



Faculty of Engineering

**Full Reference Objective Video Quality Assessment with Temporal
Consideration**

Loh Woei Tan

**Master of Engineering
2018**

Full Reference Objective Video Quality Assessment with Temporal Consideration

Loh Woei Tan

A thesis submitted

In fulfilment of the requirements for the degree of Master of Engineering

(Electronic and Computer Engineering)

Faculty of Engineering
UNIVERSITI MALAYSIA SARAWAK
2018

DECLARATION

The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Name:

Signature:

Date:

DEDICATION

Dedicated to my beloved parents, family members and friends.

ACKNOWLEDGEMENT

First and foremost, my great appreciation to my supervisor Ir Dr. David Bong Boon Liang for his excellent guidance, encouragement and advices throughout the entire duration of this project. His wide knowledge and logical way of thinking have provided excellent value for me. He had taught me a lot of things to consider especially when preparing this thesis and this helped to release some of my burden in preparing this thesis.

I would like to express appreciation to Universiti Malaysia Sarawak for providing Zamalah Siswazah UNIMAS (ZSU) Scholarship during my period of study. The allowance enables me to focus on my study without worries about living expenses. I would also like to thank the Ministry of Higher Education Malaysia for providing my study fees under MyBrain15. Moreover, my work is also supported by Ministry of Higher Education Malaysia under the research grant number RACE/c(2)/1109/2014(17).

I wish to express my deepest gratitude to my beloved family members especially my parents who had given their outmost encouragement throughout this research. They had also given me support physically and mentally when I was busy in preparing the thesis and carrying out other activities associated to this research.

Lastly, I would like to thank my friends for exchanging ideas, discussing problems and sharing useful information together. They had helped me a lot by giving advices and ideas that help to finish the research in a better manner. Without the help and supports from them, my project would not have been successful.

ABSTRACT

Video quality assessment (VQA) is an extension of image quality assessment (IQA). A video is a series of images arranged in time sequence. Therefore, IQA methods can be used to assess videos quality. A video has three dimensional data; two for the spatial dimensions and one for the temporal dimension. IQA methods assess video quality by assessing spatial effects without the need to consider the temporal effects and distortions. This makes IQA methods inappropriate and maybe inaccurate for assessing video quality. In order to apply in real time scenarios, VQA methods have to be reliable and correlated well to the judgement of human visual system (HVS). Furthermore, they have to be computationally efficient to give fast results. Current VQA methods have good correlations with subjective scores but are high in terms of computational complexity. In this thesis, two VQA methods, Index1 and Index2, with lower computational complexity are proposed. Index1 deals with Just Noticeable Difference (JND) in both spatial and temporal parts of the video. For the temporal part, JND is combined with temporal information to account for temporal distortions. For Index2, it is based on the previous work of mean difference structural similarity index (MD-SSIM). The temporal part of Index2 deals with the variation of temporal information. Both of the proposed methods are then compared with state-of-the-art VQA methods in terms of performance and computational complexity. The proposed methods were found to have acceptable performance with lower computational complexity.

Keywords: quality assessment, temporal, just noticeable difference, video, computational complexity.

Penilaian Kualiti Video Objektif Rujukan Penuh dengan Pertimbangan Dimensi Masa

ABSTRAK

Penilaian kualiti video (VQA) ialah lanjutan daripada penilaian kualiti imej (IQA). Video ialah satu siri imej disusun dalam urutan masa. Oleh itu, kaedah IQA boleh menilai kualiti video. Video ada tiga dimensi, dua untuk dimensi ruang dan satu untuk dimensi masa. Kaedah IQA mengabaikan kesan dan gangguan masa. Pengabaian ini menjadikan kaedah IQA tidak tepat dan sesuai untuk menilai kualiti video. Untuk penggunaan situasi sebenar, kaedah VQA perlu mempunyai kaitan dengan sistem visual manusia (HVS). Selain itu, kaedah VQA perlu mempunyai kerumitan pengiraan yang rendah. Walaupun kaedah-kaedah VQA yang dicadangkan baru-baru ini mempunyai korelasi yang baik dengan skor subjektif, tetapi mereka mempunyai kerumitan pengiraan yang tinggi. Dalam tesis ini, dua kaedah VQA dengan kerumitan pengiraan yang lebih rendah telah dicadangkan. Salah satu kaedah yang dicadangkan menggunakan konsep perbezaan hanya diketahui (JND) dalam dimensi ruang dan masa. Bagi bahagian masa, JND digabungkan dengan informasi masa untuk mengambil kira gangguan masa. Bagi kaedah VQA kedua, ia berdasarkan kerja perbezaan purata struktur indeks persamaan (MD-SSIM). Bahagian masa kaedah VQA ini berkaitan dengan perubahan informasi masa. Dua kaedah yang dicadangkan telah dibandingkan dengan kaedah VQA yang dicadangkan oleh penyelidik lain dari segi prestasi dan kerumitan pengiraan. Dua kaedah yang dicadangkan telah terbukti mempunyai prestasi yang lebih baik dan kerumitan pengiraan yang lebih rendah.

Kata kunci: *penilaian kualiti, dimensi masa, perbezaan hanya diketahui, video, kerumitan pengiraan.*

TABLE OF CONTENTS

	Page
DECLARATION	i
DEDICATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
ABSTRAK	v
TABLE OF CONTENTS	vi
LIST OF TABLES.....	xi
LIST OF FIGURES.....	xiii
LIST OF ABBREVIATIONS.....	xv
LIST OF SYMBOLS	xviii
CHAPTER 1: INTRODUCTION	1
1.1 Overview	1
1.2 Problem statements	8
1.3 Objectives and scopes	10
1.4 Outline of the thesis	11
CHAPTER 2: LITERATURE REVIEW	13
2.1 Introduction.....	13

2.2 Video distortions and their causes	14
2.2.1 Video transmission	14
2.2.2 Video compression	17
2.2.3 Rapid motion	20
2.3 Subjective quality measurement	21
2.3.1 Brief introduction	21
2.3.2 DSCQS	22
2.3.3 SSCQE	23
2.3.4 DCR	24
2.3.5 PCom	25
2.3.6 SAMVIQ	25
2.3.7 ACR	26
2.3.8 Advantages and disadvantages of subjective quality measurement	26
2.4 Objective quality measurement	27
2.4.1 Brief introduction	27
2.4.2 Different types of objective VQA methods	27
2.5 Related IQA methods	28
2.5.1 Brief introduction	28
2.5.2 MSE/PSNR	29
2.5.3 SSIM	30
2.6 Existing objective full reference methods	33

2.6.1 Brief introduction	33
2.6.2 VQM Index	33
2.6.3 DVQ Index	35
2.6.4 MOVIE Index	37
2.6.5 Method by Sheikh and Bovik	39
2.6.6 VSSIM Index	42
2.6.7 FMSE Index	44
2.6.8 ViMSSIM Index	46
2.6.9 ViS ₃ Index	48
2.6.10 Review of Existing Methods	51
2.7 JND	53
2.8 Summary	54
CHAPTER 3: METHODOLOGY	55
3.1 Introduction.....	55
3.2 Index1 Method	56
3.2.1 Brief Introduction to Mod-MSE	56
3.2.2 Spatial part	58
3.2.3 Temporal part	62
3.2.4 Overall index of Index1	65
3.2.5 Comparison of Index1 and Mod-MSE	65
3.3 Index2 Method	66

3.3.1 Brief Introduction to MD-SSIM	67
3.3.2 Spatial part	68
3.3.3 Temporal part	71
3.3.4 Overall index of Index2	75
3.3.5 Comparison of Index2 and MD-SSIM	75
3.4 Performance Evaluation	76
3.5 Benchmark databases	77
3.5.1 LIVE Video Database	78
3.5.2 CSIQ Video Database	78
3.6 Experimental Settings	79
3.6.1 Experiment procedures of the proposed method	80
3.6.2 Experiment procedures of the performance evaluation	82
3.6.3 Overall flow of the experiment	84
3.7 Summary	85
CHAPTER 4: RESULTS, ANALYSIS AND DISCUSSION	86
4.1 Introduction.....	86
4.2 Results, discussion, and analysis for Index1 and Index2.....	86
4.2.1 Statistical performance with Existing Methods	87
4.2.2 Statistical performance of Index1	96
4.2.3 Statistical performance of Index2	98
4.2.4 Statistical performance for videos	101

4.2.5 Computational complexity	109
4.3 Summary	110
CHAPTER 5: CONCLUSION AND FUTURE WORKS	111
5.1 Overview	111
5.2 Summary of the proposed methods	111
4.2.1 Index1.....	112
4.2.2 Index2.....	112
5.3 Contributions of the research	113
5.4 Future Works	115
REFERENCES	116
APPENDIX	128

LIST OF TABLES

	Page
Table 1.1 Consumer primary online activities.....	4
Table 3.1 Comparison of Index1 and Mod-MSE	66
Table 3.2 Comparison of Index2 and MD-SSIM.	75
Table 3.3 Details of LIVE video database.	78
Table 3.4 Details of CSIQ video database.	79
Table 4.1 Overall Performance of VQA methods for all databases.	87
Table 4.2 SROCC of all distortions in LIVE database.	88
Table 4.3 CC of all distortions in LIVE database.	88
Table 4.4 SROCC of all distortions in CSIQ database.	90
Table 4.5 CC of all distortions in CSIQ database.	91
Table 4.6 RMSE of Index1 for LIVE and CSIQ databases.	94
Table 4.7 RMSE of Index2 for LIVE and CSIQ databases.	96
Table 4.8 SROCC of Index1 related VQA methods for LIVE videos.	97
Table 4.9 CC of Index1 related VQA methods for LIVE videos.	97
Table 4.10 SROCC of Index2 related VQA methods for LIVE videos.	98
Table 4.11 CC of Index2 related VQA methods for LIVE videos.	98
Table 4.12 SROCC of Index2 related VQA methods for CSIQ videos.	99
Table 4.13 CC of Index2 related VQA methods for CSIQ videos.	100

Table 4.14	SROCC and CC for H.264 compressed videos.	101
Table 4.15	RMSE of Index1 and Index2 for H.264 compressed videos.	102
Table 4.16	SROCC and CC for compressed videos.	104
Table 4.17	RMSE of Index1 and Index2 for compressed videos.	105
Table 4.18	SROCC and CC for medium distorted videos.	107
Table 4.19	RMSE of Index1 and Index2 for medium distorted videos.	108
Table 4.20	Computation complexity for all VQA Methods.	109

LIST OF FIGURES

	Page
Figure 1.1 Global number of users from 2000 to 2016.	2
Figure 1.2 Global Internet penetration rate from 2000 to 2016.	2
Figure 1.3 Expectation growth of mobile data by Cisco.	3
Figure 1.4 Statistics of mobile video viewing form Q4 2013 to Q4 2015.	4
Figure 2.1 Scale used in DSCQS.	23
Figure 3.1 General framework of proposed methods.	55
Figure 3.2 Mod-MSE workflow.	58
Figure 3.3 Effect of applying JND.	60
Figure 3.4 Magnification of Figure 3.3	60
Figure 3.5 Workflow of spatial Index1.	62
Figure 3.6 The concept of difference on two previous frames.	63
Figure 3.7 Workflow of temporal Index1.	65
Figure 3.8 MD-SSIM workflow.	68
Figure 3.9 Effect of applying modified weight.	70
Figure 3.10 Workflow of spatial Index2.	71
Figure 3.11 Examples of variation in temporal information	73
Figure 3.12 Temporal workflow of Index2.	74
Figure 3.13 Flow of experiment.	84
Figure 4.1 Scatter plot of Index1 for LIVE videos.	94

Figure 4.2	Scatter plot of Index1 for CSIQ videos.	94
Figure 4.3	Scatter plot of Index2 for LIVE videos.	95
Figure 4.4	Scatter plot of Index2 for CSIQ videos.	96
Figure 4.5	SROCC and CC for H.264 compressed videos.	102
Figure 4.6	Scatter plots of Index1 for H.264 compressed videos.	103
Figure 4.7	Scatter plots of Index2 for H.264 compressed videos.	103
Figure 4.8	SROCC and CC for compressed videos.	104
Figure 4.9	Scatter plots of Index1 for compressed videos.	105
Figure 4.10	Scatter plots of Index2 for compressed videos.	106
Figure 4.11	SROCC and CC for medium distorted videos.	107
Figure 4.12	Scatter plots of Index1 for transmission medium distorted videos.	108
Figure 4.13	Scatter plots of Index2 for transmission medium distorted videos.	108

LIST OF ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
ACR	Absolute Category Rating
ATI	Absolute Temporal Information
BIBS	Blind Image Blur Score
CC	Pearson Linear Correlation Coefficient
CSF	Contrast Sensitivity Function
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DMOS	Difference Mean Opinion Score
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DVD	Digital Versatile Disc
DVQ	Digital Video Quality
FMSE	Foveated Mean Squared Error
FR	Full Reference
GUI	Graphic User Interface
HDTV	High-Definition Television
HVS	Human Visual System
IIR	Infinite Impulse response
IP	Internet Protocol
IQA	Image Quality Assessment

JND	Just Noticeable Difference
MAD	Most Apparent Differences
MD-SSIM	Mean Squared Error Difference Structure Similarity
Mod-MSE	Modified Mean Squared Error
MOS	Mean Opinion Score
MOVIE	Motion-Based Video Integrity Evaluation
MPEG-2	Moving Picture Experts Group type
MSE	Mean Squared Error
MSSIM	Multiscale Structure Similarity Index
NR	No Reference
PCom	Pair Comparison
PC	Phase Congruency
PS	Program Segment
PSNR	Peak Signal to Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service
QP	Quality Parameters
RR	Reduced Reference
SAMVIQ	Subjective Assessment Methodology for Video Quality
SNR	Signal-To-Noise Ratio
SROCC	Spearman Rank-Order Correlation Coefficient
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structure Similarity Index
STS	Spatiotemporal Slices

TP	Test Presentation
TS	Test Session
VQA	Video Quality Assessment
VQEG	Video Quality Expert Group
VQM	Video Quality Metrics
VIF	Visual Information Fidelity
VSSIM	Video Structural Similarity index

LIST OF SYMBOLS

$\mathcal{C}$	Constant
$\mathcal{CC}$	Pearson Linear Correlation Coefficient Value of a Method
$c(x, y)$	Contrast Value of a Pixel
$\mathcal{C}(p)$	Contrast Map at Block
$\mathcal{D}$	Distorted Frame
$\mathcal{D}(b)$	Perceived Distortions
$\mathcal{D}(x, y)$	Pixel in Distorted Frame
F	Number of Frames
f_{min}	Minimum Number of Frames
$FSIM$	Feature Similarity Index for an Image
$FSIM_c$	Feature Similarity Index for a Color Image
G	Frequency Response
$GM(x)$	Gradient magnitude value of a Pixel
K	Constant
L	Dynamic Range
$I(x, y)$	Luminance Value of a Pixel
LA	Local Amplitude
lG	Logarithm Gabor Filter Response
$Lmap$	Local Statistical Difference Map
MAD	Most Apparent Differences Index of an Image
MAX	Maximum Value of a Pixel

$MDSSIM$	Mean Squared Error Difference Structure Similarity Index of a Video
$ModMSE$	Mod-Mean Squared Error Index of a Video
MSE	Mean Squared Error Value of a Video
$MSE(f)$	Mean Squared Error Value of a Frame
$MSE(x_f, y_f)$	Mean Squared Error Value of a Pixel
$MSSIM(x, y)$	Multiscale Structural Similarity Index of a Pixel
N	Total Number of Pixels
$PC(x)$	Phase Congruency value of a Pixel
$pv(f)$	Processed Video Frame
$PSNR$	Peak Signal to Noise Value of an Image
q	Result of Mapping Objective Score
R	Reference Frame
$R(x, y)$	Pixel in Reference Frame
rd	Respective Difference
$r(f)$	Result of Averaging f frames
$s(x, y)$	Structure Value of a Pixel
$SSIM(x, y)$	Structural Similarity Index of a Pixel
$SROCC$	Spearman Rank-Order Correlation Coefficient Value of a Method
$thres$	Threshold Value
$v(f)$	Frame of a Video
θ	Orientation Angle

κ	Kurtosis
μ	Mean Intensity
ξ	Local Distortion Visibility Map
σ	Standard Deviation
$\tilde{\sigma}$	Minimum Standard Deviation
ζ	Skewness
τ	Parameters to be fitted

CHAPTER 1

INTRODUCTION

1.1 Overview

According to Oxford Dictionaries (2016), video is defined as recording, reproducing, and broadcasting of moving visual images. A video is a series of images that are presented to an observer as time passes. The movement of objects and subjects in a scene causes the variation of the images. Thus, a video is also construed as a spatial and temporal extension of an image. An image is a two dimensional (2D) signal that forms spatial contents of a scene. On the other hand, a video is a three dimensional (3D) signal that forms spatial and temporal contents of a scene. Hence, a video has an additional temporal dimension compared with an image. The existence of this third dimension makes a video analysis much more complicated than an image. There are temporal effects and distortions which could subject a video to more perception changes and errors. Some common temporal distortions are ghosting (Zhao *et al.*, 2009), jerkiness (Xue *et al.*, 2015), and mosquito effect (Seshadrinathan & Bovik, 2009).

The popularity of video increases with the rapid growth of the internet users. This is due to the ease of accessibility and increasing internet speed, especially for the mobile internet. Figure 1.1 and Figure 1.2 show the growing number of users and internet penetration rate from the year 2000 to 2014 (Internet User, 2016). The number of users and penetration rate are also estimated for the years 2015 and 2016. Figure 1.3 shows the

expected growth of transmitted mobile data (Cisco, 2015). These figures show the rise of internet users globally for mobile data.

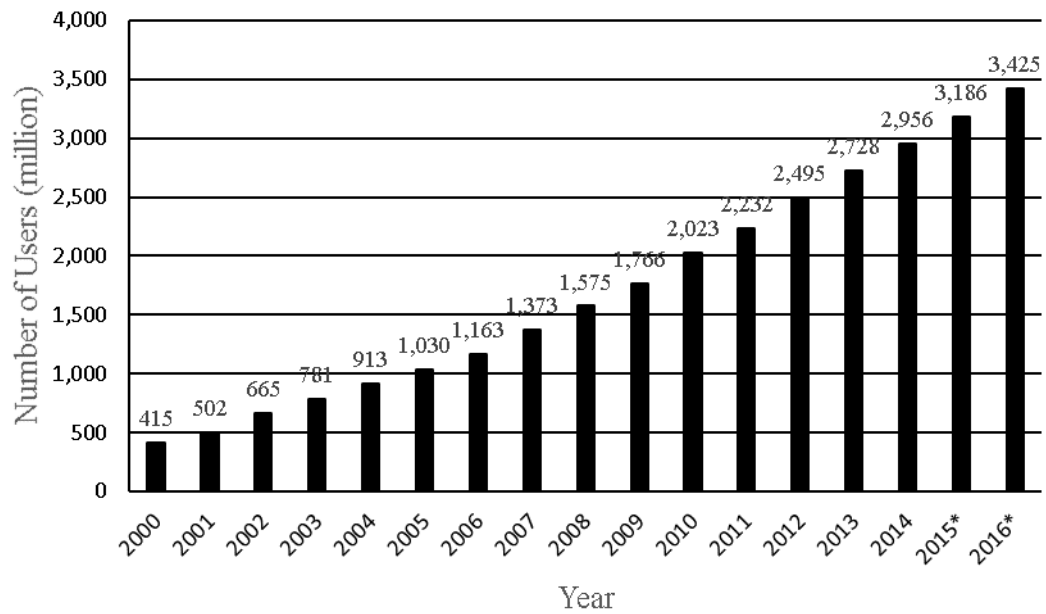


Figure 1.1: Global number of users from 2000 to 2016 (Internet User, 2016).

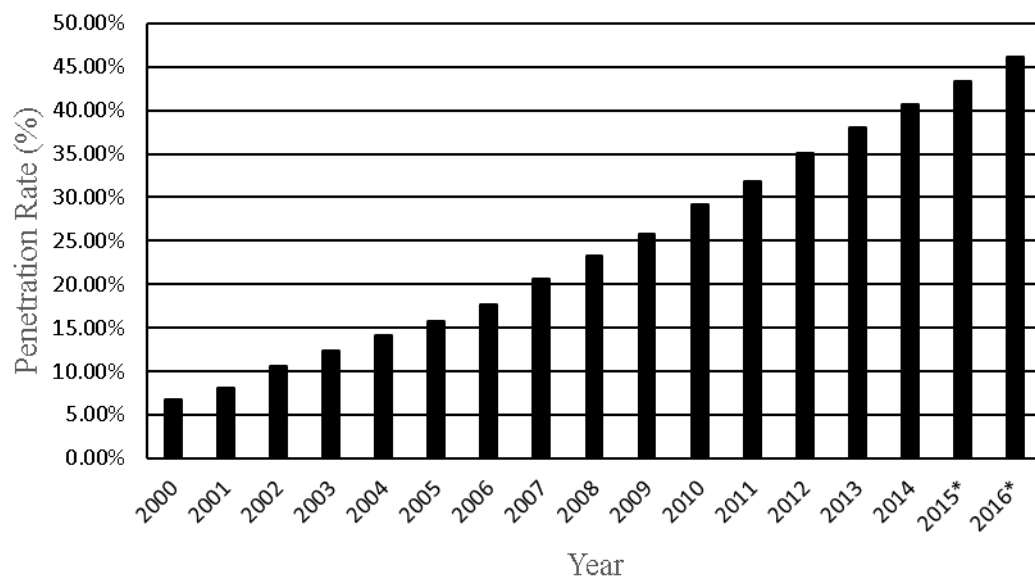


Figure 1.2: Global Internet penetration rate from 2000 to 2016 (Internet User, 2016).

