The University Of Sheffield.

NEW LEARNING FRAMEWORKS FOR BLIND IMAGE QUALITY ASSESSMENT MODEL

Thesis submitted to the University of Sheffield for the degree of Doctor of Philosophy

By

REDZUAN BIN ABDUL MANAP

Department of Electronic and Electrical Engineering

The University of Sheffield

3 Solly Street

Sheffield, S1 4DE

United Kingdom

JULY 2018

ABSTRACT

The focus of this thesis is on image quality assessment, specifically for problems of assessing the quality of an image blindly or without reference information. There are significant efforts over the last decade in developing objective blind models that can assess image quality as perceived by humans. Various models have been introduced, achieving highly competitive performances and high in correlation with subjective perceptual measures. However, there are still limitations on these models before they can be viable replacements to traditional image metrics over a wide range of image processing applications. This thesis addresses several limitations. The thesis first proposes a new framework to learn a blind image quality model with minimal training requirements, operates locally and has ability to identify distortion in the assessed image. To increase the model's performance, the thesis then modifies the framework by considering an aspect of human vision tendency, which is often ignored by previous models. Finally, the thesis presents another framework that enable a model to simultaneously learn quality prediction for images affected by different distortion types.

ACKNOWLEDGEMENTS

First and foremost, all praise to GOD for all his blessings in my life. He who I put my belief in and He who I always seek help from. I thank Him for giving me strength and courage to push myself this far for there is no power nor strength except through Him.

Second, I would like to express my deepest gratitude to two main people behind this research study: Professor Ling Shao and Professor Alejandro F. Frangi. I have benefitted a lot working under their co-supervision, not only on the research itself but also as a human being. Through their ideas, knowledge and wisdom, I identified research directions and learned how to take them forward. Their patience, dedication and commitment exemplified some of the qualities to be a better person. This thesis will not be completed if not for their continuous encouragement and motivation. For these, I thank both of you.

I would also like to acknowledge all the current and the previous members of CISTIB. Fellow students in rooms C14 and C15: Bo, Peng, Nishant, Serkan, Mohsen, Santi and the rest with whom I enjoyed having stimulating intellectual discussions. I always appreciate all the help received during our group seminars, scientific meetings or retreat. The feedbacks or even the challenges, often provide me with valuable information that worth to consider in my study. To the administration and technical staffs of CISTIB, I also thank them for trying to accommodate me within CISTIB. The support and help especially during my early time there are greatly appreciated. Not to forget my first-year friends in B28 Portobello: Fan, Liu, Mengyu, Zia, Sandipan and others. I thank you all for the assistance and for making the room an enjoyable place to be in.

Living abroad comes with its own challenges. Different language, food, culture and lifestyles can sometimes make you feel homesick. Fortunately, this is not a case for me. Thanks to the Malaysian community here, who I often referred to as “my small family”, my three-and-

a-half-year stay was very much memorable. The togetherness and the kindness shown by them especially during the organised activities: academic, religious, sports or festive, will always be remembered. May our friendship continue to the end.

There is a popular saying that there is a woman behind every man's success. In my case, there is not one but six. To my wonderful mum, Siti Fatimah, I thank you for your love, wisdom and perseverance in raising me to be who I am today. In 2006, another beautiful woman came into my life. To my lovely wife, Siti Zakiah, I thank you for your unconditional love. I appreciate all the sacrifices you had to make while I pursued this dream in the UK. Your willingness to put your career on hold just to be here with me, taking care of our family when I was too busy with the study and your continuous encouragement even when I doubted myself tell me how much you care. And then there are four beautiful little princesses whom I willing to do anything for. Irdina, Zinnirah, Amani, and Nayli, you are my happiness. To my brothers and sisters, you too are the sources of inspiration.

Next, thank you to the Ministry of Higher Education of Malaysia and the Universiti Teknikal Malaysia Melaka (UTeM) as the main sponsors of my study under Skim Latihan Akademik Institusi Pengajian Tinggi Awam (SLAI) Awards Scheme (Biasiswa (BS):790907045119). Special mention to Professor Geraint Jewell, the Head of Electronic and Electrical Engineering Department, the University of Sheffield, for providing me a one-off financial assistance during the study. Without this funding, it would be difficult to imagine how finishing this thesis could be made possible. Last but not least, I thank my viva examiners: Professor Patrick Le Callet and Dr Charith Abhayaratne for their constructive feedback and comments on the study.

CONTENTS

List of Figures

List of Tables

List of Acronyms

General Acronyms

AGGD	Asymmetric Generalised Gaussian Distribution
AGM	Accelerated Gradient Method
AUC	Area Under ROC Curve
BIQA	Blind Image Quality Assessment
CNN	Convolutional Neural Network
dB	Decibels
DCT	Discrete Cosine Transform
DMOS	Differential Mean Opinion Score
DS	Distortion Specific
FF	Fast Fading
FPR	False Positive Rate
FR-IQA	Full Reference Image Quality Assessment
GB	Gaussian Blur
GBJPEG	Gaussian Blur and JPEG
GBWN	Gaussian Blur and White Noise
GM	Gradient Magnitude
GGD	Generalised Gaussian Distribution
GLBP	Generalised Local Binary Pattern
GRNN	General Regression Neural Network
HDR	High Dynamic Range
HVS	Human Visual System
I2C	Image-to-Class
I2I	Image-to-Image
IQA	Image Quality Assessment
IQR	Inter-Quartile Range
ITU	International Telecommunication Union
JP2K	JPEG2000
LBP	Local Binary Pattern
LCC	Linear Correlation Coefficient

LDR	Low Dynamic Range
LTP	Local Ternary Pattern
MOS	Mean Opinion Score
MSE	Mean Squared Error
MTL	Multi-Task Learning
NBNN	Naïve Bayes Nearest Neighbour
NCS	Natural Colour Statistics
NR-IQA	No Reference Image Quality Assessment
NSS	Natural Scene Statistics
PC	Pair Comparison
PSNR	Peak Signal-to-Noise Ratio
QoS	Quality of Service
RBF	Radial Basis Function
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
RP	Resolving Power
RR-IQA	Reduced Reference Image Quality Assessment
SIFT	Scale Invariant Feature Transform
SROCC	Spearman Rank Order Correlation Coefficient
STL	Single-Task Learning
SVM	Support Vector Machine
SVR	Support Vector Regression
TPR	True Positive Rate
V1	Primary Visual Cortex
WN	White Noise

Specific Acronyms (IQA Model)

BLIINDS	BLind Image Integrity Notator using DCT Statistics
BIECON	Blind Image Evaluator using CONvolutional neural network
BIQI	Blind Image Quality Index
BOWSF	Bag-Of-Words using Selected Features
BRISQUE	Blind / Reference less Image Spatial QUality Evaluator

CBIQ	Codebook Based Image Quality
CORNIA	COdebook Representation for No-reference Image Assessment
DESIQUE	DErivative Statistics based Image QUality Evaluator
DIIVINE	Distortion Identification based Image Verity and INtegrity Evaluation
DIQaM	Deep Image QuAlity Measure
FSIM	Feature SIMilarity
GMLOG	Gradient Magnitude and Laplacian Of Gaussian
GMSD	Gradient Magnitude Similarity Deviation
GWH-LBP	Gradient-Weighted Histogram of Local Binary Pattern
IQVG	Image Quality based on Visual saliency and Gabor filtering
LBIQ	Learning Based Image Quality
MAD	Most Apparent Distortion
MLIQM	Machine Learning for Image Quality Metric
MP-Q	Matching Pursuit based Quality index
MTLBIQ	Multi-Task Learning based Blind Image Quality assessment
NFEQM	No-reference Free Energy based Quality Metric
NFERM	No-reference Free Energy based Robust Metric
NFSDM	No-reference Free energy and Structural degradation Distortion Metric
NSS-GS	Natural Scene Statistics Global Scheme
OSVP	Orientation Selectivity for Visual Pattern
PATCH-IQ	PATCH based Image Quality assessment
RRED	Reduced Reference Entropic Differencing
RR-SSIM	Reduced Reference Structural SIMilarity
SDIQA	Saliency guided Deep framework for Image Quality Assessment
SFLNIA	Saliency based Feature Learning for No-reference Image Assessment
SSIM	Structural SIMilarity
STAIND	STAtistical INDependence
SV-CORNIA	SuperVised CORNIA
VIF	Visual Information Fidelity
VSNR	Visual Signal-to-Noise Ratio

Symbols and Notation

Symbol	Description
$\mathbf{D}$	Labelled dataset of image patches
$\mathbf{F_I}$	Feature matrix for an image
$\mathbf{G_I}$	GM operators of an image
$\overline{\mathbf{G_I}}$	Normalised GM operators of an image
$\mathbf{I}$	Image
$\mathbf{L_I}$	LOG operators of an image
$\overline{\mathbf{L_I}}$	Normalised LOG operators of an image
$\mathbf{K}_{m,n}$	Joint empirical probability function of GM and LOG operators
$\mathbf{N_I}$	Matrix of adaptive normalisation factor for GMLOG model of an image
$\mathbf{P}$	Image patch
$\widehat{\mathbf{P}}$	Normalised image patch
$\mathbf{X}$	Feature matrix for image patches
$\mathbf{W}$	Matrix of estimated parameters from training images
$\mathbf{Z}$	Matrix of search points for gradient-based optimisation algorithm
$\mathbf{g}(\cdot,\cdot)$	Isotropic Gaussian function
$\mathbf{h}_{\mathrm{LOG}}$	LOG filter
$\mathbf{h}_x$	Gaussian partial derivative filter (horizontal)
$\mathbf{h}_y$	Gaussian partial derivative filter (vertical)
K	Window size for Gaussian weighting function (vertical)
L	Window size for Gaussian weighting function (horizontal)
M	Quantisation level for normalised GM operators
N	Quantisation level for normalised LOG operators
N_{label}	Total number of labelled images
N_{test}	Total number of test images
P	Total number of patches in an image
P_{label}	Total number of patches in a labelled image
P_{test}	Total number of patches in a test image
$P_{\overline{\mathbf{G_I}}}(\cdot)$	Marginal probability function for normalised GM operators

Symbol	Description
$P_{\overline{\mathbf{L}_\mathrm{I}}}(\cdot)$	Marginal probability function for normalised LOG operators
$Q_{\overline{\mathbf{G}_\mathrm{I}}}(\cdot)$	Independency distribution function for normalised GM operators
$Q_{\overline{\mathbf{L}_\mathrm{I}}}(\cdot)$	Independency distribution function for normalised LOG operators
R	Total number of training samples in a learning task
T	Total number of learning tasks
b	Bias parameter for linear regression function
c	Actual image distortion class
$\hat{c}$	Predicted image distortion class
c_{SVR}	Constant for SVR model
d	Euclidean distance
f_p	Feature vector for an image patch
h_I	Image height
h_p	Patch height
$k(\cdot,\cdot)$	Kernel function
$l(\cdot,\cdot)$	Loss function
k_{NN}	Number of nearest neighbour samples
p_i	Image patch indices
q_I	Quality score for an image
q_p	Quality score for an image patch
r	Combination coefficient for SVR model
s	DMOS / MOS score for an image
w_I	Image width
w_p	Patch width
$\Gamma(\cdot)$	Gamma function
ξ, ξ^*	Slack variables for SVR model
$\varepsilon_{\mathrm{SVR}}$	Deviation parameter for SVR model
ε_B	Constant for BRISQUE model
$\varepsilon_{\mathrm{GMLOG}}$	Constant for GMLOG model
η	Maximum pixel value of an image
μ	Mean

Symbol	Description
σ	Standard deviation
σ^2	Variance
σ_l^2	Left variance for AGGD model
σ_r^2	Right variance for AGGD model
σ_{G}	Scale parameter for isotropic Gaussian function
γ	Shape parameter for GGD model
ν	Shape parameter for AGGD model
γ_{RBF}	Precision parameter for radial basis kernel function
ω	General weight vector
$\omega_{\mathrm{G}}(\cdot)$	2D circularly symmetric Gaussian weighting function
ω_{P}	Estimated weight vector for a patch
ω_{t}	Estimated weight vector for a training image
λ	Regularisation parameter
τ	Step size of iteration for gradient-based optimisation algorithm
$\Omega(\cdot)$	Regularisation term of an objective function
$\nabla f(\cdot)$	Gradient of function $f(\cdot)$

Chapter 1

Introduction

1.1 Image Quality

For many people, the eyes are probably the most important among our five sense organs. We use our eyesight to obtain information and to understand the world around us. In the past, our eyes were mainly adapted to real world observations. The introduction of various digital image capture devices such as digital camera and mobile phone, however, has changed this scenario. The ubiquitous use of these devices nowadays has led to a widespread presence of digital images in our everyday life. We are now not only looking at real world environments but increasingly also at digital images. As our eyes are getting used to the high visual quality of real world environments, the same quality level is often expected when we look at digital images.

How do we define image quality? Finding the exact definition to the term can be challenging since it depends on several factors such as the process of producing the image, the features that make up image quality and the viewers of the image. Hence there is no universal, formal definition to image quality. There have been, however, some literature attempted to give proper definition to it. For example, image quality was interpreted as the integrated of perception of the overall degree of excellence of an image [1]. It can also be understood as the subjective impression on how well image content is reproduced [2]. Meanwhile, in [3], quality was a representation of the level of sufficiency to the image function for a particular application domain.

The last definition is the most suitable in this study. It essentially means that the quality is defined differently depending on application for which it is defined. For example, people working in image acquisition applications such as laser range scanning measure quality based on the imaging system aspects. Printing workers focus on tone, colour and attributes such as line and area when they determine quality. For medical imaging researchers, the quality is related to the clarity with which they can detect malfunctions or diseases from the images [4]. In computer vision applications, such as object detection, robot navigation or traffic monitoring system, the quality of an image is often associated with the determination of the failure mode of computer algorithms [5]. Meanwhile, for image communication system, the quality of an image is associated to how well the image is acquired, processed or delivered over the transmission network.

Producing digital representation of an image at the end of an image communication system involves many stages such as image acquisition, compression, transmission or storage, decoding, and display. Figure 1.1 illustrates typical processing stages of image communication system. Note that the image can also be repurposed at any of these stages, which can entail re-acquisition, re-compression, or additional transmission. In addition, an enhancement algorithm can also be applied on any stages [6].

Figure 1.1: Simplified processing stages of an image communication system.

All these processing stages may introduce various distortions into the content of the image. Image acquisition or capture stage may introduce artefacts due to optical lens, sensing elements accuracy and digitisation process. Typical artefacts includes blurring, noise,

contouring, aliasing, contrast inversion and colour artefacts. Compression with block-based coders (such as JPEG, MPEG-2 or H.264) may produce blocking and flatness artefacts, while wavelet-based coders (such as JPEG2000) can cause blurring and ringing artefacts. Meanwhile, transmitting the image data through transmission medium may generate distortions such as inter-symbol interference and multipath (ghosting) artefacts. These are due to reflections and arrival of data from different propagation path. At the receiver stage, the decoding process can also introduce artefacts such as horizontal or vertical shift of image data or DC shift due to decoding errors. Colour, luminance, interlace and flicker artefacts can also be generated at the display stage due to poor contrast range, resolution limitations, interlacing scanning or scan rate conversion.

In addition, repurposing process, which commonly aim to display image at lower resolution device such as mobile phone or tablets, involves resampling and recompression that may also produce some distortions already discussed. Image enhancement process can produce other artefacts as well. For example, deblocking and denoising can cause blurriness while sharpening can cause ringing. For further details of type of distortions in image communication system, interested readers are referred to publications in [6], [7]. Note that the term 'distortion' here refers to a general degradation introduced into the content of an image without specifying any particular type while the term 'artefact' refers to a particular distortion.

These distortions affect the original structure of the image content leading to a reduced output quality. The level of quality degradation depends on the severity and the class of artefacts generated by those stages. Measuring the quality loss introduced in any stage is crucial especially for multimedia applications. For example, visual content and service providers can use this measurement to fine-tune parameters of image transmission systems according to the quality of the transmitted images. This quality monitoring process is important to ensure that

they satisfy a given Quality of Service (QoS) so the level of quality of experience at the users' end is acceptable [8].

1.2 Image Quality Assessment

The purpose of image quality assessment (IQA) is to gauge the quality of an image using quality metrics. These metrics differ depending on the applications and they vary from the ones that assess quality of an image affected by specific distortion to the ones that measure quality globally in the presence of various impairments. For applications which the end targets are usually human observers, it would be beneficial to employ metrics that can correctly quantify the image quality as anticipated by them.

1.2.1 Conventional image metrics

Since digital image in communication or computer vision systems is presented in a pixel-based format, the traditional approach attempts to estimate the image quality on a pixel-by-pixel basis. This is done by computing the difference in a given image's pixel values to those of its associated reference image. Here, a reference image refers to a similar image of a perfect quality and contains no distortion whereby its' information is utilised to assess the quality degradations in the image. Metrics measuring image fidelity: the mean squared error (MSE) and the peak signal-to-noise ratio (PSNR) are commonly employed to this effect. The MSE between two images $\mathbf{I}_1(i,j)$ and $\mathbf{I}_2(i,j)$ can be calculated as [9]:

$$\mathrm{MSE} = \frac{1}{h_\mathrm{I} w_\mathrm{I}} \sum_{i=1}^{h_\mathrm{I}} \sum_{j=1}^{w_\mathrm{I}} [\mathbf{I}_1(i,j) - \mathbf{I}_2(i,j)]^2 \ , \tag{1.1}$$

with h_I and w_I represent the image height and the image width, respectively. The PSNR can then be computed as [9]:

$$\mathrm{PSNR} = 10 \log \frac{\eta^2}{\mathrm{MSE}} \ . \tag{1.2}$$

In Equation (1.2), η is the image's maximum pixel value.

This simple pixel-based approach is computationally efficient, explaining their continual use in monitoring system performance and system optimisation. However, it has been shown that they have low correlation with human perception of quality [10]-[11]. This example further illustrates this point. Figure 1.2 consists of a reference image plus two examples of its degraded image. The image at the centre represents the reference image. The image to the left has been compressed using JPEG2000 (JP2K) encoder while the image to the right has been subjected to artificial white noise. The computed MSE and PSNR values are given in Table 1.1.

JP2K compression **Reference** **White noise**

Figure 1.2: Reference image and examples of its degraded version

Table 1.1: The computed MSE and PSNR values

Artefact	MSE	PSNR (dB)
JP2K compression	81.33	29.06
White noise	112.42	27.66

The MSE metric with a lower value indicates high similarity between the images. In the case of PSNR, which is measured in decibels (dB), high similarity between two images is represented by a higher value metric. From the table, we can see that the lower MSE value and the higher PSNR value for JP2K compressed image may lead to a conclusion it is more similar

to the reference image and is of higher quality compared to the image affected by white noise. However, it is different when we look from human perception angle. When comparing the two distorted images, we can see that the JP2K compressed image is of lower quality than the image affected by white noise. This observation highlights the downside of conventional image metrics in relation to human perception of quality.

1.2.2 Subjective image quality assessment

Human viewers are broadly agreed to be the most accurate evaluators to image quality. Therefore, subjective quality measures that based on human perception are often considered the gold standard in perceptual assessment of image quality. These measures are normally obtained by conducting image quality experiments where participating viewers evaluate the presented images' quality using rating scales. The ratings are then averaged across all observers with the computed averages are mostly reported in the form of mean opinion score (MOS) or differential mean opinion score (DMOS). A lower MOS value indicates higher quality while a higher DMOS value indicates a lower quality image. This score represents the perceived quality metric for the image.

These subjective experiments are conducted according the specifications of the international standards. Two main standards commonly used are given by the International Telecommunication Union (ITU). The first standard Rec. BT.500-11 [12], which is produced by the radio communications sector (ITU-R), focuses on television pictures. It includes both techniques of single and double stimulus. The quality rating of the distorted stimulus is made without referring to the original stimulus in the single stimulus method. In the double stimulus method, the quality rating is performed via the double stimuli continuous quality scale involving both stimuli. Meanwhile, the Telecommunications sector (ITU-T) specifies experimental procedures for multimedia applications in the second standard Rec. P.910 [13].

The quality of a stimulus is rated in the single stimulus method via an absolute category rating. Meanwhile, in the double stimulus method, a degradation category rating is utilised to rate the stimulus' quality.

The quality ratings produced from the subjective experiments are generally accepted to be the 'ground-truth' for quality prediction. However, that these subjective ratings need to involve human observers makes them expensive, time-consuming, and unfeasible for deployment in most real world applications. An automatic IQA model that can provide image quality metrics objectively is preferred. The obtained ratings, however, are still useful in designing and validating objective quality metrics.

Several image quality databases have been developed recently for these design and validation purposes. The databases usually contain the images used in the experiments and their associated quality ratings rated by the participants. The following are the list of some of databases commonly used in previous IQA works:

- LIVE Database [14]: The database was developed at the University of Texas at Austin, USA. It comprises 29 reference images of which 779 degraded images were produced. Each reference image was subjected to 5 to 6 degradation levels in five source coding and artificial artefacts: additive white noise (WN), JPEG compression (JPEG), JPEG2000 compression (JP2K), Gaussian blur (GB) and fast fading (FF). 29 observers were involved in the experiments.
- CSIQ Database [15]: The database was developed at the Oklahoma State University, USA involving 35 participants. It comprises 866 distorted images. They were generated when 6 types of artefacts (WN, JPEG, JP2K, GB, additive pink noise and global contrast decrements) were applied to 30 reference images at 4 to 5 degradation levels.

- TID2008 Database [16]: The database was developed at the Tampere University of Technology, Finland. It contains 17 types of artefacts of different types of noise, transmission errors, compression, local distortions, denoising, blur, contrast, and luminance changes. Each artefact was applied to 25 reference images at 4 degradation levels, resulting in 1700 distorted images. The ratings were collected from 838 observers.
- TID2013 Database [17]: The database is the latest version of the TID2008 database. The same 25 reference images were subjected to 24 types of artefacts at 5 degradation levels to produce 3000 distorted images. 7 additional artefacts were included in the database: multiplicative noise, comfort noise, colour saturation change, colour quantisation error, chromatic aberrations, lossy compression, and sampling error. A total of 985 people participated with the experiments.
- IRCCyN/IVC Database [18], [19]: The database was developed at the Institut de Recherche en Communications et Cybernetique de Nantes (IRCCyN), France. It consists of 235 distorted images generated from 10 reference images. 4 types of artefacts: GB, JP2K, JPEG, JPEG and locally adaptive resolution coding; were applied to the reference image at five degradation levels. 15 subjects were involved to produce the ratings.
- A57 Database [20]: The database was developed at Cornell University, USA. It has 3 reference images that were subjected to 6 different artefacts at 3 degradation levels. The artefacts are WN, GB, JPEG, JP2K, customised JP2K via dynamic contrast-based quantization algorithm, and quantization errors of discrete wavelet transform LH sub-bands. The resulting 54 distorted images were rated by 7 observers.
- Toyoma/MICT Database [21]: The database was developed at University of Toyoma, Japan. There are 168 distorted images and 14 reference images. The distorted images

were generated based on 2 types of artefact: JPEG and JP2K at 7 degradation levels. The ratings were produced by 16 observers.

- IRCCyN_IVC_Toyoma Database [22], [23]: The database, developed at the Institut de Recherche en Communications et Cybernetique de Nantes (IRCCyN), France, improves the MICT database. It uses different protocol, different type of display device and different populations to generate further ratings for the MICT images. There were 27 subjects participated in the experiments.

These databases are developed under constraint that the images are subjected to a single type of distortion only. Taking into account that images may subjected to multiple types of distortion in more realistic life scenarios, there are recent efforts to develop database of subjective evaluation of multiply distorted images. Examples of the multiply distorted image database are as follow:

- LIVEMD Database [24], [25]: The database was developed at the University of Texas at Austin, USA. In the database, 15 reference images are first blurred at 4 levels. The blurred images are then subjected to two types of artefact, JPEG and WN, at 4 levels each. In all, 225 single / multiple distorted images are generated for each of the two cases: GBJPEG and GBWN. The ratings were collected from 19 observers.
- MDID2013 Database [26], [27]: The database was developed at Shanghai Jiao Tong University, China. The database consists of 12 reference images. These images were subjected to blurring, JPEG compression and noise successively, producing a total of 324 3-fold distorted images. A total of 25 subjects participated in the experiments.
- MDID2017 Database [28], [29]: The database was developed at Shenzhen Tsinghua University, China. There are 20 reference images. Each image was first subjected to blurring or contrast change, then compressed by JPEG or JPEG2000 and finally

subjected to additional noise. There are 4 degradation levels of each artefacts yielding a total of 1600 distorted images. The ratings were collected from 192 observers.

In addition to the above mentioned singly or multiply distorted image databases, there are other recent databases that were developed based on different motivations and targeted applications. Some differ on the nature of utilised images, some use real distortion rather than simulated ones, some collect ratings through online crowdsourcing platform rather than the typical in-lab sessions while others propose new rating representation other than MOS/DMOS. Examples of these recent databases include:

- DRIQ Database [30]: The database was developed at the Oklahoma State University, USA. Rather than associating quality with the distortion levels/types of an image, the study look on how the quality is perceived by human on enhanced images. The database consists of 26 reference images in which 78 enhanced images were generated via manual digital retouching. The ratings were collected from 9 subjects.
- CCID Database [31]: The database was developed at the Shanghai Jiao Tong University, China. Similar to the DRIQ database, CCID database is intended for enhanced image quality evaluation. It consists of 655 contrast-changed images. These image were obtained when 15 reference images were subjected to gamma transfer, cubic and logistic functions, intensity shifting, and compound function. 22 observers were involved in rating the images.
- SSID Database [32], [33]: The database was developed at Saarland University, Germany. It is intended to evaluate the quality of image in synthetic or augmented scenarios. It contains 8 reference images which were subjected to 7 different artefacts arising in image composition of real and synthetic content. The artefacts are: JPEG, JP2K, WN, GB, object scaling error, object translation error, and object rotation error.

A total of 1680 distorted images were produced and the ratings were obtained from over 200 people through crowdsourced online platform.

- LIVE Wild Image Database [34], [35]: The database was developed at the University of Texas at Austin, USA. Rather than dealing with simulated distortions as in most databases, the database is produced to evaluate image quality in the presence of real image distortions on images. A total of 1162 distorted images were captured using mobile devices without introducing extra artificial distortions beyond those occurring during capture, processing, and storage by a user's device. The subjective evaluation study where conducted via a crowdsourced platform in which more than 8,100 participants gave over 350,000 ratings to those images.
- ESPL-LIVE HDR Database [36], [37]: The database was developed at the University of Texas at Austin, USA. It comprises of 605 source images taken from modern digital SLR camera. A total of 1811 HDR-processed images were generated via tone mapper operators and multi-exposure fusion algorithms. The subjective evaluation study where conducted via a crowdsourced platform in which more than 300,000 ratings were collected from over 5,000 participants.
- PairComp TMO Database [38], [39]: The database was developed at the Institut de Recherche en Communications et Cybernetique de Nantes (IRCCyN), France to study the impact of high dynamic range (HDR) compression process to the evaluation of image quality. It consists of 10 HDR images. These images were tone-mapped via 9 sets of tone mapper operators' parameters to low dynamic range (LDR) for standard monitor display, resulting in 90 LDR images. These HDR and LDR images were observed by 20 participants to yield the ratings of the images. Most of the above databases utilised direct scaling methods and presented their ratings in terms of MOS

or DMOS. In contrast, this database use an indirect scaling method, namely paired comparison (PC), and reported the quality ratings in the form of preference scores [40].

1.2.3 Objective image quality assessment

In relation to the prediction of perceived image quality, the traditional fidelity metrics discussed in sub-chapter 1.2.1 can often be assumed as the worst-case scenario. Meanwhile, subjective experiments in sub-chapter 1.2.2 that provide accurate predictions are the best-case scenario. If we develop a hierarchy of the perceptual quality metrics' prediction capability, the fidelity metrics form the lower end while the subjective experiments represent the upper end. In this scenario, an objective IQA model should be near the top end of the hierarchy by producing image quality metric that follows human perceptual measures.

Objective IQA models can be classified into three main categories [41], [42]: full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA) and no-reference IQA (NR-IQA) / blind IQA (BIQA). FR-IQA models estimate the quality of a distorted image by comparing the entire information difference between the image and the corresponding reference image. The simplest approach to implement FR-IQA model is by measuring local pixel-wise disparity between the two images through MSE or PSNR metrics. However, they do not correlate well with subjective quality measures. Many improved FR-IQA models were then proposed through different means such as human visual system (HVS), image structure or image statistics. Primary visual cortex (V1) neural computational models are often utilised by HVS based FR-IQA metrics to estimate the image quality [43]. VSNR [20] and MAD [15] are two examples of high performance FR-IQA models from this approach. Changes in local structure such as luminance, contrast, phase or gradient are exploited to represent quality in image structure based approach. Examples of FR-IQA models for this approach include SSIM [44], FSIM [45], MP-Q [46] and GMSD [47]. Image statistics methods measure the quality based on image

statistical properties and they are often supplemented by machine learning techniques. VIF [48] and MLIQM [49] are common examples for this approach.

Meanwhile, only parts of the reference image information are necessary for RR-IQA models. A set of parameters that relate to visual perception of image quality and sensitive to various distortions are first identified and selected from the reference image. With the distorted image, these parameters are then utilised to predict its quality. Well-known examples include RR-SSIM [50], OSVP [51], and RRED [52].

While these FR-IQA and RR-IQA models produce higher correlation with human perceptual measures, in certain applications, information of reference image may not be fully or partly accessible. For example, in monitoring the quality of service of an image communication system, the original emitted signals are often unavailable at intermediate points or at the end of the system. Although RR-IQA methods could be applied by transmitting the required features of the original signals via an ancillary channel, they become impractical in a system with limited resources such as frequency spectrum in wireless communications [53], [54]. Both ends of the image communication system itself (image acquisition and image display), as illustrated by Figure 1.1, are inherently without reference. Similarly, evaluation of the quality of an image captured by or displayed on either digital camera or mobile phone need to be made without the availability of a reference image. The operation of computer vision applications such as robot navigation or autonomous driving is also based on images being continuously captured by camera. Again, evaluating the quality of those images is without reference images. In addition, for photo and film restoration application, it is possible that a degraded print is the only available record of a photo or a film [55]. Therefore, a BIQA model is preferred in such cases.

BIQA models can be further categorised into two main classes [56]: distortion-specific (DS) models and general-purpose models. In the DS BIQA cases, a particular distortion model is utilised to estimate quality using an assumption that the distortion in the image is known beforehand. For example, the quality of an image affected by motion blur is estimated by motion models introduced in [57]-[60] while the effects of blocking and noise artefacts are investigated in [61]-[62] and [63] respectively. JPEG compressed images' quality is predicted by the model in [64] whereas in [65], the quality of images compressed with JPEG2000 is estimated blindly. However, these models are only useful for specific application domains wherein the specific degradation is meaningful but cannot be used in a more general setting without substantial redesign. More fundamentally, when the distortion model is simply unknown, these models are ineffective and more general BIQA models are needed that can work across and adapt to any class of distortions.

No previous information about the distortion inside the image is needed in general-purpose BIQA models. Instead, image quality is derived solely on assumption that the image is degraded by the same distortion mechanism that affects a database of image exemplars. These image exemplars can be obtained from standard IQA databases such as the ones being listed in sub-chapter 1.2.2. Here, the models are trained to perform quality score estimation using such exemplars and their provided ratings values.

1.3 Study Scope, Aims and Objectives

1.3.1 Scope of study

As described in sub-chapter 1.2, there are various factors need to be considered in an IQA model design process such as (1) subjective approach versus objective approach, (2) the amount of reference information available to the designer, (3) the type of distortions accounted for by the metric: application-specific or general-purpose, (4) the number of distortions within

the image: singly distorted or multiply distorted, or (5) the type of image: natural, synthetic or graphics. Incorporating all these factors into the model design can lead to the development of universally applicable IQA metric but comes at the expense of highly complicated model. Therefore, it is paramount to limit and identify the scope of the study.

The study focuses on developing objective general-purpose BIQA models. The models are intended to deal only with natural images affected by a single distortion. Therefore, the models' quality estimation performance will only be benchmarked against previous models developed within the same scope. For ease of comparison, the proposed models in this study will also be tested on the same IQA databases utilised by the benchmarked models. The performance evaluation of the proposed model is also reported in the same way as reported by the benchmarked models.

1.3.2 Study aim and objectives

Most general-purpose models employ a two-stage learning framework whereby they first discover various image features that carry discriminative information about image quality. The features are then utilised as input to regression algorithms for quality prediction model learning. While most previous work focus on designing new quality-predictive image features, this study tackles BIQA from an alternative angle. This study aims to contribute to the IQA research community by introducing new learning frameworks for general-purpose BIQA models. The resulting models shall be able to perform image quality prediction accurate to human perceptual measures and have competitive prediction performance to previous models. This can be done by fulfilling these objectives:

a. Identify limitations of general-purpose BIQA models through critical analysis of their methods and corresponding performances.

b. Develop new learning frameworks for BIQA models based on potential solutions to those limitations.

c. Test the developed models according to standard experimental procedure employed by IQA research community.

d. Compare the developed models' performances to several other models through extensive analysis of the models' results on quality estimation accuracy and generalisation capabilities and speed requirements.

1.4 Study Contributions and List of Publications

With respect to limited scope of the study, the contributions of this study can be summarised as the following:

- The study first introduced a general-purpose BIQA model that employ a patch based learning framework. The main contribution of the model lies on it performing quality estimation using nearest neighbour learning techniques avoiding the need to have a prior training phase which is a prerequisite for many general-purpose BIQA models. Another key contribution is the model ability to perform image distortion identification, a useful property that is unavailable in most of previous models. Through its patch-level operation, the model further contribute by providing local quality estimation. Note that the current absence of a dataset containing ground truth quality target for each patch makes it difficult to validate the local quality estimation performance of the model. However, the model takes advantage of the distortion uniformity across images inside the utilised databases to develop such dataset in order to validate its performance.
- The study next improved the first model's prediction performance by modifying the patch extraction stage of the model's framework. The key contribution of the modified model lies on it exploring the use of two sampling strategies to extract image patches:

interest points based and saliency based, both strategy incorporate an aspect of human vision tendency which is often ignored by previous general-purpose BIQA models.

- The study then presented the third general-purpose BIQA model that integrate multi-task learning technique in its framework. As opposed to individual regression model learning by previous models, the main contribution of the third model lies on the fact it perform different regression model learning for different image distortion classes simultaneously. By exploiting a shared representation among the classes, the third model is shown to improve the prediction capability of a BIQA model in each distortion class.

Some of the thesis content have been appeared in one or more publications. The following are the list of publications associated with the study:

[1] R. A. Manap and L. Shao, ""Non-distortion-specific no-reference image quality assessment: a survey," *Information Sciences*, vol. 301, no. 1, pp. 141-160, 2015.

