BIQA method is designed by combining these two components in this study.

- Propose a joint NSS model for general-purpose BIQA.

We develop a novel comprehensive NSS model for BIQA. This new model uses multivariate Gaussian mixture model (GMM) to represent the directional log-contrast responses jointly. There are the following four advantages for the proposed method: 1) this is the first attempt to use multivariate NSS model in BIQA. By using multivariate NSS model, the space, orientation and scale correlation can be well captured; 2) this model is built by unsupervised machine learning instead of human expert observation; 3) we propose a novel way to extract quality-aware features from this model which is more accurate and efficient; 4) the proposed BIQA method delivers better performance than existing methods.

- Propose a non-intrusive SQA method.

The fundamental differences among non-intrusive SQA models are determined by the type of frame-specific (local) features, features supplied to the mapping stage (local or global) and regression model used in the mapping. Existing studies generally calculate the mean, variance, and others across all the frames as the global feature. Much information has been lost by summarizing the frame-level features using statistic moments. In contrast, we explore the BoW model for non-intrusive speech quality estimation.

1.4 Thesis Organization

The organization and major content of this thesis is illustrated in Figure 1.1. The thesis consists of eight chapters and each chapter is summarized as follows.

Figure 1.1: The organization of this thesis.

Chapter 1 gives a brief introduction to the motivation and background of this thesis, and summarizes the major contents and contributions.

Chapter 2 is the literature review to get a broad view of the current research status of the related topics. We introduce the subjective and objective quality assessment methods on image and speech signals, with an emphasis on blind methods, as they are most relevant to our works.

Chapter 3 proposes a novel distortion-specific BIQA method for multiple distortions. The new method is based on the HVS's sensitivity to structural degradation. A novel effective structural feature is proposed to capture the influence of multiple distortions on image perceptual quality.

Chapter 4 describes a general-purpose BIQA method by extracting statistical structural and luminance features. We demonstrate that complementary information provided by extracted

structural and luminance features plays an important role in image quality estimation. We then propose a novel perceptual BIQA metric for visual quality prediction based on these features.

Chapter 5 presents a general-purpose BIQA method by extracting the first- and second-order structural features in spatial domain and combining them together for quality estimation. We show that the two structural features are both affected by distortions and have significant contribution to quantify the visual quality of distorted images.

Chapter 6 investigates the effectiveness of joint NSS model in general-purpose BIQA. We make the first attempt to build joint NSS model and apply it to BIQA.

Chapter 7 proposes a non-intrusive SQA algorithm. We treat speech utterance as a document and local features extracted on per-frame basis as words and employ the bag-of-words model for speech quality-aware feature extraction.

Chapter 8 concludes the thesis with a summary of the aforementioned research works and provides the possible research directions in the future.

Chapter 2

Literature Review

Since most multimedia services are provided for human consumption, it's important to guarantee that customers receive a satisfactory experience. Humans can easily judge the quality of multimedia signals, based on the visual/auditory perception. However, it is not so straightforward for the computer to make similar judgement. The objective quality assessment that rivals the human perception is still challenging but meaningful work.

The traditional and intuitive way to quality assessment is implemented by conducting subjective tests. Following some protocols, dozens of subjects are recruited to give their opinions on the test stimuli, and their ratings are averaged to show the quality levels. Thus, subjective quality assessment produces the average of human opinions, which shows the quality level of test stimuli that most people agree on. The results of subjective tests are in the form of MOS (mean opinion score) or DMOS (differential mean opinion score). To make the subjective results usable, many conditions should be considered. Some standardized protocols on how to conduct subjective tests have been designed and recommended by ITU (International Telecommunication Union) to make the results more reliable [2, 5].

Although subjective quality assessment is supposed to provide the most accurate and reliable quality scores, it has many limitations. It's quite expensive and time-consuming, and thus many research institutes or companies cannot afford to conduct such experiments for most applications. Also, it cannot be built in many practical applications that require real-time mon-

itoring and quality prediction. Therefore, objective methods that are in high agreement with subjective ratings are more meaningful and valuable.

In this chapter, we give a literature survey on quality assessment works of image and speech signals, including subjective and objective methods.

2.1 Quality Assessment of Image Signals

2.1.1 Subjective quality assessment protocols

The ITU has standardized protocols on how to conduct subjective viewing tests on image signals [2, 5]. We give a brief introduction to the standardized subjective assessment methods here.

1) Double Stimulus Impairment Scale (DSIS): In this method, the reference images are first presented to the subjects, and then the subjects rate the degraded images' quality with respect to the reference ones. Generally, there are five discrete levels for the rating scale: 1 - very annoying, 2 - annoying, 3 - slightly annoying, 4 - perceptible but not annoying, 5 - imperceptible. This method is usually used to evaluate clearly visible distortions.

2) Double Stimulus Continuous Quality Scale (DSCQS): In this method, the reference and degraded image signals are presented to the subjects with random order. The subjects are unknown whether the presented signals are the reference or degraded ones. The rating scale is a continuous scale ranging from 0 - 100. This method is usually used to evaluate degraded signals with similar quality to the reference ones.

3) Single Stimulus (SS): In this method, only one image is presented to the subjects each time. The evaluation can be conducted on category assignment or numerical scales.

As mentioned before, the subjective viewing test is laborious, expensive and cannot be embedded in many practical applications [2, 5]. Objective IQA metrics that aim to replace human judgement of perceived visual quality have attracted quite a lot of research efforts.

2.1.2 Objective IQA methods

Generally, objective IQA metrics can be categorized into full-reference (FR), reduced-reference (RR) and no-reference (NR) algorithms according to the availability of the original images. For FR methods, both the reference and distorted images are available for quality evaluation. For RR methods, partial information of the reference images is available in the form of extracted features embedded into watermark or transmitted through auxiliary channel. For NR methods, only the distorted images is provided for visual quality estimation. NR-IQA is also called BIQA as the methods are blind to the reference images. The classification of existing objective IQA methods is illustrated in Figure 2.1.

Figure 2.1: The classification of existing objective IQA methods.

Early research works mainly focus on FR methods. Generally, the development of FR algorithms can be handled in two broad approaches: the human vision model approach and the engineering approach [2]. The first approach aims to mimic the characteristics of the HVS from the eye optic nerves to the visual cortex. The HVS models are usually established based on psychovisual and physiological experimental results. Two representative HVS models are VDP (visual difference predictor) [6] and the Sarnoff JND (just noticeable difference) [7]. The HVS models are attractive as they are built by the solid theories. However, they suffer from some drawbacks due to the high computational complexity and the current unknown knowledge of the human perceptual mechanisms [2, 3].

The engineering approach aims to calculate the structure or information content preserved in the distorted image compared with the original one. Many state-of-the-art IQA metrics fall under this category and can achieve good performance in terms of high correlation with subjective quality scores. The mean squared error (MSE) and its corresponding peak signal-to-noise ratio (PSNR) are the simplest and most widely used FR metrics due to their efficient computation and clear physical meanings. However, they are not in high agreement with human perceptual quality. In the past decades, to overcome the drawbacks of these traditional methods, various perceptual FR-IQA metrics have been proposed by using the HVS characteristics as guidance for the method design. The well-known structural similarity (SSIM) metric is designed based on the idea that the perceived quality can be quantified by the local structural changes [8]. An information fidelity criterion (IFC) metric is proposed to calculate the mutual information between the reference and distorted images in [9]. The visual information fidelity (VIF) quantifies the loss of image information to the final quality score [10]. The most apparent distortion (MAD) metric adaptively combines a detection based strategy and an appearance-based strategy to predict visual quality by including both the properties of the HVS and the structural difference calculation [11]. A feature similarity (FSIM) Index is proposed in [12] based on the philosophy that the changes of low-level features can be a good indicator of the final quality. The gradient similarity (GS) method by combining the gradient information, luminance similarity and visual masking leads to an effective scheme for quality evaluation [13]. For gradient magnitude similarity deviation (GMSD) metric [14], the standard deviation of the gradient magnitude similarity map is taken as the objective quality score.

The RR methods aim to extract several quality-aware features that are sensitive to a broad range of distortions. It serves as a compromise of NR methods when they cannot achieve satisfactory performance for practical usage. Global features are more desirable for RR methods as they are more compact and require less transmission overhead. Some RR methods are designed using the characteristics of natural images. Such methods are built based on the assumption

that natural scenes exhibit some statistical properties and most distortions disturb such statistics. In [15], the deviation of marginal wavelet coefficient distribution from reference images is used to quantify image quality. In [16], the changes of reorganized statistics in DCT coefficients are measured to estimate the impact of distortions. In [17], the correlation between wavelet subbands is analyzed and utilized to build a RR method based on multiscale geometrical analysis. In [18], the difference of scaled entropies of wavelet coefficients is used to predict the distortion severity of test images. Moreover, by incorporating the internal generative mechanism (IGM) process to separate image content into main contents and residuals, a RR metric is built to estimate visual quality in [19].

Although state-of-the-art FR and RR methods can achieve good performance in accordance with human subjective ratings, in most practical applications (e.g., image transmission, denoising, enhancement), the reference image is not available. Thus, it is necessary to develop effective BIQA methods from practical perspectives.

According to the application range, existing BIQA methods can be classified as distortion-specific and general-purpose methods. The former ones target at specific application scenarios, such as Gaussian blur [20], compression [21, 22], contrast change [23] and multiple distortions [24]. In such cases, the distortion type of the corrupted image is given a priori. This constraint would limit the application domain of such methods, which also cannot handle online images whose distortion types are unknown. By contrast, general-purpose methods can evaluate perceptual quality of images degraded by various distortion types. In this thesis, we propose one distortion-specific BIQA method for multiple distortions, and three general-purpose BIQA methods for all kinds of distortions. Thus, we will give a detailed review on BIQA methods for general-purpose as well as multiple distortions.

2.1.2.1 General-purpose BIQA methods

General-purpose BIQA methods can be classified into three categories according to the features representing image quality, i.e., NSS based methods, perceptual feature based methods

and feature learning methods. The choice of quality-aware features is the determinant factor to the success of BIQA methods. A good set of NR features should be some perceptually relevant features that are sensitive to a broad range of image distortions and robust to image content variation. NSS based methods constitute the most majority of existing general-purpose BIQA methods. The underlying philosophy of NSS based methods is that natural images possess certain statistical regularity, regardless of different contents, and distortions introduced in different image processing stages would change such regularities. Therefore, the deviation of the statistics of a query image from NSS model would indicate the distortion severity of that image. A number of NSS models in different spatial or transform domains have been incorporated into representative BIQA methods. BIQI (blind image quality index) is a wavelet NSS based method that extracts generalized Gaussian distribution (GGD) parameters from wavelet subband coefficients [25]. The exponential decay characteristics (EDC) of wavelet energies across scales have been utilized in SRNSS (sparse representation of NSS feature) method [26]. MKL (multiple kernel learning) extracts statistical features of wavelet marginal distribution, mutual information between neighboring coefficients and EDC properties as quality-aware features [27]. In DIIVINE (distortion identification-based image verity and integrity evaluation), the divisively normalized wavelet subbands are used to build GGD models in wavelet domain. Furthermore, to account for the correlation across scale and orientation, GGD models are fitted on the stacked subband coefficients with the same orientation or scale. Structural similarity between subbands is also calculated to measure their correlation. C-DIIVINE (Complex DIIVINE), a complex extension of DIIVINE, has adopted three NSS models in complex wavelet domain, including the complex GGD model of wavelet coefficient magnitudes, the GGD model of relative magnitudes and the wrapped Cauchy distribution (WCD) model of relative phases [28]. BLIINDS (blind image integrity notator using DCT statistics) extracts the contrast, structure and anisotropy features from block DCT coefficients [29]. Later, it was extended to BLIINDS2 by using GGD to fit the block DCT coefficients and some partitions of

the DCT block [30]. The mean subtracted contrast normalized (MSCN) coefficients are used to build the spatial domain NSS models, including GGD of MSCN coefficients and asymmetric GGD (AGGD) of pairwise MSCN coefficient products [31]. Distributions of neighbouring pixel differences in four color channels (hue, saturation, opponent angle and spherical angle) are used to build color NSS models to account for spatial correlation in color descriptors [32]. There are also some methods using an ensemble of NSS models from multiple domains to extract a more comprehensive feature set. In DESIQUE (derivative statistics-based image quality evaluator), GGD distributions of pixel-wise and pairwise log-derivative statistics in both spatial and log-Gabor domains are combined to form the quality-aware features [33]. ILNIQE (integrated local natural image quality evaluator) adopts the GGD of MSCN coefficients, AGGD of MSCN products, GGD of directional gradients, Weibull distribution of gradient magnitudes in luminance, log-Gabor and opponent color space, as well as the Gaussian distribution of logarithmic-scale opponent color space [34]. In FRIQUEE (feature maps based referenceless image quality evaluation engine), dozens of feature maps over several transform domains and three color spaces are used to fit the GGD, AGGD and WCD models [35].

Another kind of features used by BIQA methods are perceptual features. The gradient and phase congruency features are extracted from distorted images to assess image quality [36]. In [37], a large number of low level features, texture statistics and noise/blur estimates are combined by multiple SVR to build the LBIQ (learning based blind image quality measure) metric. In [38], the joint statistics of image gradients and Laplacian of Gaussian (LOG) responses are employed as quality-aware features. A combination of free energy principle based features and spatial NSS features has shown some impressive results in [39]. The third kind of BIQA methods directly use raw local image patches as the input and learn the quality-aware representation in an unsupervised [40] or supervised [41] way. Commonly adopted regression modules in learning based methods include neural network [36], SVR [31, 38, 42] and random forest regression [43].

Despite the prominent improvement in development of BIQA methods recently, the performance of existing methods still lags behind that of the FR-IQA methods. BIQA is an intrinsically challenging task due to the diverse distortion types and wide range of image content. There is still much room for the performance improvement before they are applied for practical applications. Although NSS models in different domains have been applied in previous BIQA methods, we can observe that the construction and utilization of NSS models follow the same framework. Generally, the marginal distributions of some bandpass coefficients are used to fit a univariate probabilistic model, then the model parameters are adopted as the quality-aware features. The most widely used natural image properties for BIQA methods include the non-Gaussian distribution of bandpass responses, and the correlation of bandpass responses across space, scale and orientation. The first kind of properties, i.e., the distribution of bandpass responses is easy to follow, and univariate distribution models have been adopted for it. For the across space correlation, the marginal distribution of pairwise products or differences has been adopted in several methods. For across scale and orientation correlation, existing BIQA methods calculate mutual information, structural correlation or energy ratio between neighboring subbands to measure the correlation.

Thus, we summarize the limitations of existing NSS based BIQA methods in four aspects: 1) only univariate probabilistic model has been adopted by previous NSS based BIQA methods; 2) there is no probabilistic model to account for the across scale and orientation correlation; 3) for measuring the correlation, only two neighboring pixels or subbands are considered each time; 4) using model parameters to represent the distribution without considering the fitting errors may result in an inaccurate representation.

Moreover, most existing BIQA methods suffer from some limitations. As one of the pioneering methods, BIQI generally gives poor performance due to its simplistic feature extraction [25]. In DIIVINE [42], the used large number of features extracted from multi-subbands in wavelet domain slows down the method greatly. Similarly, the block-based GGD fitting in

DCT domain makes BLIINDS2 difficult to predict the visual quality in real-time [30]. Feature learning based methods, CORNIA [40] and QAF [43], inevitably require a large size code-book and result in high-dimension global quality-aware features. Although GMLOG delivers promising performance on singly-distorted image database, its performance deteriorates considerably when tested on multiply-distorted images [24]. The combination of several RR and FR methods adopted by NFERM [39] leads to a too complicated method to be used in practice.

2.1.2.2 BIQA methods for multiple distortions

Currently, most IQA studies focus on predicting the visual quality of images distorted by a single distortion. In practice, images available to end users usually undergo several stages, and each stage might introduce certain specific distortion. It is very likely that images are simultaneously distorted by multiple distortions. Figure 2.2 shows the potential sources of signal distortion in an image communication system. Compared with visual quality assessment for images with a single type of distortion, it is more difficult for quality assessment for multiply-distorted images.

Figure 2.2: A typical image communication system which consists of several processes. Each process may introduce some characteristic artifacts.

The authors in [44] conducted subjective experiments to investigate the joint effects of blur and noise on image quality. They found that the perceived noise is more pronounced when added to a more blurry image. When the image is highly blurred, increasing the noise strength makes the image sharper. Moreover, they reported that the overall quality decreases with the increased blur or noise. Later on, researchers in [45] studied the joint effects of

wavelet distortion and noise. They found that with the same MSE, the perceived degradation induced by wavelet distortion is 4-5 times greater than that induced by the noise. Furthermore, they observed that adding a small amount of noise to a structurally distorted image would increase the overall quality. The authors in [46] pointed out that the distorted image with multiple distortions is a big challenge for IQA research and summarized the challenges in three aspects: the influence of individual distortions on image quality, the interaction between these distortions, and the joint effects of these distortions on the overall quality. The author also pointed out that a physical combination of individual distortion measures may not well explain the final overall quality. The two newly established subjective databases on multiply-distorted images, MLIVE (LIVE Multiply Distorted Image Quality Database) [47] and MDID2013 [48], challenge most existing objective IQA methods and confirm Chandler's statement. Thus, it is much desired to design visual quality methods for images distorted by multiple types of distortions.

Recently, there are several studies focusing on designing objective IQA methods for multiply-distorted images. The authors in [49] proposed an objective BIQA method based on several image processing blocks to mimic the perceptual process of the HVS. Specifically, the noise strength is first estimated, followed by blur and JPEG metrics deployed on the denoised image. Then the overall quality is computed by combining the estimated results of noise, blur and JPEG metrics. This work was later extended by adding a free energy term to account for the possible interaction of different distortions [48]. In [50], NSS features from three existing BIQA methods are combined to form an improved bag-of-words representation for quality prediction. Despite the progress in this research area, there are still some problems with these existing methods. These hybrid methods based on the combination of several existing IQA metrics would induce more computational overhead. Furthermore, these methods are tailored for the distortion types in the databases, and the generality to other kinds of multiple distortions is open to question.

2.2 Quality Assessment of Speech Signals

Apart from image signals, communication with sound signals is also a wide, effective and convenient means of exchanging information [51]. It has wide applications in multimedia domain, including streaming audio, communication through mobile phones, VoIP, digital audio broadcasting (DAB) and audio compression. Thus, it is also important to evaluate speech quality in multimedia systems.

2.2.1 Subjective quality assessment protocols

Subjective tests of SQA can be classified into two large categories: utilitarian test methods and analytic test methods [52]. In utilitarian test scenarios, subjects assess the speech signals using one overall rating, while in analytic test methods, subjects are asked to focus on given quality features and provide the corresponding judgment [52]. All quality feature ratings can share the same rating scale or have different rating scales.

Utilitarian test methods are designed to assess the integral quality of speech signals perceived by end users. In these tests, subjects are asked to give a one-dimension score to represent the overall impression of the speech signal under test. Thus, the overall quality score is the complete judgment of all the perceptual features of test speech signals [52].

Sometimes, using an integral number to represent the signal quality is not sufficient to compare different speech processing systems and devices. For example, two communication devices may have the same overall quality but with totally different characteristics. Analytic test methods give detailed information about different aspects of the test signals. Two representative analytical test methods are diagnostic acceptability measure (DAM) and evaluation of noise suppression methods. DAM method assesses several quality aspects of speech signals [53]. Finally, an overall quality score can be attained through a linear regression of the individual quality feature scores. The ITU-T has designed P.835 standard for subjective evaluation of speech enhancement algorithms [54]. The methodology uses three five-point scales

to assess the quality of speech signal, the background signal, and the combined signal [54]. This setting is designed to reduce the subject's uncertainty as to which component to assess and lead the subject to integrate the effects of both the speech and background distortions in making their ratings of overall quality [54].

As aforementioned, although subjective methods are the most accurate and reliable way for SQA, objective methods which use computer programs to estimate human judgment of perceived quality are more meaningful and practical [52]. A literature review of existing objective speech quality metrics is described below.

2.2.2 Objective SQA Methods

Objective SQA methods can be classified to signal based methods and parametric based methods [55]. Signal based methods estimate the speech quality through the analysis of the properties of speech signals. Parametric models use the parameters of compression and transmission protocols to estimate the final quality.

Furthermore, signal based methods can be categorized into intrusive (also called double-ended, full-reference) methods and non-intrusive (also called single-ended, no-reference) methods based on the availability of reference speech signals [51, 55]. Intrusive quality assessment methods have evolved from the calculation of physical distance to perceptual distance. We introduce several representative intrusive SQA methods here. The simplest physical distance based metric is signal-to-noise-ratio (SNR) which calculates the distance between the reference and distorted speech signals. Despite its clear physical meaning and low computational complexity, this metric doesn't correlate well with human subjective ratings. The more advanced human auditory perception based model, perceptual speech quality measure (PSQM), was standardized as the first ITU-T recommendation P.861 for intrusive speech quality assessment [56]. In the pre-processing stage, the reference and degraded speech signals are first equalized to a standard listening level, and then filtered by a intermediate reference system (IRS) filter to

mimic the response of telephone headset. The signals are aligned in time to make sure the corresponding signal parts are compared. Then they are transformed to perceptual representations to obtain the loudness spectra. Finally, the difference in loudness spectra is computed and averaged over time and frequency to obtain the objective quality score. The second ITU-T recommendation P.862, perceptual evaluation of speech quality (PESQ), has further improved the quality prediction accuracy by including the variable delay identification and asymmetric effects, i.e., components existing in reference signal while absent in the degraded signal have different scores to the opposite situation [57]. Although PESQ is the most widely used tool for speech quality estimation currently, it cannot provide accurate quality prediction in several situations, such as noise suppression and deverbation applications, low bit-rate codecs, etc. Thus, the third ITU-T recommendation P.863, perceptual objective listening quality assessment (POLQA), is proposed to address these problems [58].

Non-intrusive SQA methods are preferable in real-time applications, as the original speech signals are not always available in practical applications [55], such as VoIP, wireless communication applications, speech noise suppression process, communication systems of real-time monitoring and speech processing automation, etc. At present, most SQA metrics are intrusive and there are much less non-intrusive SQA metrics. Due to the wide applications, non-intrusive speech quality metrics have attracted much attention in the research community recently.

There are two types of non-intrusive SQA metrics: one is to construct an artificial clean reference speech sample and then treat the quality assessment problem as a comparison-based intrusive quality evaluation; while the other is to extract quality-aware features from the distorted speech signals and map the features to the subjective quality scores. For the later methods, they include the feature extraction and feature pooling subsystems, corresponding to the two processes in the human evaluation: conversion of the perceived speech signals into auditory nerve excitations, and cognitive processing in the human brain [55].