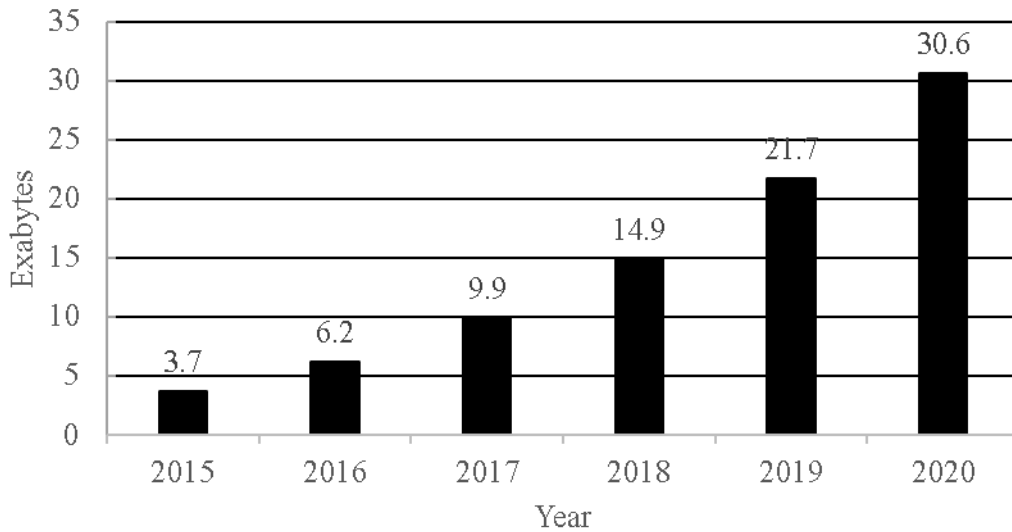


Figure 1.3: Expectation growth of mobile data by Cisco (2015).

According to Ooyala (2016), viewing of videos has also increased with the availability of internet, especially by using mobile devices. Almost half or 46 percent of the video viewing were done on mobile devices. The statistics that show the increase of mobile video viewing are shown in Figure 1.4. According to the report by Cisco (2015), 55 percent of the transmitted mobile data is accounted for video viewing in the year 2015. Cisco has forecasted that 75 percent of mobile data will be in the form of video data by 2020. This indicates that most of the internet activities are related to video viewing. Kaspersky (2015) has also showed this trend with their consumer security risk survey for the year 2015. The percentage of different online activities by consumers are shown in Table 1.1. In this table, the percentage of video viewing is highlighted. Video viewing has a high percentage as compared to other online activities. Thus, the statistics show that watching video is a major online activity.

Table 1.1: Consumer primary online activities (Kaspersky, 2015).

Activity	Any Device	Desktop/ Laptop Computer	Tablet	Smartphone
E-mail	85%	86%	77%	83%
Working	73%	76%	63%	58%
Using social media sites	67%	65%	61%	70%
Online shopping	66%	70%	48%	41%
Reading news, articles, books, etc.	66%	66%	62%	62%
Online banking	60%	61%	42%	44%
Watching movies / online videos	54%	54%	51%	67%
Downloading software / applications	51%	49%	42%	50%
Instant messaging / video calling	48%	44%	44%	51%
Education / learning	44%	46%	43%	29%
Using online payment systems / wallets	42%	44%	28%	27%
Online gaming	40%	37%	39%	37%
Uploading / sharing media	37%	36%	32%	33%
Online data storage	35%	35%	33%	27%
Visiting online dating websites	13%	12%	15%	9%
Other activities	1%	1%	1%	1%

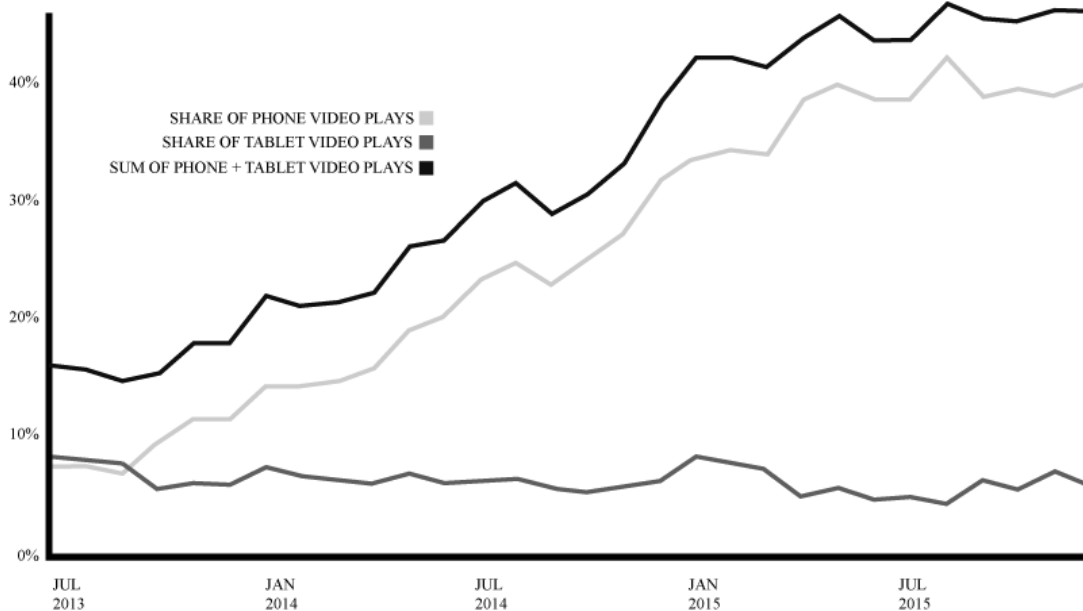


Figure 1.4: Statistics of mobile video viewing from Q4 2013 to Q4 2015 (Ooyala, 2016).

The prevalence of video activities directly trigger the need to have image and video processing tools to support these activities quantitatively and qualitatively. Compression, enhancement, watermarking, sampling, and encryption are examples of image and video processing. Compression of images and videos is needed to reduce their sizes so that they can be transmitted fast or stored easily. The reduction in size can be done by compressing some information that cannot be perceived by human visual system (HVS). Many protocols were released to standardize the compression processes. One of the famous video compression standards is Moving Picture Experts Group (MPEG) -2 (Aramvith & Sun, 2009), which is commonly used in digital versatile discs (DVD). On the other hand, enhancement process reproduces the image and video by increasing their quality. Image and video enhancements usually emphasize on contrast and illumination. Contrast is important for HVS to distinguish objects and subjects. It is perceived as sharp changes of illumination. An image or video is considered to have good contrast if the borders of all objects are sharp and clear (Bowker, 2013).

Watermarking is also an important part of image and video processing. It is usually related to copyright and authenticity. Watermarking in image and video helps to identify the authors and sources of the image and video. This is to avoid violation of intellectual property rights of the authors and sources. Meanwhile, digital sampling of image and video facilitates the processes of transmission, storage, processing, and display. A huge pile of data is hard to be transmitted at an instant due to bandwidth limitation. Thus, the data has to be split into smaller portions. Small portions of data help to mitigate the complexity in processing big data simultaneously. Encryption is used to protect the image or video from being eavesdropped by others, usually happens during the transmission. Data encryption is

important for security reason, especially for applications which use wireless transmission. The prevalence of wireless transmission brings up the importance of data encryption.

This thesis presents research works related to video quality assessment (VQA). VQA method enables a machine to measure video quality without human intervention. VQA enables machine automation with processing ability to measure video quality. This reduces the effort of getting human evaluators to manually rate the video quality. VQA methods facilitate the examination of the results after video processing. For instance, VQA methods indicate the efficiency of a transmission medium by comparing the video quality of a video before and after the transmission. The performance of a compression algorithm can also be tested with VQA methods. The VQA value of an image or a video before and after compression can be compared to indicate the efficiency of the compression algorithm.

The booming of video applications in daily life makes VQA important for enhancing the quality of service (QoS) and quality of experience (QoE) of the end users. The popularity of videos is evident by the prevalence of video sharing websites on the internet nowadays. Some of them are YouTube, Daily Motion, and Facebook. The proposed VQA methods in this thesis are objective quality assessment methods. Computational models are designed for the proposed VQA methods to predict the perceived video quality automatically by machine and without human intervention.

VQA methods are becoming more complicated nowadays. This may due to the progressive evolution of videos to higher resolutions. Many of the display devices are able to display images or videos with very fine resolution. For example, the display screen for Samsung Galaxy S7 (Samsung Electronics America, 2016) reaches resolution of 1440 x 2560 pixels. The increased capability of display screen requires better video quality to be rendered. This is because minor distortions can be easily detected by HVS.

Videos are subjected to many types of distortion. Video signals have to go through pre-processing stages before presenting to a human subject. These pre-processing stages include transmission, storing, conversion between analogue and digital signals, and compression. Each pre-processing stage would introduce different distortions to the content. The distortions would affect the perceived video quality. Therefore, VQA methods need to deal with all distortions to achieve a reliable measure of the video quality as on par with HVS (Winkler, 2005).

Besides that, VQA methods need to predict a video quality in a short period of time. This requirement ensures the practical applicability of VQA methods for real time scenarios. Computational time of a particular method is one of the main differences between different VQA or image quality assessment (IQA) methods. IQA methods normally require a shorter duration to assess image quality since they deal with only one image. On the other hand, VQA methods require hundreds or thousands fold of computational time compared to IQA methods. This is because videos are made up of multiple images in a time series. Thus, the computational time of VQA methods is directly proportional to the number of image frames in a video.

VQA methods play an important role in providing promised QoS and improving the end users' QoE (Karam *et al.*, 2009). VQA methods could predict the quality of the received video so that corrective actions can be taken. For instance, a re-transmitting process may be needed if the quality of the received video is not satisfactory.

In this thesis, two new VQA methods are proposed. The focus of the VQA methods are in their simplicity in terms of computational efficiency. Specifically, they are fast to compute and are applicable in real time scenarios. For instance, the quality monitoring for video traffic and reallocation of network resources with acceptable video quality. Most of

the existing VQA methods have high computational complexity (Narwaria *et al.*, 2012). This is corroborated in Chapter 4. Due to this reason, most of the VQA metrics are not practically applicable in a real-time scenario. A perfect VQA method shall have low computational complexity while correlates well to the subjective evaluation. However, the performance of a VQA metric is often directly proportional to its computational complexity. This is because more aspects are taken into consideration to produce a highly accurate VQA. A simpler method may neglect some important aspects that could affect the measurement of video quality. Thus, the proposed VQA methods in this thesis are trying to strike a balance between computational complexity and performance.

1.2 Problem statements

Direct method for assessing video quality is by implementing IQA on the frames of a video. The most commonly used IQA is mean squared error (MSE) or peak signal-to-noise ratio (PSNR). Video quality index of MSE or PSNR is found to have relatively low correlation with respect to subjective score (Winkler, 2005). This is because in MSE or PSNR, the behaviour of HVS is neglected while evaluating video quality (Winkler, 2005). Moreover, the temporal aspect of a video is not considered too.

All IQA methods do not take into consideration temporal aspects of a video. Thus, it is not surprising to observe that video quality indices from IQA methods do not correlate well with HVS. Thus, VQA methods need to assess video quality by considering both spatial and temporal aspects of a video.

Despite IQAs, there are also existing VQA methods for assessing video quality which are discussed in Section 2.6. Some existing VQA methods have good correlation to

subjective scores. The main drawback of the existing VQA methods is their feasibility in terms of real time application. Most of them use relatively complex methods and require more computation time to assess the quality of a video. For real-time applications, an ideal VQA should assess the quality of a video in a period of time that is equal to or shorter than the time required to transmit a video, and at the same time correlates equally well to HVS. This is difficult to achieve as more factors are needed to be taken into consideration to model a VQA method which can correlate well with HVS. Considering more factors would increase the complexity of a VQA method directly. Thus, VQA methods with acceptable performance while having low computational complexity is needed for real-time applications.

There are also attempts to accelerate the existing methods. Most of the attempts focus on the acceleration of frequency decomposition methods (Britanak *et al.*, 2010). The most common implementation is the Fast Fourier Transform (FFT) (Wang & Mersereau, 1999). There are also other methods based on algebraic properties of the decomposition methods (Hou, 1987; Yuan *et al.*, 2006). Besides that, there are also acceleration measures for specific IQA methods (Chen & Bovik, 2011; Phan, *et al.*, 2012). Other than that, hardware acceleration measures are also concerned (Wong, *et al.*, 2007; Unser, 1994). Most of them make use of general purpose graphical processing unit (GPGPU) (Fang, *et al.*, 2005; Tokdemir & Belkasim, 2011). These acceleration measures have their limitations on generalization. Frequency decomposition methods cannot be applied to methods that operate on spatial domain only. Similarly, IQA-specific acceleration measures cannot be applied on other methods that do not depends on the specific IQA method. For the case of hardware acceleration measures, their applications are limited. For instance, the mobile devices may not have capable hardware for acceleration. Therefore, it is of great value to

devise a VQA method with low computational complexity in terms of algorithm and have acceptable performance compared with existing methods.

1.3 Objectives and scopes

The objectives of this research are:

- (i) To propose new VQA methods with temporal consideration.
- (ii) To design a method with good performance and accuracy against existing methods.
- (iii) To design computational feasible VQA methods solely on software-based.

The scope of this research focuses on full reference objective VQA methods. The proposed VQA methods are divided into spatial and temporal parts. Spatial part aims to assess video quality due to spatial distortions while the temporal part is responsible for assessing video quality due to temporal distortions. The combination of spatial and temporal parts yields the final video quality index. The performance and computational complexity are the main considerations in developing the new VQA methods. The computational complexity is deemed as the measure of feasibility of the proposed method in real time applications where low computational time and acceptable performance are needed. The databases involved are LIVE video database and CSIQ video database. In total, there are nine types of distortions from both databases and one type of distortion can be found in both databases. Since the proposed methods work on the luminance part of a video frame, only the luminance part of the video frame is extracted and used to assess video quality. All computation, testing, and analysis are performed on Matrix Laboratory (MATLAB) version 2013b with Image Processing Statistics and Machine Learning toolboxes. The platform used is a laptop with Intel Core I5-4200U processor and 8GB

installed RAM. The operating system is 64-bit Window 7. No code optimization has been done to accelerate the method.

1.4 Outline of the thesis

This thesis is divided into five chapters. Chapter 1 introduces the background of this research. Some basic information about the popularity and importance of video is discussed in Section 1.1. Research fields on video processing and some relevant VQA metrics are also explained here. In Section 1.2, the needs for VQA methods are clarified. Aims and objectives of this thesis are presented in Section 1.3. Section 1.4 gives the outline of this thesis.

Chapter 2 reviews literature background of VQA and video distortions. Section 2.2 discusses common distortions found in video. The distortions are categorized to errors in video transmission, video compression, and errors due to rapid motion. Section 2.3 elaborates the subjective quality measurement. The benefits and drawbacks of subjective quality measurement are also deliberated here. Section 2.4 enlightens the objective quality measurement. Three different types of objective quality measurement are discussed here. Reviews of IQA methods for video quality assessment are presented in Section 2.5. Section 2.6 reviews some existing full reference objective VQA methods.

Chapter 3 explains the methodology of the proposed methods. There are two VQA methods proposed in this thesis. The first VQA method is deliberated in Section 3.2 while the second VQA method is deliberated in Section 3.3. The details of the proposed methods such as their formulas and concepts are explained in detail. Section 3.4 presents the metrics for evaluating the performance of VQA methods.

Chapter 4 shows the experiment results for the proposed VQA methods. The details of the testing databases are presented in Section 4.2. Two databases that involved are LIVE and CSIQ databases. The detailed analysis and discussion of the proposed VQA methods are conferred in Sections 4.3 and 4.4. Section 4.5 discusses the results of combining the two new VQA methods with existing state-of-the-art methods. Each part is discussed and analysed in detail to give a better insight about the proposed VQA methods.

Chapter 5 concludes this research. An overview of the thesis is given in Section 5.1. Section 5.2 presents a summary of the two new VQA methods. The conclusion of the thesis is presented in Section 5.3 while possible future works are suggested in Section 5.4.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Image and video signals are categorized as analogue or digital. In the early days, all the signals are in analogue form. Analogue signals are signals that are continuous with respect to time. On the other hand, digital signals are signals that are discrete with respect to time. In other words, digital signals are consisted of a series of 1s and 0s. Since the devices nowadays utilized digital signals, current technology is digitally based. Some benefits are ease of storage (Hilbert & López, 2011), security (Pal, 2011) and less bandwidth occupation (Gomez, 2002). Signals are analogue in nature. Thus, they have to be converted into digital form for further processing by existing devices.

Digital video is a time series of digital images or frames. Due to this reason, video signal processing is an extended field of technologies from image and signal processing. There are some similarities and differences among them. All types of distortion found in digital images can also be found in digital videos. Consequently, IQA can be used for VQA purpose. Some commonly used IQA metrics to measure video quality are MSE or PSNR, SSIM (Wang *et al.*, 2004), and multiscale structural similarity (MSSIM) (Wang *et al.*, 2003). These will be discussed in detail in Section 2.5. However, there are distortions that can only be found in videos. They are known as the temporal distortions. These will be further discussed in Section 2.2. The main difference between video and image is the existence of temporal distortion. Temporal distortion is the time effect of a video when one

frame transits to another frame as time passes. It causes IQA metrics unsuitable for measuring video quality due to the ignorance of the temporal effect.

IQA and VQA methods are subdivided into two categories. They are subjective and objective methods. Subjective VQA methods will be discussed in Section 2.3 while objective VQA methods will be deliberated in Section 2.4. Some commonly utilized IQA methods in assessing video quality will be discussed in Section 2.5. Section 2.6 reviews some existing state-of-the-art VQA methods.

2.2 Video distortions and their causes

This section presents the types of video distortions, especially the temporal distortion. Spatial distortions can be found in videos too. Some examples of spatial distortion are blocking, ringing, mosaic patterns, false contouring, and blur. Temporal distortions can further deteriorate the spatial distortion in the video frames. They can also cause additional visual alteration in the scenes. Temporal distortions are artifacts mainly caused by the pre-processing stages of videos, such as video acquisition, compression, filtering, and transmission. Three main causes of temporal distortions will be discussed in the Section 2.2.1, Section 2.2.2, and Section 2.2.3. They are the transmission, compression, and motion of objects. Some of these pre-processing steps can produce similar type of distortion.

2.2.1 Video transmission

Errors in the transmission of video signal across the medium can cause distortions in video. Some of them are bit errors, packet loss, and packet jitter. These errors are

common in wireless and Internet Protocol (IP) based networks. These are the common networks utilized for data transmission (Simpson & Greenfield, 2012).

Bit error (Simpson & Greenfield, 2012) are the distortions due to the unintended change of digital bits between senders and receivers. This type of errors arises due to a wide range of physical phenomena on any network. Bit errors are not only found in IP network but also in other networks such as fibre optic, satellite system, and wireless system. They affect the data utilized for picture construction. Some errors can be neglected if the errors do not cause perceivable annoyance to HVS. However, some errors cause distortions not only to one frame but also affect multiple frames of a video. Since bit errors occur randomly, there is no way to determine when bit errors would deteriorate the pictures or frames. There are various ways proposed by researchers to correct and minimize bit errors. Retransmission of the error packets is the most direct method.

Another error commonly found in IP network is packet loss (Simpson & Greenfield, 2012). There are many causes for packet loss. Some reasons are IP packet headers corrupted by bit errors, packets discarded due to congested links, and faulty networking equipment. There are multiple ways for solving packet loss. Similar to bit errors, resending of the packets can combat the issue. The technique of download and play delivery can replace the resending action. Utilization of large buffers also reduces packet loss. The remaining options are error concealment in the decoder and the employment of packet-level forward error correction (Simpson & Greenfield, 2012).

The next error commonly found in the transmission of video is packet jitter (Simpson & Greenfield, 2012). Packet jitter is caused by the disturbed order of packets arriving at the receiver. Jitter may not be an issue for other online resources other than videos such as email or Web page. This is because the information is not time-based. The

condition differs for video as video data involve temporal information. In other words, the data of a video have to be received in the order that is transmitted by the sender.

Jitter is a big problem for real time video applications such as video streaming. Jitter causes distortions to both uncompressed and compressed video data. For compressed video data, the clock information is transmitted with the packets. The clock information is the main data for a decoder to decompress the compressed video data. There can be variable influences on the video signal when the clock information is altered. Two main ways to overcome jitters are prevention and buffering. In jitter prevention, the packets are ensured not to delay during transmission. This is achieved by utilizing a better algorithm for routing and handover during transmission. Hardware improvement can be made for jitter prevention. Two examples are adding jitter suppression elements to receiving devices (Monma & Oishi, 2007) and improving clocking circuitry (Bridgewater *et al.*, 2009). For the buffering method, jitter buffers are required. A jitter buffer temporarily stores incoming packets for minimizing delay variations (Joseph & Chapman, 2009). It is limited by the tolerable delay. Packets arriving too late than the threshold are discarded.

One example of the distortions due to bit error, packet loss, and packet jitter is jerkiness. Jerkiness is an impairment caused by temporal aliasing. Possible temporal aliasing is due to delay or information loss during the transmission over an error-prone medium (Xue *et al.*, 2015). Jerkiness can cause videos to have repeated frames, frame loss, and display a frame for a long time (freezing) (Borer, 2010). Jerkiness will affect the end-user experience while watching a video because the smoothness of the video is affected.

The errors in transmission medium could also result in blocking artifact. This is due to the information loss around the borders of the objects. These artifacts are further elaborated in Section 2.2.2.