[2] R. A. Manap, L. Shao, and A. F. Frangi, "Non-parametric quality assessment of natural images," *IEEE Multimedia*, vol. 23, no. 4, pp. 22-30, 2016.

[3] R. A. Manap, L. Shao and A. F. Frangi, "PATCH-IQ: a patch based learning framework for blind image quality assessment," *Information Sciences*, vol. 420, no. 1, pp. 329-344, 2017.

[4] R. A. Manap, L. Shao and A. F. Frangi, "Blind image quality assessment via a multi-task learning framework," *IEEE Transactions on Image Processing*, pp. 1-13, 2017 (submitted).

[5] R. A. Manap, A. F. Frangi, and L. Shao, "Blind image quality assessment via a two-stage non-parametric framework," in *Proceedings of the IAPR Conference on Pattern Recognition*, Kuala Lumpur, Malaysia, 2015, pp. 796-800.

[6] R. A. Manap, L. Shao, and A. F. Frangi, "A non-parametric framework for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Signal and Information Processing*, Orlando, FL, 2015, pp. 562-566.

[7] R. A. Manap, L. Shao, A. F. Frangi, and A. M. Darsono, "Multi-task learning approach for natural images' quality assessment," in *Proceedings of the International Conference on Telecommunication, Electronic and Computer Engineering*, Melaka, Malaysia, 2017, pp. 1-6.

1.5 Thesis Layout

The rest of the thesis is structured to reflect on the objectives of the study. In Chapter 2, previous approaches in developing general-purpose BIQA models are first briefly reviewed, followed by analysis of their performances. Their limitations are then identified leading to the proposed research directions of the study. Chapter 3 addresses several limitations by introducing the first patch based learning framework for a general-purpose BIQA model. The model's framework is first described followed by the experimental results and analysis. In Chapter 4, the second patch-based learning framework is proposed based on modification made to the first framework. The modified framework is first presented before further discussions on its experimental results. Chapter 5 introduces the multi-task learning based framework to address another limitation encountered by the previous general-purpose BIQA models. Similarly, the framework is first described, followed by the experimental results and analysis. Chapter 6 concludes the thesis by summarising all the work done throughout the study and their contributions followed by discussions on the work limitations and possible works for the future.

Chapter 2

Literature Review

2.1 Chapter Introduction

To begin with, a short survey on the progress made in general-purpose BIQA models is presented whereby some of the major contributions are reviewed. However, to ensure we stay within the scope of the study, the review cannot be regarded to be exhaustive. For further discussions and references, interested readers are recommended to refer to other review publications [42], [43], and [55]. The chapter then continues with the description of common experimental procedures implemented in the previous general-purpose BIQA models followed by their performance analysis. Through these analysis, several limitations are identified leading to potential solutions that will be the basis of this study.

2.2 Previous Work

Previous general-purpose BIQA models usually follow a two-stage approach, whereby various types of features are first extracted and then used as input to a regressor. The regression algorithm is then used to model human perceptual measures based on a set of training images. At the feature extraction stage, the features generally can be classified into two types: handcrafted or machine-learned.

2.2.1 Models based on handcrafted features

BIQA models that employ handcrafted features usually design their features through the natural scene statistics (NSS) approach. The NSS models assume that certain statistical properties of natural images will be changed with the presence of distortions and the perceptual

quality of these images can be inferred by appropriately quantifying the changes. These models can be differentiated by the features used.

In [66], the Blind Image Quality Index (BIQI) performs image quality estimation utilising 18 statistical features derived from wavelet transform. The wavelet transform is first applied to an image and the resulting sub-band coefficients are parameterised by a general Gaussian distribution (GGD). The GGD model parameters are then selected as features representative of image quality. The Distortion Identification based Image Verity and INtegrity Evaluation (DIIVINE) model, which improves upon BIQI, is later proposed in [67]. Besides the wavelet sub-band coefficients' GGD model parameters, a larger set of features is also derived to account for local dependency between the coefficients across different scales and orientations. A total of 88 features are utilised in DIIVINE to compute quality prediction.

Another model that based on wavelet coefficient statistics is presented in [68]. Natural Scene Statistics Global Scheme (NSS-GS) utilises the wavelet coefficients' original marginal distribution to design its features. Because the magnitudes of wavelet coefficients are likely to be continuous across scales, the model also exploits exponential decay characteristics of the coefficients in designing additional quality predictive features. Learning Based Image Quality (LBIQ) model [69] also utilises wavelet based statistical properties whereby its features are derived from the wavelet coefficients' marginal and cross-scale joint distributions. Based on the observation that blur and noise are the two main degradation processes that occur in various distortion types, LBIQ also extracts additional features through blur and noise statistics.

In [70], another wavelet based BIQA model incorporates a divisive normalisation strategy in its feature extraction stage to reflect the non-linear behaviour of visual cortex neurons. After linear decomposition of the image, the resulting wavelet coefficients are first normalised via divisive normalisation transform. The joint distributions of the neighbouring

normalised coefficients across different scales and orientations under various types of distortions are then used to measure the statistical independence between the neighbouring coefficients. These statistical independence measurements are used as features in this STAtistical INDependence based (STAIND) model. In [71], the Sparse Representation of Natural Scene Statistics (SRNSS) model also designed their features in wavelet transform domain. The mean, variance, and entropy of the wavelet coefficients in each image sub-band over 4 scales are used to compute 24 features. The features are then encoded using sparse coding before regression.

Meanwhile, a BIQA model that build upon a local discrete cosine transform (DCT) statistical model is presented in [72]. The model, BLind Image Integrity Notator using DCT Statistics (BLIINDS), works by first applying DCT to image patch centred at each pixel in the image. It then extracts four features representing the information of image contrast and image structure and uses these features for quality estimation. Later on, using DCT statistics in BIQA task is further advanced by BLIINDS-II [73]. BLIINDS-II first divides image into blocks where the blocks are subjected to local DCT computation. The DCT coefficients of each block are then fitted by a GGD model. The resulting model parameters are utilised to retrieve relevant features. A total of 24 statistical features, extracted over three scales, are employed by BLIINDS-II to predict image quality.

A NSS based BIQA model that operates in the spatial domain is later introduced in [74]. In contrast to previous models, no transformation is required by the model called Blind / Reference less Image Spatial QUality Evaluator (BRISQUE). BRISQUE utilises the empirical distributions of the locally normalised luminance coefficients and the pairwise products of these coefficients to design 18 statistical features for image quality estimation. In [75], a new model known as DErivative Statistics based Image QUality Evaluator (DESIQUE) modifies BRISQUE feature extraction approach to include operation in the frequency domain. In the

spatial domain, DESIQUE designs the features based on the normalised luminance coefficients as in BRISQUE. In the frequency domain, DESIQUE decomposes an image using log-Gabor filters. The filter band coefficients are then subjected to log-derivative statistics to characterise the distribution of the image's high frequency components. All the statistics are then parameterised by a GGD model with the resulting parameters form another set of features. DESIQUE extracts 64 statistical features over two scales for quality analysis task.

Using statistical properties derived from local spatial contrast features is also explored in [76]. The model, coded as GMLOG, first employs the marginal distributions of the jointly normalised gradient magnitude and the Laplacian of Gaussian operators of an image as its first two features. Based on the observations that both operators are non-independent, GMLOG also introduces independency distribution indexes to measure their dependencies. These independency distributions are then selected as the other two statistical features. A BIQA model employing a general regression neural network (GRNN) architecture is later proposed in [77]. The GRNN based BIQA model measures the content of an image on three elements: gradient, phase and local information to extract four features carrying useful perceptual quality information.

Meanwhile, a BIQA model that utilises a visual codebook technique is proposed in [78]. Given an image, the model termed as Codebook Based Image Quality (CBIQ) first extracts a set of Gabor feature vectors from the randomly sampled image patches. A trained visual codebook, containing the associated Gabor features derived from training image patches, is then utilised to encode these feature vectors. CBIQ then averages the encoded feature vectors to yield the image-level features. The image-level features are used to estimate the image quality. The codebook based BIQA model is also introduced in [79]. The model, Bag-of-Word using Selected Features (BOWSF), employs three different sets of NSS features derived from BIQI, SRNSS and BRISQUE respectively to predict the image quality. A total of 78 features

are used by the model. Another Gabor filter based BIQA model is later presented in [80]. In contrast to CBIQ, the Image Quality index based on Visual saliency guided sampling and Gabor filtering (IQVG) model first incorporates the use of visual saliency map of an image to navigate its patch sampling process. IQVG then extracts similar Gabor-filter-based feature vectors as implemented in CBIQ from the sampled patches. However, rather than using a visual codebook, IQVG encodes the feature vectors directly using histograms and uses the combinations of these histograms as its image-level features.

Because the image local structures vary when an image is distorted, using low-level local structure statistics as quality predictive features is investigated by the BIQA model in [81]. The model first decomposes an image into multi-scale sub-band images. Using a generalised local binary pattern (GLBP) operator as the local structure descriptor, each sub-band image is then encoded with the operator to produce the GLBP encoding maps. The normalised histograms of the encoding maps are finally combined to produce 18 statistical features for each sub-band image. A total of 72 statistical features, extracted over four scales, are then employed by the GLBP model to predict the image quality. Similar local binary pattern (LBP) based approach is presented in [82] whereby histogram of LBP codes of sub-band of wavelet decomposed image are BIQA features.

The use of LBP-based approach for BIQA model is further presented in [83]. The model first compute the LBP operators between centre pixel and its surrounding neighbours in image gradient magnitude map. Instead of typical frequency histogram, the Gradient-Weighted Histogram of Local Binary Pattern (GWH-LBP) model then use gradient-weighted histogram of the operators as its features. A total of 40 features extracted over 5 scales is employed by the model for image quality estimation. Another BIQA model that employs statistical properties of image local structure is proposed in [84]. However, instead of using local binary pattern descriptors, the model utilises local ternary pattern (LTP) descriptors to encode the

image. Similarly, the histograms of the corresponding LTPs are then used as input features to the regression stage.

Using an assumption that colour space features are also highly correlated to human perception of image quality, a BIQA model that utilises colour channel information is introduced in [85]. The Natural Colour Statistics based (NCS) model first transforms the image into colour spaces based on four common colour models [86]: Lab, HSV, YCbCr and YIQ. Following similar approach as in BRISQUE, NCS then parameterises the locally normalised colour coefficients' empirical distribution and the empirical distribution of the pairwise product of these coefficients by GGD model. The resulting 18 model parameters are then selected as features for NCS model.

Motivated by recent advances in neuroscience, another group of BIQA models extract their quality predictive features based on free energy principle. The free energy principle is based on a hypothesis that a human brain operates in an internally generative ways to model the image we look at [87]. Using this generative model, the brain then generates the image's predictions constructively. If the differences between the image and the outputs of the internal generative model relate to visual perception, the quality of an image can then be interpreted as how close the image itself agreed with the model's output that best describes the image. The upper bound of the discrepancy between the two is given by the free energy of this cognitive process, making it possible to quantify perceptual quality using the free energy [88].

No-reference Free Energy based Quality Metric (NFEQM) [88] is the first BIQA model that employs this approach. NFEQM approximates the internal generative model by applying a linear autoregressive model to an image. Since free energy can measure the disparity between the image data and its closest representation by the generative model, NFEQM then optimises the linear autoregressive model parameters to find the minimum free energy term of the image.

NFEQM uses the term as a metric indicative of quality. The lower the free energy value, the higher the perceptual quality of the image is.

An extension to NFEQM is proposed in [89]. The No-reference Free energy and Structural degradation based Distortion Metric (NFSDM) model combines the free energy feature with 54 additional features representing structural degradation variations in the image under different scales and Gaussian low-pass filtering processes. Later on, No-reference Free Energy based Robust Metric (NFERM) [90] modifies NFSDM by incorporating NSS based features into the model. Three feature classes are considered by NFERM. The first class contains 13 features derived from the free energy and the structural degradation methods. The second class consists of 6 features that carry information on image structure, gradient and phase while the last class comprises 4 NSS based features or model parameters produced by fitting GGD to the distribution of the image's locally normalised coefficients.

2.2.2 Models based on machine-learned features

There are also general-purpose BIQA models that use features learned directly from raw image pixels. The first work along this approach can be seen in [91]. The COdebook Representation for No-reference Image Assessment (CORNIA) model first randomly samples raw patches from an image. Using a codebook, CORNIA then encodes the patches and pools the encoded patches to generate its image-level features. CORNIA is similar to CBIQ in that both utilise a visual codebook in their feature extraction stage. However, instead of employing features from Gabor-filter responses, CORNIA constructs the codebook using raw image patches in unsupervised manner. Due to a greater performance in image classification [92], CORNIA also employs max pooling in generating its features as opposed to average pooling in CBIQ.

Following the promising results achieved by unsupervised feature learning approach in CORNIA, an extended model is later proposed in [93]. The model, Saliency based Feature Learning for No-reference Image quality Assessment (SFLNIA) employs saliency detection method prior to the feature extraction stage. The salient regions are found through a saliency map calculated based on low-level local images features. Raw image patches are then extracted from these salient parts of the image. Similar feature learning strategy as in CORNIA is then employed to produce image-level features. Another model employing similar saliency detection method is presented in [94]. The model, Saliency-guided Deep framework for Image Quality Assessment (SDIQA) first detect salient regions using information divergence [95] before extracting raw image patches from the regions. SDIQA then use deep learning technique to produce image-level features.

Meanwhile, a supervised feature learning framework for general-purpose BIQA model is presented in [96]. Termed as supervised CORNIA (SV-CORNIA), the model uses a set of linear filters to encode the normalised image patches. Based on the observation that the distributions of the filter responses vary for different categories and levels of distortion and the assumption that the distributions' statistics are closely related of image quality, the model then extracts the maximal and the minimal filter responses as its features. The filters are learned in a supervised way by back-propagation method to ensure the extracted features are suitable for BIQA task.

In [97], a BIQA model that utilises convolutional neural network (CNN) is proposed. Rather than using any handcrafted features as in GRNN model, CNN learns discriminant features directly from the normalised raw image patches. Another difference to GRNN model is that the feature learning and the regression stages are integrated into one general neural network framework, making the network deeper to increase the learning capacity. In CNN model, the locally normalised image patch is first convolved with 50 filters (kernels) to produce

50 feature maps. Each feature map is pooled in the second layer of the network into one maximum value and one minimum value, reducing individual map to a two-dimensional feature vector. The third and the fourth layer of the network then train the network and the output is employed as an input to the final layer for quality prediction task (regression).

The encouraging results achieved by the CNN model leads to the introduction of other CNN-based BIQA models. A model known as Blind Image Evaluator based on CONvolutional neural network (BIECON) is presented in [98]. BIECON employs a CNN architecture which consists of 2 convolutional layers, 2 pooling layers and 5 connected layers to estimate the quality of the normalised input images patches. At the same time, BIECON utilises a perceptron with one hidden layer to regress the mean and the standard deviation values of the extracted patch-wise features for image-level quality estimation. Meanwhile, a deeper CNN architecture for BIQA task is proposed in [99]. The Deep Image QuAlity Measure (DIQaM) model first extracts quality-predictive features from a set of un-normalised image patches through a CNN architecture of 10 convolutional layers with 5 pooling layers. The extracted features are then used as inputs to two fully connected layers to perform quality prediction for the patches. The patches' quality scores are then pooled to obtain image-level quality estimate.

2.2.3 Regression

The extracted features, handcrafted or machine-learned, are then used to learn prediction models for BIQA. This is usually done by inputting the features into a regression algorithm to learn the mapping function between the features' space and the image quality score space. Given the training images' features and their associated MOS / DMOS values, kernel-based learning methods are often utilised to learn such mapping function. Usually, support vector regression (SVR) is used to this effect.

Given training data $\{(x_1, s_1), (x_2, s_2), \dots, (x_n, s_n)\}$, where $x_i, i = 1,2,3, \dots, n$ denotes the extracted feature vector and s_i represents the corresponding MOS / DMOS value, the linear function that estimate the output value from the input feature vector is given as [100]:

$$f(x) = \langle \omega, x \rangle + b \quad . \tag{2.1}$$

In Equation (2.1), ω represents the weight vector, $\langle \cdot, \cdot \rangle$ denotes the inner product, and b represents a bias parameter. In the SVR case, ω and b can be determined by minimising the later optimisation problem [101]:

$$\text{minimise } \frac{1}{2}\|\omega\|^2 + c_{\text{SVR}} \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

$$subject\ to \begin{cases} \langle \omega, x_i \rangle - (s_i - b) \leq \varepsilon_{\text{SVR}} + \xi_i \\ s_i - b - \langle \omega, x_i \rangle \leq \varepsilon_{\text{SVR}} + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} . \tag{2.2}$$

In Equation (2.2), ε_{SVR} represents the threshold / deviation parameter: all predictions must be within ε_{SVR} range of the true predictions, ξ_i and ξ_i^* are the slack variables that allow for errors while c_{SVR} represents a constant parameter for ω and ξ_i / ξ_i^* balancing. Equation (2.2) minimiser is given as [101]:

$$\omega = \sum_{i=1}^{n} r_i x_i \quad , \tag{2.3}$$

where r_i is the combination coefficient.

For non-linear cases, the input feature vector is usually mapped onto a feature space of high dimension $\Phi(x)$ prior to regression. The function for regression can then be represented as [101]:

$$f(x) = \langle \sum_{i=1}^{n} r_i \Phi(x_i), \Phi(x) \rangle + b = \sum_{i=1}^{n} r_i \langle \Phi(x_i) \Phi(x) \rangle + b \quad . \tag{2.4}$$

The term $\langle \Phi(x_i)\Phi(x) \rangle$, representing an inner product, is also recognised as a kernel function $k(x_i, x)$. This leads to:

$$f(x) = \sum_{i=1}^{n} r_i k(x_i, x) + b \quad . \tag{2.5}$$

While there are several kernel functions available, the radial basis function (RBF) kernel is often selected in BIQA task. It is given as [102]:

$$k(x_i, x) = \exp(-\gamma_{\mathrm{RBF}}(|x_i - x|)^2) \quad , \tag{2.6}$$

where γ_{RBF} is the precision parameter.

2.3 Common Experimental Procedure and Performance Analysis

Having introduced various general-purpose BIQA models' approaches in sub-chapter 2.2, their performances in predicting image quality are now briefly analysed. Prior to that, a standard experimental setup and evaluation protocols are first described to help readers obtain general ideas on how the models are typically tested and evaluated. Note that the performance results of each model are obtained from their corresponding publications. This may not represents a fair comparison between the models due to various factors such as random selection of training data, number of trials, and the choice of regression modules. However, the main purpose here is to give general observation on how well the models perform in relation to human quality perception.

2.3.1 Experimental setup

The prediction performance evaluation for a BIQA model is usually carried out using standard IQA databases. In a standard setting, the chosen database is first partitioned into two parts. 80% of the reference images and their distorted versions are randomly selected as a training set while the remaining 20% of the reference images and their distorted versions are set as a test set. This ensure no redundancy between the two sets. The training set is used to

first determine the parameters for the regression model. Once the regression model is learned, it is then used to perform the quality score prediction for the test set images. To guarantee that the obtained results are not influenced by one particular train-test partition, the experiments are normally repeated 100 to 1,000 times where different 80% train - 20% test partition is employed in each run.

Two experiments are typically performed to ascertain the overall performance and the DS performance of the model. In the overall performance experiment, the train-test run is conducted across all distorted images regardless of their classes. This is to evaluate how well the model performs across all distortion types. In the DS performance experiment, the experiment is only carried out on images in a single distortion class. This is to evaluate how well it performs for one particular distortion.

2.3.2 Metrics for performance evaluation

Evaluating the performance of a BIQA model is crucial as to provide us with information on how robust the model is, a specific failure identification and possible improvements. Inadequate evaluation can lead to false performance claims and inability to identify model weaknesses. Identifying a model's weakness is important as a model with systematic weakness may lose its interpretability, i.e. its ability to differentiate between high quality images from low quality images [103].

The performance of an objective BIQA model is usually evaluated by quantifying the differences between the predicted quality scores by the model and the ground truth ratings obtained from the subjective image quality experiments, i.e. the image databases' quality ratings. As described in ITU-T P.1401 evaluation procedure for objective metrics [104], the model's predicted scores are first mapped to MOS / DMOS values via regression before several performance metrics are utilised to analyse the model's performance. Because of this, almost

all image databases discussed in sub-chapter 1.2.2 are developed via direct scaling testing methodologies such as absolute category rating or double stimuli continuous quality scale; resulting in their ground truth ratings being reported in the form of MOS / DMOS.

For a BIQA model that is tested on a MOS / DMOS based database, the performance of the model is measured by its ability to predict the image quality score close to the MOS / DMOS value in the database. In this case, two correlation measurements are commonly used as the performance metrics. They are: the linear correlation coefficient (LCC) and the Spearman rank order correlation coefficient (SROCC). The LCC is utilised to indicate the model's prediction accuracy. It can be computed as [105]:

$$\mathrm{LCC} = \frac{\sum_{i=1}^{N_{\mathrm{test}}}(q_i-\bar{q})(\mathrm{s}_i-\bar{\mathrm{s}})}{\sqrt{\sum_{i=1}^{N_{\mathrm{test}}}(q_i-\bar{q})^2(\mathrm{s}_i-\bar{\mathrm{s}})^2}} \quad . \tag{2.7}$$

In this equation, N_{test} represents the test images number, q_i and $\bar{q}$ are the predicted quality score of the ith image and the mean of all q_i while s_i and $\bar{\mathrm{s}}$ are the subjective score of the ith image and the mean of all s_i. The second metric SROCC is used to measure the prediction monotonicity of the model. It is calculated as [105]:

$$\mathrm{SROCC} = 1 - \frac{6}{N_{\mathrm{test}}(N_{\mathrm{test}}^2-1)}\sum_{i=1}^{N_{\mathrm{test}}}(\mathrm{s}_i - q_i)^2 \quad . \tag{2.8}$$

Values closer to 1 (or -1) for both LCC and SROCC indicate higher correlation with human subjective score.

Apart from evaluating the model's performance across a wide range of quality levels, as represented by most image databases, the performance metrics computation should also be performed over a meaningful subsets of the databases. These subsets can include evaluation based upon the presence or absence of specific artefacts or sources with more or less observer variability in scores [6]. In such cases, the BIQA evaluation practice involving MOS / DMOS

based performance metrics may become less discriminative or reliable. For example, while BIQA models may achieve high correlation between the predicted MOS / DMOS and the actual MOS / DMOS when evaluated over the overall database quality range, the correlation is actually lower when we are focusing over smaller subset of the database. Recently, the MOS / DMOS based evaluation approach has been shown to be less effective in validating BIQA models that test consumer devices whereby the range of quality levels of images captured by consumer mobile devices or high-end camera is usually narrower and typically concentrated at the higher quality end [106].

Several work have been presented to address this so-called *range effect* [106]. Instead of evaluating the model's performance via subjective MOS / DMOS, these work assess the model's performance using subjective preference scores. These scores are usually generated via indirect scaling test methodologies such as paired comparison (PC). The use of PC is motivated by the observation that it has a higher discriminatory power in cases where the difference between observed images are small. In addition, it is easier for observers to identify which image is of better quality in a pair of images compared to relate image quality to a particular level on a given quality scale [40].

Evaluating the model's performance under the PC based approach requires different performance metrics to be used. The first metric that can be used for this purpose is the model Resolving Power (RP) presented in ITU-T Rec. J.149 [107]. RP measures the difference between predicted scores of image A and image B by the model ($\Delta q_{model} = q_{model}(A) - q_{model}(B)$) necessary to have 95% probability that the image A is qualitatively better than image B. This way, RP can be used to indicate the model capability to determine whether a pair of images are qualitatively different. The model with lower difference (threshold) value is considered more accurate.

The use of RP, however, does not give information about the reliability of the model classification. Therefore, another performance metric in the form of the model classification errors analysis is required. Classification errors occur when the model's evaluation on a pair of images differs from its subjective evaluation. This can happen in one of three ways [40]: (1) *False Tie* when the subjective evaluation indicates that the two images are different but the model evaluation indicates that they are identical, (2) *False Differentiation* when the subjective evaluation indicates that the two images are identical but the model evaluation indicates that they are different, and (3) *False Ranking* when subjective evaluation finds that image *A* is better than image *B* but the model evaluation finds the opposite. To evaluate the model performance, the frequencies of these errors is first computed by varying a threshold on the model's score difference of the two images, Δq_{model} while comparing the model classification outcomes to that of subjective evaluation. Analysing classification errors at $\Delta q_{model} = 0$ will show us how many times the model makes the *False Differentiation* errors. In addition, by finding the Δq_{model} point where *Correct Decision* frequency is maximised, we can also determine the highest percentage of agreement between the model and the subjective test.

Motivated by classification error analysis approach, another PC based performance metric is proposed based on receiver operating characteristic (ROC) analysis [108] – [109]. The general principle of ROC analysis is similar to classification error analysis whereby it creates a curve reflecting the correct classification when the threshold on the model's scores is varied [108]. To generate the curve, the model's True Positive Rate (TPR) and False Positive Rate (FPR) are first recorded for every threshold position. The TPR and FPR are given as [109]:

$$TPR = \frac{TP}{TP+FN}, \; FPR = \frac{FP}{FP+TN} \tag{2.9}$$

In Equation (2.9), *TP* represents true positive where positive input is correctly classified by the model, *FP* represents false positive where negative input is classified as positive by the model,

TN represents true negative where the negative input is correctly classified by the model, and *FN* represents false negative where positive input is classified as negative by the model. The ROC curve can then be generated by plotting TPR as a function of FPR. For easier comparison purpose, the Area Under the ROC Curve (AUC) is then computed as [109]:

$$AUC = \sum_{p=2}^{P} \left\{ \frac{TPR(p)+TPR(p-1)}{2} \times \frac{FPR(p)-FPR(p-1)}{2} \right\} \quad (2.10)$$

where P represents the number of threshold positions considered.

This ROC analysis can then be used to evaluate a BIQA model performance for two cases. The first case is to determine the model capability to distinguish between similar and significantly different image pairs. This is done by first computing the absolute difference of the model scores over all possible image pairs. The TPR and FPR are then recorded as the threshold of these differences' distribution is varied. The resulting AUC value is used to evaluate the model performance whereby the higher the AUC, the higher the model capability to determine whether the images are qualitatively different.

The second case is to evaluate the model capability to determine which of the images is of higher quality. Here, only image pairs that are significantly different are considered. The distributions of both image pairs with positive score differences and image pairs with negative score differences are analysed in which the TPR and FPR are again recorded as the threshold is varied. The obtained AUC value is used to indicate how well the model recognise the better image in the pair. The higher the AUC value, the more capable the model to identify the image of higher quality. Similar to classification error approach, analysing the correct and false classification in threshold equal to zero will actually shows how many times does the model correctly recognise the better image in the pair.

2.3.3 Performance results and analysis

While there are many available image quality databases, most of general-purpose BIQA models are evaluated using the LIVE IQA database [14]. Their reported results from the LIVE database are therefore utilised here for evaluation. Similarly, while there are alternative performance metrics available, the DMOS-based metrics are used here to benchmark the models. The median LCC and SROCC results of the BIQA models in both the overall and the DS experiments are tabulated in Tables 2.1 and 2.2, respectively. Due to the right-skewed distribution of the LCC and the SROCC values, median is often used in the previous works as their centre measurements. Several FR-IQA models are also included for reference. Note that the models with the highest LCC / SROCC values, in FR-IQA and both handcrafted based and machine learned based categories, are in bold.

Table 2.1: Median LCC values for different IQA models tested on the LIVE IQA database

Model	JP2K	JPEG	WN	GB	FF	ALL
PSNR	0.873	0.876	0.926	0.779	0.870	0.882
SSIM	**0.921**	0.955	**0.982**	0.893	**0.939**	0.906
FSIM	0.910	**0.985**	0.976	**0.978**	0.912	**0.960**
BIQI	0.809	0.901	0.954	0.829	0.733	0.821
DIIVINE	0.922	0.921	0.988	0.923	0.888	0.917
NSS-GS	0.947	0.933	0.963	0.950	0.942	0.926
LBIQ	-	-	-	-	-	-
STAIND	0.923	0.975	0.975	**0.972**	0.923	0.922
SRNSS	0.936	0.939	0.940	0.936	0.947	0.932
BLIINDS	-	-	-	-	-	-
BLIINDS-II	0.935	0.968	0.980	0.938	0.896	0.930
BRISQUE	0.923	0.973	0.985	0.951	0.903	0.942
GMLOG	0.934	0.974	0.990	0.935	0.921	**0.955**
GRNN	0.828	0.880	0.989	0.825	0.819	0.837
CBIQ	0.920	0.967	0.954	0.949	0.939	0.928
IQVG	0.927	0.920	0.979	0.953	0.940	0.942
GLBP	**0.956**	0.972	0.985	0.954	0.912	0.954
LTP	0.949	0.948	0.950	0.949	**0.948**	0.949
NCS	0.950	0.964	0.991	0.935	0.942	0.939
NFEQM	0.921	0.875	0.925	0.902	0.875	0.893
NFSDM	0.955	0.959	0.935	0.945	0.848	0.924
NFERM	0.955	**0.982**	**0.992**	0.937	0.888	0.946
CORNIA	0.951	0.965	**0.987**	**0.968**	0.917	0.935
SFLNIA	0.957	0.958	0.978	0.955	0.920	0.916
SV-CORNIA	0.929	0.940	0.978	0.960	0.888	0.921
CNN	0.953	0.981	0.984	0.953	**0.933**	0.953
BIECON	**0.965**	**0.987**	0.970	0.945	0.931	**0.962**

Table 2.2: Median SROCC values for different IQA models tested on the LIVE IQA database

Model	JP2K	JPEG	WN	GB	FF	ALL
PSNR	0.870	0.885	0.942	0.763	0.874	0.866
SSIM	0.939	0.946	0.964	0.907	0.941	0.913
FSIM	**0.972**	**0.984**	**0.972**	**0.971**	**0.952**	**0.965**
BIQI	0.800	0.891	0.951	0.846	0.707	0.820
DIIVINE	0.913	0.910	0.984	0.921	0.863	0.916
NSS-GS	0.931	0.915	0.971	0.939	0.935	0.930
LBIQ	0.904	0.929	0.970	0.898	0.822	0.895
STAIND	0.914	0.960	0.966	**0.973**	0.903	0.916
SRNSS	0.928	0.931	0.938	0.933	0.941	0.930
BLIINDS	0.922	0.839	0.974	0.957	0.750	0.800
BLIINDS-II	0.929	0.942	0.969	0.923	0.889	0.931
BRISQUE	0.914	0.965	0.979	0.951	0.877	0.940
GMLOG	0.928	0.966	**0.985**	0.940	0.901	**0.951**
GRNN	0.816	0.872	0.979	0.833	0.735	0.827
CBIQ	0.919	0.965	0.933	0.944	0.912	0.930
IQVG	0.919	0.900	0.962	0.943	0.938	0.942
GLBP	0.947	0.956	0.979	0.954	0.889	**0.951**
LTP	0.942	0.942	0.944	0.942	**0.942**	0.942
NCS	0.947	0.937	**0.985**	0.949	0.932	0.941
NFEQM	0.915	0.854	0.915	0.931	0.852	0.887
NFSDM	**0.951**	0.948	0.927	0.935	0.821	0.922
NFERM	0.942	**0.965**	0.984	0.922	0.863	0.941
CORNIA	0.943	0.955	0.978	**0.969**	0.906	0.942
SFLNIA	0.951	0.947	0.972	0.952	0.912	0.923
SV-CORNIA	0.924	0.928	0.962	0.961	0.879	0.920
CNN	**0.952**	**0.977**	0.976	0.962	0.908	0.956
BIECON	**0.952**	0.974	**0.980**	0.956	**0.923**	**0.961**

As shown in these tables, most of the general-purpose BIQA models consistently obtain median LCC and SROCC values close to 1. This indicates that the predicted image quality scores by those models generally have close correlation with human subjective scores and the models can reflect well the quality perception of a human observer. Compared to FR-IQA models, most of the BIQA models are already outperform PSNR and SSIM model in the overall performance experiment while approaching FSIM. They also give comparable prediction performances for individual distortion cases. For example, for images affected by noise artefacts, many BIQA models produce higher correlation values than those of FR-IQA models. The results are encouraging enough given that the FR-IQA models require additional information (in reference images) to estimate image quality.

Among the models that utilise handcrafted features, GMLOG and GLBP achieve the closest prediction performance in the overall performance experiment. When tested on JP2K compressed images, GLBP and NFSDM produce the two best correlation scores. NFERM has the highest SROCC and LCC values for JPEG compressed images while STAIND and LTP work the best in GB and FF cases, respectively. For images affected by WN artefacts, NFERM, GMLOG and NCS are the three models that achieve the highest correlation values.

It can also be seen that BIECON has the best correlation scores among the machine-learned based models in the overall performance experiment. The neural network based models also have the best correlation values for JP2K and JPEG compressed images. In FF cases, CNN and BIECON are the top two machined learned based BIQA models. Meanwhile, CORNIA has the highest correlation values when tested on blurred images.

Although these models are normally trained and tested on a single database, i.e. the LIVE IQA dataset, most models can also be database independent. Once trained, the models are capable to evaluate the quality of images over the distortions they are trained for. Specifically, these models are usually trained entirely on the LIVE IQA database and then being tested on other major databases such as CSIQ [15] and / or TID2008 [16]. The results of cross database testing for several general-purpose BIQA models are tabulated in Table 2.3 where competitive performances are produced compared to FR-IQA models. These models also maintain good correlation scores indicating their good generalisation capability.

Computational requirement is another important aspect to be considered when evaluating the performance of a BIQA model. Table 2.4 reports the average processing time required by a BIQA model in evaluating a typical 512×768 test image. BIQI appears to be the fastest model, requiring 0.08 second to predict the quality score of an image. Unfortunately, it has the worst prediction accuracy performance. BLIINDS-II, CBIQ, IQVG and NCS give more

accurate prediction than BIQI at the expense of higher processing times. Based on the table, GMLOG, BRISQUE and CNN are the best three models with high correlation scores while requiring an acceptable runtime to process an image.