Oscar et al. [59] propose a non-intrusive metric based on the analysis of spectrogram features for wireless communication scenario. The method in [60] uses the perceptual spectral

density distribution features and adaptive neuro-fuzzy network to calculate the quality score. Falk et al. [61] measure the objective quality scores of the degraded speech signals by the artificial reference model constructed using GMM. A machine learning based non-intrusive metric which can also be extended to FR method has been proposed in [62]. In [63], several speech production and auditory perception features are combined for quality estimation using GMM training. Previous works have investigated MFCC (Mel-frequency cepstral coefficient) features [64], Mel-Filtered Energies [65] and SVR for speech quality evaluation.

In 2004, the ITU-T has released the standard P.563 for non-intrusive SQA , which is composed of Psytechnics's NiQA algorithm, SwissQual's NiNA and Opticom's P3SQM [66]. P.563 is constructed as a vocal tract model, which first identifies the distortion type and then estimates the quality score by comparing the distorted signal with the reconstructed reference signal [66]. Recent studies found that P.563 works well for quality evaluation of speech communication and audio coding, but performs poorly under many other applications such as speech enhancement, dereverberation process, etc [55].

Chapter 3

Blind Quality Assessment for Multiply-Distorted Images in Gradient Domain

In this chapter, we propose a novel distortion-specific BIQA method to estimate visual quality of images degraded by multiple distortions. Generally, multiple distortions are introduced into images when they undergo more than one image processing stages. It's a practical problem since images presented to consumers are more likely distorted by multiple distortions. It's also a challenging problem as there may be interaction and joint effects among different distortions. Most existing IQA methods cannot yield good performance for multiply-distorted images, including both FR and NR methods. Therefore, we design an effective BIQA method to deal with this problem.

The proposed method operates in gradient domain, and there are two main reasons. First, gradient features are quite effective to reflect the influence of various distortions. Many IQA methods use gradient features as one essential ingredient. Second, blurring artifacts are more apparent in the gradient domain. Blurring is one of the most common distortion of multiply-distorted images, which is generally introduced in the first step (image acquisition). This method is designed based on the HVS's sensitivity to structural degradation. A novel effective structural feature has been proposed to capture the complex degradation pattern induced by

multiple distortions. Moreover, we fuse the contrast and structural features together to achieve more quality-awareness.

3.1 The Proposed Method

Proper extraction and description of image structural information play a significant role in perceptual quality assessment [67]. Here, we propose a novel effective structural feature extraction method as the gradient weighted histogram of local binary pattern (LBP) calculated on the gradient map (GWH-GLBP). First, the distorted image is filtered by Prewitt operator to extract the gradient features. Then LBP is utilized to encode the image primitive micro-structures, such as edges, lines, spots and other local features in the gradient image to form the GLBP (Gradient LBP). After that, instead of adopting the widely used frequency histogram to describe the global structural feature, we propose to accumulate the gradient magnitudes of pixels with the same GLBP pattern, which can be regarded as the gradient weighted GLBP histogram. Finally, SVR is used to map extracted perceptual features to image quality.

3.1.1 GLBP construction

Image gradient has been employed for IQA in different ways. For FR-IQA metrics, the gradient similarity between reference and distorted image is usually calculated and combined with other feature similarity for quality evaluation. FSIM combines gradient similarity and phase congruency similarity by using phase congruency based pooling [12]. GS metric combines the gradient similarity, luminance similarity and visual masking for quality evaluation [13]. GMSD directly adopts the standard deviation of the gradient similarity map as quality score [14]. The gradient information also has been employed as the weighting scheme to replace the traditional average weighting to enhance FR-IQA performance [68]. For RR-IQA, the Weibull distribution parameters of image gradients are utilized as RR feature [69]. For BIQA, the distribution

of image gradient in the form of curve parameters[34] or histogram[38] has been utilized to construct image quality-aware representation.

The gradient magnitudes of digital images are defined as the square root of image directional derivatives along two orthogonal directions. The gradient is usually computed by convolving an image with a linear filter such as the classic Prewitt, Sobel and Scharr filters. In this work, we adopt the Prewitt filters to compute image gradients due to its computational simplicity. By using other filters such as the Sobel and Scharr filters, similar results are observed. First, the gradient information is computed by convolving images with Prewitt filters defined as follows.

$$\mathbf{p}_x = \begin{bmatrix} 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \end{bmatrix}, \quad \mathbf{p}_y = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 \\ -1/3 & -1/3 & -1/3 \end{bmatrix} \quad (\text{Eq. 3.1})$$

The gradient magnitudes of a distorted image are computed by horizontal and vertical gradients as follows:

$$\mathbf{g}(i) = \sqrt{(\mathbf{d} * \mathbf{p}_x)^2(i) + (\mathbf{d} * \mathbf{p}_y)^2(i)}, \quad (\text{Eq. 3.2})$$

where symbol “*” denotes the convolution operation; $\mathbf{p}_x$ and $\mathbf{p}_y$ are the Prewitt filters; $\mathbf{d}$ and $\mathbf{g}$ denote the distorted image and its corresponding gradient magnitude map with i as the location index.

After that, we calculate the LBP for each pixel in the gradient magnitude map. We argue that the performance of LBP for IQA can be greatly improved through Prewitt operator prior to feature extraction. As LBP extracts the circular binary gradients at a micro-level, Prewitt operator provides the function of enhancing this gradient information accordingly. The LBP operator [70] is proposed to describe the relationship between the center pixel and its surrounding neighbors by computing gray-level differences. By applying the LBP operator on the gradient magnitude map, the GLBP code at one location is deduced as

$$GLBP_{P,R} = \sum_{i=0}^{P-1} s(g_i - g_c)2^i, \quad (\text{Eq. 3.3})$$

where P is the number of neighbors and R is the radius of the neighborhood. g_c and g_i are the gradient magnitudes at the center location and its neighbor.

The thresholding function $s(\cdot)$ is defined as

$$s(g_i - g_c) = \begin{cases} 1, & g_i - g_c \geq 0 \\ 0, & g_i - g_c < 0. \end{cases} \quad (\text{Eq. 3.4})$$

To achieve rotation invariance, a locally rotation invariant uniform GLBP operator can be defined as

$$GLBP_{P,R}^{riu2} = \begin{cases} \frac{\sum_{i=0}^{P-1} s(g_i - g_c)}{P + 1} & \text{if } \mathcal{U}(GLBP_{P,R}) \leq 2 \\ \text{else,} & \end{cases} \quad (\text{Eq. 3.5})$$

where $\mathcal{U}$ is the uniform measure, and superscript *riu2* denotes the rotation invariant “uniform” pattern with $\mathcal{U}$ value less than 2. The uniformity measure $\mathcal{U}$ is calculated as the number of bitwise transitions

$$\begin{aligned} \mathcal{U}(LBP_{P,R}) &= \|s(g_{P-1} - g_c) - s(g_0 - g_c)\| \\ &+ \sum_{i=0}^{P-1} \|s(g_i - g_c) - s(g_{i-1} - g_c)\|. \end{aligned} \quad (\text{Eq. 3.6})$$

It was observed that uniform LBP, which contains at most two spatial bitwise transitions, increases the discriminative ability of LBP [70]. The rotation invariant uniform GLBP would have $P + 2$ patterns, which describe the distinct local gradient structures.

GLBP describes the inter-pixel relationship in an image neighborhood, and such image microstructural patterns are effective to capture the complex degradation by various distortions. Different GLBP patterns denote different local gradient patterns. For example, in the case of $GLBP_{8,1}$, (0) stands for isolated gradient spot; (8) denotes smooth region, and (1-7) represents edges of varying positive and negative curvatures. The introduction of distortions may shift the GLBP pattern from one type to another. The first row of Figure 3.1 shows one pristine image patch and its three distorted versions; while the second row shows their corresponding GLBP maps. As can be seen in Figure 3.1, different distortions change the GLBP patterns in their own characteristic ways, making it an effective measure to describe the impact of various distortions.

Figure 3.1: GLBP maps under different multiple distortions. (a) Pristine image. (b) Image with GB+JPEG. (c) Image with GB+WN. (d) Image with GB+JPEG+WN. (e), (f), (g), (h) are the corresponding GLBP maps of images in the first row.

3.1.2 Gradient weighted GLBP histogram

GLBP is derived from the difference between center pixel and its neighbors in the gradient magnitude map, and thus it is invariant to the center pixel value (i.e., gradient magnitude at that location). Furthermore, GLBP only encodes the sign of the difference to represent local pattern, and excludes the magnitude of the difference; this mechanism makes GLBP unable to differentiate a weak contrast local pattern from a strong one. However, local contrast change has a significant impact on the perception of image quality. Image gradient magnitude is an effective measure to encode contrast information and the HVS is highly sensitive to it. To effectively capture the contrast and structural information in an image, we propose to fuse the structural and gradient information into a single representation. Specifically, we propose to accumulate the gradient magnitudes of pixels with the same GLBP pattern, which can be regarded as the gradient weighted GLBP histogram:

$$h_{glbp}(k) = \sum_{i=1}^N \omega_i f(GLBP_{P,R}(i), k), \quad (\text{Eq. 3.7})$$

$$f(x, y) = \begin{cases} 1, & x = y \\ 0, & \text{otherwise,} \end{cases} \quad (\text{Eq. 3.8})$$

where N denotes the number of image pixels, $k \in [0, K]$ is the possible GLBP pattern, and ω_i is the weight assigned to the GLBP code. In this work, we use the gradient magnitude calculated by (Eq. 3.2) as the GLBP weight of each pixel. The stronger the pixel contrast, the larger the weight assigned to the GLBP code of that pixel. In such a way, we emphasize image regions with high contrast changes and take into account the structural and contrast information in a single representation. For the traditional frequency histogram, ω_i at each location is the same.

3.1.3 Regression model for quality prediction

After extracting the quality-aware features to describe the impact of various distortions, we adopt a regression module to map these features to the subjective ratings. The SVR with radial basis function (RBF) kernel is adopted as the mapping function in this work [71]. Considering a set of training data $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\}$, where $\mathbf{x}_i \in R^n$ is the extracted quality-aware feature and y_i is the corresponding DMOS, given parameters $C > 0$ and $\epsilon > 0$, the standard form of SVR is represented as: [71]

$$\min_{\mathbf{w}, b, \xi, \xi^*} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \left\{ \sum_{i=1}^l \xi_i + \sum_{i=1}^l \xi_i^* \right\} \quad (\text{Eq. 3.9})$$

$$\text{subject to} \quad \mathbf{w}^T \phi(\mathbf{x}_i) + b - y_i \leq \epsilon + \xi_i, \quad (\text{Eq. 3.10})$$

$$y_i - \mathbf{w}^T \phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i^*, \quad (\text{Eq. 3.11})$$

$$\xi_i, \xi_i^* \geq 0, i = 1, \dots, l. \quad (\text{Eq. 3.12})$$

where $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is the kernel function. We use the RBF kernel with the kernel function of $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$. The parameters $\{C, \gamma\}$ are selected through cross validation on the training data. Once the regression model is learned, we can use it to estimate the perceptual quality of any input image.

3.1.4 Comparison with other NR features

The difficulty to find a single best representation for perceptual subjectivity is due to the diversity of image contents and variety of image distortion types. A good set of NR features should be some perceptually relevant features that are sensitive to a broad range of image distortions and robust to image content variation. To illustrate the effectiveness of the proposed feature set for BIQA, we show the 2D embedding of extracted feature by using t-SNE (stochastic neighborhood embedding) algorithm [72] and compare with the feature set used by other BQA methods. t-SNE is an unsupervised nonlinear dimension reduction method that converts high-dimensional data to two-dimensional data that can be visualized in a scatterplot. It can well preserve the significant structure of high-dimensional data, such that similar data points in the high-dimensional feature space tend to lie close together in the low-dimensional embedding. The competing features include BIQI [25], DIIVINE [42], BLIINDS2 [30], CORNIA [40], BRISQUE [31], GMLOG [38] and ILNIQE [34].

The scatter plots of the resulting 2-D embedding are shown in Figure 3.2, where each dot represents one image in the MLIVE database, which is shape coded by distortion types and color coded by DMOS values. From the scatter plots shown in Figure 3.2, we can draw several conclusions. 1) GWH-GLBP can distinguish the two combinations of multiple distortions quite well. Only when the images are mildly degraded, the two distortion types tend to mix together. Also, GWH-GLBP shows high quality relevance, with the observation that images of similar quality scores are grouped together and images of distant quality scores are separated from each other. 2) BIQI, DIIVINE, and ILNIQE cannot discriminate the two distortion combination types and show moderate quality relevance. Distorted images are sequentially grouped according to their quality in some extent. 3) GMLOG can categorize the distortion combination types roughly, however, it has low quality relevance and distorted images with different quality levels are clustered together. 4) BLIINDS2, CORNIA and BRISQUE exhibit moderate performance for both distortion type identification and quality relevance. These observations reveal

Figure 3.2: Scatter plots of t-SNE 2D embedding of feature space of representative BIQA models on MLIVE database. The distortion types are coded by shape and the quality scores are coded by color.

the advantages of the proposed GWH-GLBP method over other BIQA methods and explain the acceptable performance for BLIINDS2, CORNIA and BRISQUE.

3.2 Performance Evaluation

3.2.1 Experimental protocols

The comparison experiments are conducted on two multiply-distorted image databases (i.e., MLIVE, MDID2013). The MLIVE database [47] consists of two subsets. The first subset includes 15 reference images distorted by GB+JPEG (Blur followed by JPEG compression). The second subset includes the same 15 reference images distorted by GB+WN (Blur followed by Gaussian white noise). There are 450 distorted images in total. Subjective quality scores are in the form of DMOS ranging from 0 to 100. The MDID2013 database [48] consists of images successively corrupted by three kinds of distortions (GB+JPEG+WN). There are a total of 324 distorted images generated from 12 reference images. Subjective quality scores are in the form of DMOS ranging from 0 to 1.

It is advised in [73] that a monotonic logistic function can be used to provide a nonlinear mapping between objective scores and subjective scores.

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{\exp(\beta_2(x - \beta_3))} \right) + \beta_4 x + \beta_5 \quad (\text{Eq. 3.13})$$

where x is the original IQA score, $f(x)$ is the fitted IQA score, $\beta_j (j = 1, 2, \dots, 5)$ are regression parameters trained per database.

The performance of IQA metrics can be evaluated by four different criteria: Spearman rank order correlation coefficient (SRCC) and Kendall rank order correlation coefficient (KRCC) for prediction monotonicity; Pearson linear correlation coefficient (PLCC) and root mean squared error (RMSE) between the subjective scores and predicted scores for prediction accuracy. The latter two criteria are calculated after the monotonic logistic mapping.

SRCC is defined as:

$$\text{SRCC} = 1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)} \quad (\text{Eq. 3.14})$$

where d_i is the difference between the ranks of i th image sample in subjective and objective scores.

KRCC is defined as:

$$\text{KRCC} = \frac{n_c - n_d}{n(n-1)/2} \quad (\text{Eq. 3.15})$$

where n_c is the number of times that objective method ranks two images in agreement with subjective rankings, and n_d is the number of times when objective method ranks two images in contradiction to subjective rankings.

PLCC is defined as:

$$\text{PLCC} = \frac{\sum_{i=1}^N (w_i - \bar{w})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (w_i - \bar{w})^2 \sum_{i=1}^N (y_i - \bar{y})^2}} \quad (\text{Eq. 3.16})$$

where $\bar{w}$ is the mean of the subjective MOS w_i and $\bar{y}$ is the average of the objective prediction y_i .

RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (w_i - y_i)^2} \quad (\text{Eq. 3.17})$$

It should be noted that a better objective quality metric has higher SRCC, KRCC and PLCC values and lower RMSE value.

3.2.2 Implementation details

The principle of global precedence has revealed that the HVS perceives image edges in a coarse-to-fine strategy [74]. And distortions which interrupt such coarse-to-fine integration significantly degrade image quality [45]. Thus, the proposed features are extracted at multi-scale. Besides the original image scale, the coarser scale is formed by low-pass filtering and

Table 3.1: The SRCC comparison for different gradient operators.

Gradient operator	MLIVE	MDID2013
Prewitt	0.944	0.908
Sobel	0.932	0.894
Scharr	0.933	0.897

downsampling by a factor of 2 in each dimension. During implementation, for GLBP calculation, the number of neighbors P is 8 and the radius of the neighborhood R is 1. The proposed structural features are extracted at five scales. Thus, the extracted features are 50 dimensions in total.

In this section, we first give the parameter choice on gradient operator and number of scales for feature extraction. Then we show the contribution of gradient information by comparison analysis.

3.2.2.1 Gradient operator selection

In our proposed method, the gradient magnitude map needs to be calculated. To this end, three commonly used gradient operators (i.e., Prewitt, Sobel, Scharr) were examined, and the one providing the best results was selected. The selection criterion was that the gradient operator leading to a higher SRCC would be selected. The SRCC values obtained by the three gradient operators on the two databases are listed in Table 3.1, from which we can see that the Prewitt operator could achieve slightly better performance than the other two. Thus, the Prewitt operator was chosen to calculate the gradient image in the proposed method due to its computational simplicity and better performance.

3.2.2.2 The number of scales selection

To account for the global precedence effect of the HVS and multi-scale property of natural images, the proposed gradient weighted GLPB histograms are extracted at multiple scales. Generally, more scales would extract more detailed information from distorted image, at the

expense of a larger feature set and computational overhead. In this work, we determine the proper scale number through experimental validation. The SRCC values obtained by different scale numbers on the two databases are shown in Figure 3.3, from which we can observe that increasing the number of scales beyond five does not contribute to performance. Thus, we extract the proposed feature set at five scales and form a 50-dimension feature vector as the SVR input.

Figure 3.3: The SRCC performance of GWH-GLBP vs. the number of scales on two benchmark databases.

3.2.2.3 The contribution of gradient information

In this work, we made two statements. First, we argue that the performance of LBP for IQA can be greatly improved through Prewitt operator prior to feature extraction. As LBP extracts the circular binary gradients at a micro-level, Prewitt operator provides the function of enhancing this gradient information accordingly. Second, we assume that edge locations convey more crucial visual information than the locations within a smooth area. Since human visual cortex is sensitive to image gradients, the gradient magnitude at a location can reflect how likely it is a perceptually significant structure point. Thus, the gradient weighted GLBP histogram which combines the contrast and structure information together can provide more valid representation for IQA task. Here we examine these two statements through experimental comparison.

Table 3.2: The SRCC comparison for different feature extraction methods.

Feature	MLIVE	MDID2013
LBP + FH	0.914	0.834
GLBP + FH	0.929	0.870
GWH-GLBP	0.944	0.908

In Table 3.2, the first row “LBP + FH” refers to the feature extraction method that applying the LBP in the original gray-scale image, and using the corresponding frequency histogram (FH) as the image global feature. The second row “GLBP + FH” refers to the method that applying the LBP in the gradient image, and using the corresponding FH as the image global feature. The third row is the proposed method which calculates LBP in the gradient image and uses the gradient magnitude as the weight to construct the gradient weighted histogram as the quality-aware feature. From this table, we can observe that both GLBP and gradient weighting strategy improve the performance.

3.2.3 Performance comparison with FR-IQA methods

In this section, we compare the proposed algorithm with several state-of-the-art FR-IQA methods, including PSNR, NQM [75], SSIM [8], MS-SSIM [76], VIF [10], VSNR [77], MAD [11], ADM [78], FSIM [12], GS [13], IGM [79], VSI [80] and GMSD [14].

Since the proposed method adopts SVR learning for quality estimation, we need to divide the database into training and testing subsets. In our experiments, for each database, distorted images of 80% of the reference images are used for training, and the rest are used for testing. This training-testing split is repeated 1000 times and the median performance is reported. Although FR-IQA metrics do not require training on the database, we also report the median performance across 1000 trials on the testing subset for consistent comparison.

Figure 3.4 provides the box-plot of SRCC performance on MLIVE database over 1000 trials. On each box, the central red bar is the median; the horizontal edges of the box are the 25th and 75th percentiles; the whiskers extend to the most extreme data points not considered

outliers. For a better IQA metric, the central red bar in the box should be larger, and the vertical edges of the box should be shorter. From the box-plot, we can see that the best three FR-IQA metrics are NQM, VIF and ADM. Moreover, the proposed GWH-GLBP outperforms all competing FR-IQA metrics by delivering outstanding and consistent performance in accordance with human subjective ratings. It is worth noting that GWH-GLBP is a NR method, while other competing metrics all exploit reference image information. Moreover, as the SRCC distribution is related to different content partitions, it can also reflect the content robustness of IQA models. If one IQA method is robust to content variations, the SRCC values achieved by different partitions should be similar to each other. Thus, the smaller standard deviation of the SRCC distribution, the more robust to content variation the IQA method is. GWH-GLBP exhibits the smallest standard deviation and is more content-robust than other IQA methods. The box-plot of SRCC performance on MDID2013 database over 1000 trails is shown in Figure 3.5.

Figure 3.4: Box-plot of SRCC distributions of IQA models on MLIVE database.

The prediction performance on the two databases measured by SRCC, KRCC, PLCC, and RMSE criteria are listed in Table 3.3. We also report the SRCC on two individual distortion types (DT1 and DT2) on MLIVE database in this table. For MDID2013 database with only one distortion type, there is no need to report individual distortion performance. For each criteria,

Figure 3.5: Box-plot of SRCC distributions of IQA models on MDID2013 database.

Table 3.3: Performance comparison with state-of-the-art FR-IQA metrics. The best three IQA metrics are highlighted in boldface.

IQA model	MLIVE						MDID2013			
	SRCC	KLCC	PLCC	RMSE	DT1	DT2	SRCC	KLCC	PLCC	RMSE
PSNR	0.728	0.561	0.818	10.868	0.740	0.750	0.643	0.474	0.652	0.036
NQM [75]	0.924	0.766	0.930	6.797	0.928	0.922	0.715	0.511	0.695	0.034
SSIM [8]	0.902	0.730	0.926	6.969	0.899	0.911	0.622	0.461	0.656	0.037
MS-SSIM [76]	0.866	0.691	0.907	7.880	0.884	0.896	0.749	0.574	0.776	0.030
VIF [10]	0.913	0.748	0.930	6.794	0.908	0.925	0.905	0.725	0.915	0.020
VSNR [77]	0.828	0.639	0.880	8.881	0.838	0.814	0.636	0.469	0.652	0.037
MAD [11]	0.894	0.718	0.914	7.608	0.914	0.888	0.856	0.667	0.860	0.025
ADM [78]	0.909	0.746	0.924	7.051	0.907	0.917	0.830	0.631	0.848	0.026
FSIM [12]	0.895	0.726	0.917	7.307	0.895	0.896	0.749	0.543	0.770	0.031
GS [13]	0.887	0.713	0.913	7.429	0.887	0.885	0.785	0.585	0.804	0.029
IGM [79]	0.889	0.720	0.923	7.195	0.885	0.897	0.878	0.697	0.882	0.022
VSI [80]	0.877	0.700	0.910	7.658	0.875	0.887	0.730	0.540	0.744	0.032
GMSD [14]	0.880	0.696	0.911	7.656	0.882	0.866	0.877	0.686	0.902	0.021
GWH-GLBP	0.944	0.797	0.949	5.874	0.936	0.953	0.908	0.730	0.913	0.019

the best three IQA models are highlighted in boldface. From Table 3.3, we can see that the proposed IQA metric is the most consistent metric over different databases. It performs the best on the MLIVE, and on par with VIF on the MDID2013.