2.2.2 Video compression

For video compression stage, distortions happen due to the loss of information. This is common as the purpose of compression is to reduce unnecessary information for the ease of transmission. There are many artifacts produced in a compressed video. Each artifact and its causes are discussed individually.

The first artifact is the blocking artifact. It is caused by motion and motion compensation mismatches. Blocking artifact in a compressed video is mainly due to the partitioning of frames into macro-blocks. Macro-blocks are coded separately. However, there is a spatial correlation between them. Thus, visible edges could appear at macro-block borders. At most of the time, blocking artifact is easily spotted by the HVS (Wu & Rao, 2005). This is because a regular structure which does not belong to the image changes the visual content (Erol *et al.*, 2009). There are several ways to compensate for blocking artifact. Some examples are the method by Oosa (1998) and the method by Triantafyllidis *et al.* (2002). However, the deblocking process can cause blurring exists at borders.

There are also blurring effect in a compressed video. Blurring artifact may due to the low pass filtering in the deblocking process. Another possible cause of blurring artifact is quantization. This happens when high frequency components are quantized to zero. The results are similar to low pass filtering. Usually, quantization with high quantization parameter will cause quantization of high frequency components (Unterweger, 2012). According to Wu and Rao (2005), both quantization algorithms in MPEG-2 and H.264/AVC videos can cause blurring effect. This is due to the lack of built-in filter for compensating blurring effect.

Image or video sharpening can reduce blurring effect. However, the strength of sharpening has to be determined carefully. Over-sharpening of images or videos causes

noise. Some examples of sharpening are the method by Ng *et al.* (1999) that depends on boundary conditions and method by Biemond *et al.* (1990) that involves inverse filtering. The deblurring process may deblur the parts of an image or video that are blur due to the motion. The compensation of motion blur is different and involves deconvolution. One of the examples is the method by Nayar and Ben-Ezra (2004).

Mosquito effect is defined as a fluctuation of brightness (Seshadrinathan & Bovik, 2009) near the high-frequency components (Hosam, 2012). Some examples of high-frequency components are moving edges and contours. They are commonly found in compressed videos and are due to coded block by inter-frame prediction. Predicted moving object only constitutes part of the predicted block. The static scene changes due to the prediction and hence results in prediction error. However, the prediction from the static scene tends to reduce the error from moving object. The error reduction is more visible at the border of the objects. Consequently, high frequency noise is produced. Apart from the prediction error, quantization error also contributes to high concentration of error in the high frequency components. If the high frequency noise persists for several frames, it is known as mosquito noise (Unterweger, 2012).

Mosquito noise can be easily perceived due to variation in prediction error. The variation results in changing of frequency noise patterns. Mosquito effect may also due to the diverse coding of the same picture region across several pictures. For MPEG-2 video, mosquito noise can be caused by a drift in encoder and decoder. The drift is caused by finite precision of the floating point in the transform and inverse transform process. The imperfect reconstruction results in a drift (Wiegand *et al.*, 2003).

Another common form of artifact found in a compressed video is ringing artifact. The ringing artifact is the halos-shaped noise around sharp edges (Wu & Rao, 2005). This

is due to the quantization at steep edges. Steep edges comprised frequencies of bigger range. The sum of basic functions in quantization is inadequate for reconstruction. A less steep slope or both over- and undershooting are shaped at the samples around the original edge. It is also called Gibbs Phenomenon in one-dimensional Fourier analysis. For the case of a less steep slope, blurriness occurs as a result of the high frequency component errors. Over- and undershooting will introduce silhouette-like shade parallel to the original edge. Ringing artifacts are also present in high bit rate coded videos. This is because low quantization parameters are affected by the ringing effect. Approximation of an edge is not improved by utilizing large number of non-zero high frequency basis functions. Therefore, ringing around sharp edges is not reduced (Unterweger, 2012).

For compensating ringing artifact, there are two methods available. One of them is the sophisticated approach based on projections onto convex sets (POCS) (Zakhor, 1992). Another method depends on edge detection and adaptive filtering (Park & Lee, 1999). However, it is applicable to all DCT and quantization based video coding standards that utilize a transform size of 8x8 samples.

Staircase artifact (Unterweger, 2012) is another visual artifact that is closely related to ringing artifact. Staircase artifact refers to the inaccuracy of representation of diagonal edges by the horizontal and vertical basis. As a result, horizontal or vertical basis functions become visually prominent (Wu & Rao, 2005). Diagonal edges may look like staircase across multiple coded macro-blocks. An edge should be a smooth diagonal connection between two points. The blocking effect further impresses staircase artifacts as the staircase step size is equal to the size of a macro-block.

Basis pattern artifact (Unterweger, 2012) arises if there is single non-zero coefficient in a transformed block. The coefficient is caused by quantization. The artifacts

are very similar to staircase artifacts. However, basic pattern artifact has a single basis function with its prominent picture domain representation. This artifact arises when high quantization parameters are utilized. High quantization parameters increase the probability of only one non-zero coefficient.

2.2.3 Rapid motion

There are also distortions caused by the rapid motion in the videos. Dissimilar to the compressed video, they are motion compensation related artifacts. All the artifacts explained in this section are solely due to the motion found in videos.

Smearing in videos is defined as blurriness caused by the motion of objects (Verma *et al.*, 2014). When objects shift horizontally or vertically in a fast manner, the region between the original position and the destination position becomes blurred. HVS will actually perceive smearing as the loss of background information due to the motion of the objects.

Ghosting refers to trailing ghosts, which is the false detection of fast moving objects (Zhao *et al.*, 2009). Ghosting occurs in two conditions. One condition is the sudden change of motion velocity of a moving object. HVS will estimate the corresponding position of a moving object because it has the ability to judge the velocity of objects in a video. The sudden change in speed (slow down, increase speed or complete stop) will cause the ghosting effect. The other condition is the result of the sudden movement of stationary objects. When an object is static for some time, HVS will not be able to detect the sudden movement of the object when the object starts to move.

2.3 Subjective quality measurement

2.3.1 Brief introduction

The ultimate end user of images and videos are human or more accurately HVS. Thus, subjective quality assessment methods are methods for measuring the quality of images and videos performed by the human observers. They are usually carried out in a well-controlled environment using some standardized subjective assessment methods. Human observers are required to observe and rate the images or videos in the controlled environment. The rating results are then converted to subjective scores according to different rating standards of each method. Some subjective VQA methods are Double Stimulus Continuous Quality Scale (DSCQS) (ITU-R, 2003; Winkler, 2005), Single Stimulus Continuous Quality Evaluation (SSCQE) (Pinson & Wolf, 2003), Degradation Category Rating (DCR) (Aggarwal, 2014), Pair Comparison (PCom) (Singh & Aggarwal, 2014; ITU-T, 1999), Subjective Assessment Methodology for Video Quality (SAMVIQ) (ITU-R, 2007) and Absolute Category Rating (ACR) (ITU-T, 1999; Winkler & Mohandas, 2008). Some of these methods are discussed in this section. All the methods except DSCQS show their results in mean opinion score (MOS) which is the average of all subjective ratings. DSCQS shows its results in difference MOS (DMOS). Subjective quality assessment methods are the most reliable and accurate methods as human, the intended end user, are measuring the images or videos. However, they have limitations, such as time-consuming, costly and unable to conduct in a real-time environment (Singh & Aggarwal, 2014).

Although this thesis proposes objective VQA methods, subjective VQA methods play an important role in providing subjective scores of the videos in the benchmark database. The subjective scores are then used to measure the performance of the objective

VQA methods. This is done by utilizing rank order correlation coefficients between the subjective scores and objective scores. The understanding of subjective VQA methods is important for modelling of objective VQA methods. The analysis of subjective scores with objective scores also help in better understanding the benefits and drawbacks of the objective VQA metrics. The remaining subsections discuss the details of some subjective methods. They are DSCQS, SSCQE, DCR, PCom, SAMVIQ, and ACR.

2.3.2 DSCQS

DSCQS presents both reference and test sequences to the viewers. In a short period, assessors are shown multiple sequence pairs where each pair is consisted of a test sequence and its respective reference sequence. There are some sequence pairs consist of two reference sequences. Each sequence pair is presented twice in alternating fashion to the assessors. In each trial, sequence pairs are presented in a pseudo-random fashion. Assessors do not have the knowledge whether test sequence or reference sequence is presented. A continuous quality scale, which consists of five scales, is provided to the assessors to evaluate the sequences. The five scales are ‘Bad’, ‘Poor’, ‘Fair’, ‘Good’, and ‘Excellent’ (Winkler, 2005; ITU-R, 2003).

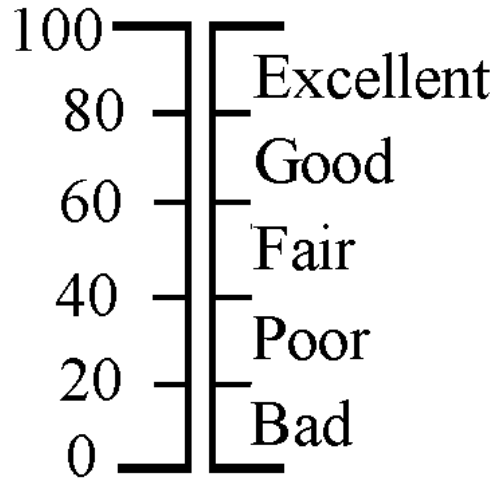


Figure 2.1: Scale used in DSCQS (Winkler, 2005).

The results of evaluation for both reference and test sequences are converted to normalized scores ranged from 0 to 100. Then, analysis based on the difference in rating for each pair is conducted. The results obtained from the DSCQS method are not the absolute scores, but differences in scores between a reference condition and a test condition. This is why the result of DSCQS is known as DMOS. DSCQS is a better method if the quality of test and reference sequence are similar. This is because this method is quite sensitive to small differences in quality (Winkler, 2005; ITU-R, 2003).

2.3.3 SSCQE

In real life scenario, human watch images or videos without references. SSCQE is proposed to mimic the situation. Instead of viewing separate short sequence pairs, the assessors need to watch a program for a longer duration. The program has been processed by the system under test. The program usually has three formats. They are program segment (PS), test session (TS), and test presentation (TP) (ITU-R, 2003). For PS, one program is being processed according to one of the quality parameters (QP) under

evaluation. The program is at least five minutes long. TS is the combination of all PS and QP in pseudo-random order with at least 30 to 60 minutes long. TP, on the other hand, is the combination of TSs. TP is used to achieve the maximum duration requirement.

In order to carry out SSCQE, an electronic recording handset with several characteristics must be used. The characteristics are fixed position, 10cm of linear range, record samples twice a second, and must have slider mechanism. Using a slider, the subjects can rate the immediately perceived quality continuously on the scale similar to DSCQS (Pinson & Wolf, 2003) shown in Figure 2.1.

2.3.4 DCR

DCR is also known as double stimulus impairment scale (DSIS) (ITU-T, 1999). The reference sequence is always shown before test sequence in DCR method. Assessor rates the test sequence with the knowledge of reference sequence. There are two possible variants for DCR. One of them is all the sequence pairs are presented to the assessors once. In another variant, assessors watch all sequence pairs twice. Assessors are being presented with a random sequence. The random sequence is distorted with random impairments to cover all the required conditions. A five-level scale is given to the assessors to rate the sequence. The scale, starts from the worst, are ‘very annoying’, ‘annoying’, ‘slightly annoying’, ‘perceptible, but not annoying’, and ‘imperceptible’. The DSIS method is suitable to measure clearly visible impairments (Pinson & Wolf, 2003).

2.3.5 PCom

In PCom method, test sequences with different conditions from the same scene are paired in all possible combinations. For example, there are three test sequences A, B, and C from the same scene to be presented. Assessors need to watch the sequence pairs of AB, BA, AC, CA, BC, and CB. Therefore, there are always $n(n-1)$ combinations where n is the total number of test sequences. Assessors need to rate the quality for each pair. Therefore, there are very fine quality differences between video clips. The number of replications is not that important for PC as there are repeated presentations of the same conditions in this method (ITU-T, 1999).

2.3.6 SAMVIQ

Assessors in SAMVIQ method are being presented several versions of both reference and test sequence. Graphic user interface (GUI) is needed in SAMVIQ method. With the help of GUI, assessors are able to pause, review and change the score of every version of a sequence. The versions of a sequence include the distorted sequences and the hidden reference sequence. Assessors are also able to assess the reference sequence any time. Thus, this method is implemented scene by scene with the provision of the reference sequence. The scale similar to the scale for rating in DSCQS is given to the assessors (ITU-R, 2007). The main idea of this method is that assessors are able to view the sequences in detail which give more accurate results.

2.3.7 ACR

For ACR method, assessors are asked to rate a sequence right after watching it. Particularly, the test sequences are presented one at a time. In order to get more accurate results, replications are made by repeating the same test conditions at different points of time in the test. Assessors are given a five-level scale similar to DSCQS for rating the sequence. However, a nine-level scale can also be used for higher discriminative power. Each sequence is rated independently on the same category scale (ITU-T,1999).

2.3.8 Advantages and disadvantages of subjective quality measurement

The main benefit of subjective quality measurement is its reliability. Subjective quality measurement is the most reliable method for assessing the video quality as the end user, human or HVS, is involved in rating the video quality. However, there are several drawbacks (Seshadrinathan *et al.*, 2010a). First of all, this method is time-consuming. A human subject is required to rate the video quality in this method for a period of time. There should be a large number of human subjects involved for higher accuracy. Moreover, this method is also costly to carry out. This is because the surrounding condition for conducting this method has to be standardized and noise free. Subjective quality measurement is also non-applicable in real-time scenario (Singh & Aggarwal, 2014). This is the main disadvantage of this method as it cannot be applied in daily life applications. Subjective quality measurement is also easily influenced by viewers' interest, type of devices, viewing conditions, methods used for evaluation and number of subjects (Pinson & Wolf, 2003). There are also other factors such as geographical location, types of people, culture, etc. that limits the applicability of this method.

2.4 Objective quality measurement

2.4.1 Brief introduction

As opposed to subjective quality measurement, human subjects are not involved in objective quality measurement. It is proposed to overcome the problems faced by subjective quality measurement such as time-consuming and unsuitable to be applied in real-time. An algorithm or mathematical model is used to automatically measure video quality. The algorithm simulates human perceptual system in order to assess the video quality. However, the accuracy of objective quality assessment methods is considered lower than the subjective methods. Since this thesis focuses on VQA, state-of-the-art objective VQA methods will be discussed instead of IQA methods. Objective VQA methods can be further divided into three types. They are full reference (FR) VQA, reduced reference (RR) VQA, and no reference (NR) VQA (Seshadrinathan & Bovik, 2009).

2.4.2 Different types of objective VQA methods

FR VQA methods are the methods that assess video quality with the condition that the reference video is available. This is the most simple and direct method as the impaired video can be compared directly to the reference video. However, this is not applicable in real life scene as reference videos are usually not available (Seshadrinathan & Bovik, 2009).

RR VQA methods measure the video quality with the provision of some features of the reference videos. This type of VQA methods is more applicable than FR VQA methods. However, the extracted features of the reference videos occupy more bandwidth during the transmission process. Moreover, more pre-processing steps are required to extract the

features (Seshadrinathan & Bovik, 2009). These steps would increase computational complexity of the VQA methods.

NR VQA methods or also known as blind VQA methods rate the video quality without any information from the reference video. This is the most practical VQA method. This is because NR VQA methods are similar to real life situation where no information of the reference video is available. However, the main drawback of these NR methods is bad accuracy. NR VQA methods usually do not correlate well with subjective scores (Seshadrinathan & Bovik, 2009).

2.5 Related IQA methods

2.5.1 Brief introduction

This section discusses IQA methods related to the proposed VQA method. Since a video is made up of a sequence of images, IQA methods can also evaluate video quality. Final video quality is taken as the average of IQA scores for every frame in the video. However, IQA methods do not consider of temporal distortion as there is no temporal effects and distortions in images. Thus, IQA methods are not accurate in assessing video quality (Winkler, 2005). There are many existing FR, RR, and NR IQA methods proposed by researchers. There are some methods focus on assessing specific distortion only such as compression-based distortion and blurriness. One of the examples is found in the work of Bong and Khoo (2014b) where an NR IQA method, Blind Image Blur Score (BIBS), was proposed to measure blurriness in images. The authors had also proposed other FR and NR IQA methods for detecting blurriness (Bong & Khoo, 2014a; Bong & Khoo, 2015). The proposed VQA methods in this thesis focus on FR method. Due the fact that VQA is the

extension of IQA, many researchers have proposed VQA methods which are based on the improvement or extension of some popular IQA methods such as MSE and SSIM.

Section 2.5.2 discusses MSE or its equivalence, PSNR. SSIM is presented in Section 2.5.3.

2.5.2 MSE/PSNR

MSE or PSNR (Wang & Bovik, 2009; Wolf, 2001) is the squared of the difference between the reference frame and the distorted frame on the pixel basis. This can be shown as:

$$MSE(f) = \frac{\sum_x \sum_y [R(x, y) - D(x, y)]^2}{N} \quad (2.1)$$

$$MSE = \frac{\sum_f MSE(f)}{F} \quad (2.2)$$

$$PSNR = 10 \log \left(\frac{MAX^2}{MSE} \right) \quad (2.3)$$

where $MSE(f)$ is the MSE value of a particular frame while MSE is the overall MSE value of a video with total number of frames, F . In Equation (2.1), $R(x, y)$ is a particular pixel in reference frame, R while $D(x, y)$ is the corresponding pixel in distorted frame, D . In Equation (2.3), $PSNR$ is the PSNR value of a video while MAX is the maximum value of a pixel.

MSE and PSNR are known for their inaccuracy in measuring video quality from the perspective of HVS. This is because the measurements ignore the fact that a particular pixel in a frame will be visually affected by its surrounding pixel (Akramullah, 2014). In

other words, they ignore the way HVS perceive (Damera-Venkata *et al.*, 2000) when rating a video quality. The disadvantages of MSE and PSNR are discussed in detail in these works (Zhang *et al.*, 2002; Wang, 2006; Fraser *et al.*, 2010; Lafferty, 2008; Channappayya, 2007). Despite their inaccuracy, MSE and PSNR are often used as benchmark methods to compare with other VQA methods. This is owing to their simplicity (Wang *et al.*, 2004) and short computation time. Through comparison with MSE and PSNR, researchers are able to have insight into the performance of their proposed VQA methods.

2.5.3 SSIM

SSIM is an IQA method proposed by Wang *et al.* (2004) which has become a de facto method nowadays. It requires short computational time as compared to other state-of-the-art metrics while maintains a good correlation with human subjective scores. The key idea of SSIM is the assumption made by the authors which stated that HVS is highly adapted to extract the structure of an image or scene. SSIM metric compares three components, i.e. luminance, signal contrast, and structure, of the reference and distorted images. According to (Wang *et al.*, 2004), the luminance comparison is defined as:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (2.4)$$

where μ_x and μ_y are defined as:

$$\mu_x = \frac{\sum_{a=1}^N x_a}{N} \quad (2.5)$$

$$\mu_y = \frac{\sum_{a=1}^N y_a}{N} \quad (2.6)$$

Parameters μ_x and μ_y are the mean intensities of two signals under comparison with N as the total number of pixels to be averaged while parameter C_1 is a small value added to Equation (2.4) to prevent instability. According to the authors, C_1 is defined as (Wang *et al.*, 2004):

$$C_1 = (K_1 L)^2 \quad (2.7)$$

where parameter K_1 is a small constant and parameter L is the dynamic range of the pixel values in the video.

The next component to be compared is the contrast. It is defined as (Wang *et al.*, 2004):

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (2.8)$$

where σ_x and σ_y are defined as (Wang *et al.*, 2004):

$$\sigma_x = \left(\frac{\sum_{a=1}^N (x_a - \mu_x)^2}{N - 1} \right)^{\frac{1}{2}} \quad (2.9)$$

$$\sigma_y = \left(\frac{\sum_{a=1}^N (y_a - \mu_y)^2}{N - 1} \right)^{\frac{1}{2}} \quad (2.10)$$

Parameters σ_x and σ_y are the standard deviations of the regions of pixels being compared.

Parameter C_2 is a small value, which is similar to C_1 . This is shown as (Wang *et al.*, 2004):

$$C_2 = (K_2 L)^2 \quad (2.11)$$

The final comparison is the structural component. It is defined as (Wang *et al.*, 2004):

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (2.12)$$

where σ_{xy} is defined as (Wang *et al.*, 2004):

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y) \quad (2.13)$$

In order to obtain the SSIM score of two images, the three comparisons are combined and is mathematically expressed as (Wang *et al.*, 2004):

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma \quad (2.14)$$

Parameters α , β and γ are numbers between zero and one for adjusting the weight of each component. For example, if luminance comparison is more important than the other two, parameter α can be adjusted to 1 while parameters β and γ can be adjusted to 0.5. This enables SSIM metric to adapt to different conditions. Wang *et al.* (2004) have proposed $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$. Therefore, the SSIM can be further simplified to:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2.15)$$

Direct implementation of SSIM on video achieves a better result than MSE but performs quite poor as compare to existing VQA methods. This may due to temporal effect is being ignored in SSIM. A proper redefinition of SSIM concept with the temporal information can help to improve the result of SSIM in assessing video quality.

2.6 Existing objective full reference methods

2.6.1 Brief introduction

Existing FR objective VQA methods are discussed in this section. Only general purpose FR objective VQA methods are discussed. Some other VQA methods that involve specific applications are excluded. Some examples are the work by Wang *et al.* (2017) on voice over internet protocol (VOIP), the work by Rodríguez *et al.* (2016) on streaming videos, and the work by Zhang *et al.* (2017) on online media. The methods discussed in this section can be sub-categorized into two parts. They are the methods improved from the IQA methods and explicit metrics proposed specifically for VQA. Section 2.6.2 to Section 2.6.5 involves non-IQA based VQA methods while Section 2.6.6 to Section 2.6.9 elaborate IQA based VQA methods.