Table 2.3: SROCC values for cross database testing

Model	CSIQ	TID2008
PSNR	0.806	0.525
SSIM	0.876	0.767
FSIM	0.924	0.881
BIQI	0.781	0.819
DIIVINE	0.857	0.889
NSS-GS	-	0.848
STAIND	0.843	0.856
BLIINDS-II	0.888	0.906
BRISQUE	0.899	0.905
GMLOG	0.911	0.920
CBIQ	0.879	-
NCS	0.854	0.844
NFERM	0.914	0.915
CORNIA	0.897	0.893
SV-CORNIA	-	0.873
CNN	-	0.920

Table 2.4: Average processing time for different BIQA models

Model	BIQI	DIIVINE	BLIINDSII	BRISQUE	GMLOG	CBIQ	IQVG	NCS	CORNIA	CNN
Runtime (s)	0.08	28.20	123.20	0.18	0.10	60.00	60.00	107.00	1.59	0.13

These analyses may to decide on which are the better BIQA models. However, it is still difficult to agree on the best BIQA model that can operate effectively for a wide range of different circumstances. The reason for this is due to their being designed based on various philosophies and having complementary features. As indicated by the results on Table 2.1 and Table 2.2, a particular model's features may carry discriminative image quality information for images with a certain type of distortion but may not be useful for other types of distorted images. In addition, some models may have excellent performance when tested on one database but do not generalise well beyond that. The choice of which BIQA model to be employed is also depends on the applications. In applications where the number of distortion types

examined can be increased, models with a modular framework, such as CBIQ and LBIQ, are preferred to cater for a higher number of distortion types. This is accomplished at the expense of higher computational load. Fast computation is essential when the model must judge the image quality instantly such as on mobile devices. In such a scenario, fast models like GMLOG and BRISQUE are the best options.

2.4 Limitations and Proposed Solutions

It can be concluded these general-purpose BIQA models generally achieve highly competitive performances regarding well-known FR-IQA models. Their quality prediction performances are highly correlated to human perception of image quality. They also serve as the state-of-the arts of BIQA work. However, there are few limitations to be further addressed.

As it can be seen in sub-chapter 2.1, two-step approach is usually employed by these models: feature extraction followed by model regression by human scores. Kernel based learning methods, in particular SVR, are often utilised by these models to develop a mapping from the image's features to its image quality score. One major drawback of this approach is that they require training phase to optimise the regression (kernel) parameters. Although the training is often considered a one-time pre-processing step, it can take a long time especially for a huge image database. These models also need to re-train their regression parameters when images of new distortion types are introduced into the training data. Therefore, developing a model that requires minimal training or no training at all would be advantageous.

To address this limitation, one requires to develop a model that requires minimal training or no training at all. This study attempts to do this by proposing the use of nearest neighbour technique in the learning framework of a general-purpose BIQA model. This is motivated by the fact that the cost of learning for this technique is virtually zero where its training process only involves storing feature vectors and labels of the training images [110], alleviating the

need of regression parameters training phase. A BIQA model that integrate a nearest neighbour technique into its learning framework is presented in Chapter 3.

Another limitation shared by these models is that their performance degrades significantly when only being trained using a few training images. While increasing the number of training samples will help alleviate this problem, collecting large amounts of training samples for IQA is expensive as it involves obtaining additional images across wide ranges of quality level, content and distortion types. A model with robust performance regardless the size of training data is preferred. In addition, the previous models accumulate features over the entire image to derive the statistics required for quality estimation. Therefore, they can only provide a global estimate of image quality. These quality scores are uninformative enough where different parts of an image are subjected to different degradation levels. In such case, a model that can predict image quality locally could be useful. For example, for an image enhancement system, we only apply enhancement where necessary.

This study attempts to address these two limitations by introducing a BIQA model that operate on patch-level. To overcome the issue of small training sample number, this study proposes to artificially augmenting the existing databases by sampling image patches from the databases' images. This helps to increase the number of training samples for the model without having to obtain additional images. This study also proposes to extract relevant statistical features from those sampled image patches. This helps the model to directly perform local quality estimation on individual patches. One may question on the model performance validation as there are no local ground truth targets available currently. Similar to implementation in CNN [97] and DIQaM [99] models, the model can address this issue by assigning the image patches with quality labels from their corresponding annotated images. For the utilised databases, this practice is acceptable since the level of distortion is uniform across

the image. The patch-based BIQA model is first introduced in Chapter 3 while the improved version is presented in Chapter 4.

As discussed in sub-chapter 2.3, a BIQA model may has great quality estimation capability for images degraded by one particular type of distortion but may suffers when tested on images with different distortion types. To address this, the study also aims to look at the possibility of integrating a multi-task learning architecture into the model's framework. Multi-task learning (MTL) represents a learning technique that utilise a shared representation to learn multiple related tasks simultaneously. Based on the assumption that the learner may find it easier to learn multiple tasks together rather than in isolation when the tasks shared what they learn, MTL has been shown to improve the learning capability of each individual task [111]. In BIQA, by treating an individual distortion class as a single task, we could employ MTL to improve the quality prediction in each distortion case. A BIQA model that is designed based on MTL framework is presented in Chapter 5.

Finally, having a BIQA model capable of identifying the distortion affecting the image could also be useful in certain application domains. For example, in the restoration stage at the receiver end of an image communication system, it is easier to repair a distorted image if the distortion afflicting the image is known beforehand. Unfortunately, this property is unavailable in most of the previous models.

The study therefore proposes to introduce a distortion identification stage into the suggested model's framework. This is motivated by the intuition that the perceived quality of an image degraded by a particular distortion would be best predicted by images of the same distortion type. Therefore, by first identifying the distortion affecting the image, more relevant training samples could be selected for quality estimation purposes. This additional property makes the model appealing for applications where the knowledge of distortion type is

necessary. Both patch-based BIQA models presented in Chapter 3 and Chapter 4 integrate distortion identification stage into their learning frameworks. The MTL-based BIQA model introduced in Chapter 5 also has the capability to identify distortion within the tested images.

At the time this thesis is being written, the author have been made aware of other works that also introduce distortion identification stage into a BIQA framework. For completeness, the works are briefly reviewed here. In [112], Chetouani et al propose to perform distortion classification prior to quality estimation. The classification is performed via linear discriminant analysis (LDA) classifier. The 8 input features to the classifier are selected to represent common image distortions such as noise (4 features), blur (1 feature), blocking (1 feature) and ringing (2 features). These features are extracted from the test image using different BIQA models depending on the distortion type. Once the classifier identifies the distortion type, appropriate BIQA models are then selected to perform quality estimation.

The work is later extended in [113]. Here, metrics from different IQA models are proposed to be employed directly as input features to the LDA classifier. This is based on the assumption that different IQA models exhibit specific response for a given degradation type. A total of 12 features are utilised. Again, depending on the identified distortion class, appropriate BIQA models are then utilised to estimate the image quality. In [114], different features are first extracted from the image to model each degradation type considered: 3 features for noise, 3 features for blur, 4 features for blocking and 3 features for ringing, respectively. The resulting models are then used to perform quality estimation for different distortion. The scores are then combined to achieve final quality score for the image.

Note that there are noticeable differences between those works and the models presented by this study. First, the models here use nearest neighbour based classifier to perform distortion identification as opposed to LDA classifier by those works. Second, the models here performs

both distortion identification and quality estimation at patch-level as opposed to image-level operation by those works. Third, all features utilised by the proposed models are general-purpose and not limited for specific distortion. In contrast, those works employ a combination of distortion-specific features to perform their operation. The utilised features will be described in details in Chapter 3, 4 and 5.

Chapter 3

Patch Based Learning Framework for Blind Image Quality Assessment Model

3.1 Chapter Introduction

At the end of Chapter 2, few limitations of BIQA models have been identified. These include intensive training phase requirements, inability to provide local quality estimation and inability to identify the distortion affecting an image. Two potential solutions were then proposed: the use of nearest neighbour techniques and local feature extraction. This chapter describes the first proposed BIQA model that integrate these solutions in its model framework.

The model, dubbed PATCH based blind Image Quality assessment (PATCH-IQ), has a five-stage framework. Given an image, PATCH-IQ first samples non-overlapped local patches. At the second stage, it then extracts spatial domain BIQA features from those patches. Rather than using the features directly for quality analysis, PATCH-IQ intuitively assumes that the perceived quality of a distorted image will be best predicted by features drawn from images of the same distortion class. Therefore, PATCH-IQ introduces a distortion identification process in the third stage. A nearest neighbour classifier is employed to perform such a task. The classifier achieves this by minimising the Image-to-Class (I2C) distance between the image's patches and a set of annotated image patches. The patches correspond to the identified distortion class are then utilised in the fourth stage to predict local image quality. This is done via a k-nearest neighbour regression that associates the local image quality with the DMOS of

the annotated patches constrained to the identified distortion class. Finally, an overall image quality score is derived by pooling the local scores of all patches in the image.

The remainder of this chapter is structured as follows. The PATCH-IQ framework will be described in details in sub-chapter 3.2. In sub-chapter 3.3, we will then look at the experimental results and later analyses. Sub-chapter 3.4 will conclude the chapter.

3.2 Patch Based Framework for Blind Image Quality Assessment

The framework for PATCH-IQ is illustrated in Figure 3.1.

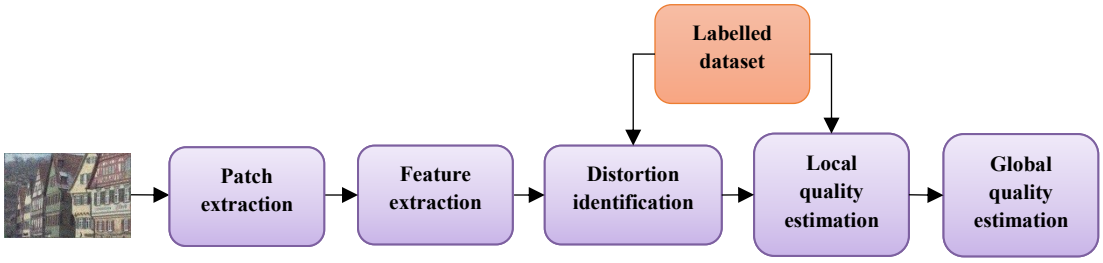


Figure 3.1: PATCH-IQ framework

3.2.1 Local feature extraction

As presented in Chapter 2, there are various statistical features can be used to perform a BIQA task. The choice of features for PATCH-IQ is affected by two main factors. First, it is crucial to employ features with low computational requirements since they are to be extracted at patch level. In this aspect, spatial domain features are chosen to avoid expensive computation normally encountered by image transform-based features. Second, the selected features should carry information not only on perceptual quality but on the distortion in the image as well. The same spatial domain features as implemented by the BRISQUE model [74] are therefore adopted.

As in BRISQUE, PATCH-IQ utilises the empirical distributions of locally normalised luminance coefficients and pairwise products of these coefficients to design 18 statistical

features for both BIQA and distortion identification tasks. Given an image **I**, PATCH-IQ first samples non-overlapped patches of $h_{\mathrm{p}} \times w_{\mathrm{p}}$ size. For a patch **P**, its locally normalised luminance coefficients are obtained by computing local mean subtraction and divisive normalisation at each location (i, j) :

$$\widehat{\mathbf{P}}(i,j) = \frac{\mathbf{P}(i,j) - \mu(i,j)}{\sigma(i,j) + \varepsilon_{\mathrm{B}}}, \tag{3.1}$$

where the local mean field $\mu(i,j)$ is defined as:

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} \mathbf{P}_{k,l}(i,j), \tag{3.2}$$

and the local variance field $\sigma(i,j)$ is given by:

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} \omega_{k,l} \left(\mathbf{P}_{k,l}(i,j) - \mu(i,j)\right)^2}. \tag{3.3}$$

In these equations, $i \in 1,2,\ldots,h_{\mathrm{p}}$ and $j \in 1,2,\ldots,w_{\mathrm{p}}$ are spatial indices with h_{p} and w_{p} being the patch height and width, respectively. Here, ε_{B} is a constant to prevent the denominator in Equation (3.1) from falling to zero while $\omega_{\mathrm{G}} = \{\omega_{k,l} | k = -K, \ldots, K, l = -L, \ldots, L\}$ is a Gaussian weighting function sampled with 3 standard deviations and rescaled to unit sum and $K = L$ is the function window size.

Figure 3.2 shows the histogram plot of the normalised luminance coefficients for a natural undistorted image and for its various distorted versions. The undistorted / reference image demonstrates a Gaussian-like distribution while different distortion changes the coefficient's distribution in its own way. For example, white noise affects the image by reducing the weight of the tail of the histogram while blur causes the image to exhibit a more Laplacian-like distribution. These observations indicate that the coefficients' statistical properties are modified by distortion. Quantifying these modifications through a statistical

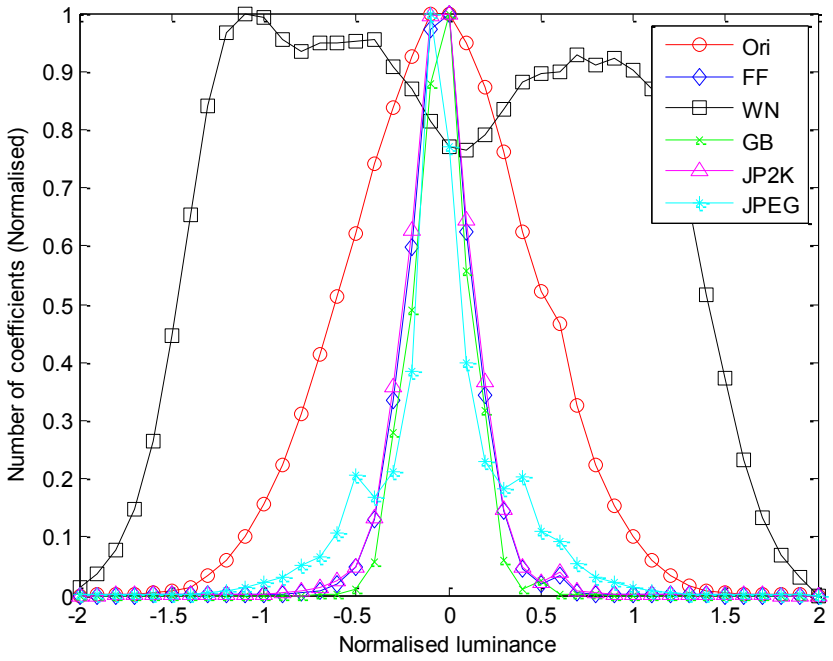


Figure 3.2: Histogram of normalised coefficients for a natural undistorted image and its various distorted versions.

model thus will make it possible for us to perform BIQA task. In agreement with BRISQUE implementation, a generalised Gaussian distribution (GGD) model is used to fit the empirical distribution of the coefficients. GGD model is chosen as it can effectively capture a broader spectrum of distorted images statistics [115]. The empirical distribution of these coefficients is fitted by a GGD model as [74]:

$$f(x;\mu,\sigma^2,\gamma) = a\exp[-(b|x-\mu|)^{\gamma}]\,, \tag{3.4}$$

with

$$a = b\gamma/2\Gamma(1/\gamma)\,, \tag{3.5}$$

$$b = (1/\sigma)\sqrt{\Gamma(3/\gamma)/\Gamma(1/\gamma)}\,, \tag{3.6}$$

and

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt \quad x > 0\,. \tag{3.7}$$

In Equation (3.4), μ, σ^2 and γ are the mean, the variance and the shape parameter of the distribution respectively, whereas $\Gamma(x)$ is the gamma function. The estimated parameters: σ^2 and γ are then chosen as the first two features.

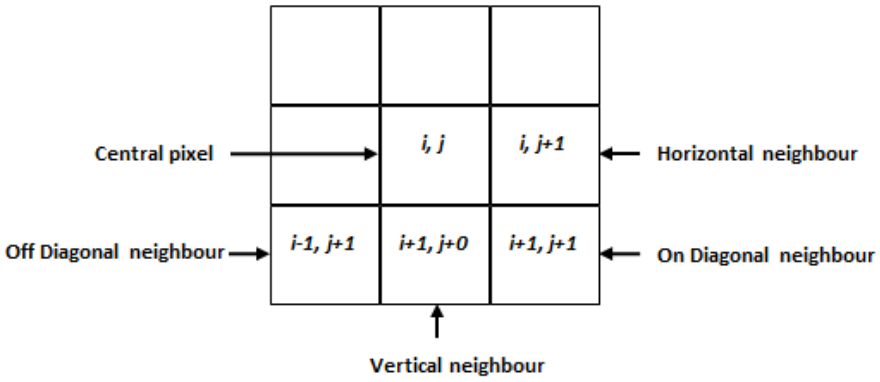


Figure 3.3: The four orientations' of the pairwise product [55]

The other 16 statistical features are next derived from the empirical distributions of the pairwise products of neighbouring luminance coefficients. The pairwise products are first computed on four orientations: horizontal, vertical, main-diagonal, and secondary-diagonal as in Figure 3.3. Instead of GGD, the distributions of these products are modelled by an asymmetric generalised Gaussian distribution (AGGD). The AGGD generalises the GGD by allowing for asymmetry in the distributions. The AGGD is defined as [74]:

$$f(x;\nu,\sigma_l^2,\sigma_r^2) = \frac{\gamma}{(b_l+b_r)\Gamma(1/\gamma)}\exp[-(-x/b_l)^{\nu}] \quad x<0\ , \tag{3.8}$$

and

$$f(x;\nu,\sigma_l^2,\sigma_r^2) = \frac{\gamma}{(b_l+b_r)\Gamma(1/\gamma)}\exp[-(x/b_r)^{\nu}] \quad x\geq 0\ , \tag{3.9}$$

where

$$b_l = \sigma_l\sqrt{\Gamma(1/\nu)/\Gamma(3/\nu)} \text{ and } b_r = \sigma_r\sqrt{\Gamma(1/\nu)/\Gamma(3/\nu)}\ . \tag{3.10}$$

In these equations, ν, σ_l^2 and σ_r^2 are the shape parameter, the left variance and the right variance of the distribution, respectively. The three parameters and the mean of the best AGGD fit are then selected at each orientation to represent those 16 features.

Since images are naturally multiscale and IQA models that incorporate multiscale information achieved better correlation with human perceptual measures of image quality [116], PATCH-IQ extracts these 18 features over two scales. A total of 36 features are used by PATCH-IQ to perform both distortion identification and quality estimation. The suitability of

the chosen features in performing both distortion identification and quality analysis will be discussed in sub-chapter 3.2.3 and sub-chapter 3.2.4, respectively. Table 3.1 summarises the extracted features.

Table 3.1: List of extracted features

Feature ID	Scale	Orientation	Feature Description
1-2	1	-	Shape parameter and variance of GGD model of normalised luminance coefficients
3-6		Horizontal	Shape parameter, mean, left variance and right variance of AGGD model of pairwise products
7-10		Vertical	
11-14		Main-diagonal	
15-18		Secondary-diagonal	
19-20	2	-	Shape parameter and variance of GGD model of normalised luminance coefficients
21-24		Horizontal	Shape parameter, mean, left variance and right variance of AGGD model of pairwise products
25-28		Vertical	
29-32		Main-diagonal	
33-36		Secondary-diagonal	

3.2.2 Labelled dataset construction

Since PATCH-IQ employs a nearest-neighbour technique to perform image distortion identification and quality estimation, a labelled dataset **D** consisting of BIQA features extracted from patch exemplars must be constructed. Most of BIQA models employ the 80:20 train-test ratio to train their regression models [117]. PATCH-IQ follows the same partition setting to build the dataset, i.e. patches from 80% of the randomly selected reference images from a standard IQA database and their distorted versions are used to extract the features for the dataset. Specifically, given a labelled image, PATCH-IQ first divides the image into P non-overlapping patches of $h_{\mathrm{p}} \times w_{\mathrm{p}}$ size. BIQA features, as discussed in sub-chapter 3.1.2, are then extracted on those patches. The extracted feature vectors are next combined over all the labelled images to form the dataset. Denote the total of labelled images by N_{label}, the size of feature matrix for the dataset is:

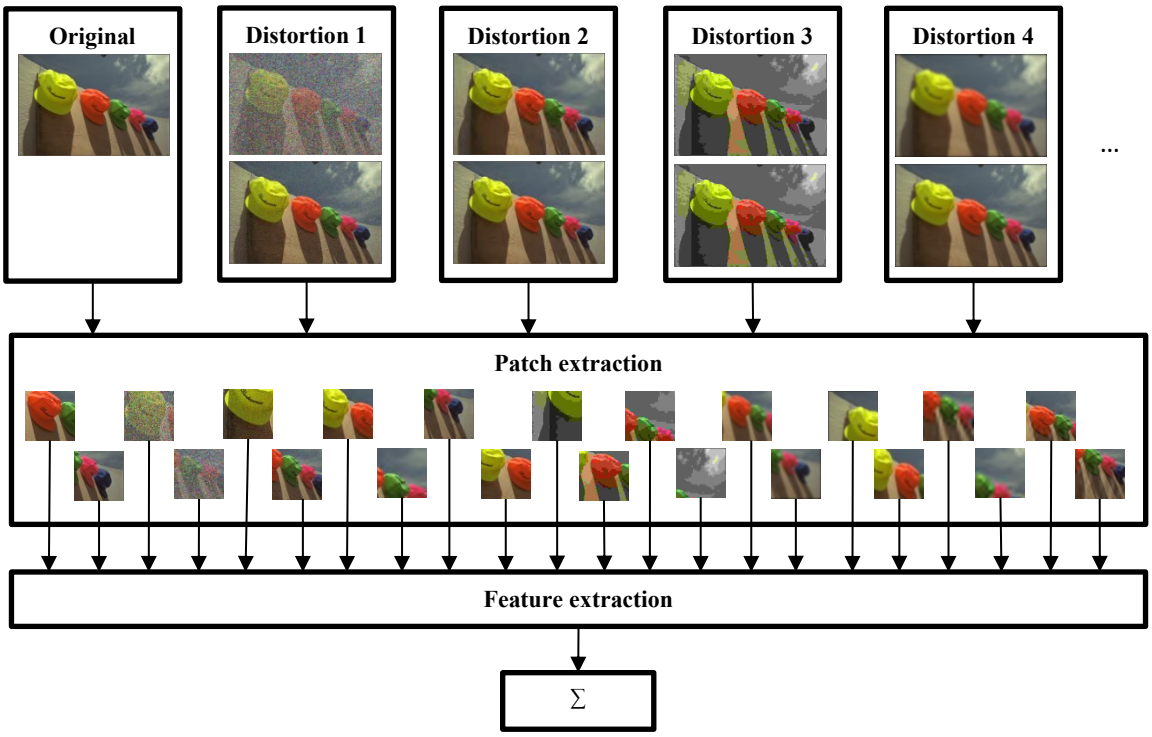


Figure 3.4: Example of labelled dataset construction

$$\mathbf{D} = \left[\left(\sum_{i=1}^{N_{\text{label}}} P_i\right) \times 36\right]. \tag{3.11}$$

PATCH-IQ assigns the patches with two labels. The first label is the distortion class. Each patch is labelled according to the distortion type in its source image. The second label is the subjective score. Each patch is assigned with its source images' subjective score, provided in the chosen IQA database. As discussed in sub-chapter 2.4, assigning the score in this way is acceptable as the distortion levels across the database images are uniform. An example of a dataset built from one image and its distorted versions is shown in Figure 3.4. There is no fixed number of distortion classes for the dataset. If the images from new distortion classes are provided, they can be added directly to the dataset.

3.2.3 Distortion identification

The third stage of the framework identifies the distortion class of the image. To show that the extracted features can capture image distortion, a 2-D scatter plot between the shape

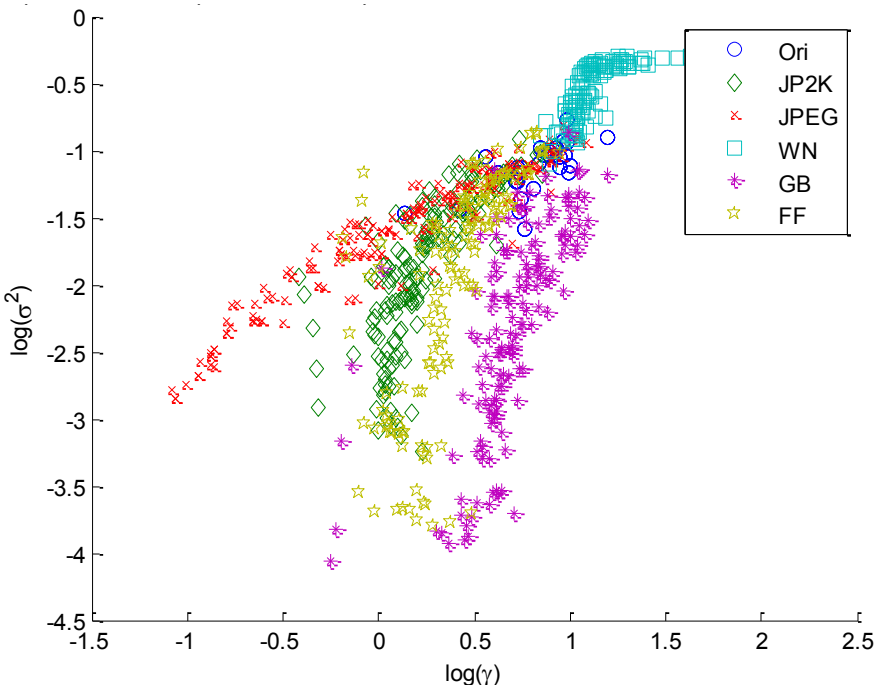


Figure 3.5: 2-D scatter plot between the shape and the scale parameters of the GGD model of the normalised luminance coefficients for the LIVE IQA database images.

and the variance parameters of the GGD model of the normalised luminance coefficients is generated. Figure 3.5 shows the results for the undistorted reference images and their corresponding distorted versions from the LIVE IQA database. It is easy to visualise from the figure that images from different distortion types are well separated in GGD parameter space showing the suitability of using these two features to perform distortion classification. WN, GB and JPEG images are well separated making them among the easiest to be identified. Meanwhile, a 3-D scatter plot of the shape parameter and both right and left variance parameters of the AGGD model of the horizontally paired products is plotted in Figure 3.6 using the same set of images. Again, it shows that different distortions occupy different regions of the parameter space. This justifies the use of these AGGD parameters as the features for distortion classification purposes. Similar patterns could be observed for features extracted on different orientations and scales.

Given a test image $\mathbf{I}_{\text{test}}$, PATCH-IQ extracts BIQA features using the same procedure as in sub-chapter 3.2.1 to form the image's feature matrix $\mathbf{F}_{\mathbf{I}_{\text{test}}}$. PATCH-IQ then identifies the

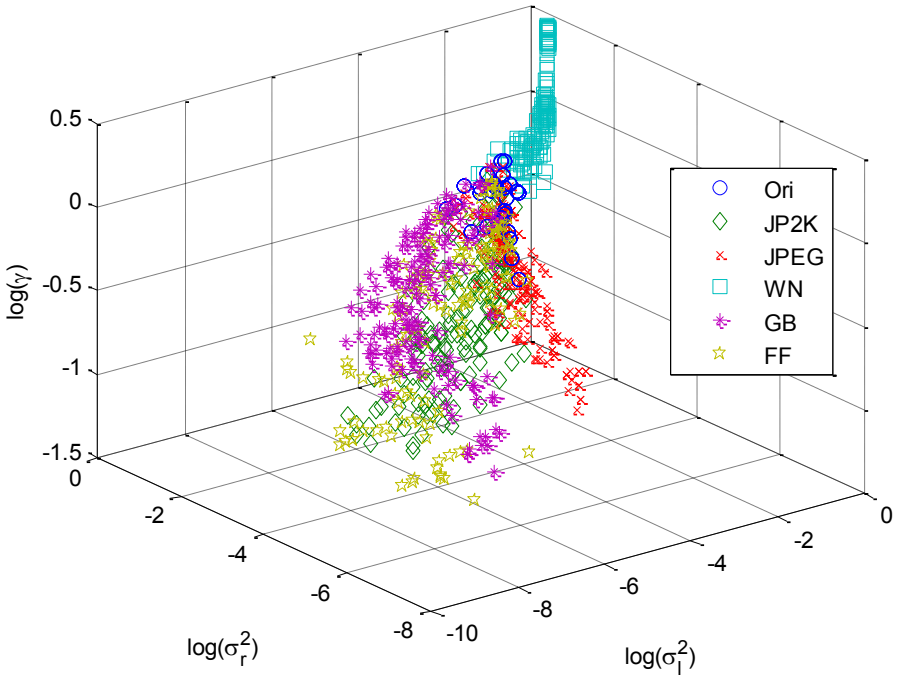


Figure 3.6: 3-D scatter plot of the shape parameter and both left variance and right variance parameters of the AGGD model of the pairwise product in horizontal orientation for the LIVE IQA database images

distortion type associated with the image by employing a nearest neighbour based classifier. In a nearest neighbour classification case, it has been shown that the optimal distance measurement is I2C distance rather than the usually used image-to-image (I2I) distance. A popular I2C based classifier, the Naïve Bayes nearest neighbour (NBNN) [118], is utilised. PATCH-IQ computes the distance between $\mathbf{F}_{\mathbf{I}_{\text{test}}}$ and the feature matrix from each of the distortion classes in the dataset $\mathbf{D}$. The predicted distortion class for the image $\hat{c}$ is then represented by the class with the minimum I2C distance value [118]:

$$\hat{c} = \arg\min_{c} \left\| \mathbf{F}_{\mathbf{I}_{\text{test}}} - NN_c\left(\mathbf{F}_{\mathbf{I}_{\text{test}}}\right) \right\|^2 , \tag{3.12}$$

where $NN_c\left(\mathbf{F}_{\mathbf{I}_{\text{test}}}\right)$ is the NN-descriptor of $\mathbf{F}_{\mathbf{I}_{\text{test}}}$ in the distortion class c.

3.2.4 Local quality estimation

The fourth stage of the framework is to estimate the quality of the image patches. To visualise the relationship between the utilised features and human perception of image quality,

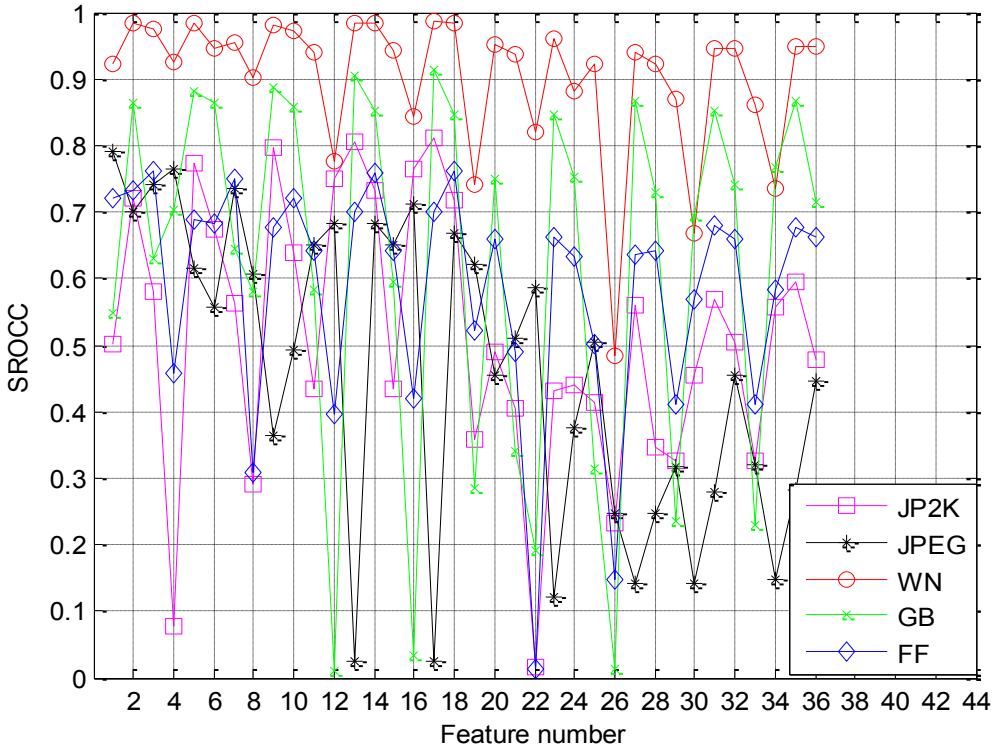


Figure 3.7: Correlation of the extracted features with the DMOS for different distorted images in the LIVE database

the SROCC values between features derived from the LIVE images and their corresponding DMOS values are plotted. The plot is shown in Figure 3.7. We can see that the way images are affected and how each feature captures quality information vary differently depending on types of distortion. The figure also indicates that the features generally correlate well with human perception of quality, particularly in WN, GB and JP2K cases, justifying their suitability for quality estimation task.

PATCH-IQ works based on the intuition that the quality of a patch would be best predicted by patches of the same distortion type. Therefore, it performs quality estimation utilising only the labelled patches within the distortion class identified in the previous stage. PATCH-IQ then assumes that patches with similar features are perceived to have the same quality. Here, better quality prediction can be achieved by selecting a set of labelled patches that are similar to the test patch in feature space. PATCH-IQ performs this through a *k*-NN regression algorithm.

For each test image patch $p_i, i = 1,2, \dots, P_{\text{test}}$ the Euclidean distances d_{ij} between the patch and the labelled patches of the identified distortion class $p_j, j = 1,2, \dots, P_{\text{label}}$ is first calculated in the feature space. The labelled patches are then rearranged in ascending order according to the computed distances. The first k_{NN} labelled patches are then utilised to estimate the patch quality. Figure 3.8 illustrates an example of this selection process.

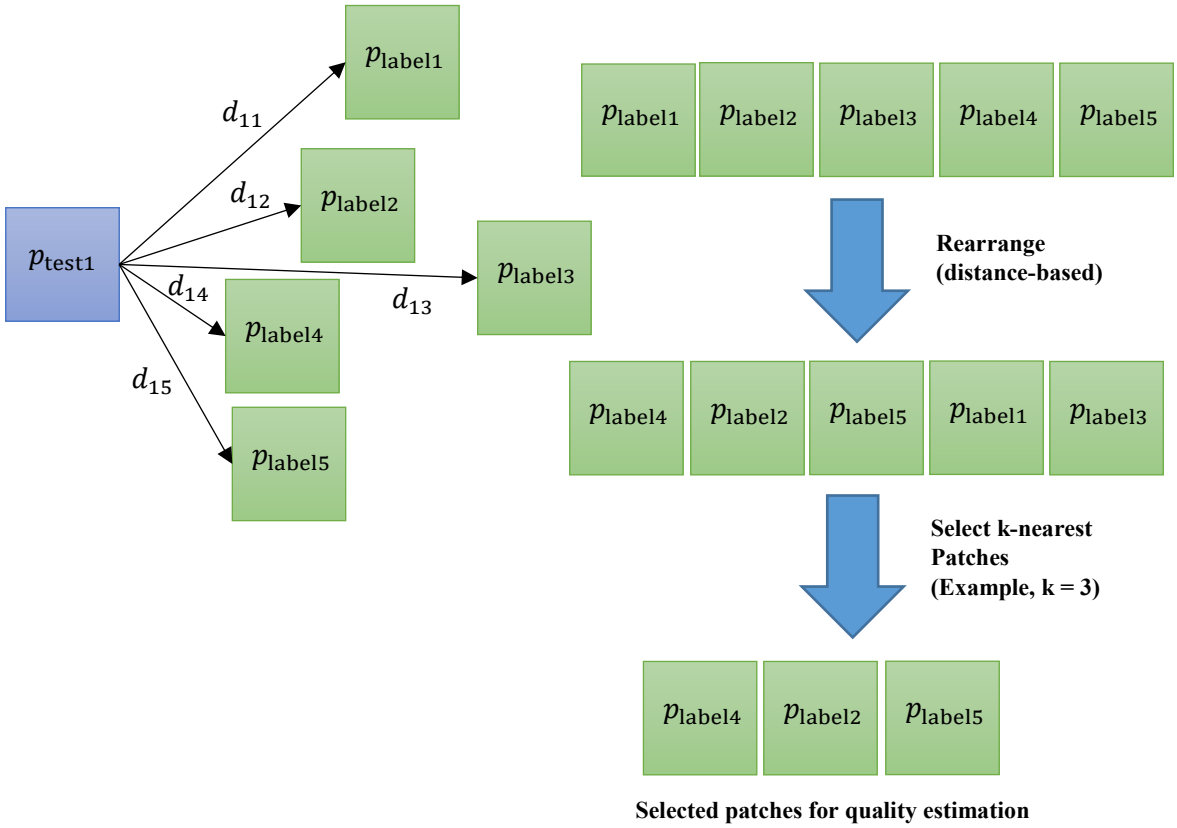


Figure 3.8: Example of k-nearest patches selection for local quality estimation

However, instead of using common inverse distance weighting scheme over the selected patches, the test patch quality is estimated through a linear regression:

$$q_{p_i} = \omega_{\text{P}}(f_{p_i}), \tag{3.13}$$

where ω_{P} are the optimised weight vector for the patch feature vector f_{p_i}.The weights can be calculated as [119]:

$$\omega_{\mathrm{P}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{s}\,, \tag{3.14}$$

where $\mathbf{X}$ is the feature matrix of the selected labelled patches and s represents their corresponding DMOS scores.