From the results presented in Table 3.3, we draw the following conclusions. 1) The best three FR methods on MLIVE database are NQM, VIF and ADM. NQM models the degradation process as linear frequency attenuation followed by additive Gaussian noise. ADM models the degradation process as detail loss and additive impairments. VIF quantifies the information

loss between reference and distorted image as quality score. These three methods can well approximate the distortion combination ways in MLIVE database, thus deliver better performance than other metrics. 2) The best three FR methods on MDID2013 database are VIF, IGM and MAD. We attribute the good performance of VIF to the decent NSS model (GSM (Gaussian scale mixtures) model in wavelet domain) and the information fidelity philosophy. IGM partitions the image into primary visual information portion and disorderly uncertainty portion and evaluates the two parts separately. With such partition, IGM can partially explain the more complex situation when three kinds of distortions are combined together. 3) The performance of all methods on MDID2013 is inferior to their performance on MLIVE database. This seems intuitive since degradation in MLIVE is formed by combining two individual distortion types and degradation in MDID2013 is formed by combining three individual distortion types. The more complex scenario makes the IQA task even more challenging. 4) The proposed BIQA method outperforms all competing FR-IQA methods on MLIVE database. On MDID2013 database, it delivers similar performance with VIF, and outperforms all other FR methods.

3.2.4 Performance comparison with BIQA methods

In this section, we compare the proposed algorithm with several state-of-the-art BIQA methods, including BIQI [25], DIIVINE [42], BLIINDS2 [30], CORNIA [40], BRISQUE [31], GMLOG [38], NIQE [81], ILNIQE [34], SISBLM [48]. We also include the performance of FR-IQA methods PSNR and SSIM [8] for reference.

From the results presented in Table 3.4, we draw the following conclusions. 1) The best two BIQA methods on MLIVE database are GWH-GLBP and SISBLM. ILNIQE, BRISQUE and CORNIA tie for the third place. 2) The best three BIQA methods on MDID2013 database are GWH-GLBP, CORNIA and SISBLM. 3) The performance of all methods on MDID2013 is inferior to their counterpart on MLIVE database. Among them, GWH-GLBP and SISBLM

Table 3.4: Performance comparison with state-of-the-art BIQA metrics. The best three BIQA metrics are highlighted in boldface.

IQA model	MLIVE						MDID2013			
	SRCC	KLCC	PLCC	RMSE	DT1	DT2	SRCC	KLCC	PLCC	RMSE
PSNR	0.728	0.561	0.818	10.868	0.740	0.750	0.643	0.474	0.652	0.036
SSIM [8]	0.902	0.730	0.926	6.969	0.899	0.911	0.622	0.461	0.656	0.037
NIQE [81]	0.789	0.601	0.858	9.489	0.894	0.832	0.614	0.440	0.645	0.037
ILNIQE[34]	0.900	0.724	0.914	7.538	0.910	0.916	0.707	0.509	0.709	0.034
BIQI [25]	0.883	0.702	0.905	7.833	0.881	0.883	0.863	0.673	0.883	0.023
DIIVINE [42]	0.866	0.680	0.898	8.257	0.865	0.881	0.836	0.632	0.848	0.026
BLIINDS2 [30]	0.887	0.707	0.904	7.981	0.898	0.892	0.808	0.593	0.844	0.027
CORNIA [40]	0.900	0.725	0.916	7.586	0.905	0.903	0.898	0.709	0.904	0.020
BRISQUE [31]	0.900	0.733	0.924	7.143	0.907	0.904	0.819	0.613	0.833	0.027
GMLOG [38]	0.833	0.647	0.872	9.164	0.867	0.819	0.824	0.629	0.830	0.027
SISBLM [48]	0.907	0.737	0.925	7.194	0.908	0.913	0.885	0.709	0.885	0.023
GWH-GLBP	0.944	0.797	0.949	5.874	0.936	0.953	0.908	0.730	0.913	0.019

are specifically designed to assess multiply-distorted images, while other methods are general-purpose BIQA methods. We also observe that BIQI achieves better performance than its extended version DIIVINE on these two databases, which is opposite to their performance when tested on single distortion database.

To further prove the superiority of the proposed method over competing BIQA methods, we calculate the statistical significance by two hypotheses tests: the two sample T-test and Wilcoxon rank-sum test. The two sample T-test tests if two independent samples come from normal distributions with the same mean, against the alternative that the means are unequal. While the Wilcoxon rank-sum test tests the equivalence of the median values of two independent samples. Both tests are performed at a significance level of 5% using the 1000 SRCC values of all pairs of BIQA methods. The null hypothesis is that the SRCC values of the pair of models are drawn from populations with equal mean (two sample T-test) or equal median (Wilcoxon rank-sum test). The alternative hypothesis is that the mean/median of one model is greater/smaller than the other.

We list the statistical significance between the proposed method and other BIQA methods in comparison in Table 3.5, where the symbol “1”, “0” or “-1” means that the proposed method is statistically (with 95% confidence) better, indistinguishable, or worse than the corresponding

Table 3.5: Results of two sample t-test and Wilcoxon rank-sum test performed between SRCC obtained by BIQA methods on two databases. 1(-1) indicates GWH-GLBP is statistically better (worse) than the method in the column. 0 indicates GWH-GLBP is statistically equivalent to the method in the column.

T-test	NIQE	ILNIQE	BIQI	DIIVINE	BLIINDS2	CORNIA	BRISQUE	GMLOG	SISBLM
MLIVE	1	1	1	1	1	1	1	1	1
MDID2013	1	1	1	1	1	1	1	1	1
W-test	NIQE	ILNIQE	BIQI	DIIVINE	BLIINDS2	CORNIA	BRISQUE	GMLOG	SISBLM
MLIVE	1	1	1	1	1	1	1	1	1
MDID2013	1	1	1	1	1	1	1	1	1

BIQA approach. The two tests have exactly the same conclusions. From Table 3.5, we can see that GWH-GLBP performs statistically better than all other BIQA methods on the two databases.

3.2.5 Run-time comparison

The running costs of competing BIQA metrics are also evaluated. Experiments are performed on a ThinkPad T430S notebook with Intel Core i7-3520M CPU@2.9GHz and 8G RAM. The software platform is MATLAB R2014b (8.4) under Windows Home Premium. The MATLAB source codes of all other IQA models are obtained from original authors. (It should be noted that whether the code is optimized may affect the running time of an algorithm.) The feature extraction time consumed by each IQA model for estimating the quality of one 512×512 image is listed in Table 3.6.

From Table 3.6, we can see that the proposed method GWH-GLBP is quite efficient with low computational complexity. More specifically, it can process 10 images per second, which meets the time requirement in most image processing applications.

3.3 Summary

In this study, we have proposed a BIQA model named GWH-GLBP to estimate the visual quality of multiply-distorted images based on structural degradation. Specifically, a novel effective

Table 3.6: Run-time comparison of BIQA models.

IQA model	Time cost (s)
NIQE [81]	0.227
ILNIQE[34]	9.840
BIQI [25]	0.050
DIIVINE [42]	15.519
BLIINDS2 [30]	61.393
CORNIA [40]	2.449
BRISQUE [31]	0.079
GMLOG [38]	0.063
SISBLM [48]	3.336
GWH-GLBP	0.105

structural feature, the gradient weighted histogram of gradient local binary pattern, is used to characterize image structural information for visual quality estimation. The proposed GWH-GLBP has been compared comprehensively with thirteen representative FR-IQA methods and ten prominent BIQA methods on two public multi-distortion databases. Experimental results demonstrate the outstanding performance in terms of both prediction accuracy and computational efficiency, making GWH-GLBP a good candidate to be used in practical applications.

Chapter 4

BIQA Using Statistical Structural and Luminance Features

In Chapter 3, we have developed one BIQA method for quality prediction of multiply-distorted images. The proposed method delivers quite competitive performance, compared to relevant state-of-the-art works. However, it is a distortion-specific method which is designed for multiple distortions, and its performance on other kinds of distortions is not guaranteed. By contrast, general-purpose BIQA methods that can handle all kinds of distortions are more valuable and versatile.

Although there are some existing general-purpose BIQA methods, their performance is still not satisfied for practical usage. Moreover, these methods have their own limitations as discussed in Section 2.1.2.1. In this chapter, we develop a novel general-purpose BIQA model by investigating the characteristics of human visual perception. The proposed method belongs to perceptual feature based BIQA methods. It has the advantages of high prediction accuracy, high generalization capability, low computational complexity and low feature dimension.

Existing studies show that image structure conveys the primary visual information of a scene, and the HVS is highly adapted to extract structural information for image perception and understanding [82, 83]. Here, we propose a novel effective structural feature as the spatial distribution of LBP based on the normalized luminance map. Also, studies from neuroscience have demonstrated that the HVS is highly sensitive to luminance change of a scene, which might

also cause visible distortions [84, 85]. Based on these findings, we extract a new feature to represent luminance changes in the proposed metric. We demonstrate that complementary information provided by extracted statistical structural and luminance features plays an important role for image quality estimation. The proposed method is denoted as NRSL (No-Reference quality assessment using statistical Structural and Luminance features). It can consistently handle both synthetically-distorted and authentically-distorted images.

Figure 4.1: Flowchart of the proposed NRSL metric.

4.1 The Proposed Method

NRSL is designed by accounting for both structural and luminance degradation. Figure 4.1 shows the framework of the proposed metric. First, the contrast normalization scheme is applied to the image to remove redundancy in the visual input. Second, the LBP descriptor is employed on the normalized image to extract the structural features. Furthermore, the distribution of normalized luminance magnitudes in the form of histogram is extracted from the same normalized image. Both structural and luminance features are calculated at three scales to account for the variation in viewing distance and image resolution. Finally, the extracted statistical structural and luminance features are combined together as the input to SVR to simulate the nonlinear mapping from feature space to subjective quality scores.

4.1.1 Local contrast normalization

Local contrast normalization has been used as a nonlinear processing stage to emulate the nonlinear masking of visual perception in many image processing applications [86]. Generally,

each coefficient is divided by the square root of Gaussian weighted combination of the squared amplitudes of its neighbors. In this study, similar model as [31, 81] has been adopted,

$$\tilde{I}(i, j) = \frac{I(i, j) - \mu}{\sigma + C} \quad (\text{Eq. 4.1})$$

where i and j are the spatial indices of the image, and

$$\mu(i, j) = \sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} I(i+k, j+l) \quad (\text{Eq. 4.2})$$

$$\sigma(i, j) = \sqrt{\sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} [I(i+k, j+l) - \mu(i, j)]^2} \quad (\text{Eq. 4.3})$$

are the local mean and standard deviation of the surrounding local patch, where $\omega = \{\omega_{k,l} | k = -K, \dots, K, l = -L, \dots, L\}$ defines a unit-volume Gaussian window. And the constant C is induced to guarantee numerical stability when the denominators are close to zero. Specifically, we choose

$$C = (\alpha L)^2 \quad (\text{Eq. 4.4})$$

where L is the dynamic range of pixel grayscale levels (255 for 8-bit grayscale image), and $\alpha \ll 1$ is a small constant.

4.1.2 The structural histogram

Natural images are highly structured with the manifestation of strong dependencies among their pixels. When these pixels are spatially neighboring, these high dependencies often carry important information about the structure of visual objects in natural scenes [8]. Image structures convey the primary visual information of a scene, and the HVS adaptively evolves to extract structural information for image perception and understanding [8, 79]. In this work, we extract image structural information as the spatial distribution of local inter-pixel relationship patterns.

After local contrast normalization, we obtain the normalized luminance image $\tilde{I}$. Assuming that the local patch of image $\tilde{I}$ can be characterized by joint distribution of normalized luminance values of successive pixels in this patch,

$$T = t(g_c, g_0, g_1, \dots, g_{P-1}) \quad (\text{Eq. 4.5})$$

where g_c is the normalized luminance value of the center pixel of the local patch, $\{g_0, g_1, \dots, g_{P-1}\}$ are the normalized luminance values of the P circularly symmetric neighbourhood.

Without losing information, we substrate the center pixel value from the neighborhood,

$$T = t(g_c, g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (\text{Eq. 4.6})$$

Analogous to [70, 87], we assume that the normalized luminance value of center pixel to be statistically independent of the differences between neighbors and center pixel, which allows for the factorization of the joint distribution,

$$T \approx t(g_c)t(g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c) \quad (\text{Eq. 4.7})$$

where the first term $t(g_c)$ is the marginal distribution of normalized luminance coefficients over the whole image $\tilde{I}$. The second term $t(g_0 - g_c, g_1 - g_c, \dots, g_{P-1} - g_c)$ is the joint distribution of differences which describes local inter-pixel relationship. By keeping the signs of the differences, Ojala et al. [70] propose the LBP operator to describe this joint distribution,

$$t(s(g_0 - g_c), s(g_1 - g_c), \dots, s(g_{P-1} - g_c)) \quad (\text{Eq. 4.8})$$

where the thresholding function $s(\cdot)$ is defined as

$$s(g_i - g_c) = \begin{cases} 1, & g_i - g_c \geq 0 \\ 0, & g_i - g_c < 0 \end{cases} \quad (\text{Eq. 4.9})$$

By assigning a binomial factor (2^p) for each neighbor from (Eq. 4.8), the LBP code of one pixel is deduced as [70],

$$LBP_{P,R} = \sum_{i=0}^{P-1} s(g_i - g_c)2^i \quad (\text{Eq. 4.10})$$

where P is the number of neighbors and R is the radius of the neighborhood.

It was observed that certain LBP patterns can capture the fundamental structural properties, composing of the vast majority of LBP patterns in natural images (sometimes over 90%) [70]. These certain LBPs share with the same attribute: the number of spatial transitions is limited by two, which are referred as the uniform LBP patterns. To achieve rotation invariance, locally rotation invariant uniform LBP operator can be defined as

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{i=0}^{P-1} s(g_i - g_c) & \text{if } \mathcal{U}(LBP_{P,R}) \leq 2 \\ P + 1 & \text{else} \end{cases} \quad (\text{Eq. 4.11})$$

where $\mathcal{U}$ measures the number of bitwise transitions and superscript *riu2* refers to the rotation invariant “uniform” patterns with $\mathcal{U}$ value less than 2. The rotation invariant uniform LBP would have $P + 2$ distinct patterns ($P + 1$ for the different uniform patterns and 1 for the non-uniform pattern).

The LBP patterns effectively detect image primitive microstructures, such as edges, lines, corners and spots. The introduction of distortion may shift the LBP pattern from one type to another. For example, blocking artifacts caused by JPEG compression may shift a flat pattern to an edge pattern, while blurring effect can change an edge pattern to flat pattern. Gaussian noise may alter the LBP pattern in a random way. Furthermore, as LBP describes the inter-pixel relationship in an image neighborhood, such micro-structural pattern is also effective to capture the complex interaction between different distortions and their joint effect on the overall image quality.

Therefore, we extract the spatial distribution of LBP patterns as an effective feature for quality perception. Since $LBP_{P,R}^{riu2}$ has a fixed set of discrete patterns ($0 \rightarrow P + 1$), no quantization is required, but the different patterns are directly mapped into a histogram of $P + 2$ bins. By generating the occurrence histogram, we effectively employ the statistical property of the structural information: the LBP operator detects microstructure whose underlying distribution is characterized by the histogram [70]. In this work, we set $P = 8$, thus there would be 10 bins

for the structural histogram. Figure 4.2(a) shows the average structural histograms of reference images and distorted images with five distortion types in LIVE database (i.e., WN (Gaussian white noise), GB (Gaussian blur), JPEG, JP2K (JPEG2000), and FF (simulated fast fading channel errors)). The average structural histograms are generated by calculating the mean of histograms of all the images with the same distortion type. Figure 4.2(b) shows the average structural histograms of reference images and distorted images with two kinds of multiple distortions in MLIVE database (i.e., GB+JPEG, GB+WN). As can be seen in Figure 4.2, different distortions alter the structural histograms in their own characteristic ways. And structural histograms are effective to describe the impact of both single distortion and multiple distortions.

Figure 4.2: Impact of distortions on structural histogram. (a) Average structural histograms on LIVE. (b) Average structural histograms on MLIVE.

4.1.3 The luminance histogram

The HVS is also highly sensitive to luminance change of an input scene, which might cause visible distortions [13, 88]. Although local contrast normalization smoothes the luminance range in a local patch, the information of global luminance variation can still be manifested by the normalized luminance image $\tilde{I}$. Thus, we extract the luminance feature from $\tilde{I}$.

Existing studies on NSS reveal that the normalized luminance coefficients of natural images follow a Gaussian-like distribution [89]. The studies [31, 81] fit a GGD to the coefficients and

use the GGD parameters as perceptual features. In this work, we directly employ the histogram to represent the image luminance information. It is advantageous to use the nonparametric histogram instead of the parametric GGD fitting as the quality-aware features. First, it frees us from making any, possibly misguided, assumptions about the underlying feature distribution. Second, it constitutes a more accurate representation to avoid any fitting error. Third, it is more computationally efficient to calculate the histogram. Moreover, after local contrast normalization, the resultant normalized luminance coefficients form a symmetric distribution centered at 0. To reduce the histogram range, we take the magnitude operation before calculating the luminance histogram.

$$\bar{I}(i, j) = |\tilde{I}(i, j)| \quad (\text{Eq. 4.12})$$

$$LH = \text{hist}(\bar{I}) \quad (\text{Eq. 4.13})$$

Figure 4.3(a) shows the average luminance histograms of the reference images and distorted images with five distortion types in LIVE database. Figure 4.3(b) shows the average luminance histograms of the reference images and distorted images with two kinds of multiple distortions in MLIVE database. As shown in this figure, different distortion types show apparently different luminance changes. For example, the luminance distribution of WN is more uniform as WN introduces random disturbs, while JP2K/GB/FF make the distribution more Laplacian-like with high peak and heavy tail as they reduce high frequency components in images. The distribution of JPEG distorted images exhibits several peaks that are caused by the blocking artifacts. The distinct distribution changes caused by different distortion types indicate that the luminance histograms are effective to reflect the quality variation. We also observe that GB+JPEG makes the distribution more peaked as there is blurring effect for both GB and JPEG, while the luminance distribution of GB+WN resembles that of clean images since GB and WN have counter-impact on this distribution. In this case, the structural histogram may be more quality-relevant than the luminance histogram.

Figure 4.3: Impact of distortions on luminance distribution. (a) Average luminance distributions on LIVE. (b) Average luminance distributions on MLIVE.

4.1.4 Analysis of two feature sets

In the previous subsection, we have made the assumption that the marginal distribution of normalized luminance coefficients is statistically independent of the differences between neighbors and center pixel. First, we validate the feasibility of the independence assumption by approximating the joint distribution by the factorized marginal distribution. As a simplification of (Eq. 4.7), we consider the case of one neighbor.

$$t(g_c, g_0 - g_c) \approx t(g_c)t(g_0 - g_c) \quad (\text{Eq. 4.14})$$

Figure 4.4(a) shows the average of distributions $t(g_c, g_0 - g_c)$ computed from all the images in LIVE database [90]. Figure 4.4(b) shows the average error between $t(g_c, g_0 - g_c)$ and $t(g_c)t(g_0 - g_c)$. As shown in Figure 4.4, the average error is relatively small in proportion to the average distribution which confirms the rationality of the independence assumption. Since the structural histogram is derived directly from the joint difference distribution, we can conclude that the structural histogram and luminance histogram are statistically independent.

In Figure 4.5, we provide a few concrete examples to show the complementary of structural and luminance features for quality assessment. Figure 4.5(a) and Figure 4.5(c) are images

Figure 4.4: Validation the independence assumption by approximating the joint distribution by the factorized marginal distribution. (a) The joint distribution $t(g_c, g_0 - g_c)$ which represents the concurrence of horizontally adjacent pixels. (b) The average error produced by factorization of $t(g_c, g_0 - g_c)$ into $t(g_c)$ and $t(g_0 - g_c)$.

with low quality, and their DMOS are 68.99 and 75.74, respectively. Figure 4.5(b) is a high-quality image with DMOS of 7.23. Their structural and luminance histograms are shown in the second row. We compare Figure 4.5(d) with Figure 4.5(e) and find that these two images share similar luminance distribution, which indicates that the luminance feature cannot distinguish the distortion differences between these two images. By contrast, their structural distributions are quite different and provide valuable information for quality estimation. The ringing and blurring artifacts introduced by JP2K to Figure 4.5(a) can be well captured by its structural histogram. The opposite phenomenon is observed when comparing Figure 4.5(e) with Figure 4.5(f). These two images share the similar structural distribution, which shows the ineffectiveness of structural feature in this case. Conversely, their luminance histograms are disparate enough to discriminate their different quality levels. Hence, we use both structural and luminance features together in this study, since they are mutually complementary for quality evaluation task. The experimental comparison in Section 4.2.2.2 also validates this conclusion.

We investigate the quality-awareness of NRSL features by performing t-SNE algorithm [72]. Figure 4.6 shows the 2D embedding of extracted features of NRSL, where each dot represents

Figure 4.5: Image samples to show the complementary of structural and luminance information. (a) Image with DMOS = 68.99. (b) Image with DMOS = 7.23. (c) Image with DMOS = 75.74. (d), (e) and (f) are the structural and luminance histograms of images in the first row.

one image in the LIVE database, and x- and y-axis denote their coordinates in the 2D plane. Figure 4.6(a) is color coded by distortion type, and Figure 4.6(b) is color coded by DMOS range. As shown in Figure 4.6(a), WN and JPEG are well grouped and separated from other distortions. By contrast, images corrupted by JP2K, GB, and FF tend to mix together. This seems reasonable as both JP2K and GB introduce the blurring artifacts to images, and FF is actually a multiple distortion of the combination of JP2K and transmission errors. The same or similar artifacts introduced by the three distortion types make them cluster together in the feature space. Although it is not a necessary requirement for good BIQA feature to handle the distortion classification problem, Figure 4.6(a) shows that the proposed feature set can also be used to identify WN and JPEG distortion types. From Figure 4.6(b), we can see that images with similar DMOS values are clustered together, no matter they are from one specific distortion type or across different distortion types. Such observation indicates that the proposed features are valuable for both distortion-specific and general-purpose BIQA task.