2.6.2 VQM index

There are many existing FR objective VQA methods. Video Quality Metrics (VQM), proposed by Pinson and Wolf (2004) and Wolf and Pinson (1998), was perhaps the first non-IQA based video quality assessment. There are five main steps in VQM. The steps in the order are sampling of the reference and distorted videos, calibration of the reference and distorted videos, extraction of perception based features, computation of video quality parameters, and calculation of VQM models. During sampling and calibration processes, reference and distorted videos are divided into regions in terms of time and space.

There are seven important parameters in VQM. They are *si_loss*, *si_gain*, *hv_loss*, *hv_gain*, *chroma_spread*, *chroma_extreme*, and *ct_ati_gain*. The first four of them, *si_loss*, *si_gain*, *hv_loss* and *hv_gain*, are extracted from spatial gradients of the

luminance component. The next two parameters, *chroma_spread* and *chroma_extreme*, are based on features extracted from chrominance components. The last parameter, *ct_ati_gain*, is the product of features that measure contrast and motion. The contrast and motion information are both extracted from the luminance component.

In overall, five perception based features are extracted from the luminance component and the chrominance component respectively based on the seven parameters. These five perception based features are features based on spatial gradients, features based on chrominance information, features based on contrast information, features based on absolute temporal information (ATI), and features based on the cross product of contrast and ATI. A set of quality indices is formed by comparing the extracted quality features.

Parameter *si_loss* detects the variation of spatial content such as blurring. The values of *si_loss* are all in negative. The *hv_loss* parameter identifies shift of edges from horizontal and vertical to diagonal orientation. It utilizes the horizontally (H) and vertically (V) filtered images from the 13 pixel spatial information filter. Two filtered images with horizontal and vertical edges (HV) and diagonal edges ($\overline{HV}$) are created. Parameter *hv_gain* is a shift of edges from diagonal to horizontal and vertical orientation. Parameter *chroma_spread* detects alteration in the spread of the distribution of two-dimensional colour samples. Parameter *chroma_spread* measures colour impairments that are nearly always present. It also performs coherent integration of colour samples over an area large enough to have a significant perceptual impact.

Parameter *si_gain* is the sole quality improvement parameter in VQM. It measures quality upgrading from edge sharpening or enhancements. Parameter *ct_ati_gain* is the product of contrast feature and temporal information feature. Distortions are more apparent

in S-T regions with lower *ct_ati_gain* value and vice versa. This is particularly true of impairments like noise and error blocks. Parameter *chroma_extreme* applies similar colour features as parameter *chroma_spread*. The only difference is the collapsing function. It distinguishes severe localized colour impairments.

After determining the indices from all parameters, all of them are united linearly to form VQM index of a video. This is mathematically represented as (Pinson & Wolf, 2004):

$$Loss = -0.2097 \times si_loss + 0.5969 \times hv_loss \quad (2.16)$$

$$Gain = 0.2483 \times hv_gain - 2.3416 \times si_gain + 0.0431 \times ct_ati_gain \quad (2.17)$$

$$Chroma = 0.0192 \times chroma_spread + 0.0076 \times chroma_extreme \quad (2.18)$$

$$VQM = Loss + Gain + Chroma \quad (2.19)$$

2.6.3 DVQ index

There are non-IQA based VQA methods designed for videos by focusing on the models of human vision. Watson *et al.* (2001) have proposed Digital Video Quality (DVQ). DVQ is based on the concept of the just noticeable difference (JND) (USD Internet Sensation & Perception Laboratory, n.d.) which states that minor distortions in a scene cannot be detected by HVS. This implies that distortions which do not exceed a certain threshold value cannot be perceived by HVS. Discrete cosine transform (DCT) based model is used in this VQA method. The DCT coefficients resulted from DCT filtering are transformed into contrast score. Then, a model of contrast sensitivity function (CSF) is utilized to filter the DCT coefficients. Lastly, JND can be computed by dividing filtering results with their respective visual threshold. The details of the DVQ are elaborated.

First, the frames from reference and distorted videos are obtained. The frames are then gone through sampling and cropping to extract a series of regions of interest. The series of regions are expressed in a perceptual colour space through colour transformation. The perceptual colour space is YOZ colour space (Peterson *et al.*, 1993). Y accounts for luminance channel, O is colour opponent channel, and Z is blue channel. The colour transformation involves gamma transformation and linear colour transformation.

After that, the processed frames go through blocking and DCT. A blocked DCT is applied to each frame in each colour channel. The results from DCT are then transformed to local contrast. Local contrast is defined as the ratio of DCT amplitude to DC amplitude. After these operations, DCT coefficient results in numbers between -1 and 1. These values express the amplitude of the basis function as a fraction of the average luminance.

The next step is temporal filtering. A second order infinite impulse response (IIR) filter is utilized to reduce the number of frames of data retained in memory. The filtered coefficients are converted to just noticeable differences. This is done by dividing their respective spatial thresholds. After conversion, the resulting reference and distorted frames are deducted to produce a difference frame sequence. The following step is the contrast masking process. A masking sequence is constructed from the reference frames. This sequence is rectified and filtered by a first-order IIR filter. This is similar to the traditional contrast masking result where contrasts below threshold have no masking effect. On the other hand, contrasts above the threshold are raised based on the power of the mask. Lastly, resulting contrasts sequence go through a pooling process to form a video quality index. Minkowski pooling is utilized. It is mathematically shown as (Watson *et al.*, 2001):

$$S = \left(\sum_{i=1}^n S_i^m \right)^{\frac{1}{m}} \quad (2.20)$$

where S is the result of Minkowski pooling, S_i is the contrasts sequence and m is the power of mask used for uplift the relative contrast if the contrast is above the threshold.

2.6.4 MOVIE Index

Seshadrinathan and Bovik (2010) proposed a well-known non-IQA based VQA metric, motion-based video integrity evaluation (MOVIE). They integrated the statistics of the natural videos into MOVIE metric. Instead of measuring spatial and temporal video quality, this metric further involves spatiotemporal quality assessment. All videos have to be pre-processed by 3D Gabor filter-bank. The filter results are used for both spatial and temporal quality measurement. A model of contrast masking is combined with the filter results to calculate the spatial impairments. On the other hand, filter outputs are pooled with motion information to account for temporal distortions. The motion information is computed using optical flow estimation. Finally, the overall MOVIE score is the arithmetic average of both spatial and temporal quality scores. The details of the process are further deliberated.

For linear decomposition, a bandpass Gabor filter is utilized to filter the frames spatiotemporally. Three dimension Gabor filter is the result of Gaussian window modulated by complex exponential wave. There are three scales of Gabor filter used. Each scale has 35 filters respectively. Nine filters are tuned to a temporal frequency of 0 radians per sample which correspond to no motion. Seventeen filters are tuned to horizontal or vertical velocities of $1/\sqrt{3}$ pixels per frame. The last nine filters are tuned to horizontal or vertical velocities of $\sqrt{3}$ pixels per frame.

Computation of spatial MOVIE index is done by utilizing the output from Gabor filters in the previous part. The Spatial MOVIE index is able to capture the contrast masking property of human vision. Gabor subband errors and DC subband errors are then calculated. The two spatial errors are then pooled to form an error index as (Seshadrinathan & Bovik, 2010):

$$E_{spa}(i_0) = \frac{1}{B} \sum_{b=1}^B E_s(i_0, b) + E_{DC}(i_0) \quad (2.21)$$

$$Q_s(i_0) = 1 - E_{spa}(i_0) \quad (2.22)$$

Parameter E_s and E_{DC} are Gabor subband errors and DC subband errors respectively. Parameter b is the filter scale while i_0 is the spatiotemporal position in the video.

For motion estimation, algorithm from Fleet and Jepson (1990) is utilized for motion estimation. The vectors from reference and distorted videos, v^r and v^d are then computed. They are defined as (Seshadrinathan & Bovik, 2010):

$$v_l^r = \frac{(r_l - \mu_r)^2 + \sum_{b=1}^B \alpha_l(b) r_l(b)^2}{(r_l - \mu_r)^2 + \sum_{b=1}^B r_l(b)^2 + C_3} \quad (2.23)$$

$$v_l^d = \frac{(d_l - \mu_d)^2 + \sum_{b=1}^B \alpha_l(b) d_l(b)^2}{(d_l - \mu_d)^2 + \sum_{b=1}^B d_l(b)^2 + C_3} \quad (2.24)$$

Parameters r_l and d_l are the low pass version of reference and distorted videos respectively. Parameters $r_l(b)$ and $d_l(b)$ are the low pass filtered reference and distorted videos of a particular Gabor frequency b respectively. Parameters μ_r and μ_d are the mean of the Gaussian filtered output of reference and distorted videos respectively. Parameter $\alpha_l(b)$ is defined as (Seshadrinathan & Bovik, 2010):

$$\alpha_i(b) = \frac{\alpha'_i(b) - \mu_\alpha}{\max_b [\alpha'_i(b) - \mu_\alpha]}, \quad b = 1, 2 \dots B \quad (2.25)$$

where

$$\mu_\alpha = \frac{\sum_{b=1}^B \alpha'_i(b)}{B} \quad (2.26)$$

$$\alpha'_i(b) = \frac{\rho(b) - \delta_i(b)}{\rho(b)} \quad (2.27)$$

Parameter $\rho(b)$ is the radius of the filter while $\delta_i(b)$ is the perpendicular distance to the plane. Then, the temporal MOVIE index can be computed using equation below (Seshadrinathan & Bovik, 2010):

$$E_{tsm}(i_0) = \frac{1}{L} \sum_{l=1}^L (v_l^r - v_l^a)^2 \quad (2.28)$$

$$Q_t(i_0) = 1 - E_{tsm}(i_0) \quad (2.29)$$

Lastly, for overall MOVIE index part, it is the product of spatial and temporal MOVIE indices. It is mathematically represented as (Seshadrinathan & Bovik, 2010):

$$MOVIE = Q_s \times Q_t \quad (2.30)$$

2.6.5 Method by Sheikh and Bovik

Based on visual information fidelity (VIF) (Sheikh *et al.*, 2005a) for spatial quality assessment, Sheikh and Bovik (2005) has proposed a non-IQA based VQA method. For the spatial part, VIF is utilized. The concept of VIF assumes that the video frames are mixtures of marginal Gaussian densities of wavelet subband. This concept applies the idea that natural image source reaches the receiver through a channel. Thus, VIF measures the

spatial image quality based on mutual information from the subband coefficients of reference and distorted videos. For the temporal quality scores, motion information is used to account for the temporal impairments. The information loss due to motion is computed by measuring the deviations in spatiotemporal derivatives of a video. There are two models and one criterion involved in this VQA method. They are source model, distortion model, and information fidelity criterion.

In this method, natural scene statistic is implied in the source model. The model of natural scene statistic utilized is Gaussian scale mixtures (GSM) in the wavelet domain. GSM radio frequency (RF) or one subband of the wavelet decomposition of an image is expressed as (Sheikh & Bovik, 2005):

$$S = X \cdot Y = \{X_i \cdot Y_i; i \in I\} \quad (2.31)$$

where

$$X = \{X_i; i \in I\} \text{ and } Y = \{Y_i; i \in I\} \quad (2.32)$$

Parameters I are spatial indices for RF and S denotes the GSM RF. Parameter S is the product of two independent stationary RFs, X and Y . RF X is the positive scalar while RF Y is the Gaussian scalar. GSM RF, conditioned on S_i and X_i is defined as (Sheikh & Bovik, 2005):

$$P_{S_i|X_i}(S_i|X_i) \sim \mathcal{N}(0, X_i^2 Y_i^2) \quad (2.33)$$

where $\mathcal{N}$ denotes the Gaussian density with the mean of zero and standard deviation of $X_i^2 Y_i^2$. GSM framework can model the marginal statistics of the wavelet coefficients of natural images.

For distortion model, it is signal attenuation and additive Gaussian noise in each subband. This is defined as (Sheikh & Bovik, 2005):

$$\mathcal{D} = \mathcal{u}\mathcal{S} + \mathcal{V} = \{u_i S_i + V_i; i \in I\} \quad (2.34)$$

where $\mathcal{D}$ is the RF of the subband from the distorted signal. Parameter $\mathcal{u}$ is the scalar attenuation and $\mathcal{V}$ is the additive noise RF. Two distortions, i.e. blur and additive noise, are captured in this model.

In this VQA, information fidelity criterion is the mutual information of the source and distortion model. The mutual information is denoted as $M(\mathcal{S}^n; \mathcal{D}^n)$, where $\mathcal{S}^n$ is a series of $\mathcal{S}$ and $\mathcal{D}^n$ is a series of $\mathcal{D}$. In conditional form, information fidelity criterion is denoted as $M(\mathcal{S}^n; \mathcal{D}^n | \mathbf{x}^n)$. Parameter $\mathbf{x}^n$ is one of the realization of $\mathcal{X}^n$. From the distorted model, information fidelity criterion can be further simplified to (Sheikh & Bovik, 2005):

$$M(\mathcal{S}^n; \mathcal{D}^n | \mathbf{x}^n) = \sum_{i=1}^N M(\mathcal{S}^n; \mathcal{D}^n | \mathcal{S}^{i-1}, \mathcal{D}^{i-2} \mathbf{x}^n) = \sum_{i=1}^N M(S_i; D_i | \mathbf{x}_i) \quad (2.35)$$

Since X_i and V_i are Gaussian with standard deviation of σ_V^2 , equation (2.35) can be further simplified to (Sheikh & Bovik, 2005):

$$M(\mathcal{S}^n; \mathcal{D}^n | \mathbf{x}^n) = \sum_{i=1}^N [h(D_i | \mathbf{x}_i) - h(D_i | S_i, \mathbf{x}_i)] \quad (2.36)$$

$$= \sum_{i=1}^N [h(u_i S_i + V_i | \mathbf{x}_i) - h(V_i)] \quad (2.37)$$

$$= \frac{1}{2} \sum_{i=1}^N \log_2 \left(1 + \frac{u_i^2 x_i^2 \sigma_V^2}{\sigma_V^2} \right) \quad (2.38)$$

Parameter h is the differential entropy. Equation (2.38) is for a single subband. Then, information fidelity criterion for all subbands can be mathematically represented as:

$$IVF = \sum_{b \in \text{subbands}} M(S^{n_b, b}; D^{n_b, b} | x^{n_b, b}) \quad (2.39)$$

where b is the RF subband. Each subband has its multiplicative constant. Thus, blur distortion could also be captured by finer scale of the subbands.

2.6.6 VSSIM index

IQA based VQA methods are discussed from this section onwards. First, Wang *et al.* (2004) adapted SSIM to detect the temporal distortions by proposing Video Structural Similarity (VSSIM) index. Instead of concentrating on the luminance component of a video, this method further applies SSIM indices to two chrominance components (C_b, C_r) of a video. However, the luminance component is given more weight. According to the authors, luminance component affects the video quality more as compared to chrominance components. The chrominance components cannot be ignored too. For the spatial part, the quality score is the combination of SSIM score of both luminance and chrominance components of a video. During the combination, luminance intensity is used to weight the overall spatial index. For the temporal part, the spatial quality score of each frame is weighted by the global motion of the successive frame to produce overall video quality score of VSSIM. The metric is discussed in more detail.

The quality of a video is measured in three levels. They are local region level, frame level, and sequence level. Before going through luminance evaluation and local SSIM index, reference and distorted frames undergo sampling process. Sampling areas are

extracted from the frames through 8x8 windows. Sampling density is represented by the number of sampling windows per video frame (R_s). This aims to reduce the computational cost while maintaining the results. SSIM method is then applied to Y , C_b and C_r colour components. The local quality index is mathematically represented as (Wang *et al.*, 2004):

$$SSIM(i, j) = W_Y SSIM^Y(i, j) + W_{C_b} SSIM^{C_b}(i, j) + W_{C_r} SSIM^{C_r}(i, j) \quad (2.40)$$

where W are the weights for each SSIM index. In this VQA method, W_Y is set to 0.8 while W_{C_b} and W_{C_r} are both 0.1.

Next, the frame level quality index needs to be determined. This is computed through the equation (Wang *et al.*, 2004):

$$Q_y = \frac{\sum_{x=1}^{R_s} w_{xy} SSIM_{xy}}{\sum_{x=1}^{R_s} w_{xy}} \quad (2.41)$$

where Q_y is the quality index of a particular frame y . Parameter w_{xy} is the weight for x^{th} sampling window at frame y . The weight, w_{xy} , is based on the luminance evaluation. The authors made an assumption that dark regions attract fewer fixations. Thus, the average luminance of the window is used to weight the local SSIM index. This is shown as (Wang *et al.*, 2004):

$$w_{xy} = \begin{cases} 0 & \mu_x \leq 40 \\ (\mu_x - 40)/10 & 40 < \mu_x \leq 50 \\ 1 & \mu_x > 50 \end{cases} \quad (2.42)$$

The overall VSSIM index, Q , can be determined from Q_y . It is mathematically represented as (Wang *et al.*, 2004):

$$Q = \frac{\sum_{y=1}^F w_y Q_y}{\sum_{y=1}^F w_y} \quad (2.43)$$

where F is the total number of frames and w_y is the weight for a particular frame y . Parameter w_y is based on the motion estimation from the reference frames. VSSIM utilizes block-based motion estimation algorithm. Parameter w_y is defined as (Wang *et al.*, 2004):

$$w_y = \begin{cases} R_s & M_y \leq 0.8 \\ [(1.2 - M_y)/0.4] R_s & 0.8 < M_y \leq 1.2 \\ 0 & M_y > 1.2 \end{cases} \quad (2.44)$$

where M_y is motion level at frame y . It is mathematically represented as (Wang *et al.*, 2004):

$$M_y = \frac{\sum_{x=1}^{R_s} m_{xy}}{R_s \times P_M} \quad (2.45)$$

Parameter m_{xy} is the motion vector length of x^{th} sampling window at frame y . Parameter P_M is a constant for normalization.

2.6.7 FMSE index

There is another IQA-based VQA method, foveated mean squared error (FMSE), proposed by Rimac-Drlje *et al.* (2010). As indicated by the name of the method, FMSE is a VQA method based on the improvement of MSE. The authors made an assumption that HVS focuses on the centre of a frame or image. Thus, the authors assign different weights to the errors according to their respective locations in the frames. The errors which are closer to the centre of a frame are given more weightage. In order to take the temporal effect of the frames into account, this method adopts the foveation-based contrast sensitivity from the method of Lee *et al.* (2002) and applied for scenes with movement. FMSE also adopts the effect of human eye contrast sensitivity changes. These changes may

be influenced by the rise of eccentricity on the retina (Geisler & Perry, 1998) and the retinal image velocity (Lee *et al.*, 2002). The details of FMSE are further described.

The difference between the reference and distorted video frames need to be obtained first. This is shown in the equations as (Rimac-Drlje *et al.*, 2010):

$$D(f) = r(f) - d(f) \quad (2.46)$$

where

$$D(f) = D(x, y), \quad r(f) = r(x, y) \quad \text{and} \quad d(f) = d(x, y) \quad (2.47)$$

Parameter $D(f)$ is the difference for a frame, f . After that, filtering is performed on parameter $D(f)$. This process of filtering is shown as (Rimac-Drlje *et al.*, 2010):

$$D_{f,k} = D * h_k, \quad k = 1, 2 \dots K \quad (2.48)$$

where h_k is the convolutional kernel of the k^{th} filter. There is a total of K filters involved. They are generated based on the Nyquist frequency, f_N , of the video. The central frequencies of the filters are distributed across the band of zero to f_N .

The filtered results are then computed with contrast sensitivity matrices to form the FMSE frame index. This process is shown in equations (2.49) and (2.50) (Rimac-Drlje *et al.*, 2010).

$$FMSE(f) = \sum_{k=1}^K \left[\frac{1}{MN} \sum_{x=1}^M \sum_{y=1}^N D'_{f,k}(x, y) \right] \quad (2.49)$$

where M is the frame width and N is the frame height. Parameter $D'_{f,k}(x, y)$ is the foveated difference matrices. It is mathematically represented as (Rimac-Drlje *et al.*, 2010):

$$D'_{f,k}(x, y) = D_{f,k}(x, y) \cdot CS_k(rD, f_k, x, y) \quad (2.50)$$

Parameter CS_k is the contrast sensitivity matrices. Parameters rD and f_k are the respective difference of particular point to foveation point and central frequencies of the filter. Contrast sensitivity matrices are defined as (Rimac-Drlje *et al.*, 2010):

$$CS_k(rD, f_k, x, y) = \quad (2.51)$$

$$\exp \left\{ -\alpha f_k \left(\tan^{-1} \left[\frac{\sqrt{(x-x_f)^2 + (y-y_f)^2} \cdot H}{N \cdot rD} \right] \cdot \frac{v_c + v_r(x, y)}{e_2 \cdot v_c} + \frac{v_r(x, y)}{v_c} \right) \right\}$$

where α is spatial frequency decay constant. Parameters x_f and y_f are the foveated position. Parameter e_2 is half-resolution eccentricity. Parameters v_c and v_r are the corner velocity and retinal image velocity.

Finally, FMSE index of a video is the average of all the FMSE frame index. This is represented as (Rimac-Drlje *et al.*, 2010):

$$FMSE = \frac{1}{F} \sum_{f=1}^f FMSE(f) \quad (2.52)$$

2.6.8 ViMSSIM Index

Vu and Deshpande (2012) had proposed an IQA-based VQA method, ViMSSIM, which is the extended version of MSSIM (Wang *et al.*, 2003). MSSIM is utilized to measure the spatial quality of a video. The authors had also proposed a modified exponential moving average procedure. This procedure is used to determine the number of frames that the spatial MSSIM frame quality index are being averaged. For assessing temporal video quality score, two differences are needed. The first difference is the

differences of distorted next frame and reference previous frame. The second difference is the differences of reference next frame and reference previous frame. These two differences are calculated and compared by using MSSIM. This FR objective VQA method performs very well but is computationally expensive.