3.2.5 Global quality estimation

The final stage of the framework is basically a pooling stage. The patches' scores are pooled to yield the global quality score for the image. Instead of typical average or max pooling, PATCH-IQ employs an inverse weighting rule to pool all the patches' scores. In this framework, PATCH-IQ assigns each local score with a weight based on their minimum Euclidean distance $d_{ij\min}$ computed in the previous local quality estimation stage. Figure 3.9 illustrates the process. The image-level quality score for the image is then given as:

$$q_{\mathbf{I}} = \frac{\sum_{i=1}^{P_{\mathrm{test}}} \omega_i q_{p_i}}{\sum_{i=1}^{P_{\mathrm{test}}} \omega_i}\,, \tag{3.15}$$

where

$$\omega_i = \frac{\sum_{i=1}^{P_{\mathrm{test}}} d_{ij\min}}{d_{ij\min}}\,. \tag{3.16}$$

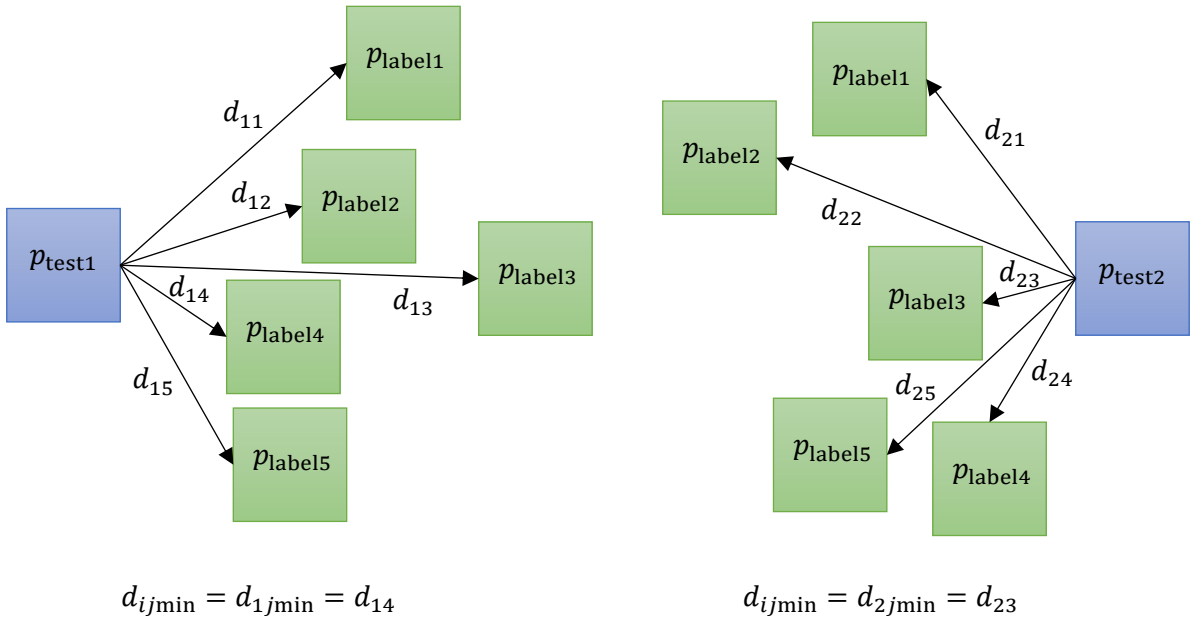


Figure 3.9: Local quality score weighting scheme

3.3 Results and Discussions

3.3.1 Experimental setup and evaluation protocol

Databases: There are several established subjective image evaluation databases within the IQA research area. Two of the widely used IQA databases were utilised to evaluate the performance of PATCH-IQ: LIVE [14] and CSIQ [15].

Framework parameters: The parameters were empirically determined. For the feature extraction stage, the local window size $K = L$ was 3 and constant ε_{B} was 1 as in the BRISQUE model while the patch size $h_{\mathrm{P}} = w_{\mathrm{P}}$ was set at 96. Meanwhile, the number of nearest neighbour patches for linear regression in the local quality estimation stage was set at 1000.

Performance metrics: There are several performance metrics available for model evaluation. Given the scope of the study and for ease of comparison, the performance evaluation of PATCH-IQ is reported in the same way as reported by the benchmarked models. Therefore, two correlation measures were utilised to evaluate the prediction performance of PATCH-IQ: LCC and SROCC. In addition, another metric the root mean square error (RMSE) was also employed. Similar to LCC, the RMSE can evaluate the prediction accuracy of a model. It is represented as [120]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{test}}}\sum_{i=1}^{N_{\mathrm{test}}}(\mathrm{s}_i - q_i)^2} \quad . \tag{3.17}$$

In Equation (3.17), N_{test} is the number of test images, q_i is the predicted score of the ith image and s_i is the image's subjective score. In contrast to LCC and SROCC, a value closer to 0 for RMSE indicates higher correlation between the predicted score and the human subjective score.

Benchmarked models: PATCH-IQ was compared against four state-of-the-art BIQA models: BIQI [66], BRISQUE [74], GMLOG [76], and CORNIA [91], whose source codes are publicly available. PATCH-IQ was also compared with three FR-IQA models: PSNR, SSIM

[44] and FSIM [45]. To train these BIQA and FR-IQA models, the databases were divided into two subsets: 80% of the reference images and their corresponding distorted versions were randomly selected to be a training set while the remaining 20% reference images and their associated distorted images were used for testing. There was no overlap between the two sets. The same training set was used to construct the labelled dataset required by PATCH-IQ. The LIBSVM [121], [122] package was utilised to perform regression for the four BIQA models: SVR with a RBF kernel for BIQI, BRISQUE and GMLOG and SVR with a linear kernel for CORNIA. For fair comparison, their SVR parameters were determined through cross validation in accordance to their respective papers.

3.3.2 Evaluation on individual databases

The same two experiments as in sub-chapter 2.2.1 were conducted to ascertain the overall performance and the distortion-specific (DS) performance of each model. The experiments were performed 1,000 times to ensure that the results are not governed by the specific train-test partition. The median results for both the overall experiment and the DS experiment are tabulated in Tables 3.2 and 3.3, respectively. For simplicity, only the SROCC results are shown for the DS performance experiment. Similar patterns can be observed for the LCC and RMSE results. Note that for the CSIQ database, only four distortions also present in the LIVE database: JP2K, JPEG, WN and GB are considered. The top FR-IQA and BIQA models are in bold.

Table 3.2: Median values across 1,000 runs of the overall performance experiment

IQA model	LIVE			CSIQ		
	LCC	SROCC	RMSE	LCC	SROCC	RMSE
PSNR	0.882	0.883	12.898	0.856	0.929	0.144
SSIM	0.946	0.949	8.804	0.935	0.936	0.099
FSIM	**0.961**	**0.964**	**7.546**	**0.968**	**0.963**	**0.071**
BIQI	0.849	0.844	15.407	0.809	0.749	0.187
BRISQUE	0.943	0.942	9.395	0.930	0.910	0.107
GMLOG	0.951	0.950	8.829	0.939	0.925	0.010
CORNIA	0.939	0.942	9.920	0.911	0.887	0.125
PATCH-IQ	***0.954***	***0.952***	***8.476***	***0.946***	***0.932***	***0.094***

Table 3.3: Median SROCC values across 1,000 runs of the DS performance experiment

IQA model	LIVE					CSIQ			
	JP2K	JPEG	WN	GB	FF	JP2K	JPEG	WN	GB
PSNR	0.895	0.881	**0.985**	0.782	0.891	0.936	0.888	**0.936**	0.929
SSIM	0.961	0.976	0.969	0.952	**0.956**	0.961	0.955	0.897	0.961
FSIM	**0.972**	**0.984**	0.972	**0.971**	0.952	**0.970**	**0.966**	**0.936**	**0.973**
BIQI	0.830	0.906	0.933	0.866	0.689	0.764	0.910	0.540	0.783
BRISQUE	0.916	0.964	0.979	0.945	0.887	0.898	0.921	0.921	**0.919**
GMLOG	0.927	0.963	0.983	0.929	0.901	0.916	0.936	0.941	0.908
CORNIA	0.921	0.936	0.961	0.952	**0.905**	0.894	0.882	0.786	0.904
PATCHIQ	***0.931***	***0.976***	***0.987***	***0.953***	*0.891*	***0.918***	***0.952***	***0.963***	*0.916*

In the overall performance experiment, PATCH-IQ produced the best values for all three performance metrics among the BIQA models when tested on the LIVE database. Similar results were obtained for the CSIQ database. In the DS performance experiment, PATCH-IQ had the highest SROCC values on both databases for images distorted by the JPEG and JP2K compression artefacts. It also yielded the best SROCC values for WN images. In the GB cases, PATCH-IQ performed the best on the LIVE database while came second on the CSIQ database. It also gave comparable prediction performance in FF cases. Compared to the FR-IQA models, PATCH-IQ achieved better overall performance compared to PSNR and SSIM while approaching FSIM. In terms of individual distortions, it outperformed PSNR and yielded competitive performance to SSIM and FSIM. It also outperformed both models for WN images. Given FR-IQA models require a reference image as their input, PATCH-IQ's performance is promising.

To access the consistency of PATCH-IQ's quality prediction performance, the inter-quartile range (IQR) results of all the SROCC and LCC obtained from the 1,000 runs of experiments on both databases are tabulated in Table 3.4. A model with low IQR value indicates that its' results are more consistent under different train-test partitions. The box plots of SROCC and LCC distributions for all tested BIQA models are also shown in Figure 3.10.

The central mark on each box is the median while the top edge and the bottom edge are the 25th and 75th percentiles, respectively.

Table 3.4: IQR values for 1,000 SROCC and LCC values obtained

Database	LIVE		CSIQ	
Metrics	LCC	SROCC	LCC	SROCC
BIQI	0.053	0.054	0.071	0.096
BRISQUE	0.020	0.020	0.036	0.039
GMLOG	0.017	0.016	0.024	0.026
CORNIA	0.018	0.018	0.041	0.052
PATCH-IQ	*0.018*	*0.019*	*0.028*	*0.027*

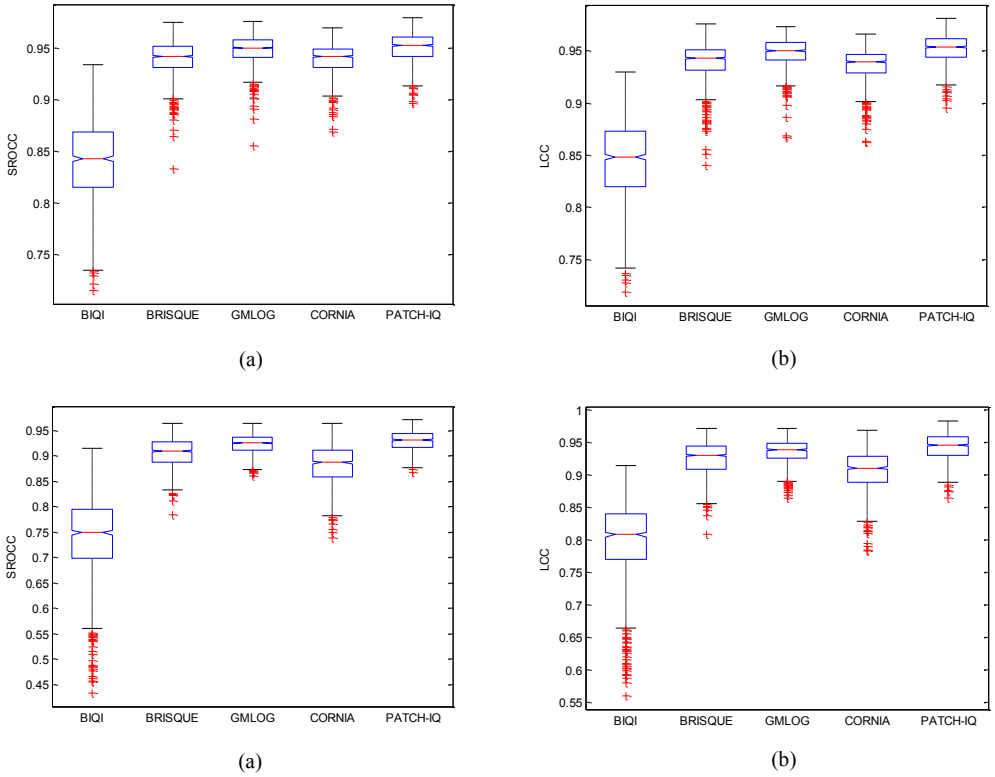


Figure 3.10: Box plots of performance metric distributions of BIQA models from 1,000 runs of experiments on the LIVE database (top row) and the CSIQ database (bottom row): (a) SROCC and (b) LCC

On the LIVE database, PATCH-IQ obtained lower IQR values than BIQI and BRISQUE but slightly higher than GMLOG and CORNIA. Similar pattern can be observed on the CSIQ database with an exception that PATCH-IQ now had lower IQR values than CORNIA. This indicates that PATCH-IQ produced more consistent prediction than both BIQI and BRISQUE

but was less consistent than GMLOG. In terms of the outliers, ideally we would like as few outliers as possible and to have them as close to the main distribution as possible. Here, it can be seen that PATCH-IQ had more compact set of outliers than most models on both databases. These IQR and outlier observations indicate that, while PATCH-IQ may not be the most consistent model, it still achieved acceptable quality prediction consistency throughout the 1000 runs of experiments. Note that PATCH-IQ predicts image quality based on the annotated patches from the previously identified distortion class. One possible factor that contributes to the prediction results variability and outliers is the PATCH-IQ's capability to classify the distortion accurately. Therefore, by improving its distortion identification accuracy, PATCH-IQ could also make its prediction performance more consistent.

3.3.3 Statistical significance and hypothesis testing

The differences in median correlations between the competing BIQA models may not be statistically significant. Therefore, a hypothesis test to evaluate the statistical significance difference between each model is conducted. As the SROCC and LCC values follow right-skewed unimodal distributions, the Wilcoxon rank-sum test is employed avoiding the normality assumption required by a typical t-test [105]. The Wilcoxon rank-sum test measures the equivalence of the median values of two independent samples. The test is performed on the SROCC values obtained from the 1,000 runs of experiments at a significance level of 0.01. The null hypothesis is that the SROCC values of the two BIQA models are drawn from the populations with equal median while the alternative hypothesis is that the median of one model is greater than the other.

The results are shown in Table 3.5. A score of '1' implies there is a statistically significant difference between both models and the model in row has a larger median than the model in column. A score of '-1' also implies there is a statistically significant difference between the

models, but the model in column now has a larger median than the model in row. A score of '0' indicates the null hypothesis cannot be rejected and there is no statistically significant difference between both row and column models. On both the LIVE and the CSIQ databases, PATCH-IQ is statistically different to all four competing models.

Table 3.5: Results of the Wilcoxon rank-sum test using the SROCC values of competing BIQA models

LIVE					
	BIQI	BRISQUE	GMLOG	CORNIA	PATCH-IQ
BIQI	0	-1	-1	-1	-1
BRISQUE	1	0	-1	1	-1
GMLOG	1	1	0	1	-1
CORNIA	1	-1	-1	0	-1
PATCH-IQ	*1*	*1*	*1*	*1*	*0*
CSIQ					
	BIQI	BRISQUE	GMLOG	CORNIA	PATCH-IQ
BIQI	0	-1	-1	-1	-1
BRISQUE	1	0	-1	1	-1
GMLOG	1	1	0	1	-1
CORNIA	1	-1	-1	0	-1
PATCH-IQ	*1*	*1*	*1*	*1*	*0*

3.3.4 Effects of labelled dataset size

The size of the labelled dataset plays an important role to ensure the model's processing time is at an acceptable level. The database size is determined by two parameters. They are: 1) the size of image patch and 2) the number of labelled images employed. In this sub-section, the study investigates how these parameters affect PATCH-IQ's prediction performance.

Since PATCH-IQ samples its image patches in non-overlapping way, a smaller patch size will lead to a larger number of samples for the labelled dataset. Having a large number of labelled samples is preferred as it normally help a model to obtain better learning capability. However, at the same time, PATCH-IQ employs BRISQUE features in its learning framework. Since BRISQUE features are designed as global IQA features [74], i.e. accumulated over the entire image, operation on a larger image patch will results in more discriminative IQA features being extracted.

To investigate the effect of patch size variation on the prediction performance, PATCH-IQ was tested on the LIVE and the CSIQ databases with 8 different patch sizes: 16, 32, 48, 64, 80, 96, 112, and 128. All other PATCH-IQ parameters were fixed at the initial values as in sub-section 3.3.1. The performance variation of PATCH-IQ is shown in Table 3.6 and Figure 3.11, respectively. A larger patch in generally will lead to higher SROCC and LCC values. For the LIVE database, there was an obvious increment in both values as the patch size increases from 16 to 64. After that, the values appear to be stabilised with the optimum values achieved at patch size of 96. As such, PATCH-IQ empirically choose image patches with the size of 96 in its framework. Similar patterns were observed for the CSIQ database testing whereby the optimum SROCC and LCC values were also achieved with patch with the size of 96. This suggests that the patch size utilised in PATCH-IQ framework is independent of the databases.

Table 3.6: LCC and SROCC comparison for different patch size

Size	16	32	48	64	80	96	112	128
LIVE								
LCC	0.581	0.824	0.941	0.950	0.951	**0.954**	0.949	0.950
SROCC	0.525	0.826	0.938	0.948	0.948	**0.952**	0.948	0.950
CSIQ								
LCC	0.577	0.818	0.934	0.942	0.944	**0.946**	0.942	0.944
SROCC	0.518	0.810	0.921	0.930	0.930	**0.932**	0.930	0.932

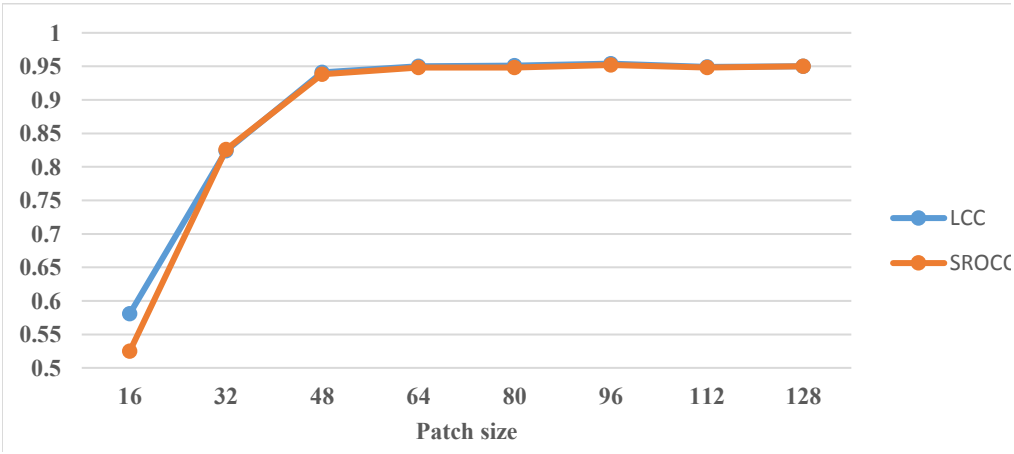


Figure 3.11: LCC and SROCC variation for different patch size tested on LIVE database

Next, to investigate the effect of the number of images in the labelled dataset, the two databases are partitioned under three labelled-test ratios: 80:20, 50:50 and 30:70. All PATCH-IQ parameters were fixed at the initial values as in sub-section 3.3.1. The four competing BIQA models are also evaluated under the same settings. The SROCC results for the overall performance experiment are shown in Table 3.7 and Figure 3.12.

Table 3.7: SROCC comparison for different training (labelled) samples ratio

Database	LIVE			CSIQ		
Ratio	80%	50%	30%	80%	50%	30%
BIQI	0.844	0.835	0.816	0.749	0.737	0.718
BRISQUE	0.942	0.927	0.903	0.910	0.895	0.872
GMLOG	0.950	0.940	0.925	0.925	0.909	0.887
CORNIA	0.942	0.937	0.929	0.887	0.881	0.873
PATCHIQ	***0.952***	***0.945***	***0.933***	***0.932***	***0.920***	***0.907***

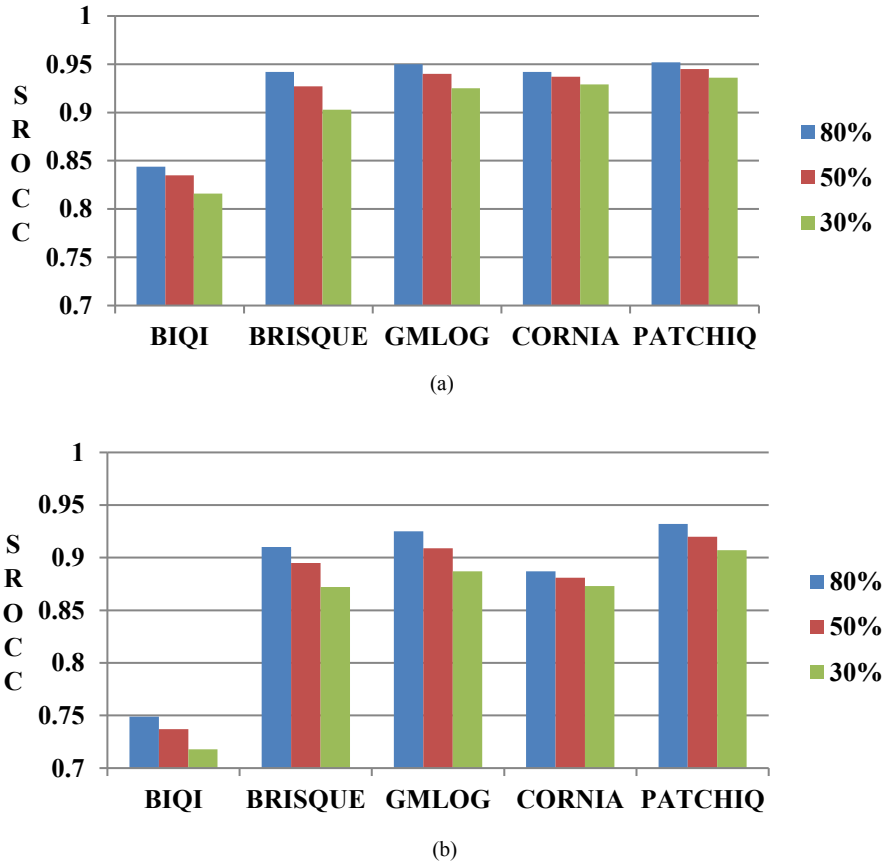


Figure 3.12: SROCC comparison for different training (labelled) ratios on: (a) LIVE and (b) CSIQ

As expected, the performances of all tested BIQA models decreased as the number of samples is reduced. We can also see that PATCH-IQ still produced the best SROCC values at all three ratios for both databases. In terms of rate of change, PATCH-IQ's performance reduced by 0.74% and 1.29% when the labelled samples were reduced from 80% to 50% on the LIVE and the CSIQ databases, respectively. When the samples were reduced from 80% to 30%, its performance degraded by 2.00% on the LIVE database and by 2.68% on the CSIQ database. Compared to the competing models, PATCH-IQ produced better rate of change than the rest except for CORNIA. The results suggest that PATCH-IQ is more robust to the number of training samples (labelled samples) than BIQI, BRISQUE and GMLOG. This also prove that PATCH-IQ works well where the number of samples is small.

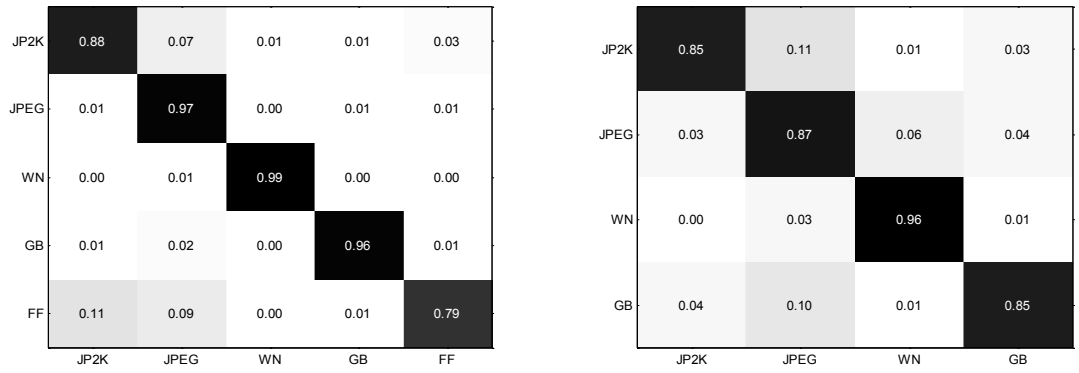
3.3.5 Distortion identification accuracy

Another useful property of PATCH-IQ is its ability to identify the distortion affecting the image. A popular NBNN classifier is employed by PATCH-IQ to perform distortion identification task. To show that the chosen classifier is capable to provide good classification performance, the median classification accuracy over 1,000 runs of experiments on both databases is reported. The results are tabulated in Table 3.8. The chosen classifier consistently achieves good performance across many distortions with the minimum accuracy value of 80%. Since the classifier uses the extracted spatial domain features as its input descriptors, the results indicate that the features are not only suitable for quality estimation but also suitable for distortion identification purposes.

Table 3.8: Median classification accuracy

LIVE	JP2K	JPEG	WN	GB	FF	ALL
Accuracy	88.57	97.19	100	96.67	80	91.92
CSIQ	JP2K	JPEG	WN	GB	FF	ALL
Accuracy	86.67	86.67	96.67	86.67	-	88.33

To allow visualisation of the classification performance of PATCH-IQ, Figure 3.13 plots the confusion matrix for each distortion classes in both the LIVE and the CSIQ databases. We can use the confusion matrix to see if PATCH-IQ is confusing two distortion classes. Each column of the matrix represents the instances in the predicted distortion class while each row represents the instances in the actual distortion class. The sum of each row is 1 and the values represent the mean percentage for the 1,000 runs of experiments. Higher value indicates greater confusion.



(a) (b)

Figure 3.13: Mean confusion matrix across 1,000 runs of experiments for distortion classification: (a) LIVE and (b) CSIQ

On the LIVE database, we can see that WN, GB and JPEG were generally well classified by PATCH-IQ and not confused with other distortion. JP2K and FF images were the worst with only 88% of JP2K images and 79% of FF images correctly classified. JP2K and FF were also most confused with each other whereby about 11% of FF images were misclassified as JP2K images and about 3% of JP2K images were predicted as FF images. This is because FF images in the database are essentially JP2K compressed images followed by packet-loss errors [14]. Meanwhile, in the CSIQ database, good classification performance was achieved by PATCH-IQ with less than 4% of the WN images were misclassified. JP2K and GB were the two most confused distortions. In JP2K cases, 12% of the images were misclassified as JPEG

or WN images while another 3% were wrongly predicted as GB images. In GB cases, 10% of the images were misclassified as JPEG images while another 5% were incorrectly predicted as either JP2K or WN images.

3.3.6 Computational complexity

Having a fast computation speed is always desirable especially for applications that require online quality assessment like adaptive coding in video streaming. In this sub-chapter, the processing time to run PATCH-IQ is analysed. There are three major stages that consume most of the processing time: (1) patch and feature extraction; (2) distortion identification; and (3) local quality estimation.

Extracting BIQA features is the most time consuming part of the model framework. This is due to the features being extracted at the patch level rather than at the image level. A higher number of patches will lead to longer extraction time. In addition, the choice of statistical features to be utilised also plays important roles in keeping acceptable processing time. On average, utilising spatial domain features described in sub-chapter 3.2.1 and using the parameter setting as in sub-chapter 3.3.1, PATCH-IQ requires 0.28 second to extract the features in a typical 512×768 image.

Processing time for the distortion identification stage is determined by the I2C distance computation. It depends on the size of the labelled dataset. The dataset size is determined by the number of labelled images and the number of patches within those images. A larger dataset will require longer time to compute the I2C distance between the test patches and their nearest neighbour labelled patches. However, as indicated by the results in Table 3.7, a larger database will lead to better prediction performance. Therefore, there is a clear trade-off between the prediction performance and the I2C distance computation time. Choosing an appropriate dataset size is essential to ensure fast computation while achieving competitive prediction

performance. At 80% ratio, PATCH-IQ requires an additional 0.04 second to perform distortion identification.

Finally, the local quality estimation processing time is directly related to the number of nearest neighbour patches selected for linear regression. Similarly, a higher number of patches will lead to longer quality estimation time. Setting the parameters as described in sub-chapter 3.3.1, an extra 0.05 second is required to perform quality estimation for all test patches. These processing times are achieved using un-optimised MATLAB R2011b code on an 8GB RAM computer with an Intel i5 3.20 GHz processor. Note that the construction time of the labelled dataset is not considered here as it is assumed that it is already available prior to the testing stage.

The average run-time comparison between PATCH-IQ and the competing BIQA models is shown in Table 3.9. BIQI is the fastest but has the worst prediction performance among all the compared models. PATCH-IQ is slower than others except CORNIA. However, given its superior performance, PATCH-IQ can be a better option for IQA applications when real time computation is not a key requirement.

Table 3.9: Average run-time

BIQA model	BIQI	BRISQUE	GMLOG	CORNIA	*PATCH-IQ*
Run-time (s)	0.05	0.10	0.07	2.43	*0.37*

3.4 Chapter Summary

In summary, this chapter introduced a new BIQA model that estimates image quality without the presence of a reference image. The model, PATCH-IQ, is based on a five-stage framework that operates in a spatial domain. In contrast to many previous BIQA models, PATCH-IQ predicts the quality of an image directly from a set of annotated patches using a nearest neighbour method. The approach alleviates the need of any prior regression parameters

training phase. PATCH-IQ also extracts its features at patch level enabling quality prediction to be performed locally, a useful property that is unavailable in most previous BIQA models. The model was tested extensively on two subject-rated image databases. The experimental results demonstrated that the image quality estimates by PATCH-IQ are highly correlated with human perceptual measures of image quality across various kinds of image distortions. PATCH-IQ also has greater performance to all competing BIQA models in quality prediction accuracy and robustness. Note that this performance analysis was only conducted through typical DMOS/MOS based performance metrics. There are extra testing that could be performed to further validate the prediction performance of PATCH-IQ. As described in sub-chapter 2.3.2, PC-based metrics such as classification error or AUC could be utilised to further benchmark PATCH-IQ to other competing models. The results will further strengthen any claim made on PATCH-IQ performance. However, this is outside the scope of the study. In addition, note that there are further steps that could be taken to improve the performance of PATCH-IQ. In the next chapter, simple modifications to its framework are implemented and examined to see if a better prediction performance could be obtained.

Chapter 4

Improving the Patch Based Learning Framework for Blind Image Quality Assessment Model

4.1 Chapter Introduction

Encouraged by the promising results reported in the previous chapter, this chapter will discuss simple modifications made to the initial PATCH-IQ's framework. Two other patch sampling strategies are studied resulting to two modified BIQA models. The first modified model, termed as PATCH-IQ2, investigates the use of an interest points based sampling strategy to extract image patches and their corresponding quality predictive features. The second model, termed as PATCH-IQ3, meanwhile incorporates visual saliency estimation method in its sampling strategy. In sub-chapter 4.2, both sampling strategies will be first described. In sub-chapter 4.3, we will then look at their experimental results and later analyses. Sub-chapter 4.4 will conclude the chapter.

4.2 Image Patch Sampling Strategy

4.2.1 Interest points based sampling strategy

In the previous chapter, PATCH-IQ samples image patches in a non-overlapping way. Although it is relatively straightforward, there exists a possibility that patches containing uniform parts of an image may be extracted. This is especially true for small-sized patches.

These patches are not useful in an IQA task as they have little effect on the evaluation results. To minimise this problem, a sampling strategy based on interest points of an image is considered next.

Interest points of an image are generally referred to points in the image detected to simplify further processing in a vision system. They are normally located at regions of interest, the regions within an image with high information content [123]. The main application of interest points in computer vision and image processing field is to find points / regions in the image domain likely to represent objects. Therefore, they are often employed in processing tasks such as object recognition and image matching. In this study, PATCH-IQ2 tries to extend interest points' application to a BIQA task. It uses interest points of an image to find image regions (patches) that contain significant information on image quality.

It has been shown that when looking at an image, most of the time human focus on object-like regions, i.e. the regions around interest points [124]. In that respect, this study assumes that any distortion applied to those regions will carry greater impact on how human perceived image quality than the distortion in any other image regions such as background. By first finding the location of interest points in an image, patches that contain more relevant information on perceptual image quality can be identified and selected. For this purpose, PATCH-IQ2 utilises an interest point detector to guide its patch sampling process.

A wide variety of interest point detectors exist in the literature such classical edge-based detectors [125]-[126], corner-based detectors [127]-[128] or blob-based detectors [129]-[130]. Edge or corner-based approaches are common choice for interest point detection when dealing with images of same scale and orientation. However, when we have images of different scales and rotations, blob-based interest point detectors are preferred. The Scale Invariant Feature Transform (SIFT) algorithm [131] which is developed based on blob detection approach is

utilised here to perform interest points detection for PATCH-IQ2. SIFT is chosen due to its ability to detect local interest points that are stable and invariant to both image scales and orientations [132].