Figure 4.6: t-SNE 2D embedding of the NRSL feature space on LIVE database. (a) Distortion types are coded by color. (b) Quality scores are coded by color.

4.1.5 Regression model for quality prediction

The SVR with RBF kernel is adopted as the mapping function for feature pooling from the feature space to quality measure [71].

4.2 Performance Evaluation

4.2.1 Database description

The comparison experiments are conducted on seven subjective image databases. Four of them are traditional synthetically-distorted image databases (i.e., LIVE [90], CSIQ [91], TID2013 [92], MLIVE [47]), which are generated by introducing graded simulated distortions onto high-quality photographs. The most commonly simulated distortions are JP2K, JPEG, WN and GB. LIVE, CSIQ and TID2013 mostly contain images corrupted by a single type of distortion. MLIVE contains images simultaneously distorted by two types of distortions. The remaining three databases are the newly emerging naturally-distorted image databases, which contain realistic distortions generated during image acquisition, processing and storage of devices (i.e., BID [93], CLIVE [94] and CID2013 [95]). For the naturally-distorted images, there are no reference images with perfect quality, and FR/RR methods are not applicable in this scenario. Also there is no definite distortion category for each image since realistic distortions are generally diverse, mixed and multipartite. These databases are summarized as follows.

- The LIVE database includes 29 reference images and 779 distorted images corrupted by five types of distortions: JP2K, JPEG, WN, GB, and FF. Subjective quality scores are provided in the form of DMOS ranging from 0 to 100.
- The CSIQ database includes 30 reference images and 866 distorted images corrupted by six types of distortions: JPEG, JP2K, WN, GB, pink Gaussian noise (PGN), and global contrast decrements (CTD). Subjective quality scores are provided in the form of DMOS ranging from 0 to 1.
- The TID2013 database includes 25 reference images and 3000 distorted images corrupted by 24 types of distortions: #01 WN, #02 WN in color components, #03 spatially correlated WN, #04 masked noise, #05 high-frequency noise, #06 impulse noise, #07

quantization noise, #08 GB, #09 image denoising, #10 JPEG, #11 JP2K, #12 JPEG transmission errors, #13 JP2K transmission errors, #14 non eccentricity pattern noise, #15 local blockwise distortion of different intensity, #16 mean shift, #17 contrast change, #18 change of color saturation, #19 multiplicative Gaussian noise, #20 comfort noise, #21 lossy compression of noisy images, #22 image color quantization with dither, #23 chromatic aberrations and #24 sparse sampling and reconstruction. Subjective quality scores are provided in the form of MOS ranging from 0 to 9.

- The MLIVE database includes 15 reference images and 450 distorted images corrupted by two types of multiple distortions: GB+JPEG and GB+WN. Subjective quality scores are provided in the form of DMOS ranging from 0 to 100.
- The BID (realistic blur image database) database includes 586 blurred images taken under a variety of lighting conditions and exposure time. It presents typical blurring scenarios in practical applications, such as out-of-focus, simple motion, complex motion and their combination. Subjective quality scores are provided in the form of MOS ranging from 0 to 5.
- The CLIVE (LIVE in the wild image quality challenge database) database includes 1162 distorted images captured using typical real-world mobile cameras. It contains various authentic image distortions, such as low-light blur and noise, motion blur, overexposure, underexposure, compression errors and their combination. Subjective quality scores are provided in the form of MOS ranging from 0 to 100.
- The CID2013 database includes 474 distorted images captured using various digital cameras. The database consists of six sets (I-VI) and each set covers around 80 images captured under six typical scenes. It represents the typical distortions that depend on camera sensor type, optics and built in image processing pipeline. Subjective quality scores are provided in the form of MOS ranging from 0 to 100.

4.2.2 Implementation details

4.2.2.1 Parameter Choice

There are two free parameters in the proposed NRSL model. The first one is the number of bins of luminance histogram. Generally, histograms with a small number of bins fail to provide enough discriminative information about the distributions. By contrast, for histograms with too many bins, the average number of entries of each bin would be too small, which may cause the histograms sparse and unstable. To investigate the effect of bin number on the quality prediction performance, we try several choices of the bin number and observe SRCC performance on each database. The results are plotted in Figure 4.7(a).

Figure 4.7: Parameter Choice. (a) The SRCC performance vs. the number of bins of luminance histogram. (b) The SRCC performance vs. constant α .

We can see that the bin number of 10 results in high and stable performance across all seven databases. Therefore, in our implementation, we set the number of bins of luminance histogram to 10.

Another free parameter to be determined is the α value in (Eq. 4.4). Apart from ensuring the numerical stability, the constant α also plays an important role in mediating the contrast saturation in low contrast regions. Figure 4.7(b) plots the SRCC curves against α by applying NRSL to the seven benchmark databases. We can see that for a wide range of choices for α ,

NRSL achieves satisfactory performance on all the databases except CID2013sub. The reason may be that there is much difference among images in CID2013 database in terms of luminance and contrast (e.g., some are captured in outdoor daylight, while others are indoor scenes with black background). Therefore, the best contrast normalization factor may vary from one scene to another. In the implementation, we set $\alpha = 0.01$ for all databases.

In addition, both the statistical structural and luminance features are calculated at three scales to account for the variations of viewing distance and image resolution [10, 76]. Thus, the extracted features are 60 dimensions in total. Besides the original image scale, the coarser scale is constructed by low-pass filtering and downsampling the image by a factor of 2.

4.2.2.2 Examination of two feature sets

In Section 4.1.4, we have provided some qualitative analysis on the complementarity of the extracted structural and luminance features. In this part, we present the experimental comparison by using only one of the two feature sets to validate our analysis. Specifically, we tested the performance on LIVE database by splitting the database into training and testing subsets 1000 times. 80% of the distorted images are used for training and the rest for testing. There is no content overlap between these two subsets. SRCC results on LIVE database [90] are shown in Table 4.1, where NRSL_S denotes the model which uses only structural feature and NRSL_L denotes the model using only luminance feature.

Table 4.1: SRCC comparison on LIVE database to validate the complementarity of structural and luminance information.

IQA model	JP2K	JPEG	WN	GB	FF	ALL
NRSL_S	0.940	0.953	0.962	0.928	0.863	0.935
NRSL_L	0.858	0.920	0.977	0.891	0.730	0.871
NRSL	0.943	0.960	0.984	0.959	0.880	0.952

From Table 4.1, we can see that structural feature is more effective for distortion types of JP2K, JPEG, GB and FF, as such distortions greatly interfere the structural information of

images; while luminance information are more effective for WN distortion. Moreover, the combination of these two feature sets always delivers higher performance than each feature set under all the cases.

4.2.3 Overall performance on individual database

First, we evaluate the overall performance of competing BIQA models on each benchmark database. As in many previous works [38, 39, 96], for CSIQ [91] and TID2013 [92] databases, we only test the four distortions that also appear in LIVE database (i.e., JP2K, JPEG, WN, GB). These four distortion types are also the most commonly encountered distortions in practical applications. We also exclude the 25th reference image with its distorted versions from TID2013 [92] as it is not a natural image. We denote the selected databases as CSIQsub and TID2013sub, which include 600 and 480 distorted images, respectively. The performance on the whole CSIQ and TID2013 databases will be presented in Section 4.2.4. Since NRSL adopts SVR learning for quality estimation, we need to divide the database into training and testing subsets. For the four synthetically-distorted image databases (i.e., LIVE, CSIQsub, TID2013sub, MLIVE), distorted images are generated by post-processing the reference images. To achieve content independence between training and testing sets, distorted images of 80% of the reference images are used for training, and the rest are used for testing. For BID and CLIVE databases, there are no reference images. Also the images are quite different from each other in terms of content properties. Thus, we randomly split the databases into two sets: 80% for training, and 20% for testing. The training-testing split is repeated 1000 times and the median performance is reported. The CID2013 database can be grouped into two parts since Sets I-III and Sets IV-VI are rated using different subjective evaluation protocols. Sets IV-VI are used in this work for comparing the performance of BIQA methods and denoted as CID2013sub. Also, each Set includes six scenes. Images from the same scene are quite similar to each other, and different from images from other scenes in terms of content properties. Thus,

we use the leave-one-out strategy by training on images from five scenes and testing on images from the remaining scene. This is repeated until all the images have their predicted scores. Then we calculate the correlation between MOS and predicted scores. Since the training and testing sets are fixed in this scheme, the experiment is conducted only once. This experimental setup on CID2013 is also in line with several recent studies [97, 98]. There is one purely black image in CID2013 database (IS_VL_C01_D14.JPG) and several BIQA methods fail to extract effective features on it, thus we also excluded this image to form the CID2013sub with 233 images.

NRSL was compared with nine state-of-the-art BIQA models, including NIQE [81], BIQI [25], DIIVINE [42], BLIINDS2 [30], CORNIA [40], BRISQUE [31], GMLOG [38], NR-GLBP [99] and NFERM [39]. The prediction performance measured by SRCC, PLCC, and RMSE criteria are listed in Table 4.2 to Table 4.4. For each criteria, the best two BIQA models are highlighted in boldface. Moreover, in order to provide an evaluation of the overall performance of the competing IQA models, we also present their weighted-average SRCC, PLCC results over all seven databases, and the weight assigned to each database is in direct proportion to the number of distorted images in that database.

From these tables, it can be observed that NRSL performs consistently well on all the benchmark databases. Particular, it delivers higher performance than other competitors on all seven databases. On the contrary, for other BIQA models, they may work well on some databases but fail to deliver good results on other databases. For example, CORNIA and BRISQUE can get encouraging results on LIVE and MLIVE, but they perform relatively poorly on CSIQsub and TID2013sub databases. GMLOG achieves quite good performance on LIVE, CSIQsub and TID2013sub databases, but it performs rather poorly on MLIVE database. Therefore, we can conclude that objective quality scores predicted by NRSL are in higher agreement with subjective ratings than all the other BIQA models examined. However, the performance of all BIQA methods drops significantly on naturally-distorted databases compared with traditional simulated databases. This is to be expected since authentic distortions are generally more

Table 4.2: SRCC comparison of BIQA models on seven benchmark databases.

IQA model	LIVE	CSIQsub	TID2013sub	MLIVE	BID	CLIVE	CID2013sub	Weighted Average
NIQE [81]	0.909	0.883	0.819	0.793	0.457	0.453	0.713	0.687
BIQI [25]	0.821	0.819	0.818	0.883	0.573	0.532	0.681	0.707
DIIVINE [42]	0.912	0.883	0.882	0.866	0.610	0.597	0.582	0.755
BLIINDS2 [30]	0.930	0.904	0.876	0.886	0.532	0.463	0.380	0.705
CORNIA [40]	0.943	0.890	0.893	0.899	0.625	0.618	0.687	0.780
BRISQUE [31]	0.943	0.903	0.900	0.900	0.581	0.607	0.759	0.778
GMLOG [38]	0.950	0.924	0.931	0.831	0.543	0.595	0.599	0.761
NR-GLBP [99]	0.938	0.916	0.920	0.891	0.628	0.612	0.625	0.780
NFERM [39]	0.938	0.929	0.929	0.899	0.586	0.540	0.600	0.757
NRSL	0.952	0.930	0.945	0.932	0.638	0.631	0.817	0.809

complex, diverse and multipartite, and thus the visual quality estimation requires more efforts in the community. The existing method NR-GLBP [99] also uses LBP for feature extraction by manually selecting several thresholds to calculate the generalized LBP on four LOG filtered images to form the quality-aware features. However, since the LBP mechanism only encodes the sign of the difference between one pixel and its neighbors, it fails to capture the impact of luminance and contrast distortions as the magnitude information is lost. Moreover, the manually selected thresholds may not well accommodate images with different content properties. On the contrary, we effectively combine the complementary features of luminance and structural histograms in our method. The bounded performance of LBP can be improved by the luminance histogram and their combination further advances the quality prediction accuracy. The experimental results in Table 4.2 to Table 4.4 show that our method delivers better performance than NR-GLBP on all the seven benchmark databases. Also, our method has a lower feature dimension and faster computation time, as shown in Table 4.10 in Section 4.2.6.

To further prove the superiority of NRSL over other compared BIQA methods, we calculate the statistical significance between NRSL and other BIQA methods. For the first six databases, experiments are conducted by splitting the database into training and testing sets 1000 times and the median performance is reported. Thus, we conduct the two sample t-test between the pair of SRCC values of 1000 training-testing loops, which measures the equivalence of the mean values of two independent samples. For the CID2013sub database, experiments are

Table 4.3: PLCC comparison of BIQA models on seven benchmark databases.

IQA model	LIVE	CSIQsub	TID2013sub	MLIVE	BID	CLIVE	CID2013sub	Weighted Average
NIQE [81]	0.909	0.892	0.831	0.860	0.467	0.509	0.715	0.713
BIQI [25]	0.838	0.873	0.856	0.905	0.598	0.557	0.684	0.735
DIIVINE [42]	0.913	0.905	0.904	0.898	0.646	0.627	0.608	0.779
BLIINDS2 [30]	0.936	0.932	0.905	0.903	0.560	0.507	0.389	0.731
CORNIA [40]	0.946	0.926	0.925	0.915	0.646	0.662	0.718	0.807
BRISQUE [31]	0.947	0.934	0.925	0.922	0.605	0.645	0.762	0.801
GMLOG [38]	0.954	0.947	0.943	0.871	0.571	0.620	0.624	0.782
NR-GLBP [99]	0.943	0.948	0.939	0.904	0.654	0.634	0.656	0.800
NFERM [39]	0.942	0.952	0.951	0.919	0.608	0.570	0.656	0.780
NRSL	0.956	0.954	0.959	0.946	0.663	0.654	0.816	0.825

Table 4.4: RMSE comparison of BIQA models on seven benchmark databases.

IQA model	LIVE	CSIQsub	TID2013sub	MLIVE	BID	CLIVE	CID2013sub
NIQE [81]	11.376	0.126	0.779	9.437	1.100	17.394	16.224
BIQI [25]	14.874	0.137	0.722	7.833	0.997	16.828	16.929
DIIVINE [42]	11.168	0.118	0.596	8.257	0.947	15.767	18.419
BLIINDS2 [30]	9.515	0.101	0.597	8.125	1.029	17.419	21.378
CORNIA [40]	8.812	0.105	0.532	7.586	0.949	15.158	16.159
BRISQUE [31]	8.794	0.099	0.533	7.273	0.990	15.450	15.032
GMLOG [38]	8.180	0.090	0.464	9.197	1.023	15.800	18.128
NR-GLBP [99]	9.075	0.089	0.479	7.932	0.942	15.622	17.510
NFERM [39]	9.101	0.084	0.436	7.413	0.986	16.641	17.517
NRSL	8.018	0.084	0.397	5.943	0.931	15.317	13.399

conducted by leave-one-subset-out strategy, and predicted quality scores at all subsets are put together before calculating the final performance. Thus, we conduct the two sample F-test to the residuals of two methods, which measures the equivalence of variances of two independent samples. Both t-test and F-test are conducted at the 5% significance level, with 1 (-1) indicating that NRSL is statistically superior (inferior) to the compared method and 0 indicating that NRSL is statistically equivalent to the compared method. From this table, we can see that on TID2013sub, MLIVE, BID, CLIVE and CID2013sub, NRSL performs statistically better than all other BIQA methods. NRSL is on par with GMLOG on LIVE, on par with GMLOG and NFERM on CSIQsub, respectively. Generally, NRSL achieves statistically better performance than other BIQA methods.

Table 4.5: Statistical significance test. 1(-1) indicates NRSL is statistically better (worse) than the method in the column. 0 indicates NRSL is statistically equivalent to the method in the column.

DB	NIQE	BIQI	DIIVINE	BLIINDS2	CORNIA	BRISQUE	GMLOG	NR-GLBP	NFERM
LIVE	1	1	1	1	1	1	0	1	1
CSIQsub	1	1	1	1	1	1	0	1	0
TID2013sub	1	1	1	1	1	1	1	1	1
MLIVE	1	1	1	1	1	1	1	1	1
BID	1	1	1	1	1	1	1	1	1
CLIVE	1	1	1	1	1	1	1	1	1
CID2013sub	1	1	1	1	1	1	1	1	1

4.2.4 Performance on individual distortion type

Here, we evaluate the performance of competing IQA models on individual distortion type. Since there is no well defined distortion category for naturally-distorted images, we only conduct experiments on the four databases with simulated distortions. For BIQA models, we train on 80% of images with various distortion types and then test on the left 20% of images with the specific distortion type. The SRCC comparison on four benchmark databases are tabulated in Table 4.6. There are 15 groups of distorted images in the four databases. The best two BIQA

models for each distortion group are shown in boldface. It should be noted that similar results can be obtained for PLCC and RMSE criteria, thus we only list SRCC here for brevity.

Table 4.6: SRCC comparison of BIQA models on individual distortion type.

	D-Type	NIQE	BIQI	DIIVINE	BLIINDS2	CORNIA	BRISQUE	GMLOG	NR-GLBP	NFERM	NRSL
LIVE	JP2K	0.923	0.780	0.902	0.928	0.922	0.915	0.925	0.933	0.937	0.943
	JPEG	0.942	0.829	0.899	0.949	0.941	0.963	0.963	0.958	0.964	0.960
	WN	0.972	0.958	0.981	0.945	0.963	0.978	0.983	0.981	0.984	0.984
	GB	0.941	0.844	0.935	0.913	0.955	0.946	0.929	0.936	0.909	0.959
	FF	0.862	0.739	0.858	0.873	0.907	0.887	0.899	0.853	0.850	0.880
CSIQsub	JP2K	0.927	0.815	0.877	0.889	0.901	0.892	0.915	0.911	0.917	0.921
	JPEG	0.883	0.856	0.887	0.916	0.886	0.918	0.936	0.926	0.927	0.937
	WN	0.837	0.842	0.901	0.899	0.797	0.919	0.939	0.925	0.934	0.955
	GB	0.906	0.837	0.898	0.919	0.907	0.913	0.906	0.921	0.924	0.926
TID2013sub	JP2K	0.906	0.848	0.892	0.916	0.900	0.905	0.927	0.927	0.938	0.940
	JPEG	0.879	0.855	0.867	0.871	0.869	0.878	0.902	0.917	0.896	0.920
	WN	0.842	0.865	0.905	0.753	0.730	0.884	0.948	0.856	0.939	0.928
	GB	0.841	0.890	0.937	0.902	0.914	0.930	0.916	0.943	0.928	0.958
MLIVE	GB+JPEG	0.899	0.881	0.864	0.892	0.904	0.905	0.865	0.897	0.919	0.928
	GB+WN	0.833	0.883	0.877	0.884	0.900	0.900	0.817	0.905	0.887	0.937
	wmean	0.893	0.849	0.897	0.899	0.896	0.916	0.913	0.919	0.923	0.938
	wstd	0.042	0.048	0.031	0.041	0.053	0.027	0.043	0.031	0.030	0.023

From the results presented in Table 4.6, we can see that NRSL is among the best two metrics 12 times, followed by NFERM (7 times) and GMLOG (5 times). We also calculate the weighted mean and standard deviation of competing IQA models across all distortion groups. Among them, NRSL is with the highest mean and the lowest standard deviation across different groups. Thus, we can draw the following conclusions. In general, when distortion is of a specific type, NRSL performs the best among all NR models. NFERM, NR-GLBP, and BRISQUE can also achieve acceptable results. Moreover, although GMLOG delivers promising results for singly-distorted images, its performance deteriorates significantly when tested on multiply-distorted images. NRSL achieves consistently better performance on most commonly encountered distortion types.

Furthermore, in order to validate the generality ability of NRSL to uncommon distortion types, we also compare NRSL with other five best performing BIQA models through 1000

training-testing experiments on the whole CSIQ and TID2013 databases. The SRCC values of different methods on the whole CSIQ and TID2013 databases are presented in Table 4.7 and Table 4.8. From this table, we can observe that NRSL is among the best two metrics 20 times, followed by NFERM (16 times) and GMLOG (11 times).

Table 4.7: SRCC comparison on the whole CSIQ database.

BIQA model	CSIQ (866)						
	WN	JPEG	JP2K	PGN	GB	CTD	ALL
CORNIA [40]	0.763	0.842	0.869	0.567	0.854	0.533	0.733
BRISQUE [31]	0.682	0.846	0.817	0.743	0.808	0.396	0.740
GMLOG [38]	0.804	0.864	0.890	0.774	0.857	0.562	0.804
NR-GLBP [99]	0.657	0.879	0.858	0.787	0.862	0.697	0.804
NFERM [39]	0.731	0.890	0.908	0.797	0.879	0.628	0.810
NRSL	0.810	0.903	0.912	0.836	0.896	0.659	0.851

Table 4.8: SRCC comparison on the whole TID2013 database.

BIQA model	TID2013 (3000)												
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
CORNIA [40]	0.550	0.209	0.717	0.360	0.797	0.585	0.727	0.840	0.721	0.806	0.800	0.595	0.654
BRISQUE [31]	0.706	0.523	0.776	0.295	0.836	0.802	0.682	0.861	0.500	0.790	0.779	0.254	0.723
GMLOG [38]	0.748	0.591	0.769	0.491	0.875	0.693	0.833	0.878	0.721	0.823	0.872	0.400	0.731
NR-GLBP [99]	0.664	0.466	0.759	0.081	0.728	0.620	0.728	0.827	0.721	0.844	0.867	0.440	0.594
NFERM [39]	0.851	0.520	0.846	0.521	0.894	0.857	0.785	0.888	0.741	0.797	0.920	0.381	0.718
NRSL	0.813	0.457	0.867	0.393	0.902	0.787	0.700	0.886	0.795	0.818	0.891	0.345	0.805
BIQA model	TID2013 (3000)												
	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24	ALL	
CORNIA [40]	0.157	0.016	0.177	0.262	0.170	0.407	0.541	0.696	0.649	0.689	0.874	0.629	
BRISQUE [31]	0.213	0.197	0.217	0.079	0.113	0.674	0.198	0.627	0.849	0.724	0.811	0.567	
GMLOG [38]	0.190	0.318	0.119	0.224	-0.121	0.701	0.202	0.664	0.886	0.648	0.915	0.679	
NR-GLBP [99]	0.226	0.204	0.105	0.123	-0.023	0.580	0.447	0.507	0.762	0.748	0.830	0.560	
NFERM [39]	0.176	0.081	0.238	0.056	-0.029	0.762	0.206	0.401	0.848	0.684	0.878	0.652	
NRSL	0.117	0.323	0.136	0.194	-0.110	0.753	0.434	0.751	0.866	0.694	0.887	0.661	

4.2.5 Cross-database validation

In the previous experiments, the training and testing images are from the same database. In this subsection, we test the generalization capability of learning based BIQA models through

cross-database validation. We train the BIQA models using all the images from one database and test the performance on another database. Since learning-based BIQA models require example images of same or similar distortions for training, here we only conduct cross-database validation on the three singly-distorted databases (i.e., LIVE, CSIQsub, TID2013sub) as they share similar distortion types (i.e., JP2K, JPEG, WN, GB, FF). The SRCC results of cross-database validation on three databases are provided in Table 4.9.