First, the reference and distorted frames are valued by using MSSIM index. The results are combined to form the spatial ViMSSIM index. This process involved a modified exponential moving average algorithm. It is mathematically expressed as (Vu & Deshpande, 2012):

$$S_1 = \frac{\sum_{f=1}^{n_0} MSSIM_i}{n_0} \quad (2.53)$$

$$S_{n+1} = \alpha MSSIM_{n+n_0} + (1 - \alpha)S_n \quad (2.54)$$

where $MSSIM$ is the MSSIM index. Parameter S is a series of value after exponential moving average and n_0 determines S_1 . In this case, it is chosen to be half of the frame rate of the video. Parameter α is the smoothing factor that is defined as (Vu & Deshpande, 2012):

$$\alpha = \frac{\eta}{n_0 + 1} \quad (2.55)$$

In ViMSSIM, η is set to 0.25 and thus α is 1/124. The spatial ViMSSIM index is the minimum of the series of S values. This is shown as (Vu & Deshpande, 2012):

$$Q_s = \min_n S_n \quad (2.56)$$

For the temporal ViMSSIM index, there are two differences that need to be computed. They are $D_{r,t}$ and $D_{d,t}$. These two differences are shown in the equations as (Vu & Deshpande, 2012):

$$D_{r,t} = r_{t+1} - r_t \quad (2.57)$$

$$D_{d,t} = d_{t+1} - r_t \quad (2.58)$$

Parameter r_t is the frame at particular instant t from the reference video while r_{t+1} is the frame after the particular instant from reference video. Parameter d_{t+1} is the frame from distorted video at instance $t + 1$. These two difference values are used in MSSIM index to produce a series of indices, $MSSIM_{t,i}$. The MSSIM indices are averaged to form the temporal ViMSSIM index. This is represented in equation as (Vu & Deshpande, 2012):

$$Q_t = \frac{\sum_{i=1}^F MSSIM_{t,i}}{F - 1} \quad (2.59)$$

where F is the total number of frames and Q_t is the temporal ViMSSIM index. The overall ViMSSIM index is the arithmetic mean of the spatial and temporal ViMSSIM indices. This is mathematically shown as (Vu & Deshpande, 2012):

$$ViMSSIM = \frac{Q_s + Q_t}{2} \quad (2.60)$$

2.6.9 ViS₃ Index

A recent VQA method, ViS₃, was proposed by Vu and Chandler (2014). This method involves spatiotemporal slices (STS). The concept of STS splits a video into frames in three different dimensions. The three dimensions are the x-y dimension, x-t dimension, and y-t dimension. ViS₃ utilizes STS for measuring the spatiotemporal dissimilarity of a video. Besides STS, ViS₃ also adopts Most Apparent Differences (MAD) (Larson & Chandler, 2010) to assess the spatial quality score. Similar to most of the VQA methods, motion magnitudes are obtained using optical flow. The motion magnitudes are combined with the MAD spatial quality score to form a single scalar ViS₁. The

spatiotemporal dissimilarity of each STS image of the reference and distorted video are then processed to give a single scalar ViS₂. ViS₃, which is the overall quality score of the video, is the geometric mean of ViS₁ and ViS₂.

In order to compute ViS₃ index, spatial ViS₃ index or ViS₁ need to be computed first. ViS₁ is computed through visible distortion maps, statistical difference map, and motion map. The visible distortion maps are computed with equation (Vu & Chandler, 2014):

$$DM_f = \xi_f \cdot MSE_f \quad (2.61)$$

where ξ is the local distortion visibility map and MSE is the local MSE map. Local distortion visibility map is the result of local contrast maps from reference and distorted frames.

Statistical difference map, SM is computed using the equation (Vu & Chandler, 2014):

$$SM_f = \sum_{j=1}^5 \sum_{\delta=1}^4 w_j [|\sigma_{j,\delta,f} - \hat{\sigma}_{j,\delta,f}| + 2|\varsigma_{j,\delta,f} - \hat{\varsigma}_{j,\delta,f}| + |\kappa_{j,\delta,f} - \hat{\kappa}_{j,\delta,f}|] \quad (2.62)$$

where w_j is the weight based on filter scale, j is the filter scale, and δ is the orientation. Parameter σ , ς , and κ are the standard deviation, skewness, and kurtosis of the filter respectively. For motion estimation, a motion map can be produced from Lucas-Kanade method. This is shown as (Vu & Chandler, 2014):

$$MM_f = \sqrt{M_{v,f}^2 + M_{h,f}^2} \quad (2.63)$$

where M_v is the motion vector with respect to vertical direction while M_h is the motion vector with respect to horizontal direction. ViS₁ is the result of the combination of the three maps. This is mathematically shown as (Vu & Chandler, 2014):

$$ViS_1 = \sqrt{\frac{1}{W \times H} \sum_{f=1}^F \Delta_{j,k}^2} \quad (2.64)$$

where W and H are the width and height of the maps. Parameter Δ_j is defined as (Vu & Chandler, 2014):

$$\Delta_{j,k} = \frac{DM_{j,k}^\alpha \times SM_{j,k}^{1-\alpha}}{\sqrt{1 + MM_{j,k}}} \quad (2.65)$$

where α is the adjustment parameter.

For temporal ViS₃ index or ViS₂, STS is involved. Spatiotemporal correlation map and spatiotemporal response difference map are needed for computing spatiotemporal dissimilarity map. The spatiotemporal correlation map, SC_f , can be computed using linear correlation coefficient, C_f , from STS images. This is shown as (Vu & Chandler, 2014):

$$SC_f = \begin{cases} 0 & C_f < 0 \\ 1 & C_f > 0.9 \\ C_f & otherwise \end{cases} \quad (2.66)$$

On the other hand, spatiotemporal response difference map can be computed using (Vu & Chandler, 2014):

$$SDR_f = \ln \left\{ 1 + A \sum_{j=1}^J \sum_{n=1}^N w_j (\rho_f)^2 \right\} \quad (2.67)$$

where A is the scaling factor and w_j are the weights based on the frame. Parameter ρ_f is the result of local mean, μ , and standard deviation, σ , of response difference. It is denoted as (Vu & Chandler, 2014):

$$\rho_f = \begin{cases} 0 & \mu_f < t \\ \sigma_f \times \sqrt{\mu_f / (t + \mu_f)} & \text{otherwise} \end{cases} \quad (2.68)$$

Spatiotemporal dissimilarity value can be computed from SC_f and SDR_f . This is shown as (Vu & Chandler, 2014):

$$ViS_2 = SDR_f \cdot \sqrt{1 - SC_f} \quad (2.69)$$

The ViS_3 index can be computed from ViS_1 and ViS_2 as (Vu & Chandler, 2014):

$$ViS_3 = \sqrt{ViS_1 \times ViS_2} \quad (2.70)$$

2.6.10 Review of Existing Methods

Most of the existing VQA methods are computationally complex. This is evidenced with the computed computational time of MOVIE and ViS_3 shown in Chapter 4. These two representative methods have long computational time and are inappropriate to be applied in real time applications. There are three main reasons for the high computational complexity for the mentioned VQA methods in this section.

First of all, some of these methods involve frequency-based decomposition. For instance, MOVIE, DVQ, and method by Sheikh and Bovik involve Gabor decomposition, DCT, and wavelet decomposition respectively. Typically, frequency based decompositions are applied on multiple orientations and scales. The processing of the resulting coefficients from different subbands further complicates the VQA method. Most of the time, it is the decomposition that requires long computational time in these methods.

Local statistic computation is another reason for high computational complexity. This is shown by the method by Sheikh and Bovik, ViMSSIM, and ViS₃. Multiple statistic computations are required for each local regions in a frame and thus longer computational time is required for this step. This is evident particularly in ViS₃ where spatiotemporal correlation, spatiotemporal response difference, and spatiotemporal dissimilarity are computed for spatiotemporal part. This is also the main reason for the good performance of ViS₃ as more factors are taken into consideration.

The last reason is the motion computation. Motion is an important factor for assessing temporal distortions in distorted videos. The most common type of motion computation used is optical flow. MOVIE utilizes the optical flow method for motion estimation. According to Peng *et al.* (2003), MOVIE spends more than 55 percent of its computational time in computing optical flow. VSSIM also involves motion computation. It utilizes block-based motion estimation algorithm. Instead of computing motion estimation, the other methods computes effect of motion directly. VQM utilizes absolute temporal information, ViMSSIM utilizes frame differencing, and FMSE and ViS₃ compute motion loss in spatiotemporal domain.

Among the VQA methods discussed in this section, ViMSSIM and FMSE seem to have lower computational complexity than the others. However, their computational time is still unsatisfactory to be applied in real time applications. For ViMSSIM, this may due to involvement of MSSIM that assesses quality in multiple scales. FMSE, on the other hand, considers HVS foveation, fixation, and eccentricity of eyeball that complicates the method. Hence, it is of great value and interest to propose a VQA method that has lower computational complexity than the existing methods while having comparable performances.

2.7 JND

Before ending this chapter, the concept of JND is briefly reviewed as it is related to the first proposed method in this thesis. JND is defined as the amount of variation of stimuli for HVS to perceive. It is originated from the field of psychophysics. This limen, or known as the onset of the psychological reaction, is well-known for signal processing, especially image processing.

In general, JND is categorized into subband and pixel based (Zhang *et al.*, 2008). Subband JND involves frequency domain. Thus, it always deals with discrete cosine transform (DCT) and wavelet transform. It takes consideration of contrast sensitivity function and luminance adaptation (Zhang *et al.*, 2008; Chen & Liu, 2014). On the other hand, pixel based JND focus on contrast and luminance masking.

JND are applied in many areas of image and video processing. In compression, Wang *et al.* (2015) have proposed a JND-based lossless compression for JPEG (Joint Photographic Experts Group) images. JND is also applied in watermarking of images and videos. Liu (2009) had proposed a method for color image watermarking based on JND. For image coding, Shang *et al.* (2013) succeeded in improving video coding efficiency by utilizing subband Foveated JND (FJND). There are also methods for image and video enhancement such as the method proposed by Zheng *et al.* (2013). Application of JND can also be found in quality assessment. One such example is the work of Ma *et al.* (2010) which applied JND in their video quality assessment method.

2.9 Summary

In Chapter 2, common distortions for video are discussed briefly. Important methods of subjective and objective VQA are also discussed. The discussed subjective methods are DSCQS, SSCQE, DCR, PCom, SAMVIQ, and ACR. The advantages and disadvantages of subjective methods are explained in Section 2.3.8. Three types of objective VQA, i.e. FR VQA, RR VQA, and NR VQA, are also presented. Some relevant IQA methods such as MSE and SSIM are discussed in detail in Section 2.5. Besides that, existing VQA methods are also briefly discussed and presented. Objective methods are categorized into IQA-based and non-IQA based. For IQA-based, there are VSSIM, FMSE, ViMSSIM, and ViS₃. VQM, DVQ, MOVIE, and method by Sheikh and Bovik are categorized as non-IQA based VQA metrics. Section 2.7 reviews the mentioned VQA methods in terms of computational time. Lastly, JND concept related to the first proposed method in this thesis is briefly discussed.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In this chapter, two new VQA methods are proposed. The first method, Index1, is based on the Just Noticeable Difference (JND) (Stern & Johnson, 2010) for spatial and temporal aspects. For temporal aspect, the temporal information (Video Quality Experts Group, 2003) is combined with JND concept to consider temporal effects and distortions. The second method, Index2, adapts MSE-SSIM for spatial part and utilizes the concept of temporal variation for the temporal part. This aims to formulate VQA methods that strike a balance between computational complexity and performance. Both of the proposed methods follow the same framework as shown in Figure 3.1.

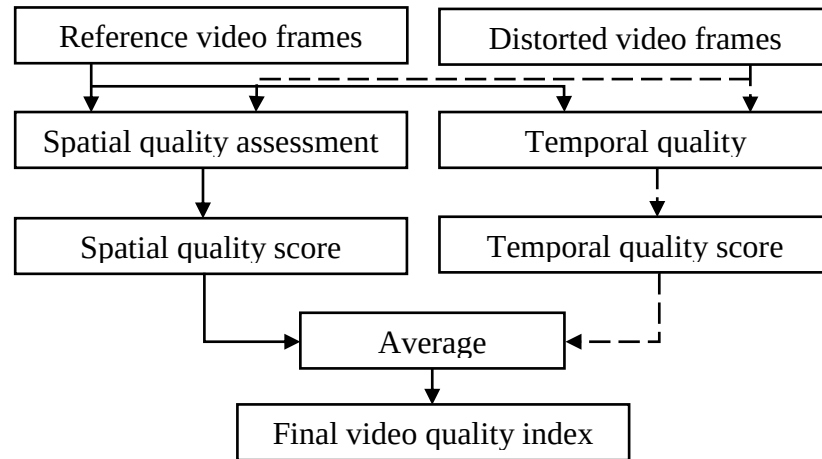


Figure 3.1: General framework of proposed methods.

Section 3.2 discusses Index1 while Section 3.3 elaborates Index2.

3.2 Index1 Method

Index1 method is based on the concept of JND (Stern & Johnson, 2010). This concept inherits from its predecessor, Mod-MSE. The main difference between Mod-MSE and Index1 is the adaption of JND with different existing IQA methods. Pixel based JND is applied to SSIM to measure the spatial quality of a video. For temporal aspect, temporal information between the reference and distorted videos are being compared. This temporal quality measurement method is inspired by the work of Vu and Deshpande (2012), Cardoso *et al.* (2014), and Oliveira *et al.* (2013). JND is applied to the temporal quality evaluation to further enhance the performance. Then, the overall video quality is calculated by computing the arithmetic mean of the indices from spatial and temporal quality measurements.

Section 3.2.1 discusses the brief introduction of Mod-MSE. Section 3.2.2 and Section 3.2.3 deliberate the spatial and temporal aspect of the Index1. Section 3.2.4 deals with the overall index of Index1. Lastly, Section 3.2.5 briefly compares Mod-MSE and Index1.

3.2.1 Brief Introduction to Mod-MSE

The Mod-MSE method can be separated into spatial and temporal parts. For the spatial part, the concept of pixel-based JND is applied with MSE. The formula for MSE is presented in equations (2.1) and (2.2). The spatial part of Mod-MSE is summarized as:

$$Selected\ MSE(x, y) = \begin{cases} 0 & MSE(x, y) \leq thres \\ MSE(x, y) & MSE(x, y) > thres \end{cases} \quad (3.1)$$

$$ModMSE(spatial) = \frac{1}{NP} \sum_x \sum_y Selected\ MSE(x, y) \quad (3.2)$$

where $thres$ is the predetermined JND threshold, NP is the total number of pixels in an image and $ModMSE(spatial)$ is the spatial quality score of Mod-MSE. The application of JND is shown in equation (3.1). MSE is first applied to reference and distorted frames to get $MSE(x, y)$. For MSE below JND threshold, they are considered as perfect quality. A value of zero for MSE means no distortion. On the other hand, for MSE index above the JND threshold, the value is maintained. The spatial quality score of Mod-MSE is the arithmetic average of the JND-applied MSE.

For the temporal part, MSE is applied to the difference of the current frame and previous frame. There are two types of differences. They are reference difference, D_1 , and distorted difference, D_2 . Reference difference is the subtraction of current frame of the reference video, $f_{r,x}$, from the previous frame of the reference video, $f_{r,x-1}$. The distorted difference is computed as the subtraction of current frame of the distorted videos, $f_{d,x}$, from the previous frame of the reference videos, $f_{r,x-1}$. The formulas are shown as:

$$D_1 = f_{r,x-1} - f_{r,x} \quad (3.3)$$

$$D_2 = f_{d,x-1} - f_{r,x} \quad (3.4)$$

$$ModMSE(temporal) = (D_1 - D_2)^2 \quad (3.5)$$

where $ModMSE(temporal)$ is the temporal quality score of Mod-MSE. The temporal quality score is in fact the MSE of D_1 and D_2 . The overall quality score of Mod-MSE, $ModMSE$, is the arithmetic mean of spatial and temporal quality scores:

$$ModMSE = \frac{[ModMSE(spatial) + ModMSE(temporal)]}{2} \quad (3.6)$$

The workflow of Mod-MSE is depicted in Figure 3.2.

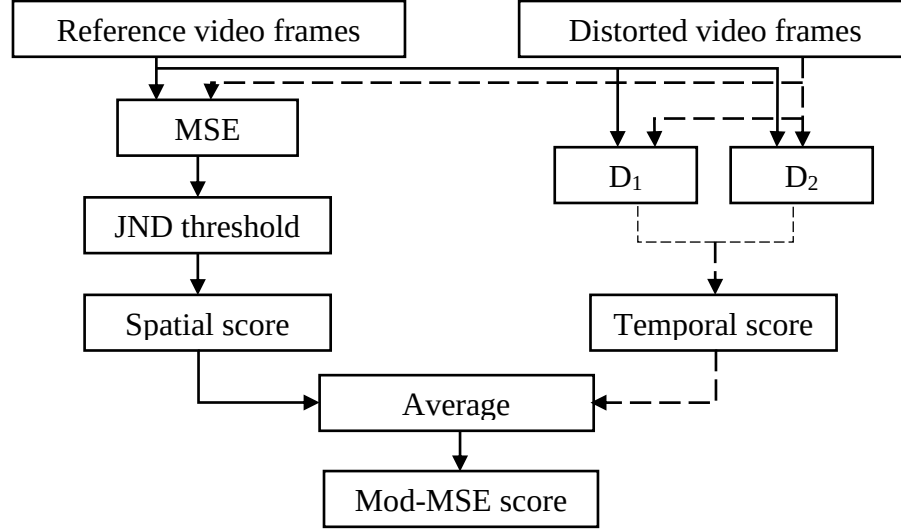


Figure 3.2: Mod-MSE workflow.

3.2.2 Spatial part

Similar to Mod-MSE, the concept of JND is applied to the spatial aspect of the Index1. In one of the existing works, JND theory has been applied in Mean-Squared Error (MSE) (Wang & Bovik, 2009; Wolf, 2001). For this new VQA method, SSIM index (Wang *et al.*, 2004) is reformulated by using a pixel based JND. The formula for SSIM is shown in equation (2.4) in Section 2.5.3.

The spatial aspect of Index1 is extended from equation (2.4). All minimal distortions that cannot be perceived by HVS are eliminated. In other words, the pixels with minimal distortions are considered as perfect or no distortion. This corresponds to the assumption that HVS tends to focus on the more severely distorted regions (Harvey *et al.*, 1985). Thus, the part with minimal distortion is ignored. Assigning perfect scores to the regions with minimal distortions helps SSIM index in giving more weightage (more severity) to the poorer quality region. Even though a distorted image or frame has a large

area of minimal distortions, it can be perceived as similar to the reference image. This is clarified through the illustrations in Figure 3.3.

Images of Figure 3.3 (a) and (b) are obtained from LIVE image database (Sheikh *et al.*, 2005b). Figure 3.3 (a) shows the undistorted image while Figure 3.3 (b) shows the same image distorted by Gaussian blurring with standard deviation of 3.08. However, there is no significant difference between Figure 3.3 (b) and its original image. The result of SSIM of the two figures is shown in Figure 3.3 (c). The SSIM figures are shown with SSIM indices of each pixel with the mask size of 3x3. SSIM indices range from 0 to 1 where 1 is the perfect score and is shown as black pixel in the figures. The severity of distortion is denoted as the gray pixel. The whiter a pixel is in the figure, the more severe it is distorted. From the figure, it is observed that the region of distortion is big and covers about three-quarters of the overall image. In other words, there are only minor areas with no distortions. By considering Figure 3.3 (c) only, the distorted image should be seen as having bad quality. However, refer back to Figure 3.3 (b), the image is perceived to be quite similar to the original image in Figure 3.3 (a) without much distortion. This is the case when almost all the distortions are minimal.

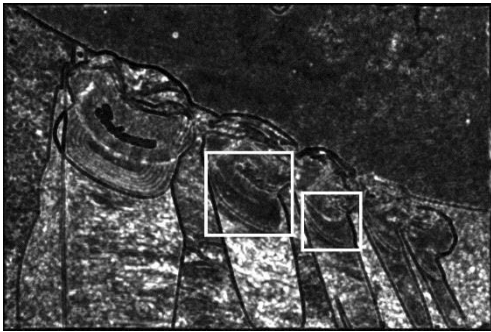
When the minimal distortions are neglected, the regions of detected distortions shrinks. Some greyish area disappears. The effect of applying JND is shown in Figure 3.3 (d). The square white boxes in Figure 3.3 (c) and Figure 3.3 (d) show the shrinking of distortion areas. All square box areas are further magnified and shown in Figure 3.4. Figure 3.4 (a) is the bigger square box before applying JND and Figure 3.4 (b) is the bigger square box after applying JND. Figure 3.4 (c) and Figure 3.4 (d) are the smaller square box before and after applying JND on SSIM results respectively.