The operation of SIFT is briefly described here. SIFT takes an image and transforms it into a large collection of local feature vectors containing descriptors that are useful to identify objects in an image. There are 4 stages involved in SIFT: 1) Scale-space extrema detection; 2) Keypoint localisation; 3) Orientation assignment; and 4) Keypoint descriptor. The first two stages aim at identifying the locations of stable keypoints at which image features / descriptors will be extracted. The third stage assigns consistent orientation to these keypoints based on local image properties while the last stage uses local gradient information to create the descriptors. Interested readers are referred to [131] for further details.

In IQA, the resulting SIFT descriptors may not be useful in estimating image quality. PATCH-IQ2, however, does not require the use of SIFT descriptors. Instead, it only utilises the first two stages of SIFT to help find the locations at which patches will be extracted. Based on the above assumption that the regions surrounding the keypoints contain greater information on image quality, PATCH-IQ2 then samples patches of $h_{\mathrm{P}} \times w_{\mathrm{P}}$ size using the provided keypoints' coordinates as centres. One may argue that an image affected by distortion can give lots of false keypoints as edges lose sharpness. These false keypoints obviously are not useful for object recognition or detection purposes. For quality assessment, these keypoints are still useful since, usually, the whole image is distorted. The extracted image patches still carry information on image quality. An example of this process is shown in Figure 4.1. Note that PATCH-IQ2 only extracts patches at the identified keypoint locations. If there is no keypoint detected at any particular image area, no patch is extracted at that area.

Figure 4.1: Patch extraction using interest point sampling strategy

PATCH-IQ2 extracts the similar spatial domain features as in Chapter 3 from the sampled patches. However, instead of using all the extracted features to construct the labelled dataset, PATCH-IQ2 only select features from P_{label} patches in each image. This is done to ensure all images contribute the same number of features to the dataset and to reduce the computational demands of the framework. The selected features are then combined over all images to form the dataset. Figure 4.2 shows an example of a dataset built from the distorted versions of an image. Denote the total number of labelled images by N_{label}, the size of feature matrix for the dataset is:

$$\mathbf{D} = [N_{\text{label}} P_{\text{label}} \times 36] \,. \tag{4.1}$$

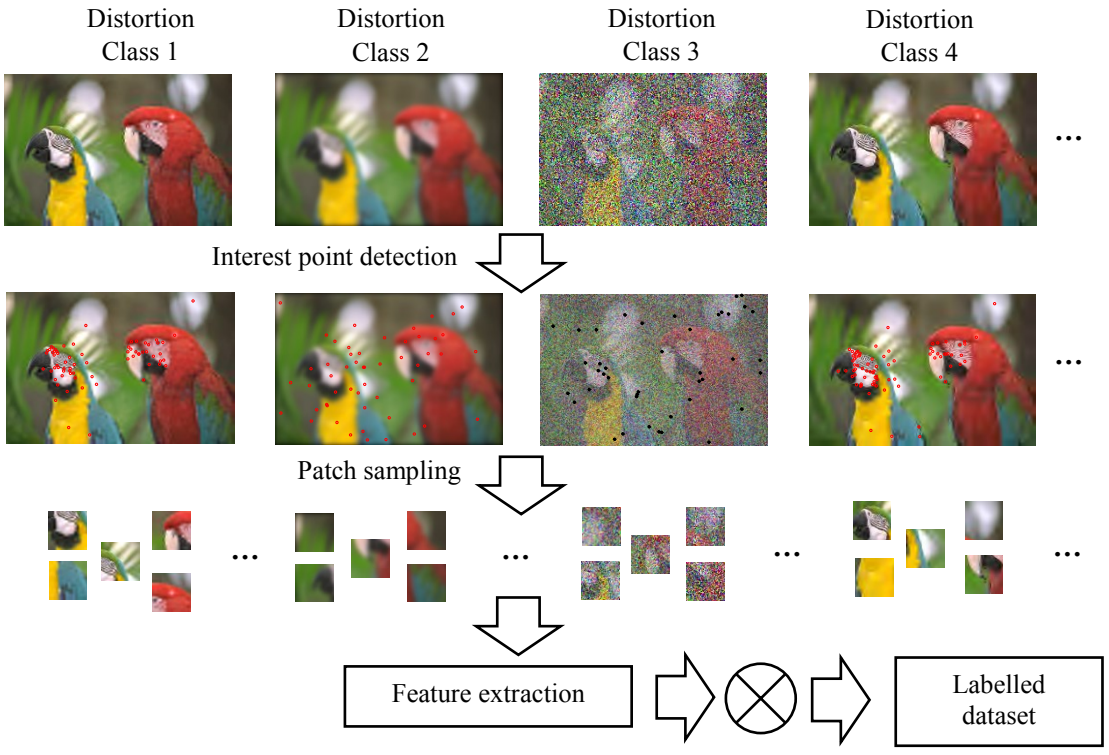


Figure 4.2: Example of labelled dataset

The remaining components of PATCH-IQ2 are unchanged from the initial framework. Given a test image $\mathbf{I}_{\text{test}}$, PATCH-IQ2 first extracts BIQA features at the identified interest points' locations. To speed-up the computation, only features from P_{test} patches are chosen. PATCH-IQ2 then employs the same NBNN classifier to perform distortion identification and the same *k*-NN regression method to predict the local quality scores. Similar inverse distance weighted pooling method as in Chapter 3 is then used to produce the global score for the image.

4.2.2 Visual saliency based sampling strategy

The second modification considered by the study is to incorporate visual saliency computation into the initial PATCH-IQ sampling strategy. In general, visual saliency is the perceptual quality that makes an object, person, or pixel stand out relative to its neighbours and thus capture human attention [133]. As human visual attention is attracted to distinctive salient features, more importance should be given to the associated regions in the image. Here, the same assumption as in sub-chapter 4.2.1 is utilised. The study assumes that any distortion applied to more salient regions will carry greater impact on how human perceived image quality than the distortion in less salient image regions. By first finding locations of higher saliency image regions, patches that contain more relevant quality information can be sampled. For this purpose, PATCH-IQ3 employs saliency detection methods to guide its sampling.

There are many saliency detection methods available in the literature that can be broadly classified as biological based, computational based, or a combination of both. All methods generally employ a low-level approach by determining contrast of image regions relative to their surroundings, using one or more features of intensity, colour, and orientation [93]. Interested readers are referred to publications in [134] – [135] for more comprehensive survey of visual saliency detection methods.

In this study, PATCH-IQ3 performs its saliency detection by adopting the spectral residual model presented in [136]. The use of the model is motivated by its simple implementation, fast computation and good detection performance in the presence of high level distortion. Specifically, given a test image $\mathbf{I}_{\boldsymbol{t}\mathrm{est}}(x)$, PATCH-IQ3 first compute the log spectrum representation of the image [136]:

$$\mathcal{L}(f) = \log A(f) \text{ with } A(f) = \mathcal{F}(\mathbf{I}_{\mathrm{test}}(x)) \tag{4.2}$$

where $A(f)$ represent the general shape of log spectra of the image and $\mathcal{F}$ denote the Fourier Transform. The spectral residual of the image is next computed [136]:

$$R(f) = L(f) - h_n(f) * L(f) \tag{4.3}$$

where $h_n(f)$ is the local average filter to approximate the shape of $A(f)$. Finally, the spectral residual is transformed using inverse Fourier Transform into spatial domain to construct the saliency map of the image [136]:

$$\mathbf{S}(x) = \mathcal{F}^{-1}[\exp(R(f) + P(f))]^2 \tag{4.4}$$

where $P(f)$ is the phase spectrum of the image.

The test image's saliency map is used by PATCH-IQ3 to guide its patch sampling process. PATCH-IQ3 chooses the patches with high mean visual saliency values since patches with small visual saliency values play little role in human perception of the image quality. Here, PATCH-IQ3 randomly sample *P* patches of which their mean visual saliency values are bigger than a threshold T. To speed-up the computation, PATCH-IQ3 follow the same implementation as in the interest point based strategy whereby PATCH-IQ3 extract features from P_{test} patches rather than the whole sampled patches . PATCH-IQ3 then employs the same NBNN classifier to perform distortion identification and the same *k*-NN regression method to predict the local quality scores. Similar inverse distance weighted pooling method as in Chapter

3 is then used to produce the global score for the image. Figure 4.3 shows an example of patch extraction based on visual saliency sampling strategy.

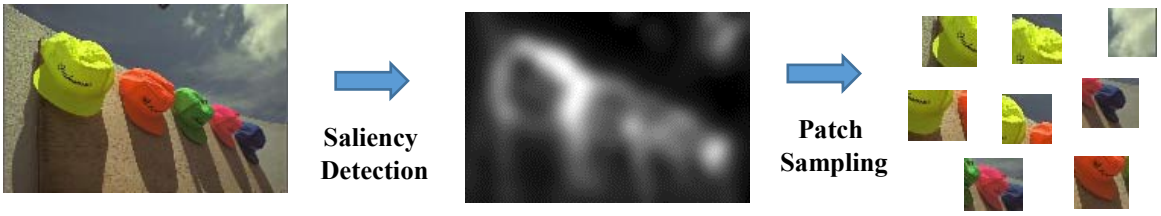


Figure 4.3: Patch extraction using saliency detection sampling strategy

4.3 Results and Discussions

4.3.1 Experimental setup and evaluation protocol

Databases: Besides the LIVE and the CSIQ databases, PATCH-IQ2 and PATCH-IQ3 were also tested on another database: the LIVEMD database [24], [25]. The LIVE and CSIQ databases contain only images distorted by a single type of artefact typically found in image communication systems such as noise, blur or compression artefacts. The LIVEMD database also provides examples of images affected by multiple types of distortions. In the database, 15 reference images are first blurred at 4 levels. The blurred images are then subjected to two types of artefact, JPEG and WN, at 4 levels each. In all, 225 single / multiple distorted images are generated for each of the two cases: GBJPEG and GBWN. Similar to the LIVE database, each image in the LIVEMD database is provided with DMOS value in the range between 0 and 100 whereby a lower DMOS value indicates a higher quality image.

Framework parameters: For both PATCH-IQ2 and PATCh-IQ3, the number of patches for each labelled image P_{label} and the number of test image patches P_{test} are empirically set at 30 and 100 respectively while the patch size $h_{\text{P}} = w_{\text{P}}$ is 256. Meanwhile, the saliency threshold T for PATCH-IQ3 is empirically set at 0.1. Other parameters remain unchanged.

Performance metrics and benchmarked models: The same three performance metrics used in the previous chapter are again employed here to measure the correlation between the predicted scores and the human subjective scores while PATCH-IQ2 and PATCH-IQ3 performances are compared with the previous four BIQA models and the initial PATCH-IQ model.

4.3.2 Evaluation on single distortion databases

The median results across the 1,000 trials for both the overall experiment and the DS experiment are tabulated in Tables 4.1 and 4.2, respectively where the top FR-IQA and BIQA models are in bold. For the overall performance experiment, PATCH-IQ3 obtained the highest values for the three performance metrics among the competing BIQA models when tested on the LIVE database. However, when tested on the CSIQ database, PATCH-IQ2 obtained the highest values. Therefore, there is no clear indication to which sampling strategy is more superior for the overall performance experiment. Both PATCH-IQ2 and PATCH-IQ3 increased the correlation values of the initial PATCH-IQ indicating that both the interest points based and saliency based sampling strategies improve the framework prediction performance.

Table 4.1: Median values across 1,000 runs of the overall performance experiment

IQA model	LIVE			CSIQ		
	LCC	SROCC	RMSE	LCC	SROCC	RMSE
PSNR	0.882	0.883	12.898	0.856	0.929	0.144
SSIM	0.946	0.949	8.804	0.935	0.936	0.099
FSIM	**0.961**	**0.964**	**7.546**	**0.968**	**0.963**	**0.071**
BIQI	0.849	0.844	15.407	0.809	0.749	0.187
BRISQUE	0.943	0.942	9.395	0.930	0.910	0.107
GMLOG	0.951	0.950	8.829	0.939	0.925	0.010
CORNIA	0.939	0.942	9.920	0.911	0.887	0.125
PATCH-IQ	0.954	0.952	8.476	0.946	0.932	0.094
PATCH-IQ2	*0.956*	*0.954*	*8.149*	***0.959***	***0.943***	***0.081***
PATCH-IQ3	***0.958***	***0.956***	***7.962***	*0.949*	*0.934*	*0.089*

For the DS performance experiment, PATCH-IQ2 produced the best SROCC values for the noisy or blurred images on both databases. It also performed the best for images affected

by the JP2K compression artefacts when tested on the CSIQ database. In the JPEG cases, PATCH-IQ2 had slightly lower SROCC value than PATCH-IQ on the LIVE database but it had a better value on the CSIQ database. For PATCH-IQ3, it improved the initial PATCH-IQ's SROCC values on JP2K and GB images. However, when tested on the WN and JPEG images, lower SROCC values were obtained. Meanwhile, direct comparison between PATCH-IQ2 and PATCH-IQ3 showed that PATCH-IQ2 had better SROCC values than PATCH-IQ3 for JPEG, WN and GB images on both database. In JP2K cases, PATCH-IQ2 had better SROCC value on the CSIQ database while PATCH-IQ3 was better on the LIVE database. These observations indicate that the SIFT interest points based sampling strategy is better than the SR visual saliency based sampling strategy for individual distortion cases.

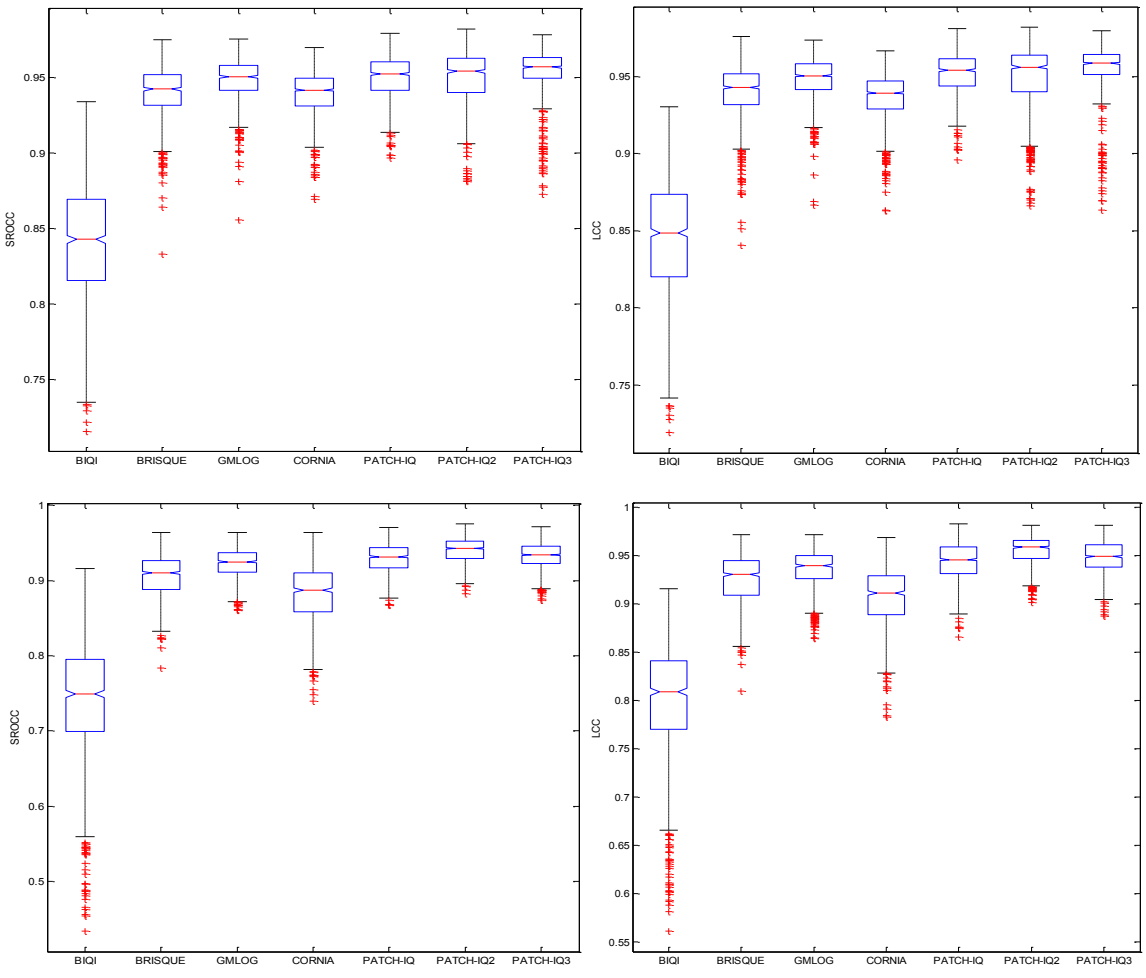
Table 4.2: Median SROCC values across 1,000 runs of the DS performance experiment

IQA model	LIVE					CSIQ			
	JP2K	JPEG	WN	GB	FF	JP2K	JPEG	WN	GB
PSNR	0.895	0.881	**0.985**	0.782	0.891	0.936	0.888	**0.936**	0.929
SSIM	0.961	0.976	0.969	0.952	**0.956**	0.961	0.955	0.897	0.961
FSIM	**0.972**	**0.984**	0.972	**0.971**	0.952	**0.970**	**0.966**	**0.936**	**0.973**
BIQI	0.830	0.906	0.933	0.866	0.689	0.764	0.910	0.540	0.783
BRISQUE	0.916	0.964	0.979	0.945	0.887	0.898	0.921	0.921	0.919
GMLOG	0.927	0.963	0.983	0.929	0.901	0.916	0.936	0.941	0.908
CORNIA	0.921	0.936	0.961	0.952	**0.905**	0.894	0.882	0.786	0.904
PATCHIQ	0.931	**0.976**	**0.987**	0.953	0.891	0.918	0.952	0.963	0.916
PATCHIQ2	*0.933*	*0.973*	***0.987***	***0.970***	*0.882*	***0.933***	***0.953***	***0.965***	***0.943***
PATCHIQ3	***0.935***	*0.970*	*0.983*	*0.966*	*0.903*	*0.922*	*0.950*	*0.959*	*0.918*

Similar patterns can be observed as in the previous chapter regarding the FR-IQA models. Both PATCH-IQ2 and PATCH-IQ3 achieved better overall performance than PSNR and SSIM while approaching FSIM. PATCH-IQ2 and PATCH-IQ3 also produced comparable SROCC values to SSIM and FSIM in most of individual distortion cases. In fact, PATCH-IQ2 also had better SROCC values than the three FR-IQA models for WN images.

Table 4.3 reports the IQR results of the 1,000 SROCC and LCC values for the competing BIQA models. The associated box plots are shown in Figure 4.4. PATCH-IQ2 produced higher

IQR values than the other models except for BIQI on the LIVE database. However, it obtained the lowest IQR values on the CSIQ database. For PATCH-IQ3, it obtained the lowest IQR values on the LIVE database and the second lowest on the CSIQ database. In terms of the outliers, both PATCH-IQ2 and PATCH-IQ3 had better set of outliers than PATCH-IQ on the CSIQ database while the reverse was true on the LIVE database. We can also see that PATCH-IQ2 had more compact set of outliers than PATCH-IQ3, thus indicating that the SIFT interest points based sampling strategy is more robust to the tested images' variations than the SR visual saliency based sampling strategy.



(a) (b)

Figure 4.4: Box plots of performance metric distributions of BIQA models from 1,000 runs of experiments on the LIVE database (top row) and the CSIQ database (bottom row): (a) SROCC and (b) LCC

Table 4.3: IQR values for 1,000 SROCC and LCC values obtained in both databases

Database	LIVE		CSIQ	
Metrics	LCC	SROCC	LCC	SROCC
BIQI	0.053	0.054	0.071	0.096
BRISQUE	0.020	0.020	0.036	0.039
GMLOG	0.017	0.017	0.024	0.026
CORNIA	0.018	0.018	0.041	0.052
PATCH-IQ	0.018	0.019	0.028	0.027
PATCH-IQ2	*0.021*	*0.022*	*0.019*	*0.023*
PATCH_IQ3	*0.013*	*0.014*	*0.023*	*0.024*

4.3.3 Evaluation on multiple distortion database

To further investigate the effectiveness of the proposed framework, all the competing BIQA models are tested on the LIVEMD database. The database is more challenging as it also contains images that underwent multiple distortions. The results are presented in Table 4.4. The first five columns show the results from the DS performance experiment while the last column represents the results from the overall performance experiment. The top models are in bold. The results suggest that both PATCH-IQ and PATCH-IQ2 generally had good prediction performance for the overall performance experiment where they consistently produced the top LCC, SROCC and RMSE values. However, the same cannot be said for PATCH-IQ3. While it obtained top three LCC value, its SROCC and RMSE values are poorer than some other competing models.

In the DS performance experiment, PATCH-IQ2 performed the best among the three proposed models for singly distorted images followed by PATCH-IQ and PATCH-IQ3. It obtained the best LCC, SROCC and RMSE values in GB cases while achieved comparable performance to PATCH-IQ in both JPEG and WN cases. In multiple distortions cases, both PATCH-IQ2 and PATCH-IQ were among the top three BIQA models for images distorted by GB and WN. For GBJPEG images, PATCH-IQ and PATCH-IQ2 produced the top two SROCC values and gave comparable LCC and RMSE values to other BIQA models. The obtained results also show that PATCH-IQ3 does not produced superior performance to other models.

Table 4.4: Median values across 1,000 iterations on the LIVEMD database

LCC						
	GBJPEG	GBWN	GB	JPEG	WN	ALL
BIQI	0.742	0.129	0.863	0.101	0.543	0.331
BRISQUE	**0.831**	0.836	0.893	0.629	0.935	0.919
GMLOG	0.812	0.780	0.771	0.674	0.845	0.869
CORNIA	0.825	**0.866**	0.854	0.530	0.803	0.913
PATCH-IQ	0.825	0.861	0.892	0.715	**0.948**	**0.931**
PATCH-IQ2	*0.821*	*0.855*	***0.895***	***0.726***	***0.948***	***0.931***
PATCH-IQ3	*0.791*	*0.837*	*0.868*	*0.683*	*0.904*	*0.921*
SROCC						
	GBJPEG	GBWN	GB	JPEG	WN	ALL
BIQI	0.752	0.062	0.859	0.083	0.551	0.357
BRISQUE	0.817	0.833	0.883	0.667	0.881	0.900
GMLOG	0.811	0.762	0.776	0.667	0.800	0.845
CORNIA	0.809	0.855	0.835	0.483	0.767	0.902
PATCH-IQ	**0.829**	0.861	0.876	**0.733**	0.867	**0.911**
PATCH-IQ2	*0.824*	***0.864***	***0.884***	*0.717*	***0.883***	***0.911***
PATCH_IQ3	*0.782*	*0.837*	*0.836*	*0.717*	*0.833*	*0.896*
RMSE						
	GBJPEG	GBWN	GB	JPEG	WN	ALL
BIQI	8.877	44.346	9.440	10.356	12.731	25.800
BRISQUE	7.999	8.482	8.719	7.280	6.338	8.428
GMLOG	8.356	9.973	12.436	7.469	9.771	10.220
CORNIA	**7.810**	**8.026**	10.117	8.178	9.353	8.681
PATCH-IQ	8.257	8.333	9.047	5.719	**5.719**	8.179
PATCH-IQ2	*8.260*	*8.409*	***8.551***	***5.741***	*5.837*	***8.143***
PATCH-IQ3	*8.593*	*8.496*	*9.379*	*5.879*	*7.305*	*8.556*

4.3.4 Statistical significance and hypothesis testing

The differences in median correlations between the competing BIQA models may not be statistically significant. Therefore, a hypothesis test to evaluate the statistical significance difference between each model is conducted. Similar to the hypothesis testing in Chapter 3, the Wilcoxon rank-sum test is employed to avoid the normality assumption required by a typical t-test. The test evaluates the median values equivalency between two independent samples. Here, the two samples are the 1,000 SROCC values obtained from a pair of BIQA models. The test was conducted by setting the significance level at 0.01 with the null hypothesis is that the SROCC values of the two models are drawn from the populations with equal median. The alternative hypothesis is that the median of one model is greater than that of the other. The results of the test are tabulated in Table 4.5.

Table 4.5: Results of the Wilcoxon rank-sum test using the SROCC values of competing BIQA models

LIVE							
	BIQI	BRISQUE	GMLOG	CORNIA	PATCH-IQ	*PATCH-IQ2*	*PATCH-IQ3*
BIQI	0	-1	-1	-1	-1	-1	-1
BRISQUE	1	0	-1	1	-1	-1	-1
GMLOG	1	1	0	1	-1	-1	-1
CORNIA	1	-1	-1	0	-1	-1	-1
PATCH-IQ	1	1	1	1	0	0	-1
PATCH-IQ2	*1*	*1*	*1*	*1*	*0*	*0*	*0*
PATCH-IQ3	*1*	*1*	*1*	*1*	*1*	*0*	*0*
CSIQ							
	BIQI	BRISQUE	GMLOG	CORNIA	PATCH-IQ	*PATCH-IQ2*	*PATCH-IQ3*
BIQI	0	-1	-1	-1	-1	-1	-1
BRISQUE	1	0	-1	1	-1	-1	-1
GMLOG	1	1	0	1	-1	-1	-1
CORNIA	1	-1	-1	0	-1	-1	-1
PATCH-IQ	1	1	1	1	0	-1	0
PATCH-IQ2	*1*	*1*	*1*	*1*	*1*	*0*	*1*
PATCH-IQ3	*1*	*1*	*1*	*1*	*0*	*-1*	*0*
LIVEMD							
	BIQI	BRISQUE	GMLOG	CORNIA	PATCH-IQ	*PATCH-IQ2*	*PATCH-IQ3*
BIQI	0	-1	-1	-1	-1	-1	-1
BRISQUE	1	0	1	0	1	-1	1
GMLOG	1	-1	0	-1	0	-1	-1
CORNIA	1	0	1	0	1	-1	1
PATCH-IQ	1	-1	0	-1	0	0	1
PATCH-IQ2	*1*	*1*	*1*	*1*	*0*	*0*	*1*
PATCH-IQ3	*1*	*-1*	*1*	*-1*	*-1*	*-1*	*0*

There is no significant differences between PATCH-IQ and PATCH-IQ2 median SROCC values on both the LIVE and the LIVEMD database. However, the median value difference between the two models are statistically significant on the CSIQ database with PATCH-IQ2 producing higher median SROCC value. Meanwhile, PATCH-IQ3 is different to PATCH-IQ on both the LIVE (PATCH-IQ3 producing higher median SROCC value) and the LIVEMD (PATCH-IQ3 producing lower median SROCC value) databases but no significant

differences observed between the two models on the CSIQ database. In addition, direct comparison between PATCH-IQ2 and PATCH-IQ3 show that the differences in the median SROCC values are not statistically significant when the models are tested on the LIVE database. However, on the CSIQ and the LIVEMD databases, the differences in the median SROCC values between the two models are statistically significant. With higher median SROCC value represents higher prediction performance, these observations indicate that the higher performance achieved by PATCH-IQ2 over PATCH-IQ3 is statistically significant.

4.4 Further Analysis on PATCH-IQ2

The results' analyses from sub-chapter 4.3.2 to sub-chapter 4.3.4 indicate that PATCH-IQ2 generally has better prediction performance than PATCH-IQ3 when tested using the chosen databases and performance metrics. In order not to burst the chapter content, the remaining performance analysis are therefore limited to PATCH-IQ2.

4.4.1 Influence of framework parameters

Several parameters in the PATCH-IQ2 framework can be varied: 1) The number of labelled images; 2) the number of patches in each labelled image; and 3) the number of NN patches selected for linear regression. In this sub-section, the study investigates how these parameters affect the performance of the framework.

To analyse the changes in prediction performance when the number of labelled images is varied, similar procedure as in sub-chapter 3.3.4 was performed. The databases were partitioned under three training (labelled) - test ratios: 80:20, 50:50 and 30:70. The numbers of selected labelled and test patches and the patch size were fixed as before. The SROCC results for the overall performance experiment are shown in Table 4.6 and Figure 4.5.

Table 4.6: SROCC comparison for different training (labelled) samples ratio

Database	LIVE			CSIQ			LIVEMD		
Ratio	80%	50%	30%	80%	50%	30%	80%	50%	30%
BIQI	0.844	0.835	0.816	0.749	0.737	0.718	0.357	0.342	0.322
BRISQUE	0.942	0.927	0.903	0.910	0.895	0.872	0.900	0.892	0.883
GMLOG	0.950	0.940	0.925	0.925	0.909	0.887	0.845	0.812	0.776
CORNIA	0.942	0.937	0.929	0.887	0.881	0.873	0.902	0.898	0.893
PATCH-IQ	0.952	0.945	0.933	0.931	0.920	0.907	**0.911**	**0.908**	0.893
PATCH-IQ2	***0.954***	***0.947***	***0.935***	***0.943***	***0.932***	***0.915***	***0.911***	*0.906*	***0.895***

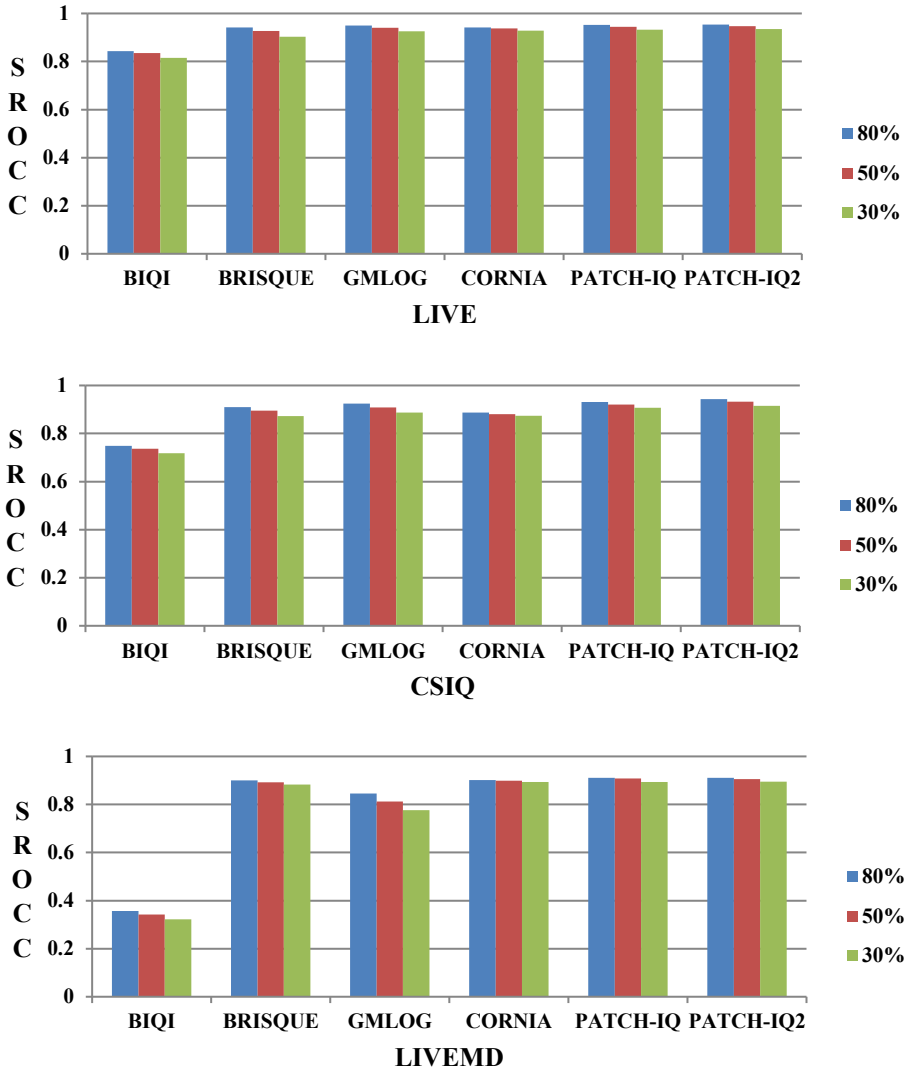


Figure 4.5: SROCC comparison for different training (labelled) sample ratio for different databases

PATCH-IQ2 had the best SROCC among the competing BIQA models for both the LIVE database and the CSIQ database. It performed well in CSIQ database whereby there was

significant increase in the produced SROCC values compared to the initial PATCH-IQ. However, when tested in the LIVEMD database, there was no significant differences between PATCH-IQ and PATCH-IQ2 with similar SROCC values were obtained at 80% labelled ratio. Compared to the remaining four BIQA models, PATCH-IQ2 consistently produced higher SROCC values across three databases at different training ratios. These observations follow the initial finding in the previous chapter that the framework has better robustness to the number of training samples and can work better where the number of training images is small.

Meanwhile, the results of varying the number of patches in each labelled image on the LIVE database at 80% training ratio are shown in Table 4.7 and Figure 4.6, respectively. A higher number of utilised patches will lead to higher SROCC and LCC values. However, it will lead to longer computation time for the identification of the distortion. Here, PATCH-IQ2 chooses the lowest number of patches that outperforms the state-of-the-art BIQA models for its framework while has acceptable processing time.

Table 4.7: LCC and SROCC comparison for different number of patches in a labelled image

Patch	10	20	30	40	50	75	100	150	200
LCC	0.948	0.951	***0.956***	0.956	0.960	0.963	0.959	0.963	0.963
SROCC	0.947	0.950	***0.954***	0.954	0.957	0.961	0.957	0.962	0.962

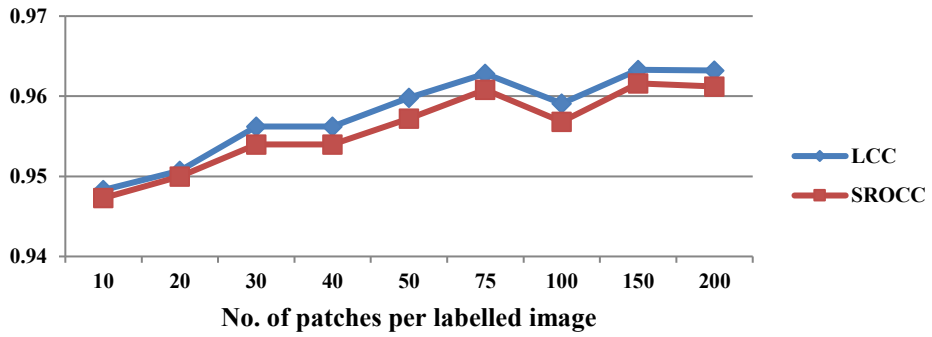


Figure 4.6: LCC and SROCC comparison for different number of patches in a labelled image on LIVE database

Next, the effect of the number of the nearest neighbour patches used for linear regression on the model performance is investigated. All other parameters were again fixed at the initial values. The performance variation of PATCH-IQ2 when tested on the LIVE database is shown in Table 4.8. Based on the results, there was a small variation on the obtained values indicating that the effect of the number of labelled patches is insignificant. The number that provides the optimum performance was empirically chosen. Here, the optimum performance was achieved when the number is set at 1,000.