It can be seen that NRSL performs quite well in this scenario. It delivers the best performance in 5 out of 6 cases, which has demonstrated the database independency and robustness of the proposed NRSL method.

Table 4.9: SRCC comparison on cross-database validation.

Train DB	Test DB	NIQE	BIQI	DIIVINE	BLINDS2	CORNIA	BRISQUE	GMLOG	NR-GLBP	NFERM	NRSL
LIVE	CSIQsub	0.869	0.762	0.871	0.901	0.898	0.890	0.897	0.911	0.907	0.917
	TID2013sub	0.811	0.821	0.865	0.855	0.879	0.878	0.907	0.916	0.913	0.916
CSIQsub	LIVE	0.905	0.755	0.801	0.894	0.920	0.919	0.903	0.917	0.918	0.921
	TID2013sub	0.811	0.738	0.862	0.765	0.852	0.874	0.879	0.893	0.904	0.921
TID2013sub	LIVE	0.905	0.742	0.828	0.894	0.907	0.877	0.889	0.899	0.838	0.896
	CSIQsub	0.869	0.669	0.823	0.864	0.859	0.861	0.794	0.821	0.863	0.875

4.2.6 Computational complexity

In many practical applications, it is much desired to have a low-complexity BIQA method which can estimate image quality in real-time. Therefore, we evaluate the computational complexity and running cost of all competing methods in Table 4.10. Experiments are performed on a ThinkPad T430S notebook with Intel Core i7-3520M CPU@2.9GHz and 8G RAM. The software platform is MATLAB R2014b (8.4) under Windows 7 Home Premium. The MATLAB source codes of all other IQA models are obtained from original authors except for NR-GLBP, which is implemented by ourselves according to [99]. The feature extraction time consumed by each IQA model for estimating the quality of one 512×512 image is listed in the second column. Assuming N is the number of pixels and d is the window size for local normalization,

we can observe that the proposed method NRSL is quite efficient with the time complexity of $\mathcal{O}(N(d^2 + 10))$, where $\mathcal{O}(d^2 N)$ is for the local contrast normalization, $\mathcal{O}(8N)$ is for the LBP calculation (8 is the number of neighbors for LBP), and $\mathcal{O}(2N)$ is for calculation of the two histograms. More specifically, it can process 10 images per second, which meets the time requirement in most image processing applications.

Table 4.10: Computational complexity and run-time comparison of BIQA models.

BIQA model	Time(s)	Time complexity	Note
NIQE	0.23	$\mathcal{O}(d^2 N)$	d: window size
BIQI	0.05	$\mathcal{O}(N)$	-
DIIVINE	15.52	$\mathcal{O}(N(\log(N) + m^2 + N + 392b))$	m: DNT neighborhood size; b: 2D histogram bin number
BLIINDS2	61.39	$\mathcal{O}((N/d^2) \log(N/d^2))$	d: window size
CORNIA	2.45	$\mathcal{O}(d^2 K N)$	d: window size; K: codebook size
BRISQUE	0.08	$\mathcal{O}(d^2 N)$	d: window size
GMLOG	0.06	$\mathcal{O}(N(h + k))$	h: filter size; k: probability matrix size
NR-GLBP	1.65	$\mathcal{O}(4N(h + 5t))$	h: filter size; t: number of threshold
NFERM	54.03	$\mathcal{O}(d^2 N \log(N))$	d: window size of AR model
NRSL	0.11	$\mathcal{O}(N(d^2 + 10))$	d: window size

4.3 Summary

In this chapter, we have proposed a novel BIQA algorithm based on structural and luminance information. Unlike previous BIQA models normally generate features by fitting transformed image coefficients to specific probability distribution, in the proposed method, we utilize two statistical distributions on normalized luminance map to characterize the HVS-sensitive features. After local contrast normalization, the structural and luminance histograms are extracted to construct the image quality-aware features. The structural histogram captures the influence of various distortions on the inter-pixel relationship patterns, while the luminance histogram describes the global variation pattern of pixel-wise luminance value. Our research findings suggest that complementary information of structural and luminance features plays an important role for quality assessment tasks. Extensive experimental results on seven large-scale public image databases have demonstrated that the proposed method NRSL is highly competitive to

the state-of-the-art BIQA methods in terms of prediction accuracy, distortion consistency, and database independency.

Chapter 5

BIQA Based on Structural Degradation

In this chapter, we propose a general-purpose BIQA method for image quality estimation based on structural degradation. This method belongs to perceptual feature based BIQA methods. It is also an extended version of previous work presented in Chapter 4. The extensions are mainly from two aspects: 1) we show that linear filter response can complement the widely used local contrast normalization response; 2) we fuse the luminance and structural information through a weighting scheme to achieve more perceptual relevance.

Local contrast normalization has been widely used as a nonlinear processing step to mimic visual masking and eliminate the influence of varying illumination. Recently, it has been incorporated into the design of spatial domain BIQA methods [40, 31, 41], including our work presented in Chapter 4. As a nonlinear image processing step, we have found that local contrast normalization response cannot well capture the variation in global contrast. Linear filter response can complement such limitation. It is a coincidence that in some visual science studies, which evidenced that human visual cortex has separate mechanisms to process first- and second-order patterns [100]. The first-order patterns are image structures detected by a linear filter which describe luminance changes. There are many linear models proposed for early visual system, such as Gabor filterbank, independent component analysis (ICA), and sparse coding. Linear filter simplifies the model complexity, but it also limits the types of visual structures that can be derived from image statistics. Moreover, the HVS is able to detect ob-

jects defined by image attributes other than luminance. The second-order patterns are those cannot be detected by linear filters, such as texture, in which there is no change in mean luminance [100, 101, 102].

Therefore, we take into account of both the first- and second-order image structures and validate their effectiveness in the IQA task. The proposed method is denoted as BSD (Blind quality assessment based on Structural Degradation).

5.1 The Proposed Method

Motivated by the research findings that the HVS has separate mechanisms to process first- and second-order structural information, we extract corresponding image features for BIQA task. These two structural features are both affected by distortions and hence they have significant contribution to quantify the quality of distorted images. The flowchart of the proposed BSD method is shown in Figure 5.1. It consists of two main stages: quality-aware feature extraction and SVR regression. There are two sets of extracted features. The first feature set is deduced from the gradient image which accounts for first-order image structures. And the second feature set is deduced from the texture image which accounts for second-order image structures. Finally, these two feature sets are combined together to characterize image structural information that is important for perceptual quality. The features are extracted at three scales to account for the multiscale properties of natural images. Apart from the original scale, the coarser scale is formed by low-pass filtering and downsampling the image by a factor of two.

5.1.1 Feature set I

The first-order image structures characterize image local luminance changes, which are generally obtained by convolution with linear filters. Considering the perceptual significance of edge information in image processing and understanding, here we adopt the simple Prewitt filters to extract first-order image structures.

Figure 5.1: Flowchart of the proposed BSD metric.

First, the gradient information is computed by convolving images with Prewitt filters defined as (Eq. 5.1).

$$\mathbf{p}_x = \begin{bmatrix} 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & 0 & -1/3 \end{bmatrix}, \mathbf{p}_y = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 \\ -1/3 & -1/3 & -1/3 \end{bmatrix} \quad (\text{Eq. 5.1})$$

The gradient magnitudes are then computed by horizontal and vertical gradients as follows:

$$\mathbf{g}(i) = \sqrt{(\mathbf{d} * \mathbf{p}_x)^2(i) + (\mathbf{d} * \mathbf{p}_y)^2(i)} \quad (\text{Eq. 5.2})$$

where symbol “*” denotes the convolution operation, $\mathbf{d}$ and $\mathbf{g}$ denote the distorted image and its corresponding gradient magnitude map with i as the location index.

The distributions of gradient magnitudes of natural images are observed to follow a two-parameter Weibull distribution and human brain responses strongly correlate with Weibull image statistics for visual perception [103]. The probability density function of Weibull distribution is given by:

$$p(x) = \frac{\gamma}{\beta} \left(\frac{x}{\beta}\right)^{\gamma-1} \exp\left(-\left(\frac{x}{\beta}\right)^\gamma\right) \quad (\text{Eq. 5.3})$$

where x is the image gradient response, γ and β are the shape and scale parameter. γ denotes the peakness of the distribution and is sensitive to local edge spatial frequency; β represents the width of the distribution and reflects the local contrast. A wide distribution indicates a texture with high contrast. Since local contrast change has a significant impact on the perception of image quality, we employ these two parameters to characterize the impact of various distortions.

Figure 5.2 shows one reference image and its four distorted versions: WN, GB, JPEG and JP2K. Figure 5.3 (a) illustrates the gradient magnitude distributions of the five images in Figure 5.2. We can make the following observations. First, the distortions make the gradient magnitude distribution deviate from that of the reference image. Second, different distortion types result in quite different variations. WN makes the distribution more uniform as it introduces random noise. GB, JPEG and JP2K lead to a more peaked distribution. Among them,

GB and JP2K are more similar due to the introduced blur effect. JPEG has a second peakedness at the moderate magnitude due to the false edges introduced by blockiness impairment. The distinct changes of gradient distribution caused by different distortions indicate that it can effectively reflect the quality variation. Figure 5.3 (b) illustrates the scatter plot of Weibull parameters of five images in Figure 5.2 and we can observe that the distorted images are clearly separated from the reference image in the Weibull parameter space.

Figure 5.2: (a) Reference image. (b) Image with WN. (c) Image with GB. (d) Image with JPEG. (e) Image with JP2K.

5.1.2 Feature set II

The second-order image structures characterize image texture attributes that cannot be detected by linear filters. Here we adopt local contrast normalization to extract second-order image structures. Local contrast normalization has been adopted as a nonlinear processing stage to emulate visual masking and reduce statistical dependencies of input visual signals [86]. In this work, we employ the same processing model as [31, 81],

$$\tilde{I}(i, j) = \frac{I(i, j) - \mu}{\sigma + C} \quad (\text{Eq. 5.4})$$

Figure 5.3: (a) The gradient magnitude distributions of images with different distortion types. (b) The Weibull parameters of gradient magnitude distributions.

where i and j are the spatial indices of the image, and μ and σ are the local mean and standard deviation of the surrounding local patch. The constant C is induced to guarantee numerical stability when the denominators are close to zero.

Figure 5.4: (a) The gradient map of Figure 5.2(a) . (b) The texture map of Figure 5.2(a).

Figure 5.4 (b) shows response image of local contrast normalization. The mean subtraction process makes the local patch of the response image have zero mean luminance and thus invisible to linear filters. Furthermore, division by local contrast suppresses high contrast region, and strengthens image regions with modest contrast. By doing so, the fine image structural details are more evident. Thus, we denote the image after local contrast normalization as tex-

ture image which readily to extract the second-order image structures. From Figure 5.4 we can observe that the image gradient map generally highlights the edge structures corresponding to object boundaries, while the texture map is to characterize fine-detail structures related to object surfaces and materials.

We extract structural features from texture image to describe the second-order structural patterns. Specifically, we calculate the rotation invariant uniform LBP for each pixel in the texture image and denote it as TLBP. After applying LBP operator at each location in the texture image, we can get one TLBP map. The local contrast normalization mimics the HVS's contrast masking process, thus the response magnitude of each pixel in the texture map represents the visibility at that location. Image regions with high visibility would make more contribution for the later perceptual processing. We extract the global structural features from this TLBP map as the visibility weighted TLBP histogram.

$$h_{tlbp}(k) = \sum_{i=1}^M \sum_{j=1}^N \omega_{i,j} f(TLBP_{P,R}(i,j), k), \quad (\text{Eq. 5.5})$$

$$f(x, y) = \begin{cases} 1, & x = y \\ 0, & \text{otherwise,} \end{cases} \quad (\text{Eq. 5.6})$$

where M and N denote the image size, $k \in [0, K]$ is the possible TLBP pattern, and $\omega_{i,j}$ is the weight assigned to the TLBP code. In this work, we use the magnitude of texture image as the TLBP weight of each pixel.

Figure 5.5 shows the TLBP map, weighting map, TLBP histogram and weighted TLBP histogram of images in Figure 5.2, respectively. We can make the following observations. 1) The TLBP map changes with respect to different distortions as the distortion may shift the TLBP from one type to another. The $TLBP_{8,1}$ have 10 distinct patterns in the texture image. WN degrades the edge information, thus TLBP patterns between 2 to 6 are greatly reduced. Besides, it introduces the non-uniform pattern 9. The blurring effect of GB widens image edge thus creates false edge pattern such as 4. JPEG increases pattern 4 and 8, where 4 is caused by

the horizontal and vertical edges of blocking effect and 8 denotes the flat patterns in the texture image. JP2K widens image edge and creates ringing effect around object boundary, thus the TLBP patterns w.r.t edges have increased. Thus, TLBP is an effective descriptor to measure the impact of distortions on image structures. 2) The weighting map is perceptually relevant. First, image structures are highly stressed in the weighting map. Second, image distortions are more apparent in the weighting map, such as the random spot of noise, false edges of blockiness and ringing effect of JP2K. 3) TLBP histogram and the weighted histogram effectively summarize the image global structural information with a compact representation. Different distortions change the histograms in their characteristic ways.

To show the perceptual significance of the weighting scheme, we calculate the histogram similarity between the reference and distorted images in Figure 5.2 before and after the perceptual weighting and compare with their MOS in Table 5.1. The histogram similarity is calculated as

$$S(R, D) = \sum_{j=0}^r \min(R(j), D(j)). \quad (\text{Eq. 5.7})$$

where R and D are the histogram representations of reference and distorted images, j is the bin index. In Table 5.1, S_{hist} is the similarity of the TLBP histogram between reference and distorted image; S_{whist} is the similarity of the weighted TLBP histogram between reference and distorted image. Distorted image that is more similar to the reference image would have higher image quality. For Table 5.1, we observe that the TLBP histogram is not in agreement with quality perception. For example, WN has lower quality than GB, but the TLBP histogram of WN is more similar to the reference image than that of GB. The same situation happens when comparing the TLBP histograms of JPEG and JP2K. After applying the weighting scheme, the weighted histograms are in high agreement with human ratings.

5.1.3 Analysis of two feature sets

Local contrast normalization has been widely used in many image processing and computer vision applications. For computer vision applications such as object recognition, it has the

Figure 5.5: The contribution of the weighting scheme. The four rows show the TLBP map, weighting map, TLBP histogram, and weighted TLBP histogram. (a*) Reference image. (b*) Image with WN. (c*) Image with GB. (d*) Image with JPEG. (e*) Image with JP2K.

Table 5.1: The effect of the weighting scheme.

	WN	GB	JPEG	JP2K
MOS	3.91	4.97	3.16	2.83
S_hist	0.87	0.84	0.71	0.81
S_whist	0.83	0.84	0.77	0.74

effect of increasing the robustness of systems to varying illumination and contrast. In IQA domain, it has been shown that coefficients after local contrast normalization of natural images obey Gaussian-like distribution. There are several BIQA methods that extract features solely based on the response image resulting from local contrast normalization [40, 31, 41]. However, the perception of image quality has been influenced by the luminance and contrast changes. Here, we would like to point out the limitations of local contrast normalization and advocate complementary component to address such limitations. In Figure 5.6, we show one image and its contrast enhanced version using linear mapping. The second row shows their response image after contrast normalization. The perceptual quality of the two images are distinguishing from each other, however, such difference cannot be captured by the normalized image. In this work, we propose to use the linear filter response as the complementary element to capture the impact of luminance and contrast changes.

The proposed two feature sets are complementary in several aspects. First, the local contrast normalization would render the response image not sensitive to local contrast change, which is also an importance factor for quality. The gradient magnitude map is an effective contrast measure to complement the loss information caused by local contrast normalization. Second, the local contrast normalization can be taken as the visual masking process. It is well accepted that the edge regions have very weak masking effect compared to the textured regions. Thus, such uniform masking process without distinguishing image local content properties would overestimate the masking effect of edge regions. Features extracted from gradient map would complement such effect.

5.1.4 Regression model for quality prediction

The SVR with RBF kernel is adopted as the mapping function for feature pooling from the feature space to quality measure [71].

Figure 5.6: (a) One image example. (b) The contrast enhanced version of (a) using linear mapping. (c) Local contrast normalization response of (a). (d) Local contrast normalization response of (b).

5.2 Performance Evaluation

5.2.1 Database description

The comparison experiments are conducted on four benchmark IQA databases (i.e., LIVE [90], CSIQ [91], TID2008 [104], MLIVE [47]). For CSIQ [91] and TID2008 [104] databases, we only test the four distortion types (i.e., JP2K, JPEG, WN, GB). Also, we exclude the 25th reference image with its distorted versions from TID2008 database [104] as it is not a natural image. The four databases are chosen as they are the most widely cited databases in IQA community, and the distortions considered are the most commonly encountered distortion types in image communication system. The summarized information of four databases is shown in Table 5.2.

Table 5.2: Benchmark databases for IQA performance validation.

Database	Source images	Distorted images	Distortion types	Subjective scores
LIVE	29	779	5	DMOS (0-100)
CSIQ	30	866	6	DMOS (0-1)
TID2008	25	1700	17	MOS (0-9)
MLIVE	15	450	2	DMOS (0-100)

5.2.2 Implementation details

During implementation, for TLBP calculation, the number of neighbors P is 8 and the radius of the neighborhood R is 1. The proposed structural features are extracted at three scales. Thus, the extracted features are 36 dimensions in total.

5.2.3 Overall performance on individual database

The comparison experiments are conducted by using nine state-of-the-art BIQA models, including NIQE [81], ILNIQE[34], BIQI [25], DIIVINE [42], BLIINDS2 [30], CORNIA [40],

BRISQUE [31], GMLOG [38] and NRSL. We also include the performance of FR-IQA methods PSNR and SSIM [8] for reference.

In this section, we test the overall performance of twelve IQA methods on individual database. Since BSD adopts SVR learning for quality estimation, we need to divide the database into training and testing subsets. In our experiments, for each database, distorted images of 80% of the reference images are used for training, and the rest are used for testing. There is no content overlap between these two sets. This training-testing split is repeated 1000 times and the median performance is reported. Although FR-IQA metrics (PSNR and SSIM) and opinion-unaware BIQA metrics (NIQE and ILNIQE) do not require training on the database, we also report the median performance across 1000 trials on the testing subset for consistent comparison. The overall performance measured by SRCC, KRCC, PLCC, and RMSE criteria is listed in Table 5.3. For each criteria, the best two BIQA models are highlighted in boldface. Moreover, in order to provide an evaluation of the overall performance of the competing IQA models, the weighted-average SRCC, KRCC and PLCC results over four databases are also presented in Table 5.4 and the weight assigned to each database is the number of distorted images in that database.

Form Table 5.3, it can be observed that BSD performs consistently well on all the benchmark databases. BSD and NRSL are the best two methods on the examined databases. The performance gain of BSD over NRSL has confirmed the effectiveness of the propose extensions. GMLOG is the third best on LIVE, CSIQ, TID2008 databases. However, its performance on MLIVE database is rather poorly. Actually, on MLIVE, GMLOG is the second worst performing BIQA model (only better than NIQUE). CORNIA is the third best on MLIVE databases, and the fourth best on LIVE database. BRISQUE is the fourth best on CSIQ, TID2008 and MLIVE databases. When all the four databases are considered together, BSD achieves the best performance, followed by NRSL. BRISQUE, GMLOG and CORNIA achieves similarly acceptable performance. Therefore, we can conclude that objective quality scores predicted

Table 5.3: Performance comparison of twelve IQA models on four benchmark databases. The best two BIQA models are shown in boldface.

IQA model	LIVE (779)				CSIQ (600)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
PSNR	0.8847	0.7016	0.8833	12.7905	0.9288	0.7636	0.8539	0.1453
SSIM [8]	0.9396	0.7912	0.9341	9.7151	0.9242	0.7727	0.9329	0.1014
NIQE [81]	0.9084	0.7333	0.9088	11.3756	0.8826	0.6998	0.8930	0.1256
ILNIQE [34]	0.9024	0.7224	0.9086	11.4064	0.8857	0.7098	0.8746	0.1330
BIQI [25]	0.8519	0.6685	0.8660	13.5893	0.8196	0.6265	0.8736	0.1367
DIIVINE [42]	0.9089	0.7450	0.9087	11.3953	0.8800	0.7005	0.8988	0.1216
BLIINDS2 [30]	0.9310	0.7773	0.9370	9.4820	0.8693	0.6934	0.9120	0.1153
CORNIA [40]	0.9446	0.7935	0.9468	8.7817	0.8933	0.7218	0.9288	0.1045
BRISQUE [31]	0.9436	0.8005	0.9475	8.6707	0.9142	0.7568	0.9396	0.0954
GMLOG [38]	0.9498	0.8105	0.9538	8.1874	0.9231	0.7659	0.9507	0.0860
NRSL	0.9517	0.8186	0.9557	8.0184	0.9302	0.7788	0.9533	0.0838
BSD	0.9618	0.8377	0.9653	7.1283	0.9330	0.7806	0.9489	0.0885
IQA model	TID2008 (384)				MLIVE (450)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
PSNR	0.8789	0.6959	0.8604	0.8046	0.7316	0.5648	0.8217	10.8397
SSIM [8]	0.9104	0.7494	0.9408	0.5351	0.9033	0.7331	0.9277	6.9740
NIQE [81]	0.8033	0.6067	0.8163	0.9130	0.7892	0.6006	0.8580	9.4891
ILNIQE [34]	0.8779	0.6895	0.8775	0.7571	0.8997	0.7235	0.9139	7.5382
BIQI [25]	0.8017	0.6153	0.8524	0.8259	0.8838	0.7034	0.9065	7.7560
DIIVINE [42]	0.8972	0.7232	0.9034	0.6770	0.8827	0.6964	0.8946	8.2898
BLIINDS2 [30]	0.8663	0.6954	0.9058	0.6736	0.8946	0.7188	0.9109	7.7735
CORNIA [40]	0.8970	0.7317	0.9305	0.5808	0.8998	0.7254	0.9156	7.5857
BRISQUE [31]	0.9049	0.7401	0.9264	0.5945	0.8964	0.7253	0.9212	7.3597
GMLOG [38]	0.9372	0.7916	0.9451	0.5165	0.8327	0.6474	0.8717	9.1642
NRSL	0.9442	0.8053	0.9596	0.4451	0.9314	0.7773	0.9469	5.9163
BSD	0.9557	0.8278	0.9673	0.4020	0.9402	0.7913	0.9506	5.7498

Table 5.4: Weighted-average performance on four benchmark databases.