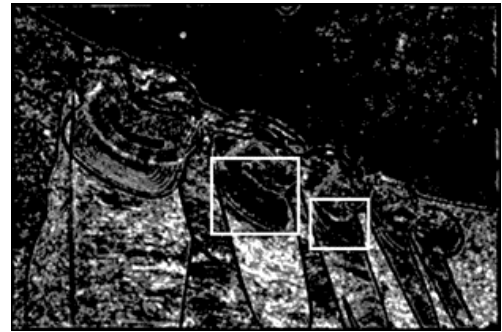
(a)



(b)

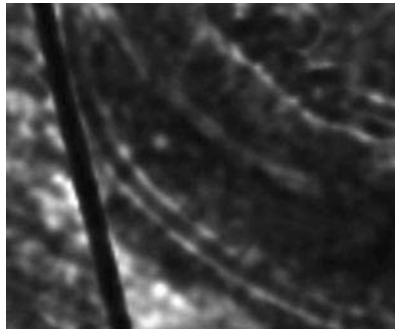


(c)

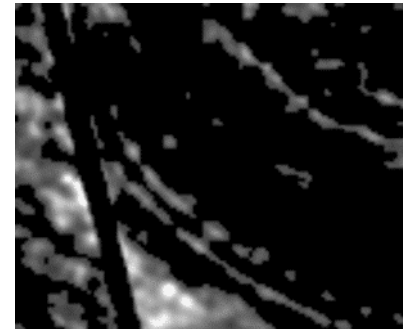


(d)

Figure 3.3: (a) Original Image (Sheikh *et al.*, 2005b). (b) Distorted Image (Sheikh *et al.*, 2005b). (c) Result of SSIM index. (d) Result of applying JND on SSIM index in Figure 3.3 (c).



(a)



(b)

Figure 3.4: (a) Magnification of big white box in Figure 3.3 (c). (b) Magnification of big white box in Figure 3.3 (d). (c) Magnification of small white box in Figure 3.3 (c). (d) Magnification of small white box in Figure 3.3 (d).

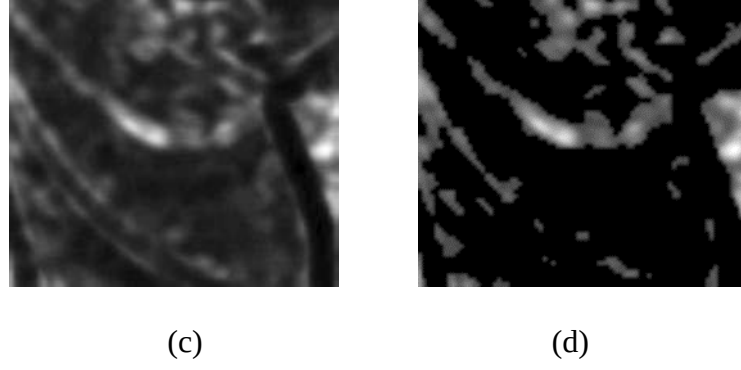


Figure 3.4 continued: (a) Magnification of big white box in Figure 3.3 (c). (b) Magnification of big white box in Figure 3.3 (d). (c) Magnification of small white box in Figure 3.3 (c). (d) Magnification of small white box in Figure 3.3 (d).

Therefore, under Index1, the pixels in the frames with near perfect quality are assumed to be undistorted since the distortion is minimal and imperceptible. The range of SSIM index is from zero to one, thus a threshold has to be determined. In this proposed method, the threshold is chosen to be five percent of the total range of SSIM index of the particular frame. This value is chosen after testing for the values range from 0 to 10 percent with interval of 1 percent. Although the performances for 4 to 8 percent show no significant difference, the best result is still achieved using 5 percent. The range is determined by the difference between the minimum and maximum values of SSIM index. Thus, the final allowable tolerance starts at 95% the maximum SSIM value. The elimination of minimal distortion is shown as:

$$SSIM'(x, y) = \begin{cases} 1 & SSIM(x, y) \geq 0.95 \max(SSIM) \\ SSIM(x, y) & otherwise \end{cases} \quad (3.7)$$

$SSIM''$ index of a frame is obtained by averaging $SSIM'$ of the two frames x and y . This is shown as:

$$SSIM''(x, y) = \frac{1}{p} \cdot \sum_x \sum_y SSIM'(x, y) \quad (3.8)$$

where p is the total number of SSIM' indices between two frames. To obtain the spatial quality index of the video, equation (3.7) and (3.8) are applied to every frames of the video. The spatial quality index of the video is attained by averaging all the $SSIM''$ indices of all frames. This is shown as:

$$Q_s = \frac{1}{N} \cdot \sum_{i=1}^N SSIM''(i) \quad (3.9)$$

where $SSIM''(i)$ is the new SSIM index for the i^{th} frame and Q_s denotes the spatial quality index.

The spatial part of Index1 is summarized in Figure 3.5

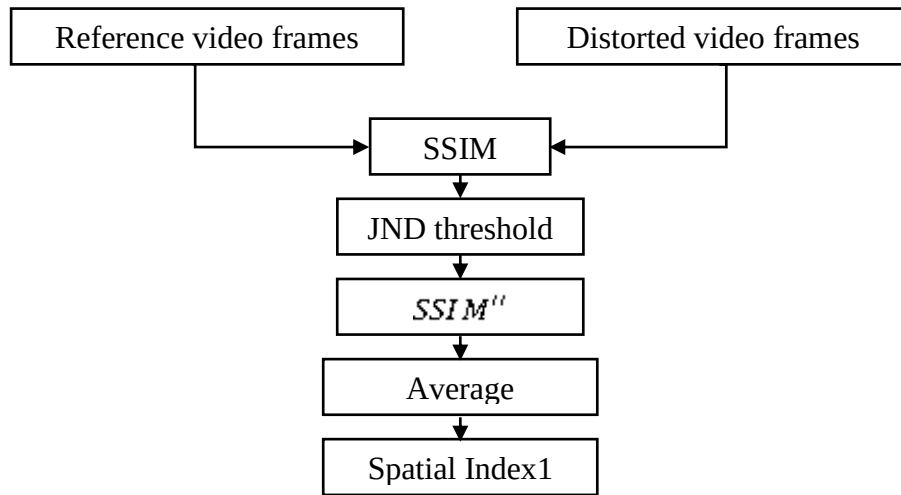


Figure 3.5: Workflow of spatial Index1.

3.2.3 Temporal part

For the temporal quality of Index1, two differences are used for comparison. The first difference is the difference of pixels between distorted frame at particular instant n , (D_n) , and reference frame at instant $n - 1$, (R_{n-1}) . The second difference is the

difference between the reference frame at particular instant n , (R_n), and reference frame at instant $n - 1$, (R_{n-1}). Instead of considering the difference at instant n and $n - 1$, this idea is also implemented at instant n and $n - 2$. This is shown in Figure 3.6.

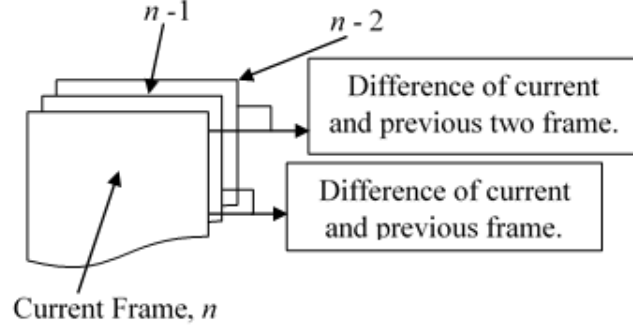


Figure 3.6: The concept of difference on two previous frames.

The idea of considering the previous frame is inspired by the work of Le Callet *et al.* (2006) which implied that not merely one previous frame will affect the quality of the current frame, but a number of previous frames. The temporal video quality is separated into two parts. One of them is the difference of D_n and R_{n-1} with R_n and R_{n-1} . Another part is the difference of D_n and R_{n-2} with R_n and R_{n-2} .

For the first part of temporal quality assessment, it is shown as:

$$D_1 = R_n - R_{n-1} \quad (3.10)$$

$$D_2 = D_n - R_{n-1} \quad (3.11)$$

where D_1 is the first reference difference and D_2 is the first distortion difference. Then, the SSIM index for D_1 and D_2 are calculated. This is shown as:

$$V_{t1} = SSIM(D_1, D_2) \quad (3.12)$$

Next, JND concept is applied. Similar to the spatial part of Index1, the threshold of five percent of the range is chosen. This is mathematically shown as:

$$V'_{t1}(x, y) = \begin{cases} 1 & V_{t1}(x, y) \geq 0.95 \max(V_{t1}) \\ V_{t1}(x, y) & \text{otherwise} \end{cases} \quad (3.13)$$

Then, the first temporal value is obtained by averaging V'_{t1} of all frames by the formula:

$$V''_{t1} = \frac{1}{N-1} \cdot \sum_{i=2}^N V'_{t1}(i) \quad (3.14)$$

where N is the total number of frames and V''_{t1} is the first temporal quality value. For the second part of the temporal quality evaluation, the process is the same as the first part except that the previous frame is changed to previous two frames. This is shown as:

$$D_3 = R_n - R_{n-2} \quad (3.15)$$

$$D_4 = D_n - R_{n-2} \quad (3.16)$$

$$V_{t2}(x, y) = SSIM(D_3, D_4) \quad (3.17)$$

$$V'_{t2}(x, y) = \begin{cases} 1 & V_{t2}(x, y) \geq 0.95 \max(V_{t2}) \\ V_{t2}(x, y) & \text{otherwise} \end{cases} \quad (3.18)$$

$$V''_{t2} = \frac{1}{N-2} \cdot \sum_{i=3}^N V'_{t2}(i) \quad (3.19)$$

where D_3 and D_4 are the second reference difference and distortion difference while V''_{t2} is the second temporal quality value. Then, temporal quality index is defined as:

$$Q_t = \frac{1}{2} (V''_{t1} + V''_{t2}) \quad (3.20)$$

where Q_t is the temporal quality index.

The workflow of temporal part of Index1 is illustrated in Figure 3.7.

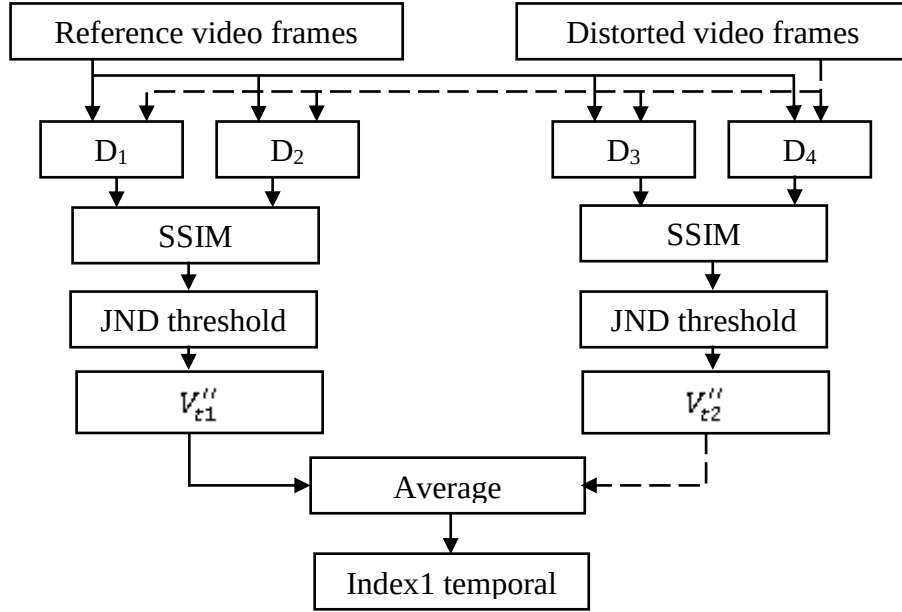


Figure 3.7: Workflow of temporal Index1.

3.2.4 Overall value of Index1

Lastly, the overall quality index of the video is the arithmetic mean of Q_s and Q_t . This is the most common method for combining spatial and temporal video quality index. This is shown as:

$$Index1 = \frac{1}{2}(Q_s + Q_t) \quad (3.21)$$

3.2.5 Comparison of Index1 and Mod-MSE

The difference and improvement of Index1 compared to Mod-MSE is listed in Table 3.1. Index1 apply SSIM instead of MSE due to the better performance of SSIM. SSIM further considers HVS characteristics in designing the algorithm. Index1 also

extends the JND threshold to temporal part of assessment methods. Lastly, instead of considering current and previous two frames only in Mod-MSE, additional frame with two frames distance from current frame is also considered.

Table 3.1: Comparison of Index1 and Mod-MSE.

Mod-MSE	Index1
Apply JND on MSE	Apply JND on SSIM
Do not apply JND in temporal part	Apply JND in temporal part
Consider two frames	Consider three frames

3.3 Index2 Method

In this section, the second proposed FR VQA method, Index2, is discussed in details. Similar to most of the existing metrics, Index2 consists of the spatial and temporal parts. The spatial part is the extension of spatial part of MSE Difference SSIM (MD-SSIM) while the temporal part is inspired by the work of Cardoso *et al.* (2014) and Oliveira *et al.* (2013) which are modified from the work in ViMSSIM (Vu & Deshpande, 2012). Then, the overall video quality score is the arithmetic mean (Galton, 1879) of the spatial and temporal quality scores.

Section 3.3.1 discusses MD-SSIM. This is important as Index2 is based on MD-SSIM method. Section 3.3.2 deals with spatial part of Index2 while Section 3.3.3 elaborates the details of the temporal part. Section 3.3.4 is a short section about the pooling of the spatial and temporal quality assessment to form a quality index for a video. The last subsection clarify the differences and improvements of Index2 compared to MD-SSIM.

3.3.1 Brief introduction to MD-SSIM

There are two parts in MD-SSIM, i.e. local part and global part. The overall video quality is the arithmetic mean of local and global quality scores. Local quality scores are the results of spatial and temporal quality scores. For the spatial part, SSIM map from SSIM index is being weighted by the error map. This is shown as:

$$MDSSIM(spatial) = \frac{\sum_x \sum_y E(x, y) S(x, y)}{\sum_x \sum_y E(x, y)} \quad (3.22)$$

where $MDSSIM(spatial)$ is the spatial quality scores of the local part of MD-SSIM, $E(x, y)$ is the error map from MSE in equations (2.1) and equation (2.2) while $S(x, y)$ is the SSIM map from SSIM from equation (2.4).

The temporal quality scores are the difference between the spatial quality scores between each frame. Then, the local quality scores are measured by weighting the spatial quality scores with the temporal scores with the equation:

$$MDSSIM(local) = \frac{\sum_{f=2}^F W_f MDSSIM(spatial)}{\sum_{f=2}^F W_f} \quad (3.23)$$

In Equation (3.23), $MDSSIM(local)$ is the local quality score of MD-SSIM, W_f is the difference between spatial quality scores of previous frame $f - 1$ and current frame f , and F is the total number of frames. Parameter W_f is defined as:

$$W_f = |MDSSIM(spatial)_f - MDSSIM(spatial)_{f-1}| \quad (3.24)$$

For the global quality scores, SSIM is applied in frame-by-frame basis. It is defined as:

$$MDSSIM(global) = \frac{1}{F} \sum_f SSIM(f) \quad (3.25)$$

where $SSIM(f)$ is the SSIM quality scores of frame f . The final MD-SSIM score of a video is the arithmetic mean of local quality score and global quality score. This is shown as:

$$MDSSIM = \frac{1}{2} [MDSSIM(local) + MDSSIM(global)] \quad (3.26)$$

MD-SSIM is summarized in Figure 3.8.

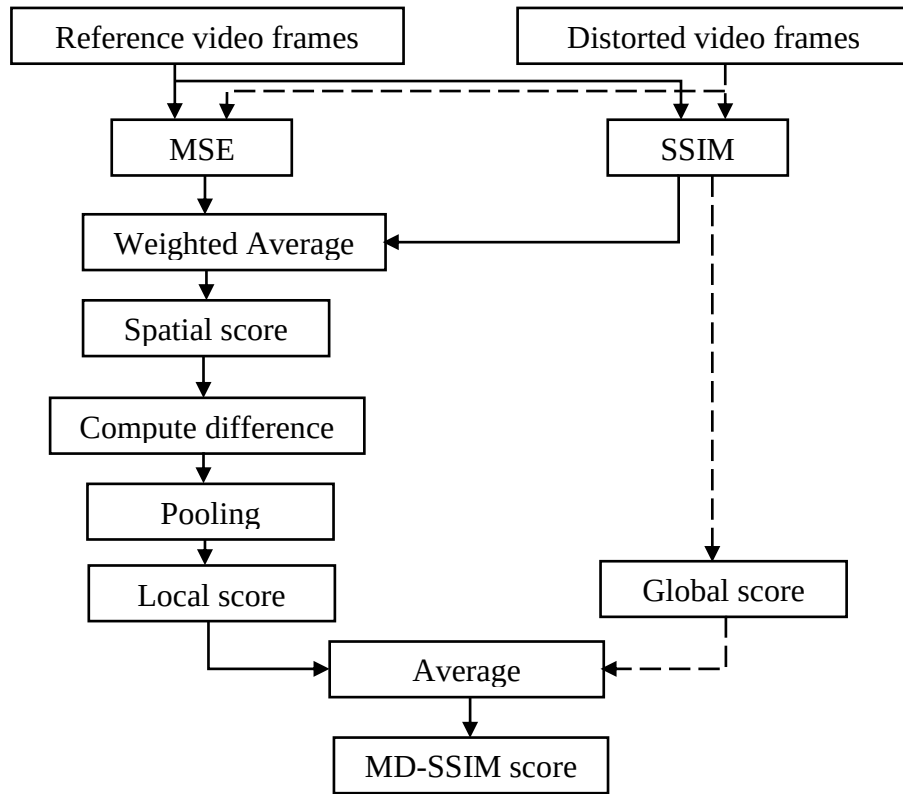


Figure 3.8: MD-SSIM workflow.

3.3.2 Spatial part

The spatial part of Index2 involves weighting of SSIM map from SSIM index. The weighting function further processes error map. Different from MD-SSIM, the error map is not the result of MSE, but is just the difference of gray intensities between two frames

without squaring. This aims to include direction factors in the weighting. Direction factors indicate whether the errors are causing the original luminance values to become brighter (positive direction) or darker (negative direction). Errors in different directions are considered the same in MSE maps due to the squaring of errors as shown in equation (2.1) (Akramullah, 2014). First, the error map needs to be measured. This is shown as:

$$E(x, y) = [R(x, y) - D(x, y)] \quad (3.27)$$

In Equation (3.27), $E(x, y)$ is the error map function while $R(x, y)$ and $D(x, y)$ are the frames from the reference and distorted videos respectively.

The error map is further processed to give better weight distribution among the pixels in a frame. The effects of surrounding pixels are taken into consideration in the weighting functions. The difference of the error in a pixel with the errors of its surrounding pixels is calculated. This is shown as:

$$W(x, y) = \frac{\sum_{m=x-1}^{x+1} \sum_{n=y-1}^{y+1} [E(x, y) - E(m, n)]^2}{N} \quad (3.28)$$

In Equation (3.28), $W(x, y)$ is the weighting function for SSIM map, parameters m and n are the pixel positions around the target pixel and N is the total number of pixels surrounding the target pixel. The squaring of values is to prevent negative values. Clarification of the surrounding error function is shown in Figure 3.9. Figure 3.9 (a) and Figure 3.9 (b) shows the first frame of the reference and distorted video, “Rushhour”, from the LIVE video database (Seshadrinathan *et al.*, 2010a; Seshadrinathan *et al.*, 2010b). Figure 3.9 (c) shows the error map of both frames. The illustrated error map is the result of inverting and modulus of results from Equation (3.27). Figure 3.9 (d) is the SSIM map from SSIM index. Figure 3.9 (e) is the weight function calculated by using Equation (3.28). The illustrated weight function is inverted and normalized for illustration purpose. The

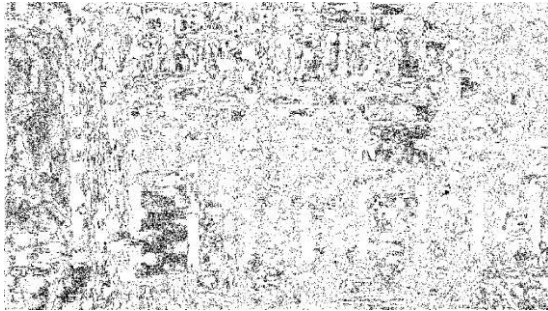
region where the distortions are more severe are emphasized in Figure 3.9 (e). Moreover, the regions with less significant errors are having less weight according to the weight function.



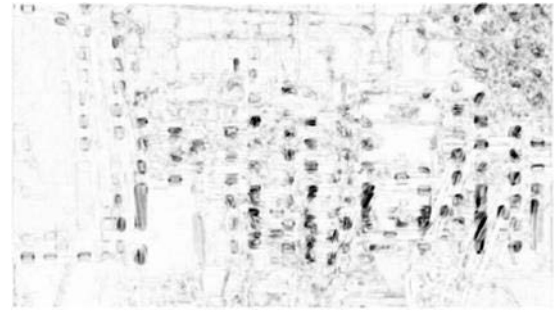
(a)



(b)



(c)



(d)



(e)

Figure 3.9: (a) Original frame (Seshadrinathan *et al.*, 2010a; Seshadrinathan *et al.*, 2010b). (b) Distorted frame (Seshadrinathan *et al.*, 2010a; Seshadrinathan *et al.*, 2010b). (c) Error map of two frames after inversion and modulus. (d) SSIM map of two frames. (e) Inverted and normalized weighted SSIM map of two frames.

SSIM map is obtained from SSIM index as shown in equation (2.4). Then, $W(x, y)$ is used to weight the SSIM map in Equation (2.4) to obtain the spatial quality score and is shown as:

$$Spatial = \frac{\sum_{x=1}^M \sum_{y=1}^N [W(x, y) \cdot SSIM(x, y)]}{\sum_{x=1}^M \sum_{y=1}^N SSIM(x, y)} \quad (3.29)$$

where M and N are the width and height of the frame from the video. The workflow of spatial part of Index2 is depicted in Figure 3.10.