Table 4.8: Performance variations for different numbers of NN patches used in regression

Patch	5	10	50	100	500	1,000	2,000	3,000	ALL
LCC	0.945	0.949	0.928	0.950	0.953	***0.956***	0.953	0.950	0.945
SROCC	0.942	0.946	0.934	0.949	0.952	***0.954***	0.951	0.949	0.944
RMSE	9.114	8.820	10.346	8.631	8.435	***8.149***	8.440	8.732	9.098

PATCH-IQ2's prediction performance also depends on how the scores from test patches are pooled. In this study, PATCH-IQ2 pools all the patches' scores by assigning weight to each score according to an inverse weighting rule. To justify this pooling approach, different pooling methods were also implemented on the model. Two other pooling methods: average pooling and max pooling were tested and the results from the LIVE and the CSIQ databases are shown in Table 4.9 and Figure 4.7. Among the three pooling methods, it can be seen that pooling method based on inverse weighting rule consistently produced the highest SROCC, LCC and RMSE values. It provided slight improvement to average pooling while better than max pooling.

Table 4.9: Performance comparison for different pooling methods

Database	LIVE			CSIQ		
Metrics	LCC	SROCC	RMSE	LCC	SROCC	RMSE
IW Rule	0.956	0.954	8.149	0.959	0.943	0.081
Average	0.954	0.951	8.481	0.957	0.933	0.085
Max	0.865	0.872	19.467	0.900	0.884	0.179

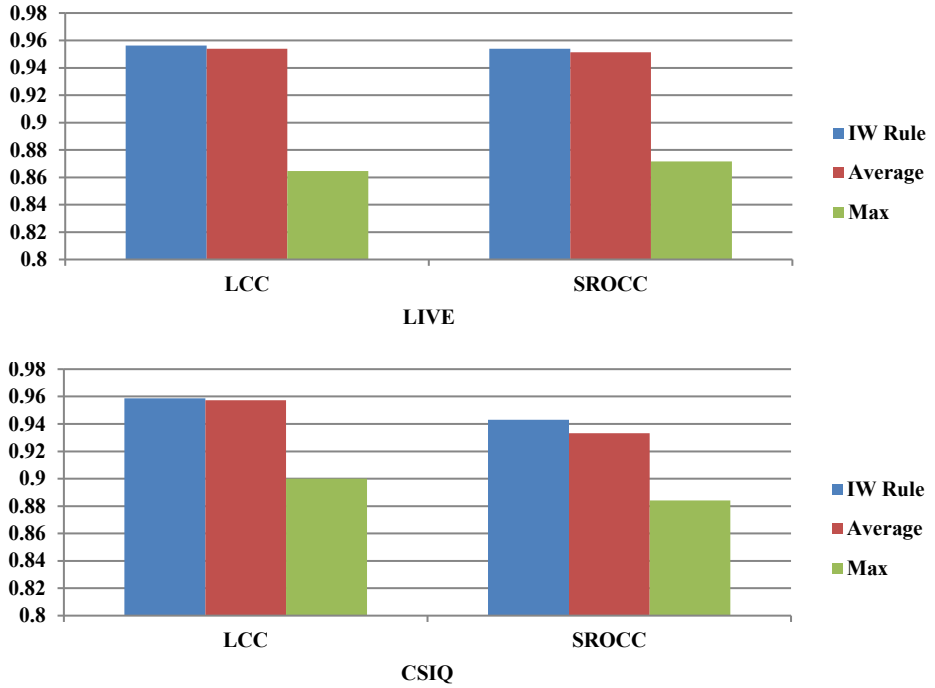


Figure 4.7: LCC and SROCC comparison for different pooling methods on LIVE and CSIQ

4.4.2 Distortion identification accuracy

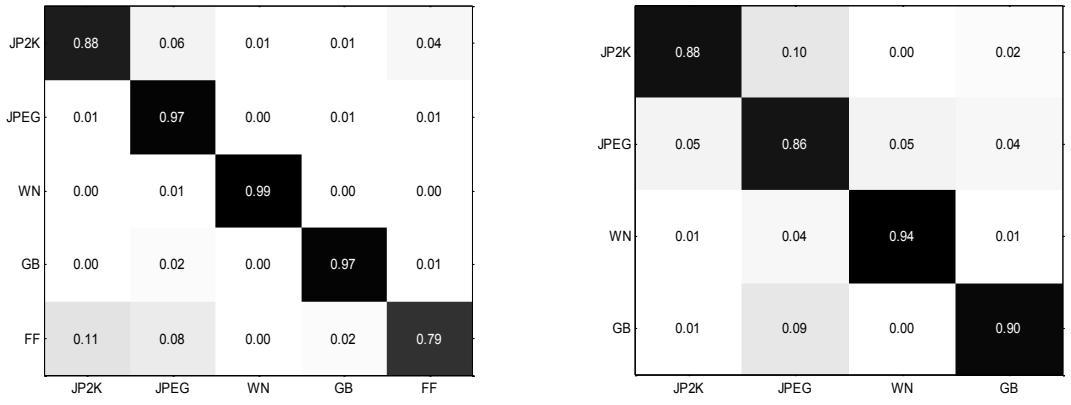
To investigate the effect of different sampling strategy to the framework's capability to perform distortion identification, the median classification accuracy over 1,000 trials of both PATCH-IQ and PATCH-IQ2 on the three databases is calculated. The results are reported in Table 4.10. Using interest points based sampling strategy contributed to small increases to the overall classification accuracy of the framework. While there were no significant increases on the LIVE and the LIVEMD databases, the increase was noticeable on the CSIQ database.

These classification observations together with the prediction performance results in sub-chapter 4.3.2 indicate that an increase in distortion classification accuracy improves the quality estimation prediction performance. This agrees with the insight obtained by this study that the quality of an image would be best predicted by images of the same distortion type. The results also indicate the suitability of the NBNN classifier in performing distortion identification

where it achieved good classification performance on different distortion across three databases.

Table 4.10: Median classification accuracy

LIVE						
	JP2K	JPEG	WN	GB	FF	ALL
PATCH-IQ	88.57	97.19	100	96.67	80	91.92
PATCH-IQ2	88.57	97.22	100	96.67	80	91.98
CSIQ						
	JP2K	JPEG	WN	GB	FF	ALL
PATCH-IQ	86.67	86.67	96.67	86.67	-	88.33
PATCH-IQ2	90	86.67	93.33	90	-	89.17
LIVEMD						
	GBJEG	GBWN	GB	JPEG	WN	ALL
PATCH-IQ	100	99.98	99.99	93.77	91.97	98.56
PATCH-IQ2	100	100	99.99	93.37	92.63	98.60



(a) (b)

Figure 4.8: Mean confusion matrix across 1,000 runs of experiments for distortion classification: (a) LIVE and (b) CSIQ

Meanwhile, the confusion matrices for each distortion classes in the LIVE and the CSIQ databases for PATCH-IQ2 are plotted in Figure 4.8. In the LIVE database, we can see that WN, GB and JPEG images were generally well classified by PATCH-IQ2 and not confused with other distortion. JP2K and FF images were most confused with each other whereby about 11% of FF images were misclassified as JP2K images and about 4% of JP2K images were predicted as FF images. Meanwhile, in the CSIQ database, good classification performance was achieved

by PATCH-IQ2 with less than 6% of the WN images were misclassified. It also achieved good performance for both GB and JP2K cases with 90% of the blurred images and 88% of the JP2K compressed images were correctly classified. JPEG was the most confused distortion class with 10% of the images were misclassified as JP2K or WN images while another 4% were wrongly predicted as GB images.

4.4.3 Feature analysis

To evaluate the contributions of the utilised features on both the distortion classification and the quality prediction performances, we can re-use the plot of the SROCC values between the features derived from the LIVE images and their corresponding DMOS (Figure 3.7). For the ease of reading, the plot is shown again below as Figure 4.9. In each distortion case, we can observe that the variance parameters of both the GGD model and the AGGD model have better correlation with subjective scores compared to the shape parameters of the models. Meanwhile, among all the utilised features, the mean parameters of the AGGD models capture quality information the least. Another observation we can made is the same features extracted in different orientations generally have similar correlation values pattern.

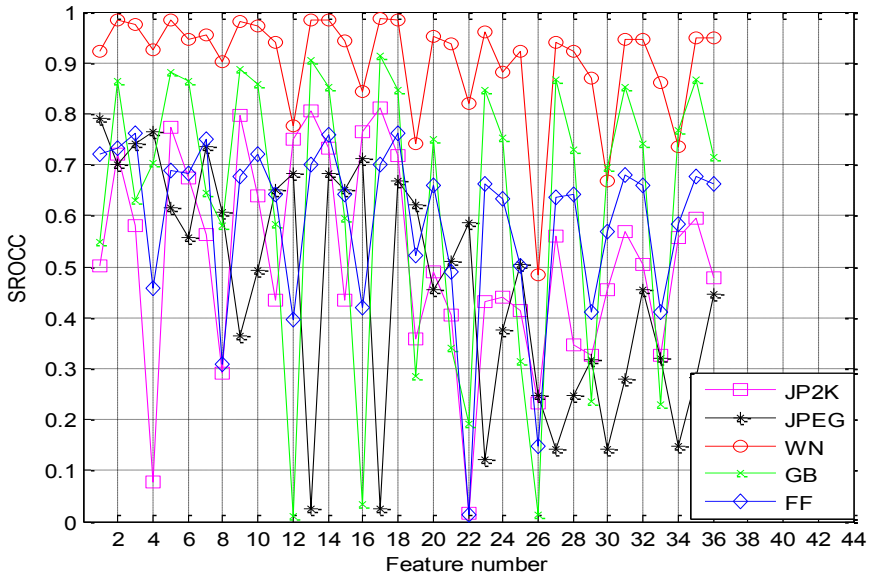


Figure 4.9: Correlation of the extracted features with the DMOS for different distorted images in the LIVE database

Five different combinations of features were tested on the LIVE database to see how they affect the performance of PATCH-IQ2. They were: 1) All features (denoted as PATCH-IQ2), 2) The GGD model-based features (denoted as PATCH-IQ2b), 3) The AGGD model-based features (denoted as PATCH-IQ2c), 4) All features except the mean parameter of the AGGD models (PATCH-IQ2d) and 5) The variance parameters of both the GGD model and the AGGD model (PATCH-IQ2e). PATCH-IQ2b should study the contribution of features derived directly from the locally normalised luminance coefficients, whereas PATCH-IQ2c was to evaluate the effects of features derived from the pairwise products of these coefficients. Meanwhile, features for PATCH-IQ2d and PATCH-IQ2e were selected based on the previous observations.

The median classification accuracy values over 1,000 runs of experiments for the five PATCH-IQ2 versions are tabulated in Table 4.11. From the table, we can see that the best classification results for both the overall and the DS experiments were achieved when all 36 features were utilised. PATCH-IQ2c had better classification accuracy than PATCH-IQ2b in both experiments, showing that the AGGD model-based features contribute more to a distortion identification task than the GGD model-based features. We can also observe that removing the mean parameters of the AGGD models as in PATCH-IQ2d had little effect to the classification performance. This indicates that the mean parameters of the AGGD models have small contributions to such a task. The classification accuracy also dropped when only variance parameters were utilised as PATCH-IQ2e's features.

Table 4.11: Median classification accuracy values for different group of features on the LIVE database

	PATCH-IQ2	PATCH-IQ2b	PATCH-IQ2c	PATCH-IQ2d	PATCH-IQ2e
JP2K	**88.57**	82.35	**88.57**	88.24	79.42
JPEG	**97.22**	88.57	**97.22**	96.92	94.29
WN	**100**	96.67	**100**	**100**	**100**
GB	**96.67**	**96.67**	**96.67**	**96.67**	93.33
FF	**80**	66.67	79.42	**80**	66.67
ALL	**91.98**	85.80	91.93	91.82	85.80

Meanwhile, Table 4.12 shows the median SROCC values over 1,000 trials obtained by the same five PATCH-IQ2 versions. Few similar observations can be made here. First, the best quality prediction performances for both experiments were produced when PATCH-IQ2 utilised all the proposed features. Second, PATCH-IQ2c had better correlation values in most distortion cases than PATCH-IQ2b. This indicates that the AGGD model-based features have better correlation to human perceptual measures than the GGD model-based features. Third, PATCH-IQ2d achieved similar prediction performances to PATCH-IQ2 for images affected by noise and compression artefacts while only suffered a slight degradation in performance for GB and FF images. In agreement to the above discussion, this shows that the mean parameters of the AGGD models contribute little to a quality prediction task. Meanwhile, PATCH-IQ2e also achieved close prediction performance to PATCH-IQ in both experiments. This suggests that, while the variance parameters of both the GGD model and the AGGD models may not be suitable features for a distortion classification task, they are still useful for image quality prediction task.

Table 4.12: Median SROCC values for different group of features on the LIVE database

	PATCH-IQ2	PATCH-IQ2b	PATCH-IQ2c	PATCH-IQ2d	PATCH-IQ2e
JP2K	**0.933**	0.909	0.924	**0.933**	0.912
JPEG	**0.973**	0.959	**0.973**	**0.973**	0.972
WN	**0.987**	0.967	**0.987**	**0.987**	**0.987**
GB	**0.970**	0.941	0.967	0.967	0.969
FF	**0.882**	0.867	0.866	0.873	0.873
ALL	**0.954**	0.932	0.948	0.953	0.947

4.4.4 Computational complexity

Similar to previous discussions in sub-chapter 3.3.6, the run-time of PATCH-IQ2 was determined by three major stages: 1) feature extraction; 2) distortion identification and 3) local quality estimation. At the feature extraction stage, PATCH-IQ2 required longer computation time than PATCH-IQ. This is due to PATCH-IQ2 having to first detect the locations of an image's interest points. On average, PATCH-IQ2 required 0.46 second to extract its features

on a typical 512×768 image using the parameter setting as in sub-chapter 4.3.1. Meanwhile, for the distortion identification stage, PATCH-IQ2 required another 0.05 second to compute the I2C distance between all the test patches and the labelled patches at 80:20 labelled-test ratios. Finally, an extra 0.08 second is required to estimate the quality scores for the test image's patches. Overall, on average, PATCH-IQ2 required 0.59 second to perform both distortion identification and quality estimation on one image. The average run-time comparison between PATCH-IQ2 and the competing BIQA models is shown in Table 4.13. PATCH-IQ2 was even slower than PATCH-IQ. However, given its superior distortion identification and quality prediction performances, PATCH-IQ2 can still be considered for applications that does not require real-time assessment.

Table 4.13: Average run-time

BIQA model	BIQI	BRISQUE	GMLOG	CORNIA	PATCH-IQ	*PATCH-IQ*
Run-time (s)	0.05	0.10	0.07	2.43	0.37	*0.59*

4.5 Chapter Summary

This chapter discusses two modifications made to the initial PATCH-IQ model. These involve investigating the effects of employing patch extraction strategies that take human visual attention into consideration. The first modified strategy is to find the image's interest point regions using SIFT algorithm prior to patch extraction. The second modified strategy is to extract patches from salient regions of the image using spectral residual based saliency detection method. These were motivated by a human observer usually focusing on the object-like or salient regions on the image. Using an assumption these regions carry greater weights on evaluating the quality of an image, both strategies were used to guide the patch sampling process in the framework. Experimental results on three major IQA databases showed that the SIFT based model, PATCH-IQ2, produced better distortion identification and quality

prediction performances than the initial PATCH-IQ and the spectral residual based model, PATCH-IQ3.

This was achieved at the expense of greater computational requirements. Longer computation time was required since the modified model, PATCH-IQ2, needs to detect the locations of the image's interest points prior to feature extraction stage. As the number of interest points based patches is generally higher than the number of non-overlapped patches, the processing times for both feature extraction and I2C distance computation will be further increased. It is necessary to reduce the number of patches in the labelled dataset while maintaining its superior performance to PATCH-IQ.

Finally, despite obtaining encouraging results, a few steps could be taken to improve these patch based models. Note that all three models rely on a labelled dataset. Introducing new types of distortions will increase the dataset size, leading to higher memory and processing time requirements. Here, the use of parallel computing or less computational expensive feature extraction methods could be explored to accelerate its speed. We could also integrate different nearest neighbour techniques [137]-[138] in the dataset construction to help dealing with an increasing number of new distortion classes. As for PATCH-IQ3, other visual saliency detection methods could also be tested to achieve better prediction performance. In addition, obtaining accurate image distortion class is essential to provide these models with better regression inputs for quality estimation stages. While they use a NBNN classifier to perform the classification, other nearest neighbour classifiers could also be tested to obtain higher classification accuracy.

Chapter 5

Multi-Task Learning Framework for Blind Image Quality Assessment Model

5.1 Chapter Introduction

In the last two chapters, the study focused on addressing several limitations encountered by the general-purpose BIQA models such as intensive training phase requirements, their inability to provide local quality estimation and their inability to perform distortion identification for an image. This is done by introducing two general-purpose BIQA models, PATCH-IQ and PATCH-IQ2 that utilise a patch based learning framework. The study now attempt to address another limitation shared by most general-purpose BIQA models.

Most models learn their prediction based on a set of training images. Given their respective features and the associated DMOS / MOS values, a regression function mapping the feature space to quality score space is learned. Since the features employed by these models are generally invariant to distortion, high prediction performances correlated with human perceptual measures are obtained by these models when tested on various types of distortions in standard IQA databases. However, it is still difficult to agree on the best general-purpose BIQA model that can work well across different distortion conditions. As discussed in Chapter 2, some observe that one BIQA model may have good prediction performance for a particular type of distortion but is less effective when tested on images with different distortion types. One reason is due to BIQA models learn their prediction for each image distortion class independently, ignoring the relationship among the learning tasks.

This scenario motivates the study to look at an alternative learning technique for a BIQA model. In this chapter, using multi-task learning (MTL) technique to simultaneously learn such regression functions for different distortion classes is explored. MTL is a learning approach that utilises a shared representation to learn multiple related tasks simultaneously. It is based on the assumption that the learner may find it easier to learn multiple tasks together rather than in isolation when the tasks share what they learn. MTL has been utilised in learning prediction models for web pages categorisation [139], disease prediction [140], therapy screening [141] and school examination scores [142]. Here, the study extends its application to BIQA tasks.

The learning task for many real-life classification or regression problems can often be divided into several related subtasks. For example, predicting the outcome of therapy may consist of predictions made based on several combinations of drugs [141] or predicting the examination scores nationally can be partitioned into predictions made based on individual schools [142]. In BIQA, we can consider these related subtasks to be the quality prediction model learning for each individual distortion cases (e.g. noise, blur, compression artefacts, etc.). Previous BIQA models typically solve these subtasks by employing single-task learning (STL) approach whereby each quality prediction model is learned independently. MTL differs from STL whereby these prediction models are learned simultaneously by exploiting relevant shared information across them. The difference between the STL approach and the MTL approach for BIQA tasks in shown in Figure 5.1. By learning simultaneously, the size of the training data for each distortion case is increased, often leading to better generalisation performance.

The proposed BIQA model is developed to utilise this advantage. The model, *Multi-Task Learning based Blind Image Quality assessment* (MTLBIQ), first extracts relevant spatial domain BIQA features from a collection of training images. These features are then utilised to simultaneously learn regression models for different distortion conditions. The training is

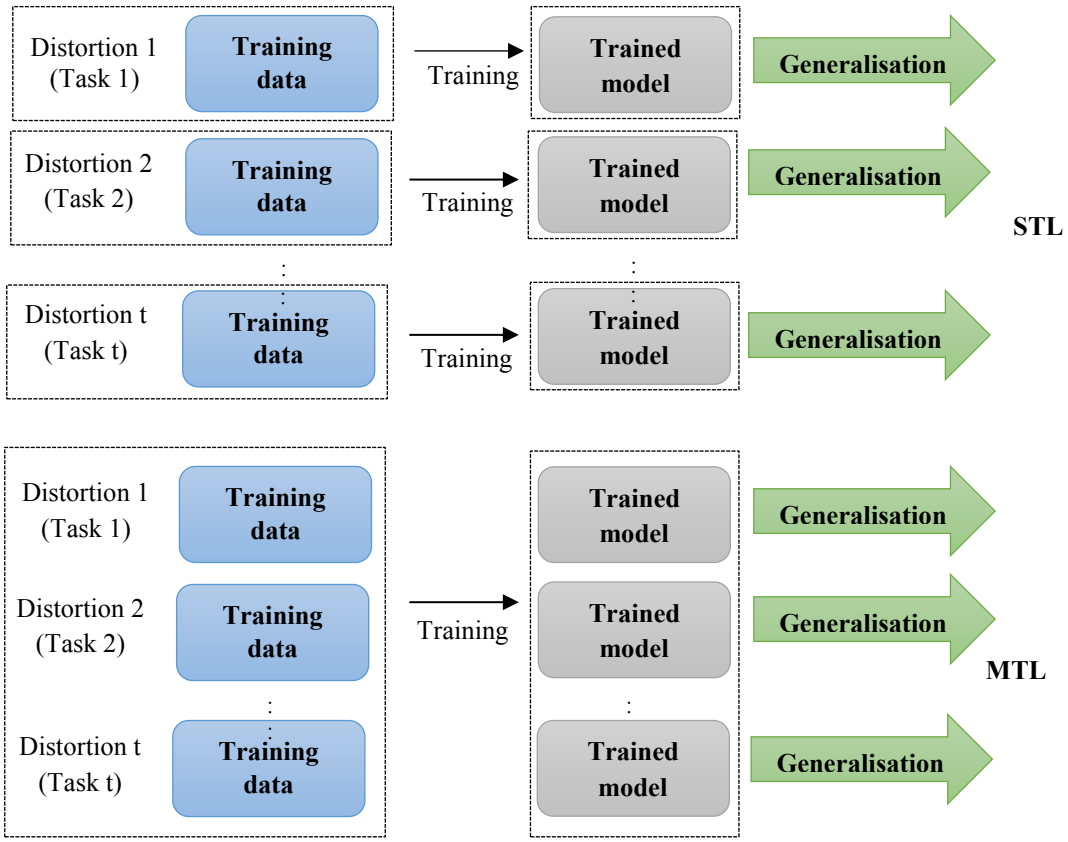


Figure 5.1: Single-task learning versus multi-task learning approaches for BIQA

performed using a trace-norm regularised MTL technique. For an image of a known distortion, MTLBIQ simply selects a specific regression model to perform the prediction of quality score. For an image of an unknown distortion, MTLBIQ estimates different distortions in the image using a support vector machine (SVM) classifier. The probability estimates from the classifier are then used to weigh the image prediction scores from different regression models. The weighted scores are then pooled to yield the final quality score.

The rest of this chapter is structured as follows. The learning framework including the utilised BIQA features for MTLBIQ is first introduced and discussed in sub-chapter 5.2. Sub-chapter 5.3 presents the experiments conducted to evaluate and validate MTLBIQ's performance. The chapter is then concluded in sub-chapter 5.4.

5.2 The Proposed Multi-Task Learning Framework

Figure 5.2 illustrates the proposed framework for MTLBIQ. It consists of feature extraction (FE), quality estimation (QE) and distortion identification (DI) stages.

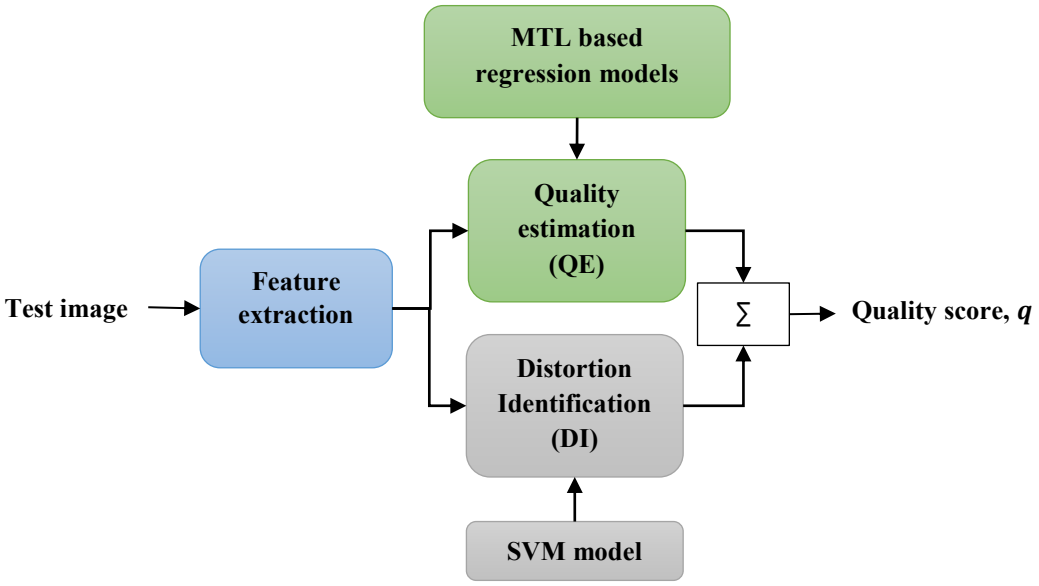


Figure 5.2: MTLBIQ framework

5.2.1 Feature extraction

The first stage of the framework is to extract BIQA features. Here, MTLBIQ proposes to extract two different set of BIQA features. The reason here is to check whether we can still achieve improvement in MTLBIQ's prediction performance when different types of BIQA features is extracted. This helps to show the benefit of MTL on BIQA evaluation regardless which sets of features are being used.

As in Chapters 3 and 4, MTLBIQ also employs spatial domain features to alleviate excessive computational load encountered by image transform based features. Two sets of spatial domain features are first extracted from an image before they are combined as BIQA features for MTLBIQ. The first set of features is similar to the ones implemented by the GMLOG model [76]. It consists of four statistical distributions derived from two image local

contrast operators: gradient magnitude (GM) and Laplacian of Gaussians (LOG). The GMLOG model showed that for a distorted image the shape of these distributions will differ from the same distributions of high quality images. As the distortion level increases, there are gradual changes in the distributions' shapes indicating they are predictive to image quality and can be features for a BIQA task.

Specifically, given an image $\mathbf{I}$, its GM map $\mathbf{G_I}$ and LOG response $\mathbf{L_I}$ are defined respectively as:

$$\mathbf{G_I} = \sqrt{[\mathbf{I} \otimes \mathbf{h}_x]^2 + \left[\mathbf{I} \otimes \mathbf{h}_y\right]^2} \tag{5.1}$$

and

$$\mathbf{L_I} = \mathbf{I} \otimes \mathbf{h}_{\mathrm{LOG}} \,. \tag{5.2}$$

In Equation (5.1), $\mathbf{h}_x$ and $\mathbf{h}_y$ are the Gaussian partial derivative filters applied along the horizontal and the vertical directions, respectively. Meanwhile, the LOG filter in Equation (5.2) is represented as:

$$\mathbf{h}_{\mathrm{LOG}}(x, y|\sigma_{\mathrm{G}}) = \frac{\partial^2}{\partial x^2}\mathbf{g}(x, y|\sigma_{\mathrm{G}}) + \frac{\partial^2}{\partial y^2}\mathbf{g}(x, y|\sigma_{\mathrm{G}}) \tag{5.3}$$

where $\mathbf{g}(x, y|\sigma_{\mathrm{G}})$ is the isotropic Gaussian function with scale parameter σ_{G}. These GM and LOG operators are then jointly normalised to achieve stable image representations. The normalised operators are given by:

$$\bar{\mathbf{G}}_\mathbf{I} = \frac{\mathbf{G_I}}{(\mathbf{N_I} + \varepsilon_{\mathrm{GMLOG}})} \;,\; \bar{\mathbf{L}}_\mathbf{I} = \frac{\mathbf{L_I}}{(\mathbf{N_I} + \varepsilon_{\mathrm{GMLOG}})} \tag{5.4}$$

where $\mathbf{N_I}$ is a local adaptive normalisation factor while $\varepsilon_{\mathrm{GMLOG}}$ is a constant that prevents numerical instability. In agreement with the GMLOG work, the normalisation factor is defined for each location (i, j) as:

$$\mathbf{N_I}(i, j) = \sqrt{\sum\sum_{(l,k)\in\Omega_{i,j}} \omega(l, k)\mathbf{F}_\mathbf{I}^2(l, k)} \tag{5.5}$$

In Equation (5.5), $\Omega_{i,j}$ represents a local window centred at (i,j), $\omega(l,k)$ is a spatially truncated Gaussian kernel weighting function rescaled to unit sum, and

$$\mathbf{F}_{\mathbf{I}}^2(i,j) = \mathbf{G}_{\mathbf{I}}^2(i,j) + \mathbf{L}_{\mathbf{I}}^2(i,j) \tag{5.6}$$

Once both operators are normalised, MTLBIQ computes their respective marginal probability functions and use them as the first two BIQA features for the image. The marginal probability functions are defined as:

$$P_{\bar{\mathbf{G}}_{\mathbf{I}}}(\bar{\mathbf{G}}_{\mathbf{I}} = g_m) = \sum_{n=1}^{N} \mathbf{K}_{m,n} \tag{5.7}$$

and

$$P_{\bar{\mathbf{L}}_{\mathbf{I}}}(\bar{\mathbf{L}}_{\mathbf{I}} = l_n) = \sum_{m=1}^{M} \mathbf{K}_{m,n} \tag{5.8}$$

In these equations, $\mathbf{K}_{m,n} = P(\bar{\mathbf{G}}_{\mathbf{I}} = g_m, \bar{\mathbf{L}}_{\mathbf{I}} = l_n)$ is the joint empirical probability function for the normalised GM and LOG operators while $m = 1,2,3,\ldots,M$ and $n = 1,2,3,\ldots,N$ represent the quantisation levels of those operators. To show that the two features ($P_{\bar{\mathbf{G}}_{\mathbf{I}}}$ and $P_{\bar{\mathbf{L}}_{\mathbf{I}}}$) are predictive of image quality, their histograms for a set of distorted images

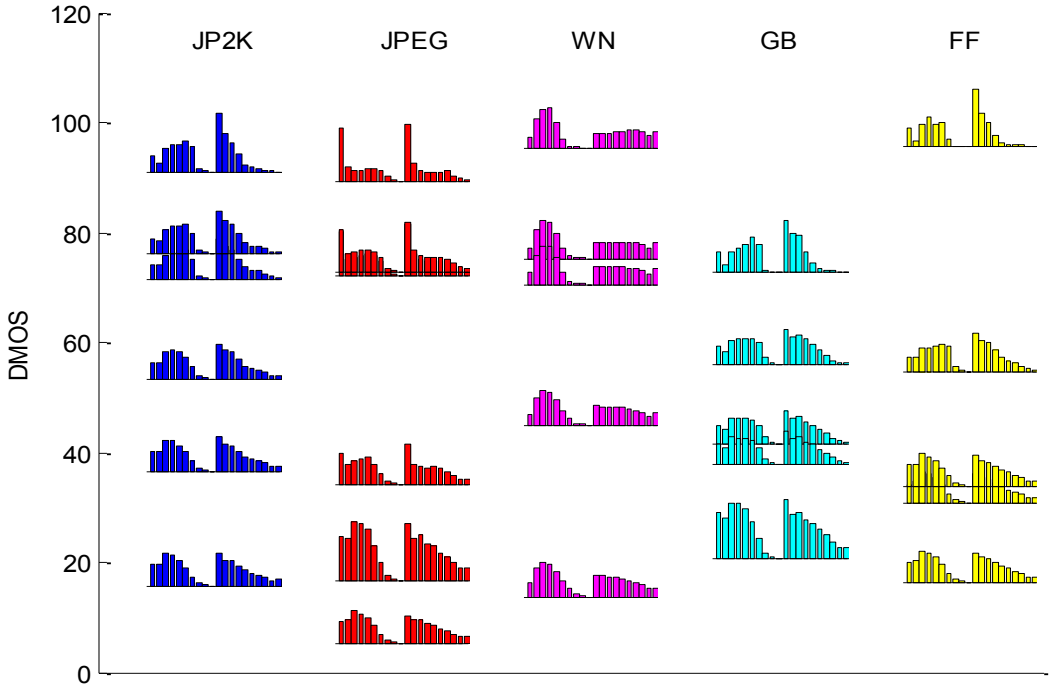


Figure 5.3: Marginal probability functions ($P_{\bar{\mathbf{G}}_{\mathbf{I}}}$ and $P_{\bar{\mathbf{L}}_{\mathbf{I}}}$) of the distorted images produced at different DMOS values for one reference image

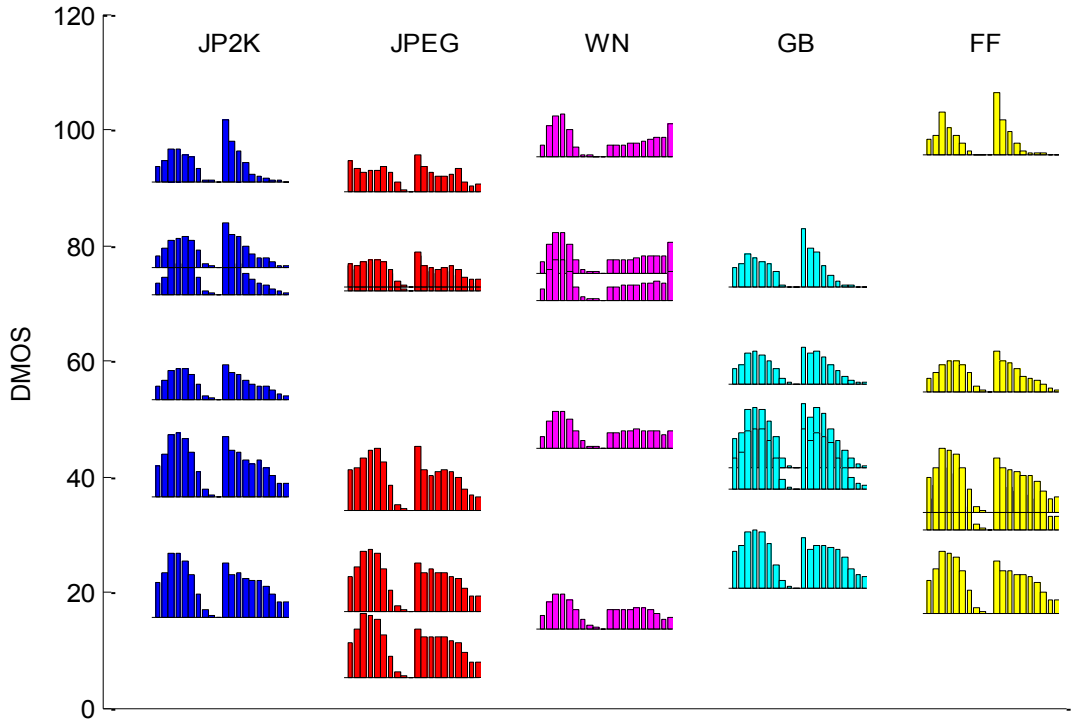


Figure 5.4: The independency distributions ($Q_{\bar{G}_I}$ and $Q_{\bar{L}_I}$) of the distorted images produced at different DMOS values for one reference image

produced from one reference image in the LIVE IQA database are plotted. The plot is shown in Figure 5.3. It can be seen that, for each type of distortion, the shape of the histogram gradually changes as the distortion level changes.