DB	PSNR	SSIM	NIQE	ILNIQE	BIQI	DIIVINE	BLIINDS2	CORNIA	BRISQUE	GMLOG	NRSL	BSD
SRCC	0.8647	0.9231	0.8594	0.8931	0.8410	0.8939	0.8960	0.9136	0.9195	0.9169	0.9406	0.9487
KRCC	0.6897	0.7673	0.6758	0.7136	0.6551	0.7195	0.7288	0.7499	0.7632	0.7624	0.7973	0.8113
PLCC	0.8590	0.9337	0.8784	0.8952	0.8738	0.9023	0.9196	0.9329	0.9365	0.9349	0.9540	0.9583

by BSD correlate consistently better with subjective ratings than all the other BIQA models examined.

Figure 5.7 provides the box-plot of SRCC performance on LIVE database over 1000 trails. From this figure, we can see that the proposed BSD outperforms all competing IQA metrics by delivering outstanding and consistent performance in accordance with human subjective ratings. Moreover, BSD exhibits the smallest standard deviation and is thus more content-robust than other IQA methods.

Figure 5.7: Box-plot of SRCC distributions of IQA models on LIVE database.

5.2.4 Performance on individual distortion type

In this section, we test the performance of ten BIQA models on individual distortion types. We train on the 80% of images with various distortion types and then test on the left 20% of images with the specific distortion type. The SRCC comparison on four benchmark databases is tabulated in Table 5.5. There are 15 groups of distortion images in the four databases. The best two BIQA models for each distortion group are shown in boldface. It should be noted that similar results are obtained for KRCC, PLCC and RMSE criteria, thus we only list SRCC here for brevity.

We can observe that BSD delivers outstanding performance for individual distortion types. Specifically, it outperforms other BIQA methods in 12 out of 15 distortion groups. For the GB distortion in LIVE database, it performs slightly worse than CORNIA and NRSL. For the JPEG distortion in CSIQ database, it achieves similar performance with NRSL and outperforms all other BIQA methods. Beyond that, for the two multiple distortion types in MLIVE database, BSD outperforms other BIQA models at a quite large margin. In general, when distortion is of a specific type, BSD achieves consistently better performance than all other BIQA models.

Table 5.5: SRCC comparison of ten BIQA models on individual distortion type. The best two models are shown in boldface.

DB	D-Type	NIQE	ILNIQE	BIQI	DIIVINE	BLIINDS2	CORNIA	BRISQUE	GMLOG	NRSL	BSD
LIVE	JP2K	0.9239	0.9002	0.8241	0.9055	0.9310	0.9233	0.9162	0.9258	0.9433	0.9532
	JPEG	0.9423	0.9439	0.8840	0.8968	0.9501	0.9419	0.9654	0.9632	0.9597	0.9673
	WN	0.9715	0.9791	0.9648	0.9822	0.9462	0.9631	0.9800	0.9831	0.9844	0.9867
	GB	0.9395	0.9244	0.8563	0.9337	0.9146	0.9560	0.9509	0.9293	0.9591	0.9511
	FF	0.8616	0.8443	0.7433	0.8538	0.8747	0.9108	0.8857	0.9012	0.8803	0.9126
CSIQ	JP2K	0.9252	0.9237	0.8178	0.8714	0.8786	0.9061	0.9006	0.9181	0.9214	0.9240
	JPEG	0.8834	0.9052	0.8594	0.8830	0.8950	0.8870	0.9239	0.9169	0.9372	0.9217
	WN	0.8356	0.8679	0.8502	0.9007	0.8683	0.7990	0.9359	0.9464	0.9554	0.9564
	GB	0.9059	0.8682	0.8444	0.8954	0.8826	0.9106	0.9164	0.9145	0.9261	0.9266
TID2008	JP2K	0.9023	0.9370	0.8549	0.8947	0.9023	0.9203	0.9005	0.9353	0.9520	0.9639
	JPEG	0.8872	0.9038	0.8481	0.8842	0.8752	0.9098	0.8932	0.9383	0.9308	0.9398
	WN	0.8165	0.8827	0.7555	0.8398	0.6848	0.5639	0.8052	0.8913	0.8993	0.9160
	GB	0.8469	0.8602	0.8992	0.8902	0.8571	0.9038	0.9018	0.8857	0.9178	0.9474
MLIVE	GB+JPEG	0.8938	0.9098	0.8831	0.8827	0.9029	0.9050	0.9086	0.8667	0.9284	0.9449
	GB+WN	0.8317	0.9159	0.8854	0.8950	0.9016	0.9029	0.8990	0.8188	0.9375	0.9467

5.2.5 Cross-database validation

In this section, we conduct the following experiments. Since LIVE, CSIQ, TID2008 only contain natural images with similar distortion types, we conduct cross-database validation across the three databases. In Table 5.6, we train BIQA models on all the images from LIVE database and test them on the other two databases. The best two results are in boldface. Again, it is notable that BSD performs well for cross-database validation. The cross-database performance results have confirmed the database independency and robustness of the proposed BSD

method.

Table 5.6: SRCC comparison on cross-database validation when models are trained on LIVE.

Test DB	NIQE	ILNIQE	BIQI	DIIVINE	BLIINDS2	CORNIA	BRISQUE	GMLOG	NRSL	BSD
CSIQ	0.8693	0.8800	0.7846	0.8765	0.9016	0.8983	0.8895	0.8985	0.9168	0.9143
TID2008	0.7949	0.8704	0.8127	0.8669	0.8635	0.8896	0.8936	0.9107	0.9190	0.9232

5.2.6 Run-time comparison

The running cost of each BIQA model is shown in Figure 5.8. Experiments are performed on a ThinkPad T430S notebook with Intel Core i7-3520M CPU@2.9GHz and 8G RAM. A 2D scatter plot of weighted average SRCC across four benchmark databases $SRCC_{wa}$ versus running time of different methods for feature extraction of one 512×512 image is shown in this figure, which is helpful to demonstrate the trade-off between quality prediction accuracy and efficiency of compared methods. From Figure 5.8, we can see that BSD is quite efficient with low computational complexity.

5.3 Summary

It's widely accepted that structures are important for IQA tasks. In this chapter, we propose a novel general-purpose BIQA method based on structural degradation. We extract image structural features from gradient and texture images to characterize the first- and second-order structural information. These two feature sets are complementary to each other and the combination of them are shown to be effective for BIQA task. Experimental results have demonstrated the superiority of the proposed method in comparison with relevant state-of-the-arts.

Figure 5.8: $SRCC_{wa}$ versus running time (log scale) of different methods.

Chapter 6

BIQA Based on Joint Log-contrast Statistics

Perceptual features used in BIQA models are designed by the HVS characteristics. Some properties such as multi-scale decomposition, visual masking, contrast sensitivity are implicitly or explicitly utilized in the feature design. There is a belief that the HVS is adapted to the statistics of natural images as visual natural environment is the main stimuli in most vision oriented tasks. Thus, NSS models also embody characteristics of the HVS.

Considering the perceptual relevance of NSS models, they are also incorporated in many existing general-purpose BIQA methods. The underlying philosophy of NSS based methods is that natural images possess certain statistical regularity, which may be violated in the presence of various distortions. Therefore, the deviation of the statistics of a query image from NSS model is indicative of the distortion severity of that image. A plethora of NSS models in different spatial or transform domains have been incorporated into representative BIQA methods.

Despite that NSS models in different domains have been applied in previous BIQA methods, we observe that the construction and utilization of NSS models follow the same framework. Generally, the marginal distributions of some bandpass coefficients are used to fit a univariate probabilistic model, then the model parameters are adopted as the quality-aware features. This framework has some limitations: 1) only univariate probabilistic model has been adopted by previous BIQA methods; 2) there is no NSS model to account for the across scale

and orientation correlation; 3) when measuring the correlation, only two neighboring pixels or subbands are considered each time; 4) using model parameters to represent the distribution without considering the fitting errors may result in an inaccurate representation.

In this work, we propose a novel NSS based general-purpose BIQA method to address these limitations. First, we build a comprehensive NSS model of joint log-contrast statistics to account for the across space and orientation correlation simultaneously. Furthermore, we employ a new approach to extract quality-aware features as the gradient of log-likelihood on the NSS model parameters. In this way, we circumvent the fitting error problems as no curve fitting is involved in the feature extraction. Finally, partial least square (PLS) regression is used to map the extracted quality-aware features to subjective quality scores.

6.1 The Proposed Method

In this work, we propose to measure across location and direction correlation using the joint directional log-contrast distribution and build a comprehensive NSS model to approximate it. We show that joint directional log-contrast responses are sensitive to degraded images of different distortion types and quality levels, and NSS model built on it is indicative of image quality. Then we extract quality-aware features as the Fisher Vector (FV) representation on the generative NSS model. Finally, PLS regression is used to map the quality-aware features to quality score.

6.1.1 Joint directional log-contrast statistics

In this work, we use the directional log-contrast responses in spatial domain to measure the impact of various distortions on natural images, which is defined as follows. First, for each image pixel, we define a squared symmetric neighbor set of P pixels placed on a square whose sides have the length $(2R + 1)$, as shown in Figure 6.1. Here P is the number of neighboring pixels, and R is the spatial resolution of the operator. Although we define the operator for a

general case based on a squared symmetric neighbor set, one can also generalize them to a circular one by interpolation.

We use the pairwise pixel differences to account for the across space correlation. Specifically, we subtract the center pixel value x_c from the P neighboring pixels on the square side $(x_1, x_2, \dots, x_P)$, yielding the feature vector $\mathbf{y} = (y_1, y_2, \dots, y_P) = (x_1 - x_c, x_2 - x_c, \dots, x_P - x_c)$. $\mathbf{y}$ is the gray-level differences vector which denotes the contrast response in different directions. To account for the human nonlinear perception on physical stimuli, we use the following logarithm function to compress the magnitudes of the gray-level differences:

$$z_i = \text{sign}(y_i) \log(|y_i| + 1) \quad (\text{Eq. 6.1})$$

The directional log-contrast responses for each pixel x_c is then represented by the vector $\mathbf{z}_c = (z_1, z_2, \dots, z_P)$, and z_i represents the log-contrast response in the i th direction. Unlike most previous works using the univariate parametric model to approximate the marginal distributions independently, in this work, we propose to jointly model the multivariate distributions by treating the directional log-contrast vector $\mathbf{z}_c$ as a whole.

The proposed method is based on the assumptions that cues to distortion type and level can be captured by the proposed joint distribution of directional log-contrast. To show that the joint distribution varies with the levels for distortions, we extract $\mathbf{z}_c$ at each pixel of distorted versions of the “painted house” image. Image blocks extracted from this group of distorted images are shown in Figure 6.2. We use settings of $(P = 24, R = 3)$ to calculate the directional log-contrast vector $\mathbf{z}_c$, then one pixel is associated with a 24-by-1 feature vector $\mathbf{z}_c$. We use the collection of $\mathbf{z}_c$ vectors from the pristine images to learn one principal component analysis (PCA) mapping matrix. Then the first principal components associated with the $\mathbf{z}_c$ computed from one distorted image can be obtained by mapping this collection using the previous PCA matrix. By fixing the PCA projecting matrix, we make the the first principal components of different distorted images directly comparable. The distributions of the first principal components of different distorted images are shown in Figure 6.3.

From Figure 6.3, we can observe that the first principal components of joint log-contrast responses are quite informative of image subjective quality. First, it is informative on the distortion type and images with different distortions have characteristic variation. Generally, WN changes the distribution to a more uniform one with larger standard deviation. GB, JPEG and JP2K generate a higher peaked distribution. GB and JP2K exhibit a smooth distribution, while JPEG has a jaggy distribution which may result from the blocking artifacts. Second, it is informative on quality levels. For the four distortion types considered here, we observe that the change of the distribution is a monotonic function of the subjective scores. For example, WN increases the standard deviation of the distribution. And we observe that the larger the standard deviation, the lower the subjective quality score. Similar observations can be made on the other three distortion types. Considering the effectiveness of the first principal components of joint log-contrast responses to represent image quality, we can safely assume that more quality-awareness can be achieved if we model the joint log-contrast responses altogether.

Figure 6.1: Squared symmetric neighborhood for different (P, R)

6.1.2 Joint NSS model

In this work, we use the GMM to represent the joint distribution of log-contrast response of natural images. Given a feature space $\mathcal{F} \in \mathcal{R}^D$, a GMM is defined as:

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{k=1}^K \omega_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (\text{Eq. 6.2})$$

Figure 6.2: Examples of image blocks with different types and levels of distortions. The distortion types for the four rows are: WN, GB, JPEG, JP2K. (Block size = 128×128 .)

Figure 6.3: The first principal components of joint log-contrast response of distorted images. (a) WN. (b) GB. (c) JPEG. (d) JP2K. MOS is the subjective quality score of each image.

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)^T} \quad (\text{Eq. 6.3})$$

with prior probability $\omega_k \in \mathcal{R}$, mean $\boldsymbol{\mu}_k \in \mathcal{R}^D$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathcal{R}^{D \times D}$. The GMM distribution is a linear combination of Gaussian densities, with the ω_k as the probabilistic contribution of the k th component. Thus, it can be taken as a probabilistic codebook, with each Gaussian component as one codeword, and the priors ω_k as the probabilistic soft assignment on each codeword.

In this work, we build a universal GMM model to represent the joint distribution of log-contrast response for natural images. We extract a large set of log-contrast vector $\mathbf{z}_c$ computed from natural images, and build the GMM model on it. The GMM parameters are learned using the Expectation Maximization (EM) algorithm. For computational simplicity, we use GMM with diagonal covariances, and apply PCA on the feature vector before GMM learning.

6.1.3 Quality-aware feature extraction

After building the GMM model to represent the joint distribution of log contrast responses of natural images, we extract quality-aware features from each image based on the learned joint NSS model. The traditional way to extract features from NSS model is to train the NSS model for each distorted image and use the model parameters as features for quality assessment. However, for the joint NSS model, training the GMM for each image would be quite time-consuming, especially when the components of GMM increase. Thus, we adopt a new method to extract quality-aware features from NSS model by using the FV encoding [105].

Given one GMM model, the FV calculates the gradient with respect to the model parameters, thus it describes how the set of descriptors extracted from one image deviates from the average distribution represented by the universal GMM model [105]. Previous works observed that the gradient with respect to the weight parameter ω_k does not provide significant information. Thus, in this work, the gradient is calculated with respect to the mean $\boldsymbol{\mu}_k$ and

the covariance Σ_k of the GMM. Given a set of low-level features from an image, $z_c^1, \dots, z_c^N$, where N is the number of features extracted from one image, we first apply the PCA to reduce the correlations and get $z_1, \dots, z_N$. The log-likelihood of z_i from this GMM is given by:

$$\mathcal{L}(z_i) = \log \sum_{k=1}^K \omega_k \mathcal{N}(z_i; \mu_k, \Sigma_k) \quad (\text{Eq. 6.4})$$

Assuming that low-level features in an image are independent, the log-likelihood of a set $\mathbf{Z}$ of N features is then represented as:

$$\mathcal{L}(\mathbf{Z}) = \sum_{i=1}^N \log \left(\sum_{k=1}^K \omega_k p_k(z_i) \right) \quad (\text{Eq. 6.5})$$

Fisher encoding [105] calculates the gradient of this log-likelihood with respect to GMM model parameters for all the feature vectors from one image:

$$\nabla \mathcal{L}(\mathbf{Z}) = \sum_{i=1}^N \nabla \log \left(\sum_{k=1}^K \omega_k \mathcal{N}(z_i; \mu_k, \Sigma_k) \right) \quad (\text{Eq. 6.6})$$

These gradients can be derived as follows for Gaussian model [105]:

$$\mathcal{G}_k^\mu = \frac{1}{\sqrt{\omega_k}} \sum_{i=1}^N \gamma_{ik} \left(\frac{(z_i - \mu_k)}{\sigma_k} \right) \quad (\text{Eq. 6.7})$$

$$\mathcal{G}_k^\sigma = \frac{1}{\sqrt{2\omega_k}} \sum_{i=1}^N \gamma_{ik} \left(\frac{(z_i - \mu_k)^2}{\sigma_k^2} - 1 \right) \quad (\text{Eq. 6.8})$$

where γ_{ik} is the probability that local descriptor z_i has been generated by the k -th Gaussian component.

$$\gamma_{ik} = \frac{\omega_k \mathcal{N}(z_i; \mu_k, \Sigma_k)}{\sum_{j=1}^K \omega_j \mathcal{N}(z_i; \mu_j, \Sigma_j)} \quad (\text{Eq. 6.9})$$

The final Fisher vector representation for the set of descriptors $\mathbf{Z}$ of one image is the concatenation of $\mathcal{G}_k^\mu$ and $\mathcal{G}_k^\sigma$.

6.1.4 Partial least square regression

By applying FK encoding to a K-component GMM with D-dimensional features, the final image representation is of length 2KD. For a GMM with 1024 components and features of 24 dimensions, this yields descriptors of 49,152 dimensions. Considering the high-dimensional image representation, partial least square (PLS) regression is adopted here to map from this high-dimensional feature space to quality scores due to its low-complexity and remarkable capability to handle high-dimensional data [106]. PLS regression works by reducing the input high-dimensional features to several uncorrelated latent components and performing least squares regression on these components.

Let $\mathcal{X} \in \mathbb{R}^m$ denote an m -dimensional space of input variables and $\mathcal{Y} \in \mathbb{R}^n$ represent an n -dimensional space of response variables. PLS models the relationships between these two data sets by means of score vectors. Let the number of samples in the data sets be N , PLS decomposes the zero-mean variables $\mathbf{X} \in \mathbb{R}^{N \times m}$ and $\mathbf{Y} \in \mathbb{R}^{N \times n}$ into

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad (\text{Eq. 6.10})$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \quad (\text{Eq. 6.11})$$

where $\mathbf{T} \in \mathbb{R}^{N \times p}$ and $\mathbf{U} \in \mathbb{R}^{N \times p}$ are matrices of the p extracted components (score vectors, latent scores), $\mathbf{P} \in \mathbb{R}^{m \times p}$ and $\mathbf{Q} \in \mathbb{R}^{n \times p}$ are matrices of loadings, and $\mathbf{E} \in \mathbb{R}^{N \times m}$ and $\mathbf{F} \in \mathbb{R}^{N \times n}$ are matrices of residuals. The classical form of PLS method uses the nonlinear iterative partial least squares (NIPALS) algorithm to iteratively decompose $\mathbf{X}$ and $\mathbf{Y}$ to find the weight vectors $\mathbf{w}, \mathbf{c}$ such that

$$[\text{cov}(\mathbf{t}, \mathbf{u})]^2 = [\text{cov}(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2 = \max_{|\mathbf{r}|=|\mathbf{s}|=1} [\text{cov}(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})]^2 \quad (\text{Eq. 6.12})$$

where $\mathbf{t}$ and $\mathbf{u}$ are the column vectors of matrices $\mathbf{T}$ and $\mathbf{U}$, and $\text{cov}(\mathbf{t}, \mathbf{u}) = \mathbf{t}^T \mathbf{u} / N$ is the sample covariance between latent vectors $\mathbf{t}$ and $\mathbf{u}$. After the extraction of the latent vectors $\mathbf{t}_1$ and $\mathbf{u}_1$, the matrices $\mathbf{X}$ and $\mathbf{Y}$ are deflated by subtracting their rank-one approximations based

on t_1 and u_1 . This process is repeated until the residuals are small enough or the desired number of latent vectors is obtained. More details on PLS analysis can be referred to [106].

In this work, we adopt the PLS regression as the mapping module from quality-aware features to quality scores. We use $\mathcal{X} \in \mathbb{R}^m$ to represent the feature space of extracted quality-aware features (i.e. Fisher vector representation), and $\mathcal{Y} \in \mathbb{R}$ to denote the quality score space of images.

6.2 Performance Evaluation

6.2.1 Implementation details

In this work, we have used three neighborhood settings as shown in Figure 6.1, and denote them as M1, M2, and M3. We fix the number of GMM components as 1024, and also analyze the performance of our methods using different choices of GMM component number. Before PLS regression, we also adopt power normalization followed by L2 normalization to further boost the performance of Fisher vector representation [107], and the power normalization parameter is 0.2.

The proposed method is examined on four large-scale image databases (i.e., LIVE [90], CSIQ [91], TID2013 [92] and CLIVE [95]). These four databases are selected for the following reasons: 1) LIVE and CSIQ are the most widely used benchmark databases which contain commonly encountered distortion types. 2) TID2013 the the largest IQA database so far and contains the highest number of image distortions. 3) CLIVE is the largest database on realistic distortions. The performance is evaluated by four criteria: SRCC, KRCC, PLCC and RMSE.

6.2.2 Parameter choice

The number of GMM components is a key factor to the performance. As the number of GMM components increases, the feature dimension increases and more discriminative information is

retained in the extracted features. However, it would also induce more memory and computational costs. In Figure 6.4, we show the performance measured by SRCC on four benchmark databases with different GMM sizes. From this figure, we observe that the performance of the proposed methods increase with the number of GMM components on all the four databases. However, the performance gradually flattens with the increase of GMM components. Therefore, to achieve a tradeoff between performance and computational complexity, we fix the number of GMM components as 1024 in the following experiments.

6.2.3 Overall performance on individual database

We compare our methods with several state-of-the-art BIQA methods, including NIQE [81], ILNIQE [34], BRISQUE [31], GMLOG [38], NRSL, and BSD. We split each database into training and testing sets with 80% for training and 20% for testing. For LIVE, CSIQ and TID2013, the split is conducted according to reference images to make sure that the two sets are content disjoint. For CLIVE, the database is split randomly to the training and testing sets. The split is repeated 1000 times and the median performance is reported. The 25th reference image and its distorted versions are excluded from TID2013 as it is not a natural image.

First, we show the box-plots of all examined methods on four databases. The box-plot shows the SRCC distributions of 1000 times of training-testing split. On each box, the central red bar is the median; the left and right edges of the box are the 25th and 75th percentiles; the whiskers extend to the most extreme data points not considered outliers. It can be clearly seen that the proposed methods achieve better performance than other competitors.

Moreover, the overall performance of examined IQA methods measured by SRCC, KRCC, PLCC and RMSE on four image databases is tabulated in Table 6.1. From this table, we observe that the proposed methods deliver better performance than other representative BIQA methods. For CSIQ, TID2013 and CLIVE databases, the proposed methods advance other methods at a significant margin. The CLIVE is a database with realistic distortions, which poses a new

(a) M1

(b) M2

(c) M3

Figure 6.4: The SRCC performance vs. the number of GMM components.

challenge to IQA research, however, our methods achieve quite promising performance on this database. Also, the three settings M1, M2, M3, using different sizes of neighborhood, achieve similar performance. More specifically, CSIQ and TID2013 achieves the best results with 3×3 neighborhood, CLIVE achieves the best results with 5×5 neighborhood, and LIVE achieves the best results with 7×7 neighborhood.