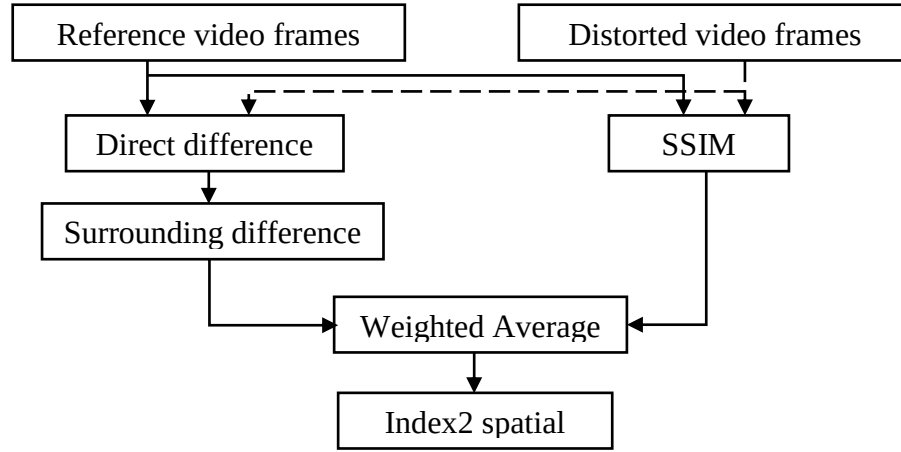


Figure 3.10: Workflow of spatial Index2.

3.3.3 Temporal part

For the temporal quality score of the Index2, temporal information has to be computed. The most direct method of obtaining the temporal information is by measuring the difference between two frames. This is shown as (ITU-T, 1999):

$$TI = \max\{std_{x,y}[A_{f-1}(x, y) - A_f(x, y)]\} \quad (3.30)$$

where TI is the temporal information, max is the maximum operator and $std_{x,y}$ is the standard deviation over space dimension. Parameter $A_f(x, y)$ is the current frame from a video while $A_{f-1}(x, y)$ is the previous frame from the same video.

Instead of computing the temporal information directly from the same video, the temporal information is computed from frames of different videos too. One of the frames is the frame from a distorted video at particular instant f , $D_f(x, y)$ and the other frame is the frame from the reference video frame at instant $f - 1$, $R_{f-1}(x, y)$. The computed temporal information is defined as distorted temporal information. This was first proposed by Vu and Deshpande (2012). The distorted temporal information is compared with the reference temporal information. The reference temporal information is computed by only utilizing frames from the reference video. Dissimilar to the works by Vu and Deshpande (2012), SSIM index is used instead of MSSIM index.

Temporal information alone is not good enough to measure the temporal distortions. Variations of temporal information are also computed to enhance temporal information. This is illustrated in Figure 3.11. Figure 3.11 (a), Figure 3.11 (b), and Figure 3.11 (c) are the first three frames from the video “Rushhour” in LIVE video database. Parameter TI of the second and third frame is illustrated in Figure 3.11 (d). Figure 3.11 (e) is the variation of TI . For better illustration, Figure 3.11 (d) and Figure 3.11 (e) are the results of inversion and modulus. In this case, the level of variation is shown in black. The darker is a pixel, the more variations it experiences. It is clear that variation of TI in Figure 3.11(e) shows more temporal information than solely TI in Figure 3.11 (d). The calculation of $map(x, y)$ as the SSIM map function for the variation of distorted temporal information, $VTI_D(x, y)$, and the variation of reference temporal information, $VTI_R(x, y)$, is shown by the equations:

$$VTI_D(x, y) = [D_f(x, y) - R_{f-1}(x, y)] - [D_{f-1}(x, y) - R_{f-2}(x, y)] \quad (3.31)$$

$$VTI_R(x, y) = [R_f(x, y) - R_{f-1}(x, y)] - [R_{f-1}(x, y) - R_{f-2}(x, y)] \quad (3.32)$$

$$map(x, y) = SSIM(VTI_D(x, y), VTI_R(x, y)) \quad (3.33)$$



(a)



(b)



(c)



(d)



(e)

Figure 3.11: (a) First frame from the video “Rushhour”. (b) Second frame from the video “Rushhour”. (c) Third frame from the video “Rushhour”. (d) *TI* of second and third frame after inversion and modulus. (e) Variation of *TI* for three frames after inversion and modulus.

Instead of averaging $map(x, y)$, a weight function is used to improve the map function. Similar to the spatial part of Index2, the surrounding error weighting function is utilized to weight the SSIM map. This is shown as:

$$e(x, y) = VTI_D(x, y) - VTI_R(x, y) \quad (3.34)$$

$$w(x, y) = \frac{\sum_{m=x-1}^{x+1} \sum_{n=y-1}^{y+1} [e(x, y) - e(m, n)]^2}{N} \quad (3.35)$$

where $e(x, y)$ is the error map of $VTI_D(x, y)$ and $VTI_R(x, y)$, $w(x, y)$ is the weighting function for SSIM map of $VTI_D(x, y)$ and $VTI_R(x, y)$, and the parameters m, n and N are the same as defined in Equation (3.28). The temporal quality score is defined as:

$$Temporal = \frac{\sum_{x=1}^M \sum_{y=1}^N [w(x, y) \cdot map(x, y)]}{\sum_{x=1}^M \sum_{y=1}^N map(x, y)} \quad (3.36)$$

where parameters M and N are defined in Equation (3.29). The assessment flow of temporal part of Index2 is depicted in Figure 3.12.

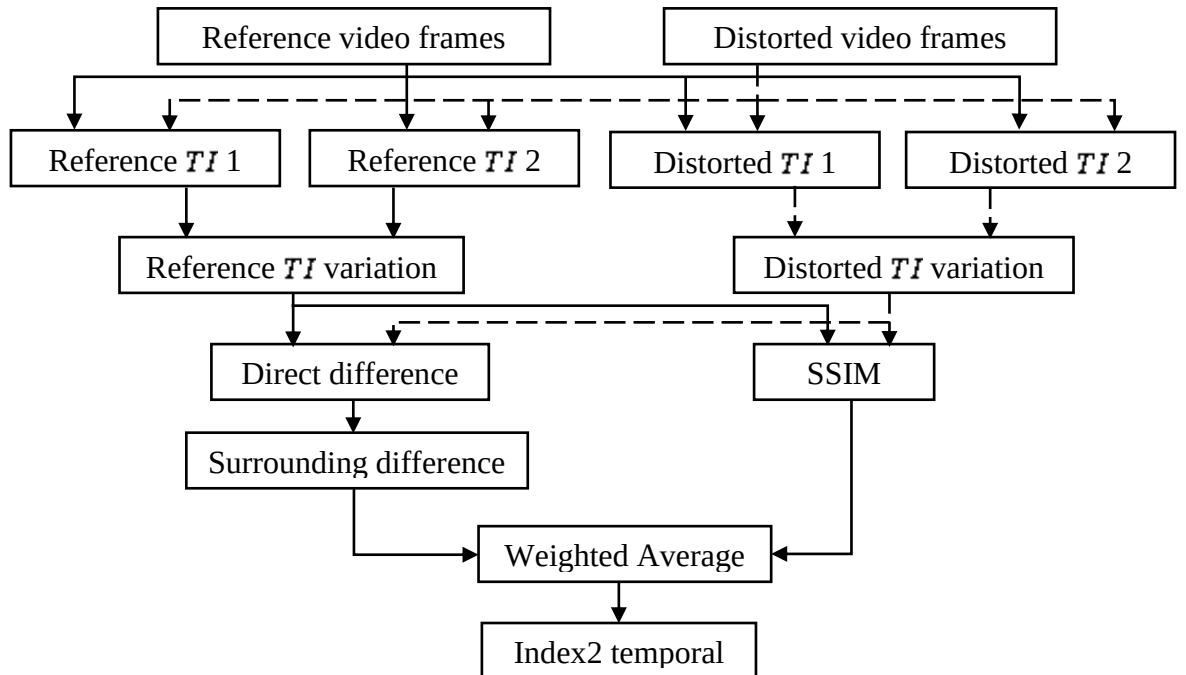


Figure 3.12: Temporal workflow of Index2.

3.3.4 Overall value of Index2

The overall video quality index, *Index2*, is the result of arithmetic mean of the spatial and temporal quality scores as:

$$Index2 = \frac{1}{2} \cdot (Spatial + Temporal) \quad (3.37)$$

3.3.5 Comparison of Index2 and MD-SSIM

The difference and improvement of Index2 compared to MD-MSSIM is listed in Table 3.2. Both of the methods employ different framework. MD-SSIM involves global and local scores while Index2 involves spatial and temporal scores. For the spatial part, MD-SSIM weights SSIM with MSE. Meanwhile, direct differences are used to weight SSIM. The direct difference are further processed to consider the effect surrounding pixels. This is not applied in MD-SSIM. In Index2, variation of *TI* is employed for considering temporal distortions while MD-SSIM considers temporal effect through difference of spatial scores. Weighting similar to the spatial part is applied in temporal part of Index2. MD-SSIM uses difference in spatial scores to weight the spatial scores to generate local scores.

Table 3.2: Comparison of Index2 and MD-SSIM.

MD-SSIM	Index2
Utilize MSE for weighting	Utilize direct difference for weighting
Does not consider surrounding pixels	Consider surrounding pixels
Utilize difference of spatial scores	Utilize variation of <i>TI</i>
Apply weighting in local part	Apply weighting in temporal part
Pooling of global and local scores	Pooling of spatial and temporal scores

3.4 Performance Evaluation

The performance of the proposed methods is measured and compared with other state-of-the-art VQA methods. Through standardized processes, the calculated objective scores of the VQA methods are correlated to the subjective scores produced from HVS. The correlation result serves as performance indicator of the proposed VQA methods compared with other existing methods. Two types of correlation coefficients as recommended by Video Quality Experts Group (2003) are used to measure the performance of the proposed metric. The correlation coefficients are Spearman rank-order correlation coefficient (SROCC) and Pearson linear correlation coefficient (CC). SROCC is used to measure prediction monotonicity, degree to which the predictions agree with actual score, while prediction accuracy, the correlation between prediction and actual score, is measured by CC (Video Quality Experts Group, 2003). SROCC is mathematically expressed as:

$$SROCC = 1 - \frac{6 \times \sum_{i=1}^n d_i^2}{n(n-1)^2} \quad (3.38)$$

where parameter d_i is the difference between i^{th} image's rank in the subjective and objective evaluations. CC, on the other hand, is defined as:

$$CC = \frac{\sum_{i=1}^n (s_i - \bar{s})(q_i - \bar{q})}{\sqrt{\sum_{i=1}^n (s_i - \bar{s})^2} \sqrt{\sum_{i=1}^n (q_i - \bar{q})^2}} \quad (3.39)$$

where parameter s_i is the subjective score of i^{th} image and parameter q_i is the result of mapping x_i , the objective score of the i^{th} image with the following fitting equation (Video Quality Experts Group, 2003; Seshadrinathan *et al.*, 2010a; Seshadrinathan *et al.*, 2010b):

$$q(x) = \tau_2 + \frac{\tau_1 - \tau_2}{1 + \exp^{-\left(\frac{x - \tau_3}{|\tau_4|}\right)}} \quad (3.40)$$

where parameters τ_1, τ_2, τ_3 , and τ_4 are parameters to be fitted, x is the objective scores, and $q(x)$ is the predicted score. Both SROCC and CC range from 0 to 1, where 1 indicates the predicted scores are the same with subjective scores and vice versa.

Besides that, root mean square error (RMSE) is also utilized for measuring the average distance of the predicted scores to the curve fitted by logistic function in equation (3.40). It also measures the deviation distance of the prediction from the actual score. RMSE is mathematically represented as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^I [q(x) - x]^2}{I}} \quad (3.41)$$

where I is the total number of objective scores. Parameter $q(x)$ and x are the same with Equation (3.39) and Equation (3.40). RMSE value ranges from 0 to infinity, where 0 indicates the predicted scores are the same with subjective scores. For the case of infinity, the predicted scores is very different and has long distance to the subjective scores.

3.5 Benchmark databases

Two benchmark video databases are utilized to compare the performances of the proposed VQA methods with existing VQA methods. The databases are LIVE video database (Seshadrinathan *et al.*, 2010a; Seshadrinathan *et al.*, 2010) and CSIQ database (Vu & Chandler, 2014). The quality scores in the database are obtained from human subjects that viewed and rated the video under controlled conditions. For instance, the subjects viewed and rated the videos in the database under similar lighting condition and viewing distance. The ratings from the subjects are then converted to subjective scores through subjective quality measurement methods as discussed in Section 2.3.

3.5.1 LIVE Video Database

LIVE video database (Seshadrinathan *et al.*, 2010a; Seshadrinathan *et al.*, 2010b) was released by the University of Texas, Austin in the year 2010. All video files have planar YUV 4:2:0 format without any headers. The spatial resolution of the videos is 768x432 pixels. There are four types of distortions for every video in the LIVE video database. The distortions are wireless distortions, IP distortions, H.264 compressions, and MPEG-2 compressions. There are four levels of distortions for each type of distortions except for the IP distortion with only three levels. Specifically, there are 15 distorted videos for each reference video. There are altogether 10 reference videos and 150 distorted videos in LIVE video database. The 10 videos involved are Blue Sky, River Bed, Pedestrian area, Tractor, Sunflower, Rush hour, Station, Park run, Shields, and Mobile & Calendar. These videos consist of indoor and outdoor scenes with fast and slow movements. There are also camera movements such as circular movement and panning of the camera. Table 3.3 summarizes the details of LIVE video database.

Table 3.3: Details of LIVE video database.

Type of Distortion	Number of Videos	Number of Image Frames
Wireless	40	12,868
IP	30	9,651
H.264 compression	40	12,868
MPEG-2 compression	40	12,868
Total	150	48,255

3.5.2 CSIQ Video Database

CSIQ (Vu & Chandler, 2014) video database was released by 318 Advanced Technology Research Center, School of Electrical and Computer Engineering, Oklahoma

State University. All videos are in the YUV 4:2:0 format at a resolution of 832x480. The duration of the video is 10 seconds each. There are 12 reference videos and 216 distorted videos. The videos involved are Chipmunks, Kimono, ParkScene, Carving2, Flowervase, Keiba, Timelapse, BasketballDrive, Cactus, PartyScene, BQMall, and BQTerrace. Each reference video is distorted by six types of distortions. They are H.264/ Advanced Video Coding (AVC) compression, packet loss rate (plr), Motion Joint Photographic Experts Group (MJPEG) compression, wavelet compression, white noise, and High Efficiency Video Coding (HEVC) compression. There are three levels of the distortions for each distortion type. Particularly, each reference video has 18 distorted videos. These videos consist of indoor and outdoor scenes with fast and slow movements. There are also camera movements such as circular movement and panning. Some videos have a main object such as a human but some videos only consist of scenery. There are also videos with many subjects such as BasketballDrive. The details of CSIQ database is listed in Table 3.4.

Table 3.4: Details of CSIQ video database.

Type of Distortion	Number of Videos	Number of Image Frames
H.264/AVC	36	13,710
plr	36	13,710
MJPEG compression	36	13,710
Wavelet compression	36	13,710
White noise	36	13,710
HEVC	36	13,710
Total	216	82,260

3.6 Experimental Settings

In this section, the experiment settings are presented. Section 3.6.1 explains the functions used in the two proposed methods while Section 3.6.2 elaborates the functions involved in performance calculations.

All computations were performed using a laptop with 64-bit Window 7 operating system. The laptop is installed with Intel Core i5-4210U processor and 8GB of RAM. The whole process of testing and evaluation are performed by using the software MATLAB version 2013b. The toolboxes used are Image Processing Statistics and Machine Learning toolboxes. There are also built-in functions from MATLAB which had been used during the experiment.

3.6.1 Experiment procedures of the proposed method

For Index1, the experiment procedures are elaborated here. First of all, distorted and reference frames are read and stored into MATLAB through *readyuv* function. The code for this function is shown as follows:

```
function [Y U V]=read_yuv(filename,height,width,frames,format)
%read_yuv    Read the yuv file frame by frame and save the data into
matrix Y,U,V respectively
%filename    The name of the yuv file
%width       The width of the yuv file
%height      The height of the yuv file
%frames      The total frame of the yuv file

fid=fopen(filename,'r');
Y=zeros(height,width,frames,'uint8');
U=zeros(height/2,width/2,frames,'uint8');
V=zeros(height/2,width/2,frames,'uint8');
for frame=1:frames
    Y(:,:,frame)=fread(fid,[width height],'uint8=>uint8');
    U(:,:,frame)=fread(fid,[width/2 height/2],'uint8=>uint8');
    V(:,:,frame)=fread(fid,[width/2 height/2],'uint8=>uint8');
end
Y = double(Y);
U = double(U);
V = double(V);
fclose(fid);
```

Index1 and Index2 methods only deal with the luminance part of the video. Since all videos in the databases are in YUV format, only Y frames are utilized. U and V frames are of the chrominance part of the video and are not involved in getting the desired results. Then, the computation of spatial part of Index1 starts by generating SSIM map using the

function *ssim* from Image Processing toolbox. JND threshold is applied using a comparison operator. All pixels from SSIM map which are greater than five percent of the maximum SSIM index are assigned with the perfect value of 1. The coding is shown as:

```
%ssim_map is the SSIM map from eq 2.15
ssim_map(ssim_map>=0.95*max(ssim_map(:))) = 1;
```

Then the spatial quality of Index1 is computed through *mean2* function which is built-in in MATLAB to calculate the mean of a 2D matrix. For the temporal part of Index1, video frames from reference and distorted videos are read and compared by using *read_yuv* and *ssim* functions. Similar procedures for the JND threshold from the spatial part are applied here. Two temporal quality indices can be obtained by averaging two SSIM map from the two parts in the temporal part of Index1. The coding is shown as:

```
%ssim_map1 is the SSIM map of the first part
%ssim_map2 is the SSIM map of the second part
ssim_map1(ssim_map1>=0.95*max(ssim_map1(:))) = 1;
ssim_map2(ssim_map2>=0.95*max(ssim_map2(:))) = 1;
```

The total temporal quality index of Index1 is the average of these two indices. This is done by simple addition and division. The overall video quality is then the average of spatial and temporal quality indices of Index1 which also involves simple addition and division.

For index2, the frames from reference and distorted videos are first read and compared using *read_yuv* and *ssim* functions. Then, the SSIM map is weighted by a weight. The weight is calculated by direct subtraction of reference and distorted frames. Then, using nested *for* loops, each pixel is compared to its surrounding eight pixels. For the comparison, it is the squaring of the values after subtraction. This is shown as:

```
% W(x,y) is weight at particular pixel position (x,y)
% E(x,y) is the error at particular pixel position (x,y)
% W(m,n) is the error at particular pixel position (m,n) which is
the surrounding of (x,y)
for m = x-1:x+1
    for n = y-1:y+1
        W(x,y) = W(x,y) + (E(x,y)-E(m,n)).^2/8;
    end
end
```

The spatial index of Index2 is the weighted average of the SSIM map. In this case, SSIM map is pixel-wise multiplied by the weight and pixel-wise divided by the sum of the weight.

This is shown as:

```
% spatial_index is the spatial quality index of Index1
spatial_index = sum(sum(W.*ssim_map))./sum(sum(W));
```

For the temporal part, temporal information is computed by direct difference of frames at different time instant. The variation of temporal information is the difference of temporal information calculated. This is shown as:

```
% TI1 is the first temporal information of frame f and f-1
% TI2 is the second temporal information of frame f-1 and f-2
% VTI is the variation of temporal information
TI1 = Dis(:, :, f) - Ref(:, :, f-1);
TI2 = Dis(:, :, f-1) - Ref(:, :, f-2);
VTI = TI2 - TI1;
```

The reference and distorted variation of temporal information are then compared using *ssim* function. The calculation of weight for the temporal part follows the procedure of the spatial part of Index2. The temporal index of Index2 is the weighted average of resulting SSIM map. Lastly, the overall video quality index is the average of spatial and temporal quality index which can be done by simple addition and division.

After computing the predicted scores for all videos, the data are stored as a column vector with the length depends on the number of videos. The data for each database are stored as different vectors. After assigning names for the data, they are saved as the MATLAB data file format with .mat extension using *save* function. These data can be reloaded using another MATLAB built-in function *load*.

3.6.2 Experiment procedures of the performance evaluation

The results of comparing the objective indices of Index1 and Index2 methods are also compared to subjective indices from the database by using the correlation formulas

presented in Section 3.4. The subjective results are imported from Excel files into MATLAB using built-in *xlsread* function. SROCC and CC are computed using *corr* function from Statistics and Machine Learning Toolbox. This process is shown as:

```
% performance evaluation
% sub_score is subjective score from databases
% pred_score is the objective score from proposed methods
SROCC = corr(sub_score, pred_score, 'type', 'spearman');
CC = corr(sub_score, pred_score, 'type', 'Pearson');
```

Before computing CC and RMSE, fitting of equation (3.40) is done using *nlinfit* and *nlpredci* function from Statistics and Machine Learning Toolbox. This is shown as:

```
% method_score is objective score from the VQA methods
[bayta,ehat,J] = nlinfit(method_score,mos, @eq_3.40,beta);
[pred_score,~] = nlpredci(@eq_3.4, method_score,bayta,ehat,J);
```

The correlation results are controlled and approximated up to four decimal points. For computing the computational complexity, it is measured in terms of computational time. The function of stopwatch timer (*tic* and *toc*) in MATLAB is used to calculate the exact time needed to perform the VQA methods in the unit of seconds. For loop is utilized to compute the computational time for ten times in order to get accurate results of average computational time. This is shown as:

```
% Computational Complexity Computation
tic
for i = 1:10
    run method for ten times here
end
time = toc;
time = time/10;
```

3.6.3 Overall flow of the experiment

Figure 3.13 shows the overall flow of the experiment. The experiment starts with the design of methods. After that, the methods are run on both databases to obtain the objective scores. Besides that, the existing methods are also evaluated on the databases to obtain objective scores respectively. Then, SROCC, CC, and RMSE are computed using

equations stated in Section 3.4. Before computing CC and RMSE, nonlinear fittings are done using *nlinfit*. Next, all methods are run on the same video for ten times. The average of the running time is the computation time for each method. Finally, the performance of the proposed methods is compared with existing methods and discussed.