The next two BIQA features are derived because both the GM and LOG operators are inter-related. MTLBIQ measures the statistical interaction between both operators by computing the reliance of one particular value $\bar{\mathbf{G}}_\mathbf{I} = g_m$ over all potential values of $\bar{\mathbf{L}}_\mathbf{I}$ and vice-versa. The computations can be represented as:

$$Q_{\bar{\mathbf{G}}_\mathbf{I}}(\bar{\mathbf{G}}_\mathbf{I} = g_m) = \frac{1}{N}\sum_{n=1}^{N} P(\bar{\mathbf{G}}_\mathbf{I} = g_m | \bar{\mathbf{L}}_\mathbf{I} = l_n) \tag{5.9}$$

and

$$Q_{\bar{\mathbf{L}}_\mathbf{I}}(\bar{\mathbf{L}}_\mathbf{I} = l_n) = \frac{1}{M}\sum_{m=1}^{M} P(\bar{\mathbf{L}}_\mathbf{I} = l_n | \bar{\mathbf{G}}_\mathbf{I} = g_m) \tag{5.10}$$

Equations (5.9) and (5.10) can be the sum of conditional probabilities for one particular value of $\bar{\mathbf{G}}_\mathbf{I}$ over $\bar{\mathbf{L}}_\mathbf{I}$ and vice-versa. The distributions of both $Q_{\bar{\mathbf{G}}_\mathbf{I}}$ and $Q_{\bar{\mathbf{L}}_\mathbf{I}}$, known as

independency distributions, are plotted in Figure 5.4. Similar patterns can be observed whereby there are gradual changes in both $Q_{\bar{\mathbf{G}}_\mathbf{I}}$ and $Q_{\bar{\mathbf{L}}_\mathbf{I}}$ as the distortion level is varied, indicating their suitability for BIQA tasks.

Table 5.1: List of MTLBIQ's second set of features

Feature ID	Scale	Orientation	Feature Description
1-2	1 (S1)	-	Shape parameter and variance of GGD model of normalised luminance coefficients
3-6		Horizontal (H)	Shape parameter, mean, left variance and right variance of AGGD model of pairwise products
7-10		Vertical (V)	
11-14		Main-diagonal (MD)	
15-18		Secondary-diagonal (SD)	
19-20	2 (S2)	-	Shape parameter and variance of GGD model of normalised luminance coefficients
21-24		Horizontal (H)	Shape parameter, mean, left variance and right variance of AGGD model of pairwise products
25-28		Vertical (V)	
29-32		Main-diagonal (MD)	
33-36		Secondary-diagonal (SD)	

Table 5.2: Overall BIQA features extracted for MTLBIQ

ID	Length	Notations
1	M = N	$P_{\bar{\mathbf{G}}_\mathbf{I}}$
2	M = N	$P_{\bar{\mathbf{L}}_\mathbf{I}}$
3	M = N	$Q_{\bar{\mathbf{G}}_\mathbf{I}}$
4	M = N	$Q_{\bar{\mathbf{L}}_\mathbf{I}}$
5 - 6	2	$\gamma_{\mathrm{GGD(S1)}}, \sigma^2_{\mathrm{GGD(S1)}}$
7 - 10	4	$\nu_{\mathrm{AGGD(S1-H)}}, \mu_{\mathrm{AGGD(S1-H)}}, \sigma^2_{l\,\mathrm{AGGD(S1-H)}}, \sigma^2_{r\,\mathrm{AGGD(S1-H)}}$
11 - 14	4	$\nu_{\mathrm{AGGD(S1-V)}}, \mu_{\mathrm{AGGD(S1-V)}}, \sigma^2_{l\,\mathrm{AGGD(S1-V)}}, \sigma^2_{r\,\mathrm{AGGD(S1-V)}}$
15 - 18	4	$\nu_{\mathrm{AGGD(S1-MD)}}, \mu_{\mathrm{AGGD(S1-MD)}}, \sigma^2_{l\,\mathrm{AGGD(S1-MD)}}, \sigma^2_{r\,\mathrm{AGGD(S1-MD)}}$
19 - 22	4	$\nu_{\mathrm{AGGD(S1-SD)}}, \mu_{\mathrm{AGGD(S1-SD)}}, \sigma^2_{l\,\mathrm{AGGD(S1-SD)}}, \sigma^2_{r\,\mathrm{AGGD(S1-SD)}}$
23 - 24	2	$\gamma_{\mathrm{GGD(S2)}}, \sigma^2_{\mathrm{GGD(S2)}}$
25 - 28	4	$\nu_{\mathrm{AGGD(S2-H)}}, \mu_{\mathrm{AGGD(S2-H)}}, \sigma^2_{l\,\mathrm{AGGD(S2-H)}}, \sigma^2_{r\,\mathrm{AGGD(S2-H)}}$
29 - 32	4	$\nu_{\mathrm{AGGD(S2-V)}}, \mu_{\mathrm{AGGD(S2-V)}}, \sigma^2_{l\,\mathrm{AGGD(S2-V)}}, \sigma^2_{r\,\mathrm{AGGD(S2-V)}}$
33 - 36	4	$\nu_{\mathrm{AGGD(S2-MD)}}, \mu_{\mathrm{AGGD(S2-MD)}}, \sigma^2_{l\,\mathrm{AGGD(S2-MD)}}, \sigma^2_{r\,\mathrm{AGGD(S2-MD)}}$
37 - 40	4	$\nu_{\mathrm{AGGD(S2-SD)}}, \mu_{\mathrm{AGGD(S2-SD)}}, \sigma^2_{l\,\mathrm{AGGD(S2-SD)}}, \sigma^2_{r\,\mathrm{AGGD(S2-SD)}}$

These four distributions ($P_{\bar{\mathbf{G}}_\mathbf{I}}, P_{\bar{\mathbf{L}}_\mathbf{I}}, Q_{\bar{\mathbf{G}}_\mathbf{I}}, Q_{\bar{\mathbf{L}}_\mathbf{I}}$) are then concatenated to represent the first set of BIQA features for MTLBIQ. For the second set of its BIQA features, MTLBIQ utilised the same spatial domain features as described in Chapters 3 and 4. I.e. the second set of MTLBIQ's

features consists of 36 features extracted over two scales. For ease of reading, the features are listed again in Table 5.1. The two sets of features are then concatenated to produce the final feature vector. Table 5.2 summarises all the features used by MTLBIQ.

5.2.2 Quality estimation via multi-task learning

Given a set of training images, the extracted feature vectors are then utilised to learn quality prediction models under different distortion conditions. Previous BIQA approaches find these models by employing STL whereby each quality prediction model is treated as a single task and learned independently. MTLBIQ, meanwhile, learns these models (tasks) simultaneously by employing a MTL technique. MTL techniques generally aim to minimise this objective function:

$$\min_{\mathbf{W}} F(\mathbf{W}) = f(\mathbf{W}) + \Omega(\mathbf{W}) \ , \tag{5.11}$$

with $f(\mathbf{W})$ represents the empirical loss function of the training set while $\Omega(\mathbf{W})$ represents the regularisation term that captures the relationship among the tasks. For a BIQA case, $f(\mathbf{W})$ is represented by a loss function $\ell(\cdot,\cdot)$ as:

$$f(\mathbf{W}) = \sum_{i=1}^{T} \sum_{j=1}^{R_i} \ell\left(s_i^j, \omega_{\mathrm{t}i}^{\mathrm{T}} x_i^j\right) , \tag{5.12}$$

with T is the number of BIQA learning tasks, R_i represents the number of samples for the ith task, x_i^j and s_i^j are the jth feature vector and the associated DMOS value in the ith task, respectively and $\mathbf{W} = [\omega_{\mathrm{t}1}, \omega_{\mathrm{t}2}, \dots, \omega_{\mathrm{t}T}]$ where ω_{t} represents the estimated parameter of the training samples.

Depending on the assumptions made on the task relatedness, there are many formulations of $\Omega(\mathbf{W})$ [143]-[146]. Because the utilised features are high dimensional and under the assumption that all BIQA tasks are inter-related, MTLBIQ employs a trace-norm regularised based MTL technique. The technique is chosen because of its well performance when dealing

with high dimensional MTL data [147]. The technique captures the task relatedness through low dimensional subspace learning whereby a common low-rank structure is shared among the models of various tasks. Figure 5.5 illustrates the trace-norm regularised training framework for MTLBIQ.

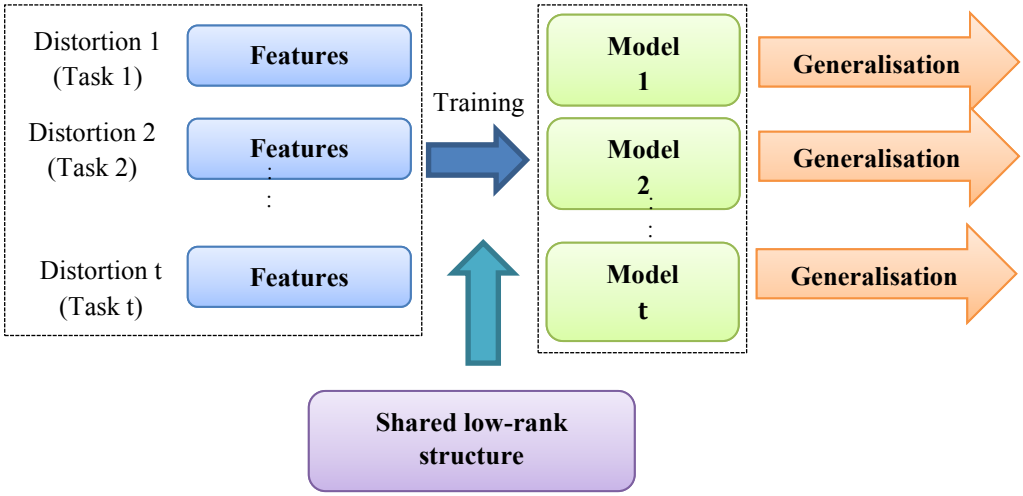


Figure 5.5: Trace-norm regularised MTL training framework

To capture the low-rank structure shared by the tasks, the trace-norm regularised technique treats the objective function as a matrix rank minimisation problem. Equation (5.11) can now be rewritten as [147]:

$$\min_{\mathbf{W}} F(\mathbf{W}) = f(\mathbf{W}) + \lambda[\text{Rank}(\mathbf{W})] \quad . \tag{5.13}$$

Minimising the matrix rank is a NP-hard problem. To solve this, the rank function Rank($\mathbf{W}$) is often approximated through convex relaxation methods. A trace-norm relaxation method is widely used to this effect as it has been shown theoretically to be a good approximation for Rank($\mathbf{W}$) [148]. Therefore, the problem can now be approximated as a trace-norm minimisation problem whereby Equation (5.13) is rewritten as [147]:

$$\min_{\mathbf{W}} F(\mathbf{W}) = f(\mathbf{W}) + \lambda\|\mathbf{W}\|_* \quad , \tag{5.14}$$

where λ is positive regularisation parameter and $\|\cdot\|_*$ denotes the trace norm defined as the sum of singular values. Equation (5.14) can be solved typically by a sub gradient method [149]. However, for faster convergence rate, MTLBIQ employs an accelerated gradient method (AGM) [150] to find the optimised values of $\mathbf{W}$:

$$\mathbf{W} = \arg\min_{\mathbf{W}} \frac{\tau}{2} \left\| \mathbf{W} - \left(\mathbf{Z} - \frac{1}{\tau} \nabla f(\mathbf{Z}) \right) \right\|_{\mathrm{F}}^{2} + \lambda \|\mathbf{W}\|_* , \tag{5.15}$$

where $\mathbf{Z}$ represents the search point on the ongoing iteration, τ stands for the current step size while $\nabla f(\cdot)$ represents the gradient for $f(\cdot)$. The optimised values are then used to represent the trained model for each distortion case. Further details on AGM can be found in [150].

5.2.3 Distortion identification

The trained models are then used to predict the quality score of a test image. For a test image of a known distortion type, MTLBIQ simply selects the trained regression model associated with the distortion. For a test image of unknown distortion, MTLBIQ first estimates different distortion types present in the image. The process is performed using the extracted feature vector as an input to an SVM classifier. SVM is chosen here due to its good generalisation capabilities and excellent performance in high dimensional spaces [151]. In this work, a multiclass SVM with a kernel of radial basis function (RBF) is utilised. Note that the aim is not to perform hard classification but to estimate each distortion class present in the image. These estimates are given by the probabilities provided by the classifier. These probability values are then used to weigh the image prediction scores from different MTL models. The weighted scores are then pooled to yield the final quality score for the image.

5.3 Results and Discussions

5.3.1 Experimental setup and evaluation protocol

Databases: Three publicly available subjective image databases were utilised to analyse the performance of MTLBIQ: LIVE, CSIQ and TID2008. As in the previous chapters, for both the CSIQ and the TID2008 databases, only 4 types of distortion present in the LIVE database are considered by MTLBIQ: JP2K, JPEG, WN and GB.

Parameter setting: MTLBIQ's features combine two sets of spatial domain features as implemented in GMLOG and BRISQUE. Here, MTLBIQ's parameters are set to follow their implementation. The filters' scale parameter σ_{G} to compute GM and LOG operators was set at 0.5 while the quantisation level $M = N$ is 10. The local window size $K = L$ to compute the locally normalised luminance coefficients was set at 3. Both constant $\varepsilon_{\mathrm{GMLOG}}$ and ε_{B} are 1.

Regression model learning: To investigate the effects of using different feature sets on MTLBIQ's prediction performance, three MTLBIQ models were learned. The first model, denoted by MTLBIQ1, was trained using only the first set of features (GMLOG) while the second model MTLBIQ2 was trained using only the second set of features (BRISQUE). The third model, MTLBIQ3, utilised both sets of features in its training.

Performance metrics and benchmarked models: To evaluate MTLBIQ's performance, three metrics as in Chapter 4 were used to measure the consistency between the quality scores predicted from the experiments and the subjective DMOS/MOS values. They were: LCC, SROCC and RMSE. The benchmarked models were similar whereby the three MTLBIQ models were compared against six BIQA models: BIQI, BRISQUE, GMLOG, CORNIA, PATCH-IQ and PATCH-IQ2. The MTLBIQ models were also compared with two FR-IQA models: SSIM and FSIM. The train-test partition is set at 80:20 ratio. The trace-norm regularised MTL technique to train the three MTLBIQ models was implemented using the MALSAR package [152]. In the package, the loss function $\ell(\cdot,\cdot)$ is set as a least squares function. Meanwhile, regression for the competing models were performed using the LIBSVM

package as before. The same LIBSVM package was used to train the SVM classifier required by MTLBIQ models in the DI stage.

Experiments: Two experiments were performed to evaluate the performance: the overall performance experiment and the DS performance experiment. Note that MTLBIQ contains different trained models for different distortion classes. For the DS performance experiment in which the distortion type is known beforehand, MTLBIQ can directly select a specific trained model for the QE stage without having to perform the DI stage.

5.3.2 Overall performance comparison

Table 5.3: Median values across 1,000 runs of the overall performance experiment

IQA model	LIVE			CSIQ			TID2008		
	LCC	SROCC	RMSE	LCC	SROCC	RMSE	LCC	SROCC	RMSE
SSIM	0.946	0.949	8.804	0.935	0.936	0.099	0.909	0.903	0.662
FSIM	**0.961**	**0.964**	**7.546**	**0.968**	**0.963**	**0.071**	**0.954**	**0.956**	**0.471**
BIQI	0.849	0.844	15.407	0.809	0.749	0.187	0.870	0.844	0.787
BRISQUE	0.943	0.942	9.395	0.930	0.910	0.107	0.914	0.908	0.700
GMLOG	0.951	0.950	8.829	0.939	0.925	0.100	0.926	0.929	0.625
CORNIA	0.939	0.942	9.920	0.911	0.887	0.125	0.912	0.884	0.711
PATCHIQ	0.954	0.952	**8.476**	0.946	0.932	0.094	0.939	0.930	0.572
PATCHIQ2	**0.956**	**0.954**	**8.149**	**0.959**	**0.943**	**0.081**	0.946	0.932	0.536
MTLBIQ1	***0.960***	***0.957***	*8.806*	*0.948*	*0.926*	*0.092*	***0.947***	***0.934***	***0.522***
MTLBIQ2	*0.955*	*0.949*	*9.452*	***0.949***	***0.934***	***0.090***	***0.957***	***0.951***	***0.468***
MTLBIQ3	***0.963***	***0.958***	***8.643***	***0.966***	***0.950***	***0.074***	***0.966***	***0.961***	***0.424***

The median results across 1,000 iterations for the overall performance experiment are reported in Table 5.3. The best three BIQA models and the top FR-IQA model are in bold. MTLBIQ1 and MTLBIQ3 are among the top three models on the LIVE database while MTLBIQ2 and MTLBIQ3 are among the top three models on the CSIQ database. All three MTLBIQ models produced the top three LCC, SROCC and RMSE values on the TID2008 database. We can also observe that MTLBIQ1 improved upon GMLOG and MTLBIQ2 improved upon BRISQUE, respectively. This implies that MTL generally can improve the overall prediction performance of a BIQA model. MTLBIQ1 obtained better performance

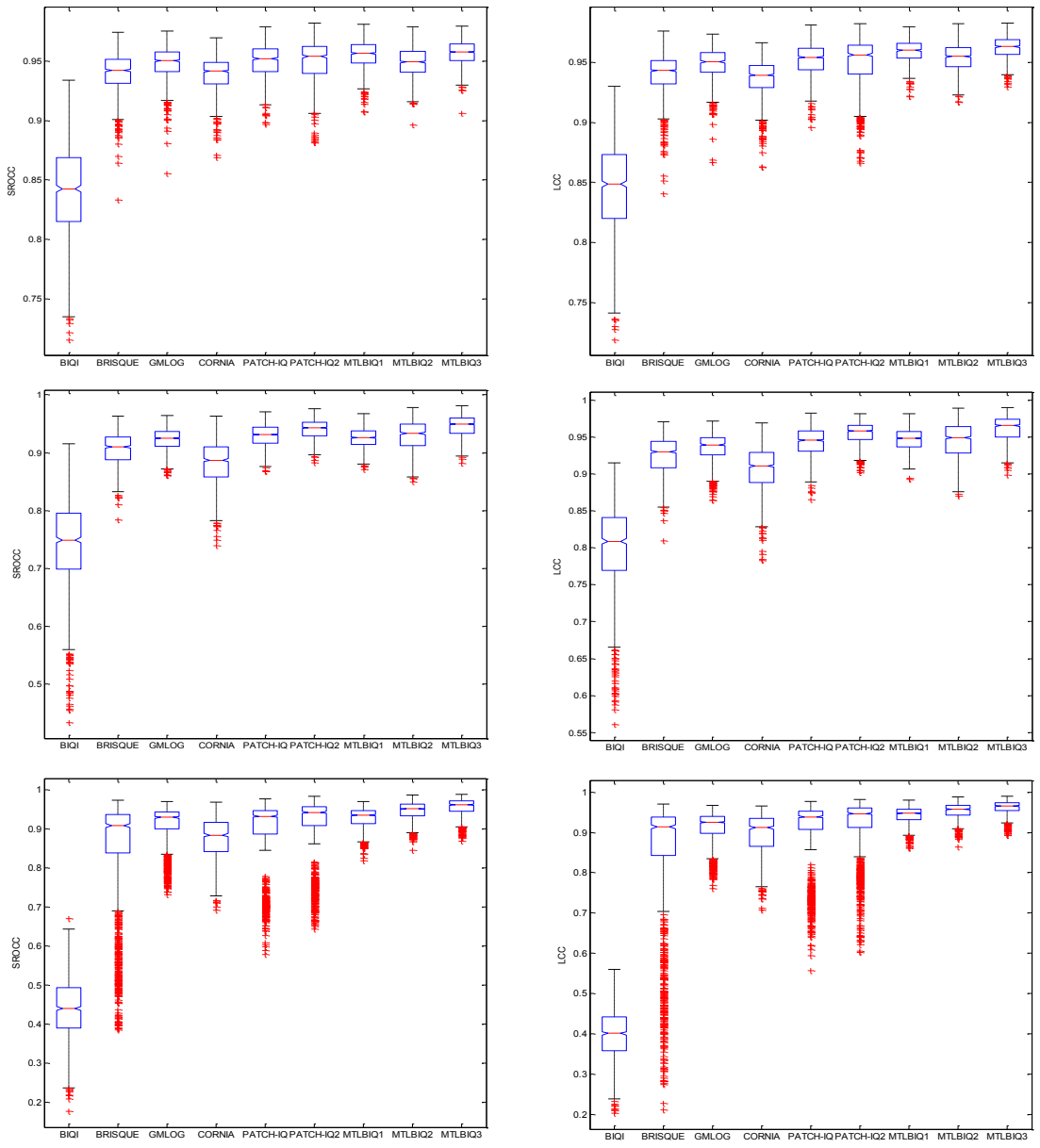
metrics' values than MTLBIQ2 on the LIVE database. In reverse, MTLBIQ2 outperformed MTLBIQ1 on the CSIQ and the TID2008 databases. Therefore, there is no clear indication to which set of features is more discriminative of image quality. The best metric values are achieved when we utilised both sets of features as in MTLBIQ3. Compared to FR-IQA models, MTLBIQ models outperformed SSIM while approaching FSIM. This is promising since MTLBIQ requires no reference image information. We can also compare both PATCH-IQ and PATCH-IQ2 with MTLBIQ2 since they employs the same set of features. Among the three models, PATCH-IQ2 produced the best metric values on the LIVE and the CSIQ databases while MTLBIQ2 had better metrics values on the TID2008 database. Thus, there is no clear indication to which learning framework is better for BIQA evaluation. We should also note that MTLBIQ2 extracts its features on image level. Better comparison could be made provided MTLBIQ2 is performed using features that are extracted on patch level as in both PATCH-IQ and PATCH-IQ2 operation.

Table 5.4: IQR values for the overall performance experiment

BIQA model	LIVE		CSIQ		TID2008	
	LCC	SROCC	LCC	SROCC	LCC	SROCC
BIQI	0.053	0.054	0.071	0.096	0.084	0.104
BRISQUE	0.020	0.020	0.036	0.039	0.095	0.099
GMLOG	0.017	0.017	0.024	0.026	0.043	0.043
CORNIA	0.018	0.018	0.041	0.052	0.069	0.076
PATCHIQ	0.018	0.019	0.028	0.027	0.047	0.060
PATCHIQ2	0.021	0.022	**0.019**	**0.023**	0.049	0.048
MTLBIQ1	***0.012***	***0.015***	***0.020***	***0.023***	***0.027***	***0.032***
MTLBIQ2	***0.016***	***0.017***	*0.036*	*0.037*	***0.023***	***0.030***
MTLBIQ3	***0.012***	***0.014***	***0.024***	***0.027***	***0.021***	***0.027***

The IQR value of the 1,000 SROCC and LCC results obtained by each BIQA model are also computed. The values are recorded in Table 5.4 with the best three models are in bold. We can see that MTLBIQ1 and MTLBIQ3 were among the top three models across all three databases while MTLBIQ2 was also in the top three for the LIVE and the TID2008 databases. These observations suggest that the first set of features (GMLOG features) produces more

consistent prediction results than the second set of features (BRISQUE features). These also indicate that MTLBIQ framework generally produces more consistent prediction results compared to PATCH-IQ and PATCH-IQ2 frameworks.



(a) (b)

Figure 5.6: Box plots of performance metric distributions of BIQA models for 1,000 experiment trials on the LIVE database (top row), the CSIQ database (middle row) and the TID2008 database (bottom row): (a) SROCC and (b) LCC

To visualise the IQR for each model, the box-plots of the SROCC and the LCC distributions were also generated as in Figure 5.6. We can see that MTLBIQ models have more

compact outlier distributions than other competing BIQA models. The IQR and outlier observations indicate that MTLBIQ models have better quality prediction consistency and more robust to variation of training samples.

Table 5.5: The Wilcoxon rank-sum test results based on the BIQA models SROCC values

LIVE							
	BIQI	BRISQUE	GMLOG	CORNIA	MTLBIQ1	MTLBIQ2	MTLBIQ3
BIQI	0	-1	-1	-1	-1	-1	-1
BRISQUE	1	0	-1	1	-1	-1	-1
GMLOG	1	1	0	1	-1	0	-1
CORNIA	1	-1	-1	0	-1	-1	-1
MTLBIQ1	1	1	1	1	0	-1	0
MTLBIQ2	1	1	0	1	-1	0	-1
MTLBIQ3	1	1	1	1	0	1	0
CSIQ							
	BIQI	BRISQUE	GMLOG	CORNIA	MTLBIQ1	MTLBIQ2	MTLBIQ3
BIQI	0	-1	-1	-1	-1	-1	-1
BRISQUE	1	0	-1	1	-1	-1	-1
GMLOG	1	1	0	1	0	-1	-1
CORNIA	1	-1	-1	0	-1	-1	-1
MTLBIQ1	1	1	0	1	0	-1	-1
MTLBIQ2	1	1	1	1	1	0	-1
MTLBIQ3	1	1	1	1	1	1	0
TID2008							
	BIQI	BRISQUE	GMLOG	CORNIA	MTLBIQ1	MTLBIQ2	MTLBIQ3
BIQI	0	-1	-1	-1	-1	-1	-1
BRISQUE	1	0	-1	1	-1	-1	-1
GMLOG	1	1	0	1	-1	-1	-1
CORNIA	1	-1	-1	0	-1	-1	-1
MTLBIQ1	1	1	1	1	0	-1	-1
MTLBIQ2	1	1	1	1	1	0	-1
MTLBIQ3	1	1	1	1	1	1	0

Next, the statistical significance testing was performed via the Wilcoxon rank-sum test. The test was conducted as in Chapters 3 and 4. The results of the test for MTLBIQ models against the four other competing models are tabulated in Table 5.5. Observations on the results demonstrate that the differences between the MTLBIQ3 model and the rest of BIQA models were statistically significant on all three databases. MTLBIQ2 also differed from the rest with an exception to GMLOG on the CSIQ database in which no statistically significant differences in the prediction performance is observed. For MTLBIQ1, it also differed from the rest with an exception to GMLOG on the LIVE database.

Meanwhile, the test results for MTLBIQ models against PATCH-IQ and PATCH-IQ2 are tabulated in Table 5.6. We can see that the differences in prediction performance between MTLBIQ3 and PATCH-IQ and between MTLBIQ and PATCH-IQ2 are statistically significant when tested over the three databases. This is also the case when comparing MTLBIQ1 to both PATCH-IQ and PATCH-IQ2. For MTLBIQ2, it also differed from those two models with an exception to PATCH-IQ on the CSIQ database in which no statistically significant differences in the prediction performance is observed.

Table 5.6: The Wilcoxon rank-sum test results for MTLBIQ models versus PATCH-IQ models

LIVE					
	PATCH-IQ	PATCH-IQ2	MTLBIQ1	MTLBIQ2	MTLBIQ3
PATCH-IQ	0	0	-1	1	-1
PATCH-IQ2	0	0	-1	1	-1
MTLBIQ1	1	1	0	1	0
MTLBIQ2	-1	-1	-1	0	-1
MTLBIQ3	1	1	0	1	0
CSIQ					
	PATCH-IQ	PATCH-IQ2	MTLBIQ1	MTLBIQ2	MTLBIQ3
PATCH-IQ	0	-1	1	0	-1
PATCH-IQ2	1	0	1	1	-1
MTLBIQ1	-1	-1	0	-1	-1
MTLBIQ2	0	-1	1	0	-1
MTLBIQ3	1	1	1	1	0
TID2008					
	PATCH-IQ	PATCH-IQ2	MTLBIQ1	MTLBIQ2	MTLBIQ3
PATCH-IQ	0	-1	-1	-1	-1
PATCH-IQ2	1	0	-1	-1	-1
MTLBIQ1	1	1	0	-1	-1
MTLBIQ2	1	1	1	0	-1
MTLBIQ3	1	1	1	1	0

5.3.3 Distortion specific performance comparison

The median results for the DS performance experiment are tabulated in Table 5.7. For simplicity, only the SROCC results are reported. Similar patterns can be observed from the LCC and the RMSE results. Again, the top three BIQA models are in bold. We can see that MTLBIQ3 obtained the highest correlation with human perceptual measures for each distortion case in all three databases except for GB images on the LIVE database. Direct comparison

between MTLBIQ1 and MTLBIQ2 showed that MTLBIQ1 had higher prediction performance for images affected by JP2K compression artefacts while MTLBIQ2 performed better prediction in WN and GB cases. In JPEG cases, MTLBIQ1 is slightly better on the LIVE database while MTLBIQ2 is better when tested on the CSIQ and TID2008 databases. Compared to FR-IQA models, MTLBIQ3 produced better prediction performance than SSIM and FSIM for noisy images. It also obtained comparable performance for other distortion cases.

Table 5.7: Median SROCC values across 1,000 runs for the DS performance experiment

IQA model	LIVE					CSIQ				TID2008			
	JP2K	JPEG	WN	GB	FF	JP2K	JPEG	WN	GB	JP2K	JPEG	WN	GB
SSIM	0.961	0.976	0.969	0.952	**0.956**	0.961	0.955	0.897	0.961	0.963	0.925	0.811	0.954
FSIM	**0.972**	**0.984**	0.972	**0.971**	0.952	**0.970**	**0.966**	**0.936**	**0.973**	0.978	0.929	0.876	0.947
BIQI	0.830	0.906	0.933	0.866	0.689	0.764	0.910	0.540	0.783	0.796	0.894	0.508	0.888
BRISQUE	0.916	0.964	0.979	0.945	0.887	0.898	0.921	0.921	0.919	0.889	0.908	0.868	0.853
GMLOG	0.927	0.963	0.983	0.929	0.901	0.916	0.936	0.941	0.908	0.902	0.922	0.905	0.877
CORNIA	0.921	0.936	0.961	**0.952**	**0.905**	0.894	0.882	0.786	0.904	0.915	0.880	0.566	0.892
PATCHIQ	0.931	**0.976**	0.987	0.953	0.891	0.918	0.952	0.963	0.916	0.910	0.923	0.904	0.905
PATCHIQ2	0.933	**0.973**	**0.987**	**0.970**	0.882	**0.933**	**0.953**	**0.965**	**0.943**	0.922	0.931	0.900	0.921
MTLBIQ1	**0.936**	0.969	0.984	0.929	**0.904**	**0.928**	0.929	0.945	0.918	**0.947**	**0.935**	**0.907**	**0.929**
MTLBIQ2	**0.933**	0.966	**0.990**	0.945	0.891	0.926	**0.961**	**0.981**	**0.944**	**0.944**	**0.964**	**0.951**	**0.955**
MTLBIQ3	**0.948**	**0.976**	**0.990**	**0.949**	**0.914**	**0.946**	**0.965**	**0.981**	**0.951**	**0.955**	**0.968**	**0.954**	**0.964**

Direct comparison between MTLBIQ2 and BRISQUE and between MTLBIQ1 and GMLOG can investigate whether MTL can improve a BIQA model's prediction performance for individual distortion category. On all three databases, we can see that MTLBIQ1 achieved higher SROCC values than GMLOG in all tested distortion cases with an exception for GB images on the LIVE database in which both achieved the same SROCC values. The same pattern can be observed in MTLBIQ2 versus BRISQUE cases. These observations validate the use of MTL to achieve better prediction performance for BIQA tasks.

As in the overall performance experiment, we can also compare both PATCH-IQ and PATCH-IQ2 with MTLBIQ2 since they employs the same set of features. We can see that MTLBIQ2 obtained better SROCC values across for all distortion concerned when tested on the TID2008 database. When tested on the LIVE database, MTLBIQ2 produced higher SROCC values for JP2K, WN and FF images but lower SROCC values for JPEG and GB

images. However, the reverse happened when the models were tested on the CSIQ database. MTLBIQ2 now obtained higher SROCC values for JPEG and GB images but lower SROCC values for JP2K images. These observations across the three tested databases suggest that MTL based model can produced higher prediction performance than PATCH-IQ and PATCH-IQ2 models for individual distortion case particularly for WN images. Again, fairer comparison could be made if MTLBIQ2 operates at patch level as in PATCH-IQ and PATCH-IQ2 framework.

5.3.4 Distortion identification accuracy

Introducing a distortion identification stage for an unknown test image (overall experiment) brings additional property to MTLBIQ: it is capable to provide information on the distortion affecting the image. This property, which is unavailable in most of the previous BIQA models, could be useful in certain application domains. For example, it is easier to repair a distorted image at the receiver end of an image communication system when the distortion affecting the image is known beforehand. MTLBIQ utilises the SVM classifier for this purpose. To show that the classifier has a good classification performance, the mean classification accuracy over 1,000 runs of experiments on all three databases is recorded. The results are tabulated in Table 5.8. A good classification performance with minimum accuracy of 78% was consistently achieved by the classifier when tested on all four types of distortion shared by the three databases. Slight degradation in the classifier's classification performance could be observed when it was tested on FF images on the LIVE database. This is to be expected as the FF images are essentially multiple distorted images whereby the images are first compressed by JP2K encoder before subjected to packet loss error. The results also validate the suitability of the utilised features for distortion identification purposes.