Figure 6.5: Box-plot of SRCC distributions of IQA models on four databases. (a) LIVE. (b) CSIQ. (c) TID2013. (d) CLIVE.

6.2.4 Performance on individual distortion type

In this section, we compare the performance of examined methods on individual distortion types. Table 6.2 shows the performance on LIVE and CSIQ databases, and Table 6.3 shows

Table 6.1: Overall performance comparison of BIQA methods on four benchmark databases. The best two methods are shown in boldface.

IQA model	LIVE (779)				CSIQ (866)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
NIQE	0.9084	0.7333	0.9088	11.3756	0.6299	0.4665	0.7253	0.1787
ILNIQE	0.9024	0.7224	0.9086	11.4064	0.8209	0.6323	0.8160	0.1482
BRISQUE	0.9436	0.8005	0.9475	8.6707	0.7403	0.5590	0.8053	0.1559
GMLOG	0.9498	0.8105	0.9538	8.1874	0.8035	0.6189	0.8465	0.1394
NRSL	0.9517	0.8186	0.9557	8.0184	0.8454	0.6681	0.8849	0.1199
BSD	0.9618	0.8377	0.9653	7.1283	0.8192	0.6351	0.8607	0.1331
M1	0.9580	0.8264	0.9608	7.5614	0.8793	0.7036	0.9138	0.1060
M2	0.9610	0.8318	0.9638	7.2626	0.8770	0.6972	0.9091	0.1090
M3	0.9651	0.8415	0.9685	6.8228	0.8779	0.6987	0.9102	0.1088
IQA model	TID2013 (2880)				CLIVE (1162)			
	SRCC	KRCC	PLCC	RMSE	SRCC	KRCC	PLCC	RMSE
NIQE	0.3238	0.2208	0.4200	1.1249	0.4496	0.3066	0.5084	17.4015
ILNIQE	0.5251	0.3730	0.6475	0.9461	0.4393	0.2990	0.5126	17.3462
BRISQUE	0.5739	0.4149	0.6682	0.9246	0.6072	0.4291	0.6445	15.4497
GMLOG	0.7100	0.5256	0.7694	0.7989	0.5972	0.4199	0.6235	15.8269
NRSL	0.6706	0.4951	0.7548	0.8185	0.6290	0.4472	0.6510	15.3444
BSD	0.6622	0.4875	0.7548	0.8143	0.6194	0.4398	0.6515	15.3417
M1	0.7641	0.5726	0.8176	0.7144	0.6694	0.4789	0.7017	14.3952
M2	0.7484	0.5602	0.8072	0.7325	0.7040	0.5106	0.7375	13.6803
M3	0.7405	0.5524	0.7994	0.7479	0.6919	0.5010	0.7246	13.9423

Table 6.2: SRCC comparison of BIQA methods on individual distortion type for LIVE and CSIQ. The best two methods are shown in boldface.

IQA model	LIVE (779)					CSIQ (866)					
	JP2K	JPEG	WN	GB	FF	WN	JPEG	JP2K	PGN	GB	CTD
NIQE	0.9239	0.9423	0.9715	0.9395	0.8616	0.8366	0.8826	0.9267	0.3275	0.9062	0.2786
ILNIQE	0.9002	0.9439	0.9791	0.9244	0.8443	0.8681	0.9043	0.9239	0.8835	0.8676	0.5447
BRISQUE	0.9162	0.9654	0.9800	0.9528	0.8857	0.6816	0.8465	0.8126	0.7410	0.8106	0.3912
GMLOG	0.9258	0.9632	0.9831	0.9293	0.9012	0.7992	0.8639	0.8910	0.7731	0.8611	0.5739
NRSI	0.9413	0.9586	0.9840	0.9572	0.8770	0.8004	0.9006	0.9097	0.8309	0.8919	0.6481
BSD	0.9546	0.9669	0.9849	0.9515	0.9070	0.8205	0.8702	0.8994	0.7422	0.9057	0.6246
M1	0.9478	0.9633	0.9818	0.9362	0.9139	0.8457	0.9026	0.9230	0.8861	0.8822	0.7242
M2	0.9455	0.9615	0.9815	0.9415	0.9253	0.8252	0.8968	0.9181	0.8705	0.8664	0.7293
M3	0.9538	0.9663	0.9813	0.9484	0.9286	0.8139	0.8941	0.9190	0.8745	0.8808	0.7212

the performance on TID2013 database. Since images degraded by realistic distortions has no definite distortion category, we exclude CLIVE database in this experiment. We only list the SRCC results in this section for brevity, KRCC, PLCC and RMSE would follow similar trend. On LIVE database, BSD ranks the best two methods 4 times, followed by M3 (3 times). We found that our methods achieve quite good performance on FF distortion, which is a difficult case for most IQA methods. On CSIQ database, M1 ranks the best two methods 5 times, followed by ILNIQE (4 times). On TID2013 database, ILNIQE ranks the best two methods 12 times, followed by M1 (9 times). The good performance of ILNIQE may be attributed to the diversity of distortions in TID2013. Since ILNIQE is an unsupervised method, it does not need training on distorted images. Thus, the compound of distortions with different characteristics would have no effect on its performance. For the supervised methods, we train the IQA model using distorted images with all kinds of distortions and test its performance on one specific distortion type. Since TID2013 contains diverse distortion types, and some distortions have different or even opposite characteristics, thus training on these distortions would deteriorate the model when tested on one kind of distortion.

Table 6.3: SRCC comparison of BIQA methods on individual distortion type for TID2013. The best two methods are shown in boldface.

IQA model	TID2013 (2880)											
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12
NIQE	0.8492	0.6976	0.7155	0.8100	0.8613	0.8162	0.8746	0.8385	0.6577	0.8808	0.9077	0.0104
ILNIQE	0.9034	0.8288	0.9377	0.6962	0.8843	0.8299	0.8867	0.8646	0.8019	0.8832	0.9209	0.3577
BRISQUE	0.6650	0.4812	0.7569	0.3178	0.8072	0.7714	0.6520	0.8768	0.5572	0.7942	0.7797	0.2931
GMLOG	0.7823	0.5845	0.7923	0.5155	0.8834	0.7218	0.8419	0.8900	0.7767	0.8546	0.8907	0.4977
NRSL	0.7823	0.4618	0.8465	0.4413	0.8871	0.7980	0.6541	0.8958	0.8207	0.8315	0.8954	0.4292
BSD	0.6700	0.4269	0.8615	0.3524	0.8394	0.7321	0.6469	0.9162	0.8078	0.8477	0.9136	0.6201
M1	0.7631	0.4393	0.7792	0.4900	0.8103	0.6182	0.8870	0.8754	0.8643	0.8569	0.9192	0.7854
M2	0.7669	0.4996	0.8100	0.4839	0.8681	0.6751	0.8054	0.8346	0.8662	0.8809	0.9215	0.7250
M3	0.7351	0.4762	0.7785	0.4308	0.8679	0.5896	0.7354	0.8373	0.8485	0.8660	0.9185	0.7062
IQA model	TID2013 (2880)											
	#13	#14	#15	#16	#17	#18	#19	#20	#21	#22	#23	#24
NIQE	0.6004	-0.1762	-0.0465	-0.1335	0.0433	-0.2250	0.7503	0.1662	0.8431	0.8095	0.7131	0.8738
ILNIQE	0.6600	-0.1772	-0.0569	0.2512	0.0623	-0.1187	0.7438	0.4285	0.8820	0.7977	0.8169	0.9054
BRISQUE	0.7119	0.1994	0.1982	0.2116	0.0842	0.0904	0.5986	0.2418	0.6436	0.8435	0.7613	0.8142
GMLOG	0.7673	0.1966	0.3800	0.1362	0.2893	-0.1296	0.7139	0.2250	0.7510	0.8962	0.6987	0.9254
NRSL	0.7815	0.0925	0.3279	0.1389	0.2923	-0.1529	0.7062	0.4933	0.7368	0.8562	0.7387	0.8938
BSD	0.7746	0.1496	0.2547	0.1531	0.0292	0.2746	0.5685	0.6162	0.7859	0.7592	0.7940	0.8677
M1	0.7865	0.2647	0.2718	0.1408	0.4988	-0.1254	0.7136	0.6754	0.8192	0.8696	0.7938	0.9100
M2	0.7315	0.2262	0.3016	0.2131	0.3696	-0.0008	0.7046	0.6304	0.8570	0.8523	0.7101	0.9012
M3	0.6954	0.1733	0.2994	0.1719	0.3926	-0.1415	0.6373	0.5958	0.8204	0.8677	0.7243	0.8960

6.2.5 Cross-database validation

We conduct cross-database validation to show the generalization ability of the BIQA methods. We train the methods using images from LIVE database and test on CSIQ and TID2013 databases. Only the four distortions shared by these databases are used (i.e., JP2K, JPEG, WN, GB). From this table, we can see that the proposed methods have good generalization ability to unseen distorted images.

Table 6.4: SRCC comparison on cross-database validation when models are trained on LIVE.

Test DB	NIQE	ILNIQE	BRISQUE	GMLOG	NRSL	BSD	M1	M2	M3
CSIQ	0.8693	0.8800	0.8895	0.8985	0.9168	0.9143	0.9124	0.8900	0.8904
TID2013	0.8106	0.8768	0.8776	0.9072	0.9160	0.9094	0.9152	0.9258	0.9172

6.3 Summary

In recent years, quality-aware features extracted from NSS models achieve great success in BIQA algorithms. Generally, the marginal distributions of bandpass coefficients are used to fit a parametric probabilistic model and the model parameters are adopted as the quality-aware features. However, the correlation across space, direction and scale of natural images cannot be well exploited by such NSS models as it is hard to capture such dependencies using univariate marginal distribution. In this chapter, we build a novel NSS model of joint log-contrast distributions to take into account the space and direction correlation of natural images. Furthermore, we provide a new approach to extract quality-aware features as the gradient of log-likelihood on the NSS model parameters. Finally, we develop a novel BIQA method which delivers outperforming quality prediction performance. Extensive experiments on four large-scale image databases have validated that objective quality scores predicted by the proposed BIQA method are in high accordance with subjecting ratings generated by human observers.

Chapter 7

Bag-of-Words Representation for Non-Intrusive SQA

Apart from image signals, speech signals are also an important part in multimedia systems. Research on non-intrusive SQA aims to estimate the perceptual quality of speech signals accurately and automatically, without requirement of the reference speech signals. Since speech signals are piecewise stationary, speech features are extracted on per-frame basis. By contrast, the quality score is a scalar rated on the whole speech utterance. How to obtain the global representation of the whole utterance from local segments is a useful problem with little investigation. In existing studies, statistical properties such as mean, variance, skewness and kurtosis across all frame-level features are calculated and combined together as the global features [65, 61, 108, 63]. However, much information is lost by summarizing the frame-level features using several kinds of statistic moments. In this study, we explore the BoW model for non-intrusive speech quality estimation. In the BoW representation, a speech utterance is treated as a text document and features extracted from local segments as words.

7.1 The Proposed Method

In this section, we describe the proposed non-intrusive SQA metric based on BoW model. The BoW representation elides the temporal sequence of local segments within a speech utterance

and represents the signal as a histogram of codewords. Figure 7.1 shows the flowchart of the proposed BoW metric. First, the speech signals are segmented into overlapped frames of 20 ms length and the 5th order PLP (perceptual linear prediction) coefficients are extracted as local features. Then, all PLP features extracted from the training speech signals are clustered by k-means algorithm to create a codebook with the cluster centers as codewords. Next, the extracted local PLP features are assigned to codewords with the minimum distance, and speech utterance is represented as a histogram of the codewords. After that, the BoW representation of speech signals as well as MOS are used as input to SVR to model the complex non-linear relationship between the BoW representation of speech signals and the perceived quality scores.

Figure 7.1: Flowchart of the proposed BoW metric for non-intrusive SQA.

7.1.1 Local PLP feature extraction

The PLP coefficients proposed by Hermansky [109] are used to represent the speech spectrum by a compact set of linear prediction coefficients after warping the frequency axis into the Bark frequency scale. Three psycho-acoustic concepts are considered to estimate the auditory spectrum: critical band spectral analysis, the equal loudness curve and the intensity power law. Here, the 5th order PLP coefficients are computed to suppress the speaker-dependent information. Figure 7.2 shows the flowchart of PLP computation.

Figure 7.2: Flowchart of the PLP feature extraction.

7.1.2 Codebook construction

The next stage aims to build a codebook of local speech segments, which is used for feature space quantization. Given each speech utterance, we segment it into overlapped frames of 20 ms length. PLP feature vectors are calculated for all the speech segments. By repeating this operation on all speech signals in the training set, we can obtain a large set of PLP feature vectors Θ . A codebook $\mathbb{C}$ is then created from this set by using k-means clustering algorithm with each learned cluster center as one codeword.

7.1.3 BoW representation of speech signals

Speech utterances are represented as histograms on the codebook obtained in the previous stage. After the codebook is constructed, an effective speech utterance representation is obtained by the following steps: 1) extract local PLP descriptors, 2) assign the local segment to

Table 7.1: The ITU-T P.Supplement-23 database.

DB	language	No. of utterances	No. of conditions	Distortion type
Exp.1(A)	French	176	44	codecs interworking
Exp.1(D)	Japanese	176	44	
Exp.1(O)	English	176	44	
Exp.3(A)	French	200	50	channel errors and background noise
Exp.3(C)	Italian	200	50	
Exp.3(D)	Japanese	200	50	
Exp.3(O)	English	200	50	

the codeword with minimum distance, 3) calculate the count of each codeword occurring in the speech utterance, and 4) normalize the histogram to obtain the BoW model which represents the probability of the occurrence of different codewords. Therefore, the codebook histogram H_S is used as the input to SVR regression for estimating the quality score of utterance S .

7.1.4 SVR mapping

We adopt the SVR with RBF kernel to learn the complex nonlinear mapping between codebook histogram and subjective quality score [71]. The SVR parameters are selected through cross validation on the training set.

7.2 Performance Evaluation

7.2.1 Experimental setup

We use the ITU-T P.Supplement-23 database [110] to validate the performance of the proposed metric. Experiment-1 and Experiment-3 are ACR-labeled and used in our experiments. Experiment-1 is divided into three data sets (A, D and O), each having 176 speech utterances at 44 different conditions. Experiment-3 is divided into four data sets (A, C, D, and O), each having 200 speech utterances at 50 different conditions. Therefore, there are seven databases including 1,328 speech utterances at 332 conditions. We summarize the database in Table 7.1.

We adopt the cross database strategy for performance evaluation, i.e., training is performed on six databases and validation on the remaining database.

The performance of SQA metrics can be evaluated by PLCC and RMSE between subjective and predicted scores. A third-order monotonic polynomial mapping recommended by [66] is applied before calculating the performance measures.

7.2.2 Experimental results and analysis

To validate the advantage of using BoW representation for non-intrusive speech quality estimation, we also build a non-intrusive metric M1 using traditional feature combination method. The M1 consists of three steps: 1) extract local PLP features, 2) calculate statistical properties mean, variance, skewness and kurtosis of these per-frame features and combine them together as the global features, and 3) adopt SVR for nonlinear mapping from the global features to subjective ratings. We can see that M1 and the proposed BoW metric share the same local feature extraction and nonlinear mapping steps, and the performance difference would be directly related to the utilization of BoW representation. Also, the results of ITU-T standard P.563 are included for performance comparison as it represents the state-of-the-art non-intrusive SQA algorithms.

The PLCC comparison for these three algorithms over the seven databases of ITU-T P. Supplement-23 is presented in Table 7.2. From this table, we observe that the proposed BoW metric outperforms M1 algorithm in all the seven databases, which demonstrates that the BoW model is an effective representation for non-intrusive SQA. Other than that global statistical properties across all the frames such as mean and variance can only get a rough description of the speech signals, both local and global information has been included in the BoW representation. The local information is coded into the codebook entries, and the global information is described as codebook histogram. Besides, the M1 method performs the worst when tested on Exp.3 (C) database. The reason might be that there is no Italian language in training set (other six

databases). However, by using BoW model, the proposed algorithm can predict the quality of the unknown language. The BoW metric outperforms P.563 on four out of seven databases and by average. For three databases in Experiment 1, the proposed BoW metric achieves outstanding performance as compared to other two methods, which shows that BoW is quite suitable for distortions caused by various codes interworking. The RMSE comparison for these three algorithms is shown in Table 7.3.

Figure 7.3 shows the scatter plot of condition-averaged subjective scores versus condition-averaged objective scores predicted by the proposed BoW metric on Exp.1 (A) database. Each point denotes one of the 44 distortion conditions. From Figure 7.3, we can see that all data points line on the 45° line with few deviations from this line. It indicates that the proposed metric can accurately predict quality of speech signals in accordance with human perception.

Figure 7.3: Scatter plot of the condition-averaged MOS versus condition-averaged objective scores predicted by the BoW. Each point denotes one of the distortion conditions in the database.

7.3 Summary

In this study, we have proposed a novel non-intrusive SQA metric based on BoW representation. The proposed BoW method treats each speech utterance as a document and per-frame

Table 7.2: PLCC comparison on the seven databases of ITU-T P.Supplement-23.

Database	P.563	M1	BoW
Exp.1(A)-French	0.895	0.852	0.933
Exp.1(D)-Japanese	0.809	0.797	0.902
Exp.1(O)-English	0.913	0.866	0.949
Exp.3(A)-French	0.869	0.783	0.825
Exp.3(C)-Italian	0.840	0.239	0.849
Exp.3(D)-Japanese	0.931	0.849	0.888
Exp.3(O)-English	0.905	0.835	0.902
Mean	0.880	0.746	0.893

Table 7.3: RMSE comparison on the seven databases of ITU-T P.Supplement-23.

Database	P.563	M1	BoW
Exp.1(A)-French	0.335	0.393	0.271
Exp.1(D)-Japanese	0.372	0.382	0.273
Exp.1(O)-English	0.313	0.384	0.242
Exp.3(A)-French	0.334	0.420	0.382
Exp.3(C)-Italian	0.452	0.810	0.441
Exp.3(D)-Japanese	0.245	0.354	0.308
Exp.3(O)-English	0.306	0.395	0.310
Mean	0.337	0.746	0.318

PLP features extracted from local segments as words. The speech utterance is then represented as a codebook histogram. Experimental results have demonstrated that the BoW model is an effective representation for non-intrusive SQA as it can capture both local and global structural information.

Chapter 8

Summary and Future Work

8.1 Summary

Multimedia signals are subject to various distortions during the process of acquisition, compression, transmission and reproduction. To provide satisfactory services that meet user expectation, it is necessary to measure the human perception of multimedia quality objectively. In this thesis, we design several objective quality assessment methods by using computer programs to estimate human perceived quality of image and speech signals. We develop methods in three application domains, one is distortion-specific BIQA, which targets at multiple distortion types. The second one is general-purpose BIQA, which applies for all kinds of image distortions. The third one is non-intrusive SQA, to estimate the quality of speech utterances. This thesis targets at developing objective methods to evaluate the quality of multimedia signals.

The main technical contributions of this thesis are highlighted in Section 1.3 and detailed from Chapter 3 to Chapter 7. Here, we give a brief summary of works in this thesis. Following that, we would give some further possible directions for the related research topics.

8.1.1 Distortion-specific BIQA for multiple distortions

Image signals would undergo several processing stages before they are consumed by human subjects. Thus, it is common that images are simultaneously distorted by multiple types of

distortions. In Chapter 3, we propose a BIQA method to predict the visual quality of multiply-distorted images. In the proposed method, a novel structural feature is extracted as the gradient weighted histogram of LBP calculated on the gradient map, which is effective to describe the complex degradation pattern introduced by multiple distortions. This method works in the gradient domain, and it is designed based on the importance of gradient information to quality perception. Also, LBP describes the inter-pixel relationship, which we believe is important to capture the complex degradation by multiple distortions. Previous BIQA methods adopt the marginal distribution of bandpass transformed coefficients as quality-aware features, which describes the pixel-level statistics [25, 42]. Such features are insufficient to characterize the complex degradation pattern of multiple distortions. For example, white noise and Gaussian blur tend to widen and narrow these marginal distributions [47], and images simultaneously corrupted by these two distortions would resemble the marginal distributions of reference image. Under such a circumstance, these marginal distributions would fail to describe the impact of multiple distortions. On the other hand, the structural features used in this work aim to describe the inter-pixel relationship in an image neighborhood, and such image micro-structural patterns are more effective to capture the complex interaction between different distortions and their joint effect on the overall image quality.

8.1.2 General-purpose BIQA based on structural and luminance features

In Chapter 4, we introduce a general-purpose BIQA method by structural and luminance information, based on the characteristics of human visual perception for distorted image. We extract the perceptual structural features of distorted image by the LBP distribution. Besides, the distribution of normalized luminance magnitudes is extracted to represent the luminance changes in distorted image. After extracting the features for structure and luminance, SVR is adopted to model the complex nonlinear relationship from feature space to quality measure. First, we use the luminance histogram to represent the image luminance information, instead

of the GGD parameters as quality-aware features. It is advantageous in three aspects: 1) it frees us from making any, possibly misguided, assumptions about the underlying feature distribution; 2) it constitutes a more accurate representation to avoid any fitting error; 3) it is efficient to calculate the histogram. Second, we use the LBP histogram computed on the normalized luminance map to represent the image inter-pixel patterns, which is effective to capture the impact of various distortions. Finally, we have shown the independence and complementarity of these two features and combined them together to construct a novel BIQA method.

8.1.3 General-purpose BIQA based on structural degradation

Local contrast normalization has been widely adopted as a nonlinear processing step to mimic visual masking and increase the robustness to illumination and contrast variation. Although it has been widely used in IQA research, we have found that it cannot capture the variations in image global contrast. On the other hand, gradient magnitude is an effective contrast measure and a good indicator to image quality. Gradient calculation is one linear filter operation, and local contrast normalization is one nonlinear processing. Studies in visual science have evidenced that human visual cortex has separate mechanisms to process first- and second-order patterns. The image structures detected by linear filters are the first-order patterns which describe luminance changes, while patterns that cannot be detected by linear filters are often referred to as the second-order patterns. Therefore, we propose one BIQA method by taking account of both the first- and second-order image structures. Specifically, the Prewitt linear filters are used to extract first-order image structures and the local contrast normalization is employed to extract second-order image structures. Quality-aware features are extracted from these two image structural maps and used as the input to SVR to learn the nonlinear mapping from feature space to quality measure.