Table 5.8: Mean classification accuracy value over 1,000 iterations

LIVE						
	JP2K	JPEG	WN	GB	FF	ALL
MTLBIQ1	83.32	96.12	98.20	94.79	78.72	90.24
MTLBIQ2	82.51	87.53	97.23	92.67	83.48	88.45
MTLBIQ3	82.74	89.86	99.29	95.26	84.26	89.66
CSIQ						
	JP2K	JPEG	WN	GB	FF	ALL
MTLBIQ1	91.15	87.59	95.36	90.39	-	91.12
MTLBIQ2	82.41	78.00	94.39	88.00	-	85.70
MTLBIQ3	86.17	79.02	94.24	90.61	-	87.51
TID2008						
	JP2K	JPEG	WN	GB	FF	ALL
MTLBIQ1	97.29	99.36	97.56	92.90	-	96.78
MTLBIQ2	92.86	97.83	99.14	93.07	-	95.73
MTLBIQ3	93.32	97.80	99.10	93.43	-	95.92

5.3.5 Cross database test

The experiments performed in sub-chapters 5.3.2 and 5.3.3 used training and test sets taken from the same database. A BIQA model is said to be robust and has good generalisation capability if the model, trained on one database, can still obtain good prediction results when tested on another database. Therefore, a cross database testing was performed in this sub-chapter. A BIQA model was first trained with the LIVE database before the model was tested on both the TID2008 and the CSIQ databases. The same model was then trained on the CSIQ database and it was tested using the TID2008 and the LIVE databases. Finally, the model was trained using the TID2008 database before being tested on the remaining two databases. Again, SROCC were used for evaluation. The results for the cross database test are presented in Table 5.9. We can see that MTLBIQ1 achieved higher SROCC values than GMLOG in 4 out of the 6 tests. Similar patterns can be observed between MTLBIQ2 and BRISQUE whereby MTLBIQ2 outperformed BRISQUE in 5 out of the 6 tests. These show that MTL can improve the generalisation capability of a BIQA model. We can also see that MTLBIQ3 produced the best SROCC values in 4 tests and has close results in the other 2 tests.

Table 5.9: SROCC values for cross database test

Train	Test	BIQI	BRISQUE	GMLOG	CORNIA	MTLBIQ1	MTLBIQ2	MTLBIQ3
LIVE	CSIQ	0.781	0.899	0.911	0.897	0.909	0.900	**0.916**
LIVE	TID2008	0.819	0.905	0.920	0.893	0.930	0.926	**0.931**
CSIQ	LIVE	0.454	0.931	**0.946**	0.928	0.931	0.933	0.938
CSIQ	TID2008	0.698	0.899	0.905	0.870	**0.925**	0.908	0.918
TID2008	LIVE	0.763	0.929	0.934	0.909	0.940	0.922	**0.943**
TID2008	CSIQ	0.801	0.867	0.839	0.833	0.855	0.876	**0.877**

5.3.6 Computational complexity

Fast computation is another crucial aspect to consider in any BIQA model evaluation. The processing time required to run the MTLBIQ models is analysed in this sub-chapter. The average run-time comparison between MTLBIQ models and the competing BIQA models for a typical image of 512×768 size is shown in Table 5.10. These processing times are achieved using un-optimised MATLAB R2011b code on an 8GB RAM computer with an Intel i5 3.20 GHz processor. Note that the training time is not considered here as it is assumed that the models are already trained prior to the testing stage.

Table 5.10: Average run-time comparison

Model	BIQI	BRISQUE	GMLOG	CORNIA	MTLBIQ1	MTLBIQ2	MTLBIQ3
Run times	0.05	0.10	0.07	2.43	0.08	0.11	0.19

All MTLBIQ models are faster than CORNIA. Both MTLBIQ1 and MTLBIQ2 are slower by 0.01 seconds than GMLOG and BRISQUE, respectively. This is due to the distortion identification requirement. The differences are negligible and the MTLBIQ models can process up to 12 images per second (in MTLBIQ1 case), addressing real-time applications.

5.4 Chapter Summary

In this chapter, a simple yet effective BIQA model that employs a trace-norm regularised MTL technique in its learning framework is presented. The model, dubbed as MTLBIQ, utilises

a shared representation among differently distorted training samples to simultaneously learn prediction models for each distortion class. Experimental results on three standard IQA databases showed that MTLBIQ correlates highly with human perceived quality measures across various types of image distortions. MTLBIQ also achieves improved prediction performances when compared to several well-known BIQA models. Besides, MTLBIQ also can provide information on the distortion affecting an image, which is a useful property unavailable in most of the previous models. There are a few steps that could be taken to improve the MTLBIQ model. Further validation of its performance could be performed using different quality predictive features and databases. The MTL technique itself could also perform distortion identification for an image of unknown distortion. In addition, other MTL techniques can be tested for faster computation.

Chapter 6

Conclusion and Future Work

The chapter begins with a summary of the work performed throughout the study, including its contributions. It is followed by further discussions on several limitations of the work, which can be research topics worth to be pursued. The thesis is then concluded with some final remarks.

6.1 Summary and Contributions

The study focuses on image quality assessment (IQA) research area with specific aim of developing automatic prediction models that can provide image quality metric consistent with human perceptual measures. To begin with, Chapter 1 of the thesis provides a general overview on image quality and its traditional quality metrics. The chapter continues with discussions on the downside of traditional metrics and the needs for developing perceptual IQA models. This is followed by a general classification of the current perceptual IQA models whereby the scope of the work is set on one class: general-purpose blind IQA (BIQA).

Chapter 2 reviews design philosophies and approaches on the general-purpose BIQA models. The review shows there are large number of distinct image features that can be considered regarding developing a successful BIQA model. We could incorporate all features into one model design. However, this approach can results in a highly complicated model that would be difficult for implementation in image communication systems. Rather than introducing new quality-predictive features, which is the focus of most current models, the study takes a different direction to perform BIQA. Specifically, the study aims to contribute to the IQA research community by developing new learning frameworks for BIQA models. This

is motivated by several limitations identified through the review and the corresponding performance analyses.

The first proposed learning framework for BIQA model is presented in Chapter 3. The corresponding model, termed as PATCH-IQ, operates on a five-stage patch based framework. The main contribution of PATCH-IQ lies on it performing quality prediction using nearest neighbour learning methods avoiding the need to have a prior training phase, which is a prerequisite in many previous models. Other key contributions include its ability to provide local quality estimation and to perform image distortion identification, two useful properties that are unavailable in most of BIQA models. The reported experimental results show that PATCH-IQ performs competitively compared to some state-of-the-art models.

In Chapter 4, a simple modification is proposed to the PATCH-IQ's framework to further improve the quality prediction performance. This second learning framework considers the fact that, when presented with an image, human observer mostly concentrate on the object-like regions. The first modified models, PATCH-IQ2 employs interest points based sampling strategy in its framework whereby the utilised image patches are extracted at the locations of an image's interest points. The second modified models, PATCH-IQ3 utilises image saliency map to guide its sampling strategy whereby patches are extracted at the image regions of high saliency values. Upon testing on three common IQA databases, the results show that higher distortion identification and quality estimation accuracy can be obtained at the expense of a slight increase in the computation time. Further analyses also show that both PATCH-IQ2 and PATCH-IQ3 have close correlation to subjective quality measures and they generalise well across different databases including the one with multiple distorted images.

The third learning framework is next and is proposed in Chapter 5. Motivated by the observation that BIQA models may perform well in one particular type of distortion but is less

effective on others, the study presents a BIQA model that integrate multi-task learning (MTL) technique in its framework. The model, termed as MTLBIQ, consists of different sub-models for different distortion classes. Instead of individually trained as in the previous BIQA approaches, these sub-models are trained simultaneously by exploiting shared information across them. Using a set of spatial domain image features as training input, experimental results on three standard IQA databases show that MTLBIQ produces higher prediction performance than the BIQA model of the same image features across different distortion classes. MTLBIQ can also identify distortion of the image, which can be beneficial for different image processing applications particularly in image enhancement and image restoration systems.

6.2 Limitations and Future Work

Despite the promising results obtained by the three presented models, there are other steps that could be taken to allow for further models' extension and future research. Besides limitations identified at the end of Chapters 4 and 5, other limitations that the study know of are highlighted below:

- As most general-purpose BIQA models, the models presented here are developed through the luminance values analysis. Colour [153] is often a neglected factor which will further advance IQA research. Colour artefacts are among the most significant artefacts in image and video sequences [154], and cannot be ignored. Here, future work could include testing the models with colour images or developing an extended model that also incorporate colour information.

- The proposed models are developed to handle only images subjected to a single type of distortion. Certain applications can cause images to be concurrently subjected to multiple types of distortions. In such cases, we should consider the collective effects of these distortions on the image and the effects of these

distortions on each other. As demonstrated in previous studies on the joint effects of these distortions on image quality [155] - [156], multiple types of distortions can intuitively interact with each other when added to an image. Interaction with the image itself is also possible in ways that might be difficult to predict based on their physical combinations. Although PATCH-IQ and PATCH-IQ2 have been tested against an image of multiple distortions, both models treat these distortions as one type of distortion and do not take the interaction between the distortions into account. There were previous attempts [157] – [158] to determine the quality of a multiply distorted image. The BIQA models were developed based on combination of several image processing blocks to mimic the HVS image perceiving process. Competitive performance is reported as compared to single distortion based models though higher performance can be produced by replacing one or several blocks with more powerful models. We can take similar approach to incorporate these multiple distortions' factors into the presented models, making them more generally applicable.

- The presented models use handcrafted features and operate on a shallow learning architecture where a massive cost of computational time and expert knowledge can incur. An alternative way to do this is by automatically learn features and perform quality estimation through deep architectures. One advantage of a deep architecture has over a shallow architecture is that some highly non-linear functions can be expressed more compactly in terms of the number of parameters with deep architectures. The curse of dimensionality affecting shallow architectures is also addressed by deep architectures through distributed representations [159]. Deep learning has been successfully applied to other application domains such as audio classification [160] and image retrieval [161]. The proposed models could be

modified to incorporate deep learning architectures such as deep belief network and deep convolutional network. While this thesis is being written, initial work utilising deep architecture in designing BIQA model can already be seen in the two latest papers [162] - [163].

- The three models are designed for distorted images where they are based on the assumption that a good quality image is most identical to the original image. Yet, the concept of similarity is less applicable in the images such as artwork images [164], fused images [165] or user-generated images [166]. Although these images are outside the scope of this study, establishing different databases and learning strategies for accessing quality would be valuable for further understanding on how human observe and rate quality when dealing with such images.

6.3 Conclusions

The research on IQA, in particular BIQA, has seen tremendous improvements over the past few years. The increased usage of BIQA metrics in image processing applications indicate that the metrics are gradually accepted as substitutes to traditional image metrics. While these advances and efforts should be applauded, the IQA researchers generally agreed to the fact that BIQA metrics that can perform reliably under a different range of situations are yet to be produced. Recognising that BIQA research is far from finished, the work in this thesis made another contribution to the research by introducing three BIQA models that operate under differently designed learning frameworks. The results demonstrate that the presented models have high correlation with subjective perceptual measures and have better prediction performances than some, in not all, BIQA models. However, the models are not without its own limitations. Given a massive range of image processing applications and the subjectivity

of human perception on image quality, further works are still needed before they can be accepted as reliable and universally applicable quality metrics.

References

[1] P. G. Engeldrum, "A theory of image quality: The image quality circle," *Journal of Imaging Science and Technology*, vol. 48, no. 5, pp. 446-456, 2004.

[2] G. Ciocca, S. Corchs, F. Gasparini, C. Batini, and R. Schettini, "Quality of images," in *Data and Information Quality: Dimensions, Principles and Techniques*, C. Batini and M. Scannapieco, Eds. Switzerland: Springer International Publishing, 2016, pp. 113-135.

[3] G. Ciocca, S. Corchs, F. Gasparini, and R. Schettini, "How to assess image quality within a workflow chain: an overview," *International Journal on Digital Libraries*, vol. 15, no. 1, pp. 1-25, 2014.

[4] A. Mittal, *Natural scene statistics-based blind image quality assessment in the spatial domain*, Master Thesis, The University of Texas, USA, 2011.

[5] H. Yin, *Multipurpose image quality assessment for both human and computer vision systems via convolutional neural network*, Master Thesis, University of Waterloo, Canada, 2017.

[6] S. S. Hemami and A. R. Reibman, "No reference image and video quality estimation: applications and human motivated design," *Signal Processing: Image Communication*, vol. 25, no. 1, pp. 469-481, 2010.

[7] A. Punchihewa and D. G. Bailey, "Artefacts in image and video systems: classification and mitigation," in *Proceedings of International Conference on Image and Vision Computing*, Auckland, New Zealand, 2002, pp. 197-202.

[8] International Communication Union, "New definitions for inclusion in recommendation ITU-T P.10/G.100," ITU-T, Rec. P.10 / G.100, 2008.

[9] A. K. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1989.

[10] B. Girod, “What’s wrong with mean-squared error?” in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 207-220.

[11] Z. Wang and A. C. Bovik, “Mean squared error: Love it or leave it? A new look at signal fidelity measures,” *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98-117, 2009.

[12] International Communication Union, “Methodology for the subjective assessment of the quality of television pictures,” ITU-R, Rec. BT.500-11, 2002.

[13] International Communication Union, “Subjective video quality assessment methods for multimedia applications,” ITU-T, Rec. P.910, 1999.

[14] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, *LIVE Image Quality Assessment Database Release 2*. Online: http://live.ece.utexas.edu/research/quality

[15] E. C. Larson and D. M. Chandler, “Most apparent distortion: Full-reference image quality assessment and the role of strategy,” *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 1-21, 2010.

[16] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008 – A database for evaluation of full-reference visual quality assessment metrics,” *Advanced Modern Radioelectronic*, vol. 10, no. 4, pp. 30-45, 2009.

[17] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. Jay Kuo, “Image database TID2013: peculiarities, results and perspective,” *Signal Processing: Image Communication*, vol. 30, no. 1, pp. 57-77, 2015.

[18] A. Ninassi, P. L. Callet, and F. Autrusseau, “Pseudo no reference image quality metric using perceptual data hiding,” *Human Vision and Electronic Imaging, Proceedings of SPIE*, vol. 6057, pp. 146-157, 2006.

[19] P. L. Callet and F. Autrusseau, *Subjective Quality Assessment IRCCYN/IVC Database*, Online: http://www.irccyn.ec-nantes.fr/ivcdb/

[20] D. M. Chandler and S. S. Hemami, “VSNR: a wavelet-based visual signal-to-noise ratio for natural images,” *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284‑2298, 2007.

[21] Z. M. P. Sazzad, Y. Kawayoke, and Y. Horita, *Image Quality Evaluation Database*, Online: http://mict.eng.u-toyama.ac.jp/mictdb.html

[22] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, “Which semilocal visual masking model for wavelet based image quality metric?”, in *Proceedings of the IEEE Conference on Image Processing*, San Diego, California, 2008, pp. 1180-1183.

[23] S. Tourancheau, F. Autrusseau, Z. M. P. Sazzad, and Y. Horita, “Impact of subjective dataset on the performance of image quality metrics,” in *Proceedings of the IEEE Conference on Image Processing*, San Diego, California, 208, pp. 365‑368.

[24] D. Jayamaran, A. Mittal, A. K. Moorthy, and A. C. Bovik, “Objective quality assessment of multiply distorted images,” in *Proceedings of the IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, 2012, pp. 1693-1697.

[25] D. Jayamaran, A. Mittal, A. K. Moorthy, and A. C. Bovik, *LIVE Multiply Distorted Image Quality Database*, Online: http://live.ece.utexas.edu/research/quality/live_multidistortedimage.html

[26] K. Gu, G. Zhai, X. Yang, and W. Zhang, “Hybrid no-reference quality metric for singly and multiply distorted images,” *IEEE Transactions on Broadcasting*, vol. 60, no. 3, pp. 555-567, 2014.

[27] K. Gu, G. Zhai, X. Yang, and W. Zhang, *Multiply Distorted Image Database*, Online: https://sites.google.com/site/guke198701/home.

[28] W. Sun, F. Zhou and Q. M. Liao, “MDID: a multiply distorted image database for image quality assessment,” *Pattern Recognition*, vol. 61, no. 1, pp. 153-168, 2017.

[29] W. Sun, F. Zhou and Q. M. Liao, *Multiply Distorted Image Database*, Online: http://www.sz.tsinghua.edu.cn/labs/vipl/mdid.html

[30] C. Vu, T. Phan, P. Singh, and D. M. Chandler, *Digitally Retouched Image Quality (DRIQ) Database*, Online: http://vision.okstate.edu/driq/

[31] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: from quality assessment to automatic enhancement," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 284-297, 2016.

[32] C. Haccius and T. Herfet, "An image database for design and evaluation of visual quality metrics in synthetic scenarios," *Journal of Communication and Computer*, vol. 13, no. 1, pp. 351-365, 2016.

[33] C. Haccius and T. Herfet, *SSID – A Synthetic Image Database*, Online: http://www.nt.uni-saarland.de/SSID/

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

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

[36] D. Kundu, D. Ghadiyaram, A. C. Bovik and B. L. Evans, "Large-scale crowdsourced study for high dynamic range images," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4725-4740, 2017.

[37] D. Kundu, D Ghadiyaram, A. C. Bovik, and B. L. Evans, *ESPL-LIVE HDR Image Quality Database*, Online: http://signal.ece.utexas.edu/~debarati/HDRDatabase.zip

[38] L. Krasula, M. Narwaria, K. Fliegel and P. Le Callet, "Influence of HDR reference on observers preference in tone-mapped images evaluation," in *Proceedings of the IEEE Workshop on Quality of Multimedia Experience*, Pylos-Nestoras, Greece, 2015, pp. 1-6.

[39] L. Krasula, M. Narwaria, K. Fliegel and P. Le Callet, *PairComp TMO Database*, Online: http://ivc.univ-nantes.fr/en/databases/PairCompTMO/

[40] P. Hanhart, L. Krasula, P. Le Callet, and T. Ebrahimi, "How to benchmark objective quality metrics from paired comparison data?" in *Proceedings of the IEEE Workshop on Quality of Multimedia Experience*, Lisbon, Portugal, 2016, pp. 1-6.

[41] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: a survey," *Majlesi Journal of Electrical Engineering*, vol. 9, no. 1, pp. 55-83, 2014.

[42] W. Lin and C. –C. J. Kuo, "Perceptual visual quality metrics: a survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297-312, 2011.

[43] D. M. Chandler, "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Processing*, vol. 2013, no. 1, pp. 1-53, Nov. 2013.

[44] Z. 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*, vol. 13, no. 4, pp. 600-612, 2004.

[45] L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, 2011.

[46] R. Hong, J. Pan, S. Hao, M. Wang, F. Xue, and X. Wu, "Image quality assessment based on matching pursuit," *Information Sciences*, vol. 273, no. 1, pp. 196-211, 2014.

[47] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: a highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684-695, 2014.

[48] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430-444, 2006.

[49] C. Charrier, O. l'ezoray, and G. Lebrun, "Machine learning to design full-reference image quality assessment algorithm," *Signal Processing: Image Communication,* vol. 27, no. 3, pp. 209-219, 2012.

[50] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3378-3389, 2012.

[51] J. Wu, W. Lin, G. Shi, L. Li and Y. Fang, "Orientation selectivity based visual pattern for reduced reference image quality assessment," *Information Sciences*, vol. 351, no. 1, pp. 18-29, 2016.

[52] R. Soundararajan and A. C. Bovik, "Rred indices: reduced reference entropic differencing for image quality assessment," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 517-526, 2012.

[53] Z. Wang, "Applications of objective image quality assessment methods," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 137-142, 2011.

[54] M. Shahid, A. Rossholm, B. Lovstrom, and H. J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," EURASIP Journal on Image and Video Processing, vol. 40, no. 1, pp. 1-32, 2014.

[55] R. A. Manap and L. Shao, "Non-distortion-specific no-reference image quality assessment: a survey," *Information Sciences*, vol. 301, no. 1, pp. 141-160, 2015.

[56] R. A. Manap, A. F. Frangi, and L. Shao, "Blind image quality assessment via a two-stage non-parametric framework," in *Proceedings of the IAPR Conference on Pattern Recognition*, Kuala Lumpur, Malaysia, 2015, pp. 796-800.

[57] Q. Lu, W. Zhou and H. Li, "A no-reference image sharpness metric based on structural information using sparse representation," *Information Sciences*, vol. 369, no. 1, pp. 334-346, 2016.

[58] N. D. Narvekar and L. J. Karam, "No-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2678-2683, 2011.

[59] L. Li, D. Wu, J. Wu. H. Li, W. Lin, and A. C. Kot, "Image sharpness assessment by sparse representation," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1085-1097, 2016.

[60] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, "No-reference and robust image sharpness evaluation based on multi-scale spatial and spectral features," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1030-1040, 2017.

[61] Y. Zhan and R. Zhang, "No-reference image quality assessment based on blockiness and luminance change," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 760-764, 2017.

[62] L. Li, Y. Zhou, W. Lin, J. J. Wu, X. F. Zhang, and B. J. Chen, "No-reference quality assessment of deblocked images," *Neurocomputing*, vol. 177, no. 1, pp. 572-584, 2016.

[63] E. Cohen and Y. Yitzhaky, "No-reference assessment of blur and noise impacts on image quality," *Signal, Image and Video Processing*, vol. 4, no. 3, pp. 289-302, 2010.

[64] S. A. Golestaneh and D. M. Chandler, "No-reference quality assessment of JPEG images via a quality relevance map," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 155-158, 2014.

[65] J. Zhang and T. M. Le, "A new no-reference quality metric for JPEG2000 images," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, pp. 743-750, 2010.

[66] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513-516, 2010.

[67] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: from natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350-3364, 2011.

[68] 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 Network and Learning Systems*, vol. 24, no. 12, pp. 2013-2026, 2013.

[69] H. Tang, N. Joshi, and A. Kapoor, "Learning a blind measure of perceptual image quality," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, Colorado, 2011, pp. 305-312.

[70] Y. Chu, X. Mou, H. Fu, and Z. Ji, "Blind image quality assessment using statistical independence in the divisive normalization transform domain," *Journal of Electronic Imaging*, vol. 24, no. 6, pp. 1-20, 2015.

[71] L. He, D. Tao, X. Li and X. Gao, "Sparse representation for blind image quality assessment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2012, pp. 1146-1153.

[72] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 583-586, 2010.

[73] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: a natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3339-3352, 2012.

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

[75] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging*, vol. 22, no. 1, pp. 1-10, 2013.

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

[77] C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Transactions on Neural Network*, vol. 22, no. 5, pp. 793-799, 2011.

[78] P. Ye and D. Doermann, "No-reference image quality assessment using visual codebooks," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3129-3138, 2012.

[79] Y. Lu, F. Xie, T. Liu, Z. Jiang, and D. Tao, "No-reference quality assessment for multiply-distorted images based on an improved bag-of-words model," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1811-1815, 2015.

[80] Z. Gu, L. Zhang, and H. Li, "Learning a blind image quality index based on visual saliency guided sampling and Gabor filtering," in *Proceedings of the IEEE Conference on Image Processing*, Melbourne, Australia, 2013, pp. 186-190.

[81] M. Zhang, C. Muramatsu, X. Zhou, T. Hara, and H. Fujita, "Blind image quality assessment using joint statistics of generalized local binary pattern," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 207-210, 2015.

[82] F. Rezaie, M. S. Helfroush, and H. Danyali, "No reference image quality assessment using local binary pattern in the wavelet domain," *Journal of Multimedia Tools and Applications*, vol. 76, no. 1, pp.1-13, 2017.

[83] Q. Li, W. Lin and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 541-545, 2016.

[84] P. G. Freitas, W. Y. L. Akamine, and M. C. Q. Farias, "No-reference image quality assessment based on statistics of local ternary pattern'" in *Proceedings of Conference on Quality of Multimedia Experience*, Lisbon, Portugal, 2016, pp. 1-6.

[85] Q. Wang, J. Chu, L. Xu, and Q. Chen, "A new blind image quality framework based on natural colour statistics, *Neurocomputing*, vol. 173, no. 3, pp. 1798-1810, 2016.

[86] J. D. Foley, A. Van Dam, S.K. Feiner and J.F. Hughes, *Introduction to Computer Graphics*, Boston, MA: Addison-Wesley Publishing, 1994.

[87] K. Friston, J. Kilner, and L. Harrison, "A free-energy principle for the brain," *Journal of Physiology Paris*, vol. 100, no. 1, pp. 70-87, 2006.

[88] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 41-52, 2012.

[89] K. Gu, G. Zhai, X. Yang, W. Zhang, and L. Liang, "No-reference image quality assessment metric by combining free energy theory and structural degradation model," in *Proceedings of the IEEE Conference on Multimedia and Expo*, San Jose, CA, 2013, pp. 1-6.

[90] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50-63, 2015.

[91] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Providence, Rhode Island, 2012, pp. 1098-1105.

[92] L. Liu, L. Wang, and X. Liu, "In defence of soft-assignment coding," in *Proceedings of the IEEE Conference of Computer Vision*, Barcelona, Spain, 2011, pp. 2486-2493.

[93] Z. Hong, F. Ren, and Y. Ding, "Saliency-based feature learning for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Green Computing, Communications, Internet of Things, Cyber, Physical and Social Computing*, Beijing, China, 2013, pp. 1790-1794.

[94] W. Hou and X. Gao, "Saliency-guided deep framework for image quality assessment," *IEEE Multimedia*, vol. 22, no. 2, pp. 46-55, 2015.

[95] W. Hou, X. Gao, D. Tao and X. Li, "Visual saliency detection using information divergence," *Pattern Recognition*, vol. 46, no. 10, pp. 2658-2669, 2013.

[96] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 987-994.

[97] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, Ohio, 2014, pp. 1733-1740.

[98] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal on Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206-220, 2017.

[99] S. Bosse, D. Maniry, K. R. Muller, T. Weigand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206-219, 2018.

[100] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning: with applications in R*, New York: Springer, 2013.

[101] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199-222, 2004.

[102] J. P. Vert, K. Tsuda, and B. Scholkopf, "A primer on kernel methods," in *Kernel Methods on Computational Biology*, Cambridge, MA: MIT Press, 2004, pp. 35-70.

[103] F. M. Ciaramello and A. R. Reibman, "Systematic stress testing of image quality estimators," in *Proceedings of the IEEE Conference on Image Processing*, Brussels, Belgium, 2011, pp. 3101-3104.

[104] International Communication Union, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," ITU-T, P.1401, 2012.

[105] D. Sheskin, *Handbook of parametric and nonparametric statistical procedures*, London: Chapman & Hall, 2004.

[106] M. A. Saad, P. Le Callet, and P. Corriveau, "Blind image quality assessment: unanswered questions and future directions in the light of consumers' needs," *VQEG eLetter*, vol. 1, no. 2, pp. 62-66, 2014.

[107] International Communication Union, "Method for specifying accuracy and cross-calibration of video quality metrics," ITU-T, J. 149, 2004.

[108] L. Krasula, P. Le Callet, K. Fliegel, and M. Klima, "On the accuracy of objective image and video quality models: new methodology for performance evaluation," in *Proceedings of the Conference on Quality of Multimedia Experience*, Lisbon, Portugal, 2016, pp. 1-6.

[109] L. Krasula, P. Le Callet, K. Fliegel, and M. Klima, "Quality assessment of sharpened images: challenges, methodology, and objective metrics," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1496-1508, 2017.

[110] J. Kim, B. –S. Kim, and S. Savarese, "Comparing image classification methods: K-nearest-neighbour and support-vector-machines," in *Proceedings of the WSEAS Conference on Computer Engineering and Applications*, Cambridge, MA, 2012, pp. 133-138.

[111] R. Caruana, "Multi-task learning," *Machine Learning*, vol. 28, no. 1, pp. 41-75, 1997.

[112] A. Chetouani, A. Beghdadi, A. Bouzerdoum, and M. Deriche, "A new scheme for no-reference image quality assessment," in *Proceedings of the Conference on Image Processing Theory, Tools and Applications*, Istanbul, Turkey, 2012, pp. 270-274.

[113] A. Chetouani, A. Beghdadi, and M. Deriche, "A hybrid system for distortion classification and image quality evaluation," *Signal Processing: Image Communication*, vol. 27, no. 9, pp. 948-960, 2012.

[114] A. Chetouani, "Neural learning-based image quality metric without reference," in *Proceedings of the Conference on Image Processing Theory, Tools and Applications*, Paris, France, 2015, pp. 1-6.

[115] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems Video Technology*, vol. 5, no. 1, pp. 52-56, 1995.

[116] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proceedings of the IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, 2003, pp. 1398-1402.

[117] R. A. Manap, L. Shao, and A. F. Frangi, "A non-parametric framework for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Signal and Information Processing*, Orlando, FL, 2015, pp. 562-566.

[118] O. Boiman, E. Shechtman, M. Irani, "In defense of nearest-neighbor based image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008, pp. 1-8.

[119] A. Ng, *Supervised learning*, CS229: Machine Learning Lecture Notes, Stanford University, USA, 2014.

[120] R. A. Manap, L. Shao, and A. F. Frangi, "Non-parametric quality assessment of natural images," *IEEE Multimedia*, vol. 23, no. 4, pp. 22-30, 2016.

[121] C. –C. Chang and C. –J. Lin, *LIBSVM: A library for support vector machines* [Online]. Available: https://www.csie.ntu.edu.tw/~cjlin/libsvm

[122] C. –C. Chang and C. –J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1-27, 2011.

[123] T. Tuytelaars, "Dense interest points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, 2010, pp. 2281-2288.

[124] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the IEEE Conference on Computer Vision*, Kyoto, Japan, 2009, pp. 2106-2113.

[125] I. Sobel, "An isotropic 3x3 gradient operator," in *Machine Vision for Three-Dimensional Scenes*, H. Freeman, Ed. New York, NY: Academic Press, 1990, pp. 376-379.

[126] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679-698, 1986.

[127] C. Harris and M. Stephens. "A combined corner and edge detector," in *Proceedings of the Fourth Alvey Vision Conference*, Manchester, UK, 1988, pp.147-151.

[128] S. Smith and J. Brady, "SUSAN—A new approach to low level image processing," *International Journal of Computer Vision*, vol. 23, no.1, pp.45-48, 1997.

[129] L. G. Minor and J. Sklansky, "Detection and segmentation of blobs in infrared images," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 11, no. 3, pp. 194-201, 1981.

[130] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp.79-116, 1998.

[131] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Journal on Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[132] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, "A comparative study of SIFT and its variants," *Measurement Science Review*, vol. 13, no. 3, pp. 122-131, 2013.

[133] E. Niebur and C. Koch, “Computational architectures for attention,” in *The Attentive Brain*, R. Parasuraman, Ed. Cambridge, MA: MIT Press, 1998, pp. 163-186.

[134] A. Toet, “Computational versus psychophysical bottom-up image saliency: a comparative evaluation study,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2131-2146, 2011.

[135] A. Borji and L. Itti, “State-of-the-art in visual attention modelling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185-207, 2013.

[136] X. Hou and L. Zhang, “Saliency detection: a spectral residual approach,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1-8.

[137] R. Raj Kumar, P. Viswanath, and C. S. Bindu, “An approach to reduce the computational burden of nearest neighbour classifier,” *Procedia Computer Science*, vol. 85, no. 1, pp. 588-597, 2016.

[138] L. I. Kuncheva, “Reducing the computational demand of the nearest neighbour classifier,” in *School of Informatics Symposium on Computing*, Aberystwyth University, Aberystwyth, UK, 2001, pp. 61–64.

[139] J. Chen, L. Tang, J. Liu, and J. Ye, “A convex formulation for learning shared structures from multiple tasks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1025-1038, 2013.

[140] D. Zhang and D. Shen, “Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer’s disease,” *Neuroimaging*, vol. 59, no. 2, pp. 895-907, 2012.

[141] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer, “Multi-task learning for HIV therapy screening,” in *Proceedings of the International Conference on Machine Learning*, Helsinki, Finland, 2008, pp. 56-63.

[142] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243-272, 2008.

[143] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 2004, pp. 109-117.

[144] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," *Advances in Neural Information and Processing Systems*, vol. 19, no. 1, pp. 41-48, 2006.

[145] P. Gong, J. Ye, and C. Zhang, "Multi-stage multi-task feature learning," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2979-3010, 2013.

[146] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," *Advances in Neural Information and Processing Systems*, vol. 24, no. 1, pp. 702-710, 2011.

[147] J. Zhou, J. Chen, and J. Ye, "Multi-task learning: theory, algorithms and applications," *SDM Tutorial, SIAM Conference on Data Mining*, Anaheim, CA, 2012, pp. 1-83.

[148] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American Control Conference*, Arlington, VA, 2001, pp. 4734-4739.

[149] Y. Nesterov, *Introductory lectures on convex optimization: a basic course*, Kluwer Academic Publishers, Boston, 2004.

[150] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proceedings of the International Conference on Machine Learning*, Montreal, Canada, 2009, pp. 457-464.

[151] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Journal on Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[152] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-task learning via structural regularization*, Arizona State University, 2012 [Online]. Available: http://www.public.asu.edu/~jye02/software/MALSAR

[153] M. Fairchild, *Color Appearance Models*, 2nd ed., Wiley IS&T, Chichester, UK, 2005.

[154] J. Xia, Y. Shi, K. Teunissen, and I. Heynderickx, "Perceivable artifacts in compressed video and their relation to video quality," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 548–556, 2009.

[155] V. Kayargadde and J. B. Martens, "Perceptual characterization of images degraded by blur and noise: experiments," *Journal of the Optical Society of America A*, vol. 13, no. 6, pp. 1166-1177, 1996.

[156] D. M. Chandler, K. H. Lim, and S. S. Hemami, "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," *SPIE Human Vision and Electronic Imaging,* vol. 6057, no. 1, pp. 1-15, 2006.

[157] K. Gu, G. Zhai, M. Liu, X. Yang, and W. Zhang, "FISBLIM: a five-step blind metric for quality assessment of multiply distorted images," in *Proceedings of the IEEE Workshop on Signal Processing*, Taipei City, Taiwan, 2013, pp. 241-246.

[158] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Transactions on Broadcasting*, vol. 60, no. 3, pp. 555-567, 2014.

[159] L. Arnold, S. Rebecchi, S. Chevallier and H. P. Moisy, "An introduction to deep learning," in *Proceedings of the European Symposium on Artificial Neural Network*, Bruges, Belgium, 2011, pp. 477-488.

[160] H. Lee, Y. Largman, P. Pham, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Advances in Neural Information Processing Systems*, vol. 22, no, 1, pp. 1096-1104, 2009.

[161] J. Wan, D. Wang, S.C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: a comprehensive study," in *Proceedings of the ACM Conference on Multimedia*, Orlando, FL, 2014, pp. 157-166.

[162] Y. Li, L. M. Po, L. Feng, and F. Yuan, "No-reference image quality assessment with deep convolutional neural networks," in *Proceedings of the IEEE Conference on Digital Signal Processing*, Beijing, China, 2016, pp. 685-689.

[163] H. Oh, S. Ahn, J. Kim, and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4923-4936, 2017.

[164] M. Wang, R. Hong, X. T. Yuan, S. Yan, and T. S. Chua, "Movie2Comics: towards a lively video content presentation," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 858-870, 2012.

[165] R. Hong, W. Cao, J. Pang, and J. Jiang, "Directional projection based image fusion quality metric," *Information Sciences*, vol. 281, no. 1, pp. 611-619, 2014.

[166] Y. Yang, X. Wang, T. Guan, J. Shen, and L. Yu, "A multi-dimensional image quality prediction model for user-generated images in social networks," *Information Sciences*, vol. 281, no. 1, pp. 601-610, 2014.