8.1.4 General-purpose BIQA based on joint NSS model

In recent years, quality-aware features extracted from NSS models achieve great success in the development of BIQA algorithms. Generally, the marginal distributions of bandpass coefficients are used to fit a parametric probabilistic model and the model parameters are adopted as the quality-aware features. However, the interlocation, interdirection and interscale correlation of natural images cannot be well exploited by such NSS models as it is hard to capture such dependencies using univariate marginal distributions. Here, we build a novel NSS model of joint log-contrast distribution to take into account the interlocation and interdirection correlation of natural images. This new model uses multivariate GMM to model the directional log-contrast response jointly. It is novel in several aspects. 1) To the best of our knowledge, this is the first multivariate NSS model to be used in BIQA. Previous BIQA methods generally adopt univariate probabilistic model (such as 1-D GGD, AGGD, WCD) to approximate the marginal distribution of pixel-wise coefficients or pairwise pixel differences. 2) This model is built by unsupervised machine learning instead of human expert observation. The general way to developing a NSS model is to collect a quantity of natural images as well as their distorted versions, and visualize their marginal distribution to show whether certain probabilistic model can well capture such distribution. This process would require the human expert as the designer to find the appropriate model. In our work, we use the unsupervised machine learning to design the appropriate NSS model. 3) By using multivariate GMM to jointly model the directional log-contrast response, we incorporate the correlation of interlocation and interdirection simultaneously. All these factors are considered together and their interaction can be well captured by the joint NSS model. While in previous BIQA works, the correlation can only be analyzed by two subbands at each time in a limited way.

8.1.5 Non-intrusive SQA based on bag-of-words representation

The architecture of objective SQA models generally consists of signal preprocessing, feature extraction (typically on a frame-by-frame basis) and mapping from the features to the quality ratings [111]. Thus, the fundamental differences among non-intrusive SQA models are determined by the type of frame-specific (local) features, features supplied to the mapping stage (local or global) and regression model used in the mapping [111]. In existing studies, to construct the global features which represent the characteristics of the whole speech utterance, statistical properties such as mean, variance, skewness and kurtosis across all frame-level features are calculated and combined together as a whole feature vector [108]. However, much information is lost by summarizing the frame-level features using several kinds of statistic moments. In this study, we explore the BoW model for non-intrusive speech quality estimation. In the BoW representation, a speech utterance is treated as a text document and features extracted from local segments as words. BoW constitutes an more accurate and comprehensive representation for speech utterances, thus it is more discriminative than statistic moments for quality assessment. The proposed non-intrusive SQA method yields quite promising performance.

8.2 Future Work

This study proposes several methods for multimedia quality assessment. Multimedia quality evaluation is still a challenging task and there are many problems existing. In this section, we would like to point out several problems and indicate the possible future directions.

- **Unsupervised BIQA**

Most existing BIQA methods consist of two steps. The first is to extract quality-aware features that are sensitive to a broad range of distortions, and the second step is to map these features to the subjective ratings. However, in practice, there are many kinds of distortions, and some of them have quite different impact on extracted features. Thus,

models training on some kinds of distortions may not generalize to other distortion types. One way to solving this problem is to employ unsupervised methods which do not need training on distorted images and subjective quality scores. Both supervised and unsupervised methods have their own advantages and disadvantages. The former ones generally achieve better performance due to insightful feature design and intelligent machine learning techniques. However, they require similar distortion types with test images present in the training set. Such requirement would limit their generalization ability to unknown distortion types. Clean model based unsupervised methods which are built by clean images only have more potential for quality assessment of signals with various distortion types. For now, there are several works [81, 34] focusing on proposing unsupervised BIQA methods, however, their performance is not quite satisfactory. Further research efforts are required in this direction.

- **Quality assessment on images with realistic distortions**

Most existing BIQA methods are designed and validated on images with synthetic distortions. The distortions are generated by post-processing an ideal clean reference images in a controlled manner. Realistic distortions are generated during image acquisition, processing and storage of devices. For the naturally-distorted images, there are no reference images with perfect quality, and FR/RR methods are not applicable in this scenario. Also, there is no definite distortion category for each image since authentic distortions are generally diverse, mixed and multipartite. The performance of existing BIQA methods deteriorates when tested on images with realistic distortions. This is to be expected since realistic distortions are generally more complicated. Considering its wide applications in practice, more research efforts are called to this direction.

- **Joint Audiovisual Quality Assessment**

Most studies on joint audiovisual quality assessment focus on conducting subjective assessment experiments and determining the relationship of overall audiovisual quality

with the audio-only and video-only quality [112]. There are also some works focusing on finding the impact of different artifacts on the overall quality. They define the quality as linear or nonlinear combination of compression artifacts, transmission artifacts and other distortions [113]. It is straightforward as existing speech and visual quality assessment models can be easily embedded in this framework to define the final quality. However, this is not the reflection of human quality assessment procedure. When humans are presented with audiovisual signals, they do not focus on audio-only and video-only quality separately, but to obtain the perceived impression of the hybrid signals. In the future, we will investigate the audiovisual quality models in a different framework: 1) to extract audio and visual features from the audiovisual signals and combine these features together and map to the overall quality; 2) to learn a multi-modal representation using audio and visual signals, and predict the overall quality based on this multi-model representation.

Publication List

- [1] **Qiaohong Li**, Weisi Lin, Jingtao Xu, Yuming Fang, Blind image quality assessment using statistical structural and luminance features. in IEEE Transactions on Multimedia, vol. 18, no. 12, pp. 2457-2469, Dec 2016. (Chapter 4 in this thesis is based on this paper)
- [2] **Qiaohong Li**, Weisi Lin, Yuming Fang, No-reference quality assessment for multiply-distorted images in gradient domain, in IEEE Signal Processing Letters, vol. 23, no. 4, pp. 541-545, April 2016. (Chapter 3 in this thesis is based on this paper)
- [3] **Qiaohong Li**, Weisi Lin, Yuming Fang, BSD: Blind image quality assessment based on structural degradation, Neurocomputing, vol. 236, pp. 93-103, May 2017. (Chapter 5 in this thesis is based on this paper)
- [4] **Qiaohong Li**, Weisi Lin, Yuming Fang, Blind image quality assessment based on joint log-contrast statistics, to be submitted to IEEE Transactions on Image Processing. (Chapter 6 in this thesis is based on this paper)
- [5] **Qiaohong Li**, Weisi Lin, Yuming Fang, Xinfeng Zhang, NRLRS: No-reference image quality assessment based on local region statistics, 2016 IEEE Visual Communications and Image Processing (VCIP).
- [6] **Qiaohong Li**, Weisi Lin, Yuming Fang, No-reference image quality assessment based on high order derivatives, 2016 IEEE International Conference on Multimedia and Expo (ICME). (oral)

- [7] **Qiaohong Li**, Weisi Lin, Jingtao Xu, Yuming Fang, Daniel Thalmann, No-reference image quality assessment based on structural and luminance information, 2016 International Conference on Multimedia Modeling (MMM). Springer International Publishing, 2016: 301-312. (oral)

- [8] **Qiaohong Li**, Yuming Fang, Weisi Lin, Daniel Thalmann, Bag-of-words representation for non-intrusive speech quality assessment, 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), Chengdu, 2015, pp. 616-619. (oral) (Chapter 7 in this thesis is based on this paper)

- [9] **Qiaohong Li**, Yuming Fang, Weisi Lin, Daniel Thalmann, Gradient-weighted structural similarity for image quality assessments, 2015 IEEE International Symposium on Circuits and Systems (ISCAS), Lisbon, 2015, pp. 2165-2168.

- [10] **Qiaohong Li**, Yuming Fang, Weisi Lin, Daniel Thalmann, Non-intrusive quality assessment for enhanced speech signals based on spectro-temporal features, 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Chengdu, 2014, pp. 1-6.

References

- [1] Kalpana Seshadrinathan and A. C. Bovik. Automatic prediction of perceptual quality of multimedia signals: a survey. *Multimedia Tools and Applications*, 51(1):163–186, 2011.
- [2] Zhou Wang and A. C. Bovik. Modern image quality assessment (synthesis lectures on image, video, and multimedia processing). *San Rafael, CA: Morgan Claypool*, 2006.
- [3] Weisi Lin and C-C Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.
- [4] Kalpana Seshadrinathan and A. C. Bovik. Video quality assessment. *The essential guide to video processing*, pages 417–436, 2009.
- [5] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Transactions on Broadcasting*, 57(2):165–182, 2011.
- [6] Scott J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. volume 1666, pages 2–15. 10.1117/12.135952.
- [7] Jeffrey Lubin and D Fibush. Sarnoff JND vision model, 1997.
- [8] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [9] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, 2005.

- [10] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.
- [11] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006–21, 2010.
- [12] Lin Zhang, D. Zhang, Xuanqin Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, Aug 2011.
- [13] Anmin Liu, Weisi Lin, and M. Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, April 2012.
- [14] Wufeng Xue, Lei Zhang, Xuanqin Mou, and A.C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, Feb 2014.
- [15] Zhou Wang and Eero P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model, 2005.
- [16] L. Ma, S. Li, F. Zhang, and K. N. Ngan. Reduced-reference image quality assessment using reorganized dct-based image representation. *IEEE Transactions on Multimedia*, 13(4):824–829, Aug 2011.
- [17] X. Gao, W. Lu, D. Tao, and X. Li. Image quality assessment based on multiscale geometric analysis. *IEEE Transactions on Image Processing*, 18(7):1409–1423, July 2009.
- [18] R. Soundararajan and A. C. Bovik. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2):517–526, Feb 2012.
- [19] J. Wu, W. Lin, G. Shi, and A. Liu. Reduced-reference image quality assessment with visual information fidelity. *IEEE Transactions on Multimedia*, 15(7):1700–1705, Nov 2013.

REFERENCES

- [20] E. Ong, W. Lin, Z. Lu, X. Yang, S. Yao, F. Pan, L. Jiang, and F. Moschetti. A no-reference quality metric for measuring image blur. In *2003. Proceedings. Seventh International Symposium on Signal Processing and Its Applications*, volume 1, pages 469–472 vol.1, July 2003.
- [21] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *2002 International Conference on Image Processing (ICIP).*, volume 1, pages I–477. IEEE, 2002.
- [22] Hamid Rahim Sheikh, Alan Conrad Bovik, and Lawrence Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005.
- [23] Yuming Fang, Kede Ma, Zhou Wang, Weisi Lin, Zhijun Fang, and Guangtao Zhai. No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Processing Letters*, 22(7):838–842, 2015.
- [24] Q. Li, W. Lin, and Y. Fang. No-reference quality assessment for multiply-distorted images in gradient domain. *IEEE Signal Processing Letters*, 23(4):541–545, April 2016.
- [25] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5):513–516, 2010.
- [26] L. He, D. Tao, X. Li, and X. Gao. Sparse representation for blind image quality assessment. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1146–1153, June 2012.
- [27] X. Gao, F. Gao, D. Tao, and X. Li. Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning. *IEEE Transactions on Neural Networks and Learning Systems*, 24(12):2013–2026, Dec 2013.
- [28] Yi Zhang, Anush K. Moorthy, Damon M. Chandler, and Alan C. Bovik. C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes. *Signal Processing: Image Communication*, 29(7):725 – 747, 2014.

- [29] M. A. Saad, A. C. Bovik, and C. Charrier. DCT statistics model-based blind image quality assessment. In *2011 18th IEEE International Conference on Image Processing*, pages 3093–3096, Sept 2011.
- [30] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012.
- [31] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.
- [32] D. Lee and K. N. Plataniotis. Toward a no-reference image quality assessment using statistics of perceptual color descriptors. *IEEE Transactions on Image Processing*, 25(8):3875–3889, Aug 2016.
- [33] Yi Zhang and Damon M. Chandler. An algorithm for no-reference image quality assessment based on log-derivative statistics of natural scenes, 2013.
- [34] Lin Zhang, Lei Zhang, and A.C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, Aug 2015.
- [35] Deepti Ghadiyaram and Alan C. Bovik. Feature maps driven no-reference image quality prediction of authentically distorted images, 2015.
- [36] Chaofeng Li, Alan Conrad Bovik, and Xiaojun Wu. Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks*, 22(5):793–799, 2011.
- [37] H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 305–312, June 2011.
- [38] Wufeng Xue, Xuanqin Mou, Leiqi Zhang, Alan C Bovik, and Xiaowei Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014.

REFERENCES

- [39] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1):50–63, Jan 2015.
- [40] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1098–1105. IEEE, 2012.
- [41] Le Kang, Peng Ye, Yi Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740, June 2014.
- [42] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011.
- [43] Lin Zhang, Zhongyi Gu, Xiaoxu Liu, Hongyu Li, and Jianwei Lu. Training quality-aware filters for no-reference image quality assessment. *IEEE MultiMedia*, 21(4):67–75, Oct 2014.
- [44] Vishwakumara Kayargadde and Jean-Bernard Martens. Perceptual characterization of images degraded by blur and noise: experiments. *J. Opt. Soc. Am. A*, 13(6):1166–1177, Jun 1996.
- [45] Damon M. Chandler, Kenny H. Lim, and Sheila S. Hemami. Effects of spatial correlations and global precedence on the visual fidelity of distorted images, 2006.
- [46] Damon M. Chandler. Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, 2013:53, 2013.
- [47] D. Jayaraman, A. Mittal, A.K. Moorthy, and A.C. Bovik. Objective quality assessment of multiply distorted images. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 1693–1697, Nov 2012.

- [48] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Transactions on Broadcasting*, 60(3):555–567, Sept 2014.
- [49] Ke Gu, Guangtao Zhai, Min Liu, Xiaokang Yang, Wenjun Zhang, Xianghui Sun, Wanhong Chen, and Ying Zuo. FISBLIM: A five-step blind metric for quality assessment of multiply distorted images. In *2013 IEEE Workshop on Signal Processing Systems (SiPS)*, pages 241–246, Oct 2013.
- [50] Yanan Lu, Fengying Xie, Tongliang Liu, Zhiguo Jiang, and Dacheng Tao. No reference quality assessment for multiply-distorted images based on an improved bag-of-words model. *IEEE Signal Processing Letters*, 22(10):1811–1815, Oct 2015.
- [51] Wongeun Oh and Sung-Keun Lee. *Quality Assessment of Sound Signals in Multimedia and Communication Systems*, volume 353 of *Communications in Computer and Information Science*, book section 8, pages 57–64. Springer Berlin Heidelberg, 2012.
- [52] Nicolas Ct. *Speech Quality Measurement Methods*, volume 1 of *T-Labs Series in Telecommunication Services*, book section 2, pages 37–85. Springer Berlin Heidelberg, 2011.
- [53] INC Dynastat. Diagnostic acceptability measure (DAM): a method for measuring the acceptability of speech over communication systems. specification DAM-IIC, Dynastat. Inc., Austin, TX, 1995.
- [54] ITU-T Rec. P. 835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. *Recommendation P*, 835, 2003.
- [55] S. Moller, Chan Wai-Yip, N. Cote, T. H. Falk, A. Raake, and Marcel Waltermann. Speech quality estimation: Models and trends. *IEEE Signal Processing Magazine*, 28(6):18–28, 2011.
- [56] John G. Beerends and Jan A. Stemerding. A perceptual speech-quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc*, 42(3):115–123, 1994.

REFERENCES

- [57] JG Beerends, AP Hekstra, AW Rix, and MP Hollier. Pesq, the new itu standard for objective measurement of perceived speech quality–part ii: perceptual model. *Journal of the Audio Engineering Society*, 50:765 – 778, 2002.
- [58] ITUT Rec. P. 863, perceptual objective listening quality assessment (POLQA). *International Telecommunication Union, CH-Geneva*, 2011.
- [59] O. C. Au and K. H. Lam. A novel output-based objective speech quality measure for wireless communication. In *1998 Fourth International Conference on Signal Processing Proceedings (ICSP)*, pages 666–669 vol.1, 1998.
- [60] Guo Chen and Vijay Parsa. *Objective Speech Quality Evaluation Using an Adaptive Neuro-Fuzzy Network*, volume 83 of *Studies in Computational Intelligence*, chapter 5, pages 97–116. Springer Berlin Heidelberg, 2008.
- [61] T. H. Falk, Xu Qingfeng, and Chan Wai-Yip. Non-intrusive gmm-based speech quality measurement. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 1, pages 125–128, 2005.
- [62] T. H. Falk and Chan Wai-Yip. Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1935–1947, 2006.
- [63] RajeshKumar Dubey and Arun Kumar. Non-intrusive speech quality assessment using several combinations of auditory features. *International Journal of Speech Technology*, 16(1):89–101, 2013.
- [64] Manish Narwaria, Weisi Lin, IanVince McLoughlin, Sabu Emmanuel, and ChiaLiang Tien. *Non-intrusive Speech Quality Assessment with Support Vector Regression*, volume 5916 of *Lecture Notes in Computer Science*, chapter 34, pages 325–335. Springer Berlin Heidelberg, 2010.
- [65] M. Narwaria, Lin Weisi, I. V. McLoughlin, S. Emmanuel, and Chia Liang-Tien. Non-intrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1217–1232, 2012.

REFERENCES

- [66] L. Malfait, J. Berger, and M. Kastner. P.563 The ITU-T standard for single-ended speech quality assessment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):1924–1934, 2006.
- [67] Jinjian Wu, Weisi Lin, Guangming Shi, and Long Xu. Reduced-reference image quality assessment with local binary structural pattern. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 898–901, June 2014.
- [68] Qiaohong Li, Yuming Fang, Weisi Lin, and D. Thalmann. Gradient-weighted structural similarity for image quality assessment. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 2165–2168, May 2015.
- [69] Wufeng Xue and Xuanqin Mou. Reduced reference image quality assessment based on weibull statistics. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6, June 2010.
- [70] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.
- [71] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.
- [72] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.
- [73] Video Quality Experts Group. Final report from the video quality experts group on the validation of objective models of video quality assessment. *VQEG*, Mar, 2000.
- [74] Howard C. Hughes, George Nozawa, and Frederick Kitterle. Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience*, 8(3):197–230, 1996.
- [75] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, Apr 2000.

REFERENCES

- [76] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004.*, volume 2, pages 1398–1402 Vol.2, Nov 2003.
- [77] D.M. Chandler and S.S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, Sept 2007.
- [78] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan. Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13(5):935–949, Oct 2011.
- [79] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. Perceptual quality metric with internal generative mechanism. *IEEE Transactions on Image Processing*, 22(1):43–54, Jan 2013.
- [80] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, Oct 2014.
- [81] A. Mittal, R. Soundararajan, and A.C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, March 2013.
- [82] HC Nothdurft. Sensitivity for structure gradient in texture discrimination tasks. *Vision Research*, 25(12):1957–1968, 1985.
- [83] Andrew B Watson and Joshua A Solomon. Model of visual contrast gain control and pattern masking. *JOSA A*, 14(9):2379–2391, 1997.
- [84] Robert A Frazor and Wilson S Geisler. Local luminance and contrast in natural images. *Vision research*, 46(10):1585–1598, 2006.
- [85] Valerio Mante, Robert A Frazor, Vincent Bonin, Wilson S Geisler, and Matteo Carandini. Independence of luminance and contrast in natural scenes and in the early visual system. *Nature neuroscience*, 8(12):1690–1697, 2005.

- [86] Siwei Lyu and Eero P Simoncelli. Nonlinear image representation using divisive normalization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- [87] Timo Ojala, Kimmo Valkealahti, Erkki Oja, and Matti Pietikinen. Texture discrimination with multidimensional distributions of signed gray-level differences. *Pattern Recognition*, 34(3):727 – 739, 2001.
- [88] Jinjian Wu, Weisi Lin, Guangming Shi, Xiaotian Wang, and Fu Li. Pattern masking estimation in image with structural uncertainty. *IEEE Transactions on Image Processing*, 22(12):4892–4904, Dec 2013.
- [89] Daniel L Ruderman. The statistics of natural images. *Network: computation in neural systems*, 5(4):517–548, 1994.
- [90] Hamid R Sheikh, Zhou Wang, Lawrence Cormack, and Alan C Bovik. LIVE image quality assessment database release 2, 2005.
- [91] Eric C Larson and DM Chandler. Categorical image quality (CSIQ) database. *Online*, <http://vision.okstate.edu/csiq>, 2010.
- [92] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.
- [93] A. Ciancio, A. L. N. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Transactions on Image Processing*, 20(1):64–75, Jan 2011.
- [94] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, Jan 2016.
- [95] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Hkkinen. CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 24(1):390–402, Jan 2015.

REFERENCES

- [96] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 995–1002, June 2013.
- [97] M. A. Saad, P. Corriveau, and R. Jaladi. Objective consumer device photo quality evaluation. *IEEE Signal Processing Letters*, 22(10):1516–1520, Oct 2015.
- [98] Mikko Nuutinen, Toni Virtanen, and Pirkko Oittinen. Image feature subsets for predicting the quality of consumer camera images and identifying quality dimensions. *Journal of Electronic Imaging*, 23(6):061111, 2014.
- [99] M. Zhang, C. Muramatsu, X. Zhou, T. Hara, and H. Fujita. Blind image quality assessment using the joint statistics of generalized local binary pattern. *IEEE Signal Processing Letters*, 22(2):207–210, Feb 2015.
- [100] Jonas Larsson, Michael S. Landy, and David J. Heeger. Orientation-selective adaptation to first- and second-order patterns in human visual cortex. *Journal of Neurophysiology*, 95(2):862–881, 2006.
- [101] Mihran Tuceryan, Anil K Jain, et al. Texture analysis. *Handbook of pattern recognition and computer vision*, 2:207–248, 1993.
- [102] Timothy Ledgeway and Andrew T. Smith. Evidence for separate motion-detecting mechanisms for first- and second-order motion in human vision. *Vision Research*, 34(20):2727 – 2740, 1994.
- [103] H Steven Scholte, Sennay Ghebreab, Lourens Waldorp, Arnold WM Smeulders, and Victor AF Lamme. Brain responses strongly correlate with weibull image statistics when processing natural images. *Journal of Vision*, 9(4):29–29, 2009.
- [104] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, M Carli, and F Battisti. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.
- [105] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

REFERENCES

- [106] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop*, pages 34–51, Berlin, Heidelberg, 2006.
- [107] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *2010 11th European Conference on Computer Vision (ECCV)*, pages 143–156, Berlin, Heidelberg, 2010.
- [108] Qiaohong Li, Yuming Fang, Weisi Lin, and Daniel Thalmann. Non-intrusive quality assessment for enhanced speech signals based on spectro-temporal features. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.
- [109] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.
- [110] ITU-T Rec. P. supplement 23: ITU-T coded-speech database. *International Telecommunication Union*, 1998.
- [111] Petko Nikolov Petkov. Improving quality of service in baseband speech communication. 2014.
- [112] Junyong You, Ulrich Reiter, Miska M. Hannuksela, Moncef Gabbouj, and Andrew Perkis. Perceptual-based quality assessment for audiovisual services: A survey. *Signal Processing: Image Communication*, 25(7):482–501, 2010.
- [113] M. N. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. Raake. Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2. In *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, pages 482–487.