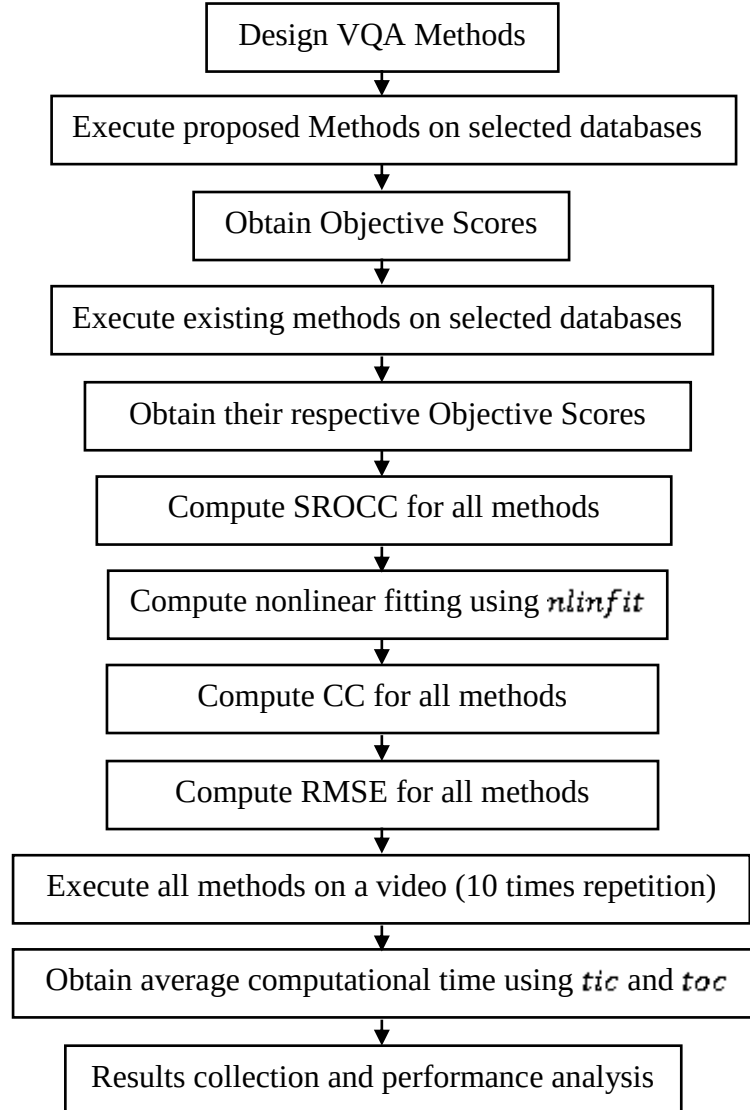


Figure 3.13: Flow of experiment.

3.8 Summary

Chapter 3 explained the proposed VQA methods. There are two new VQA methods, Index1 and Index2, proposed in this thesis. Index1 is based on JND. The JND concept is applied on the spatial and temporal aspects of Index1. Index2 is an extension of MD-SSIM. The concept of error weighting is applied on both spatial and temporal quality evaluations. For the temporal part of Index2, variation of TI is applied instead of TI of the video. The metrics used for performance evaluation are also introduced. Benchmark databases involved and the experimental settings are also elaborated. The overall flow of the experiment is also discussed briefly in this chapter. The performance of both metrics is evaluated in Chapter 4.

CHAPTER 4

EXPERIMENT RESULTS, ANALYSIS AND DISCUSSION

4.1 Introduction

In this thesis, there are two proposed VQA methods, Index1 and Index2. Both of the methods are compared separately to other state-of-the-art VQA methods in the performance evaluation. They are also compared to each other in Section 4.6. State-of-the-art VQA methods that are compared with the proposed VQA methods are MOVIE (Seshadrinathan & Bovik, 2010) and ViS₃ (Vu & Chandler, 2014). ViS₃ is applied in its default implementation. On the other hand, MOVIE requires long computational time, thus the results of MOVIE (Vu & Chandler, 2014; Seshadrinathan & Bovik, 2010) are used directly for comparison purpose.

Analysis, and discussion of the proposed VQA methods are shown in Sections 4.2. Lastly, a brief summary is presented in Section 4.3.

4.2 Results, discussion, and analysis for Index1 and Index2

In this section, the results and analysis of Index1 and Index2 are presented. Both methods are compared with existing state-of-the-art methods. They are also being compared to their predecessor methods which are Mod-MSE and MD-SSIM. In Section 4.2.4, both methods are compared among different types of videos from the two databases. The computational complexity is also analysed and compared in Section 4.2.5.

4.2.1 Statistical performance with Existing Methods

The results of SROCC and CC for all VQA methods of all database are shown in Table 4.1. In the table, two best performing VQA methods are highlighted. The results are discussed according to the databases. For LIVE video database, Index1 and Index2 perform well. For SROCC, ViS₃ perform the best for LIVE videos. This is followed by the Index2. They achieve 0.8160 and 0.8013 respectively. Index1 also performs pretty well, which attains 0.7967. MOVIE, on the other hand, has the worst result that is 0.7890. For CC, the best result is still attained by ViS₃ while Index1 achieves the second best result. They achieve 0.8300 and 0.8156 respectively. Index2 achieves 0.8129, which is slightly lower than Index1. Similar to SROCC, MOVIE attains the worst result for CC, which is 0.8116. However, it is shown that the Index1 and Index2 have an almost comparable performance for LIVE videos.

Table 4.1: Overall Performance of VQA methods for all databases.

Metric	LIVE		CSIQ	
	SROCC	CC	SROCC	CC
MOVIE	0.7890	0.8116	0.8060	0.7880
ViS ₃	0.8160	0.8300	0.8410	0.8300
Index1	0.7967	0.8156	0.6434	0.6963
Index2	0.8013	0.8129	0.8676	0.8704

For CSIQ database, Index1 has the worst result in terms of SROCC and CC. It achieves 0.6434 and 0.6963 respectively. Conversely, Index2 has the best performance in both SROCC and CC. It attains 0.8676 and 0.8704 for SROCC and CC respectively. The results are much higher than the second best results achieved by ViS₃. ViS₃ achieves 0.8410 for SROCC and 0.8300 for CC. Similar to the results in LIVE, MOVIE achieves moderate results, which are 0.8060 and 0.7880 for SROCC and CC respectively.

Table 4.2 and Table 4.3 show the results of all VQA methods for all videos with different types of distortions in LIVE database. For wireless distorted videos, ViS₃ has the highest correlation with subjective scores in terms of SROCC. It attains 0.8450 of SROCC. For CC, the best result is achieved by the Index1. It achieves 0.8522 while ViS₃ achieves 0.8460. The results show that ViS₃ and Index1 have comparable performance in evaluating videos distorted by wireless distortions. For Index2, it has the poorest performance for SROCC and CC for wireless distorted videos. It attains 0.8066 and 0.8127 for SROCC and CC. The correlation results are very close to MOVIE. MOVIE attains 0.8109 and 0.8386 for SROCC and CC respectively. When considering the two proposed VQA methods, Index1 performs better than Index2 in wireless distorted videos.

Table 4.2: SROCC of all distortions in LIVE database.

Distortion	MOVIE	ViS ₃	Index1	Index2
Wireless	0.8109	0.8450	0.8261	0.8066
IP	0.7157	0.7880	0.7206	0.7976
H.264	0.7664	0.7570	0.8026	0.7769
MPEG-2	0.7330	0.7300	0.7480	0.7720

Table 4.3: CC of all distortions in LIVE database.

Distortion	MOVIE	ViS ₃	Index1	Index2
Wireless	0.8386	0.8460	0.8522	0.8127
IP	0.7622	0.8160	0.7613	0.8260
H.264	0.7902	0.7730	0.8418	0.7907
MPEG-2	0.7596	0.7460	0.7783	0.8030

For videos with IP distortion, Index2 achieves the best results for both SROCC and CC. It achieves 0.7976 for SROCC and 0.8260 for CC. The second best result is achieved by ViS₃. ViS₃ attains 0.7880 and 0.8160 for SROCC and CC respectively. For the performance of IP distorted videos in terms of SROCC, MOVIE performs the worst. It

only achieves SROCC of 0.7157 while Index1 attains slightly higher, which is 0.7206. The results in CC is the opposite case of the results in SROCC. Index1 has the poorest performance, which is 0.7613. Yet, MOVIE attains CC of 0.7622. In terms of IP distorted videos, Index2 has better performance compared to Index1.

The next category of videos to be analysed is H.264 compressed videos. Index1 and Index2 are the best two performing methods. Index1 has the highest correlation, which is 0.8026 for SROCC and 0.8418 for CC. Although Index2 has the second best results, the results are much lower than the best results. Index2 attain 0.7769 and 0.7907 for SROCC and CC respectively. Dissimilar to the results in wireless and IP distorted videos, ViS₃ has the worst performance in H.264 compressed videos. It merely achieves 0.7570 for SROCC and 0.7730 for CC. MOVIE has the performance which is slightly lower than Index2. For SROCC, it attains 0.7664. Meanwhile, it achieves 0.7902 for CC. In overall, Index1 has better performance in H.264 compressed videos compared to Index2.

Lastly, the results for MPEG-2 compressed videos are quite similar to the results for H.264 compressed videos. The two proposed VQA methods have the highest correlation results for both SROCC and CC. Index2 has the best performance in terms of SROCC and CC. It achieves 0.7720 and 0.8030 respectively. Index1 has much lower results than the best performance although it achieves the second best result. This condition is opposite for the results of videos distorted by H.264 compression. Index1 attains 0.7480 for SROCC and 0.7783 for CC. Akin to the results in H.264 compressed videos, ViS₃ has the poorest performance that achieves 0.7300 and 0.7460 for SROCC and CC respectively. MOVIE has analogous results to ViS₃ but is slightly better. It achieves 0.7330 for SROCC and 0.7596 for CC. Index2 performs better than the Index1 for MPEG-2 compressed videos.

For LIVE database, MOVIE has the worst performance in videos distorted by all distortion types. ViS₃ has poor performance for the compressed videos. This is because it has the poorest correlation results in H.264 and MPEG-2 compressed videos. Index1 and Index2 perform similarly for all videos. They have good performance in evaluating compressed videos. For Index1, it has a better performance in wireless distorted videos and a poorer performance in assessing the quality of videos with IP distortions. The condition is reversed for the case of Index2. It has a better performance in IP distorted videos while lower correlation results in videos distorted by wireless distortions.

The results of the proposed and existing VQA methods for videos in CSIQ database are shown in Table 4.4 and Table 4.5. For H.264 compressed videos, Index1 has the best performance. It achieves 0.9205 for SROCC and 0.9291 for CC. In terms of SROCC, ViS₃ has the second best result, which is 0.9200. Index2 attains 0.9117, which is slightly lower than ViS₃. However, Index2 has the second best results in terms of CC. It achieves 0.9197 while ViS₃ only attains 0.9180. MOVIE has the worst results. It achieves 0.8970 and 0.9040 for SROCC and CC respectively. For H.264 compressed videos in CSIQ database, Index1 has better performance than Index2.

Table 4.4: SROCC of all distortions in CSIQ database.

Distortion	MOVIE	ViS ₃	Index1	Index2
H.264	0.8970	0.9200	0.9205	0.9117
plr	0.8860	0.8560	0.7503	0.8690
MJPEG	0.8870	0.7890	0.4803	0.8870
Wavelet	0.9000	0.9080	0.8728	0.8952
White noise	0.8430	0.9280	0.9107	0.9375
HEVC	0.9330	0.9170	0.9362	0.9542

Table 4.5: CC of all distortions in CSIQ database.

Distortion	MOVIE	ViS ₃	Index1	Index2
H.264	0.9040	0.9180	0.9291	0.9197
plr	0.8820	0.8500	0.7711	0.8858
MJPEG	0.8820	0.8000	0.4288	0.8911
Wavelet	0.8989	0.9080	0.8964	0.9144
White noise	0.8550	0.9160	0.9297	0.9685
HEVC	0.9370	0.9330	0.9504	0.9635

The next videos are videos with plr distortion. MOVIE has the best performance in terms of SROCC that attains 0.8860. This is followed by Index2 which achieves 0.8690. Meanwhile, Index2 has the best correlation results in terms of CC. It attains 0.8858 while MOVIE achieves the second best result, which is 0.8820. Index1 has the poorest performance in plr distorted videos. It only attains 0.7503 and 0.7711 for SROCC and CC respectively. ViS₃ achieves results better than Index1 but worse than MOVIE and Index2. It achieves 0.8560 for SROCC and 0.8500 for CC. When considering only the two proposed VQA methods, Index2 has much better performance than Index1.

Index2 and MOVIE achieve the same result in MJPEG compressed videos in terms of SROCC. Both of them achieve 0.8870. For CC, Index2 has better performance than MOVIE. It attains 0.8911 while MOVIE achieves 0.8820. Index1 has the performance that is much lower than the other three VQA methods. It only attains 0.4803 for SROCC and 0.4288 for CC. This is the main reason that Index1 has bad performance in CSIQ database. ViS₃, on the other hand, achieves moderate results, which are 0.7890 and 0.8000 for SROCC and CC respectively. In overall, Index2 has the best performance in assessing the quality of MJPEG compressed videos. Meanwhile, Index1 does not perform well for MJPEG compressed videos.

Next, the results of VQA methods involving videos distorted by wavelet compression in CSIQ database are discussed. ViS₃ has the best correlation results in terms of SROCC that is 0.9080. This is followed by MOVIE, which attains 0.9000. Index2 has slightly lower performance than MOVIE that achieves 0.8952. This is not the case in terms of CC. Index2 has the best performance that achieves 0.9144. ViS₃ attains 0.9080, which is faintly lower correlation result than Index2. This is followed by MOVIE that attains 0.8989. For Index1, it has performance that is somewhat lower than MOVIE. It achieves 0.8728 and 0.8964 for SROCC and CC respectively. For wavelet compressed videos, Index2 outperforms Index1.

For white noise distorted videos in CSIQ database, Index2 has the best performance. It achieves 0.9375 for SROCC and 0.9685 for CC. ViS₃ has the second best correlation results in terms of SROCC and attains 0.9280. Index1 has the result that is slightly lower than ViS₃ at 0.9107. In terms of CC, the results are the reversed of the results for SROCC. Index1 has the second best performance. It attains 0.9297 while ViS₃ only achieves 0.9160. MOVIE has the worst performance for both SROCC and CC. It achieves 0.8430 and 0.8550 for SROCC and CC respectively. Index2 still has better performance for white noise distorted videos than Index1.

The last category of videos to be considered in CSIQ database is HEVC compressed videos. The proposed VQA methods are the two leading methods in terms of SROCC and CC. Index2 has the highest correlation results to the subjective scores. It achieves 0.9542 for SROCC and 0.9635 for CC. On the other hand, Index1 achieves the second best result, which attains 0.9362 and 0.9635 for SROCC and CC respectively. ViS₃ has the poorest performance among the four VQA methods. It only achieves 0.9170 for SROCC and 0.9330 for CC. MOVIE has the results that are slightly better than ViS₃ at

0.9330 for SROCC and 0.9370 for CC. In terms of proposed VQA methods, Index2 has better results than Index1 for HEVC distorted videos.

In overall, Index1 does not perform well for videos in CSIQ database. This is mainly due to its incompetence to assess the quality of MJPEG compressed videos. From the tables, Index1 has low performance in evaluating the quality of videos distorted by MJPEG compression, plr, and wavelet compression. The performance of Index1 in videos distorted by plr and wavelet compression is still acceptable as the correlation results are close to other VQA methods. Index1 can perform well for LIVE database because there are no videos distorted by MJPEG compression. Thus, the correlation results of Index1 for CSIQ databases is lower than the LIVE database. ViS₃ has good performance in videos distorted by H.264 compression, wavelet compression, and white noise. However, MOVIE has a moderate performance for all videos. It has better results in plr and MJPEG distorted videos. Index2 has the best overall performance. It is able to assess video quality well for all distortion types of CSIQ database.

Scatter plots of Index1 are shown in Figure 4.1 and Figure 4.2. Figure 4.1 is the scatter plot for LIVE database while the scatter plot for CSIQ database is shown in Figure 4.2. When the points are closer to the curve fitted with logistic function, the VQA method will have better correlation results. In other words, the VQA method can predict the video quality with good correlation to the subjective scores. The distance of the points to the curve is calculated by using RMSE. The results are shown in Table 4.6. From Figure 4.1, the curve line by Index1 shows good correlation. This is observed from the points in the scatter plot which are fairly close to the curve. The result from Table 4.6 also shows that the points only have average RMSE of 6.3478 from the curve function.

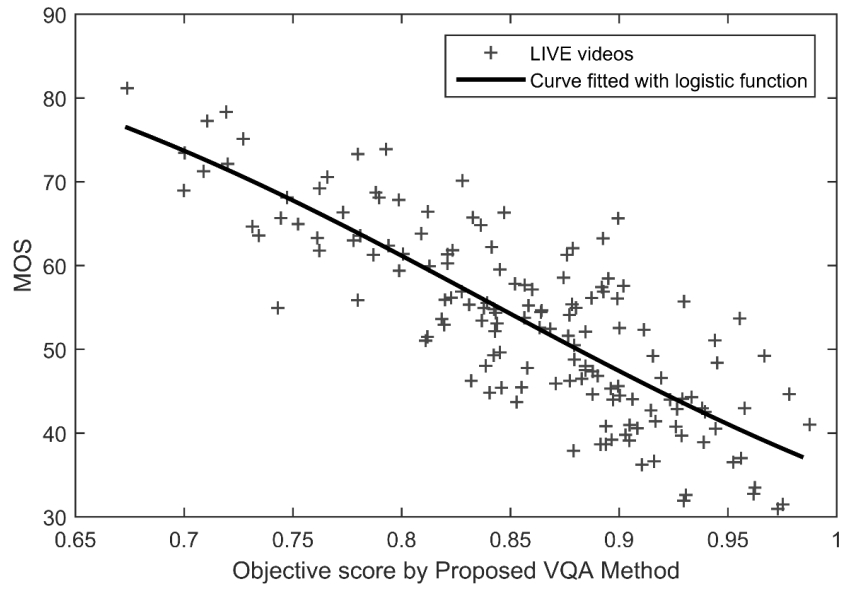


Figure 4.1: Scatter plot of Index1 for LIVE videos.

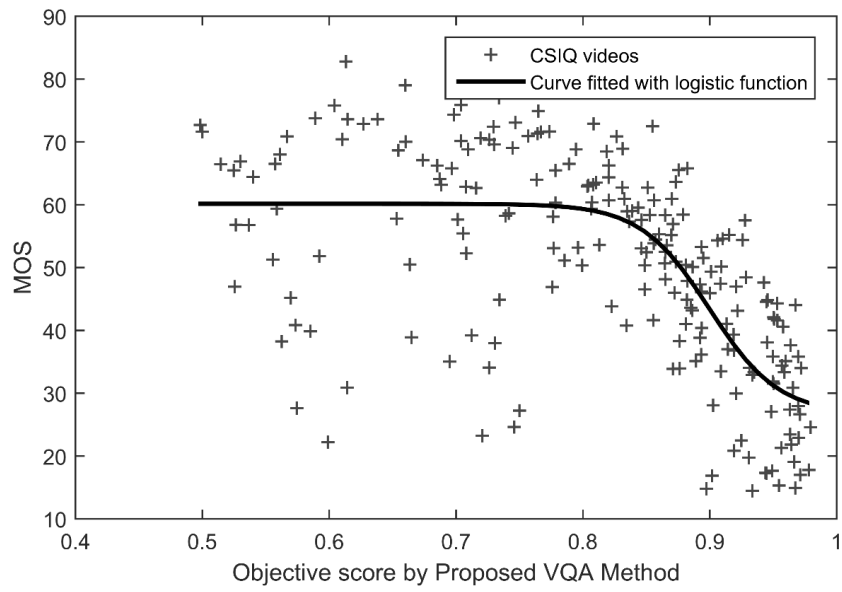


Figure 4.2: Scatter plot of Index1 for CSIQ videos.

Table 4.6: RMSE of Index1 for LIVE and CSIQ databases.

Database	RMSE
LIVE	6.3478
CSIQ	11.9396

For CSIQ database, Index1 has lower performance than for the LIVE database. The points in Figure 4.2 are moderately far to the curve fitted by the logistic function in equation (3.34). Index1 has RMSE of 11.9396 unit, which is almost double of the RMSE value for LIVE database. The points below the curve function before it bends have longer distances than other points. These points belong to the video distorted by MJPEG compression. This also proves that Index1 has poorer performance for MJPEG compressed videos.

The scatter plots of Index2 for LIVE and CSIQ databases are shown in Figure 4.3 and Figure 4.4 respectively. The results of RMSE are shown in Table 4.7. From the figures, Index2 has a good correlation to the subjective scores. The Index2 points in LIVE videos are close to the curve fitted by the logistic function. Index2 has RMSE of only 6.3934, which is quite small. Akin results are found in CSIQ database. Index2 has RMSE of 8.1878. The points in Figure 4.4 have longer distances to the curve compared to the points for LIVE database. However, Index2 still has good performance for CSIQ database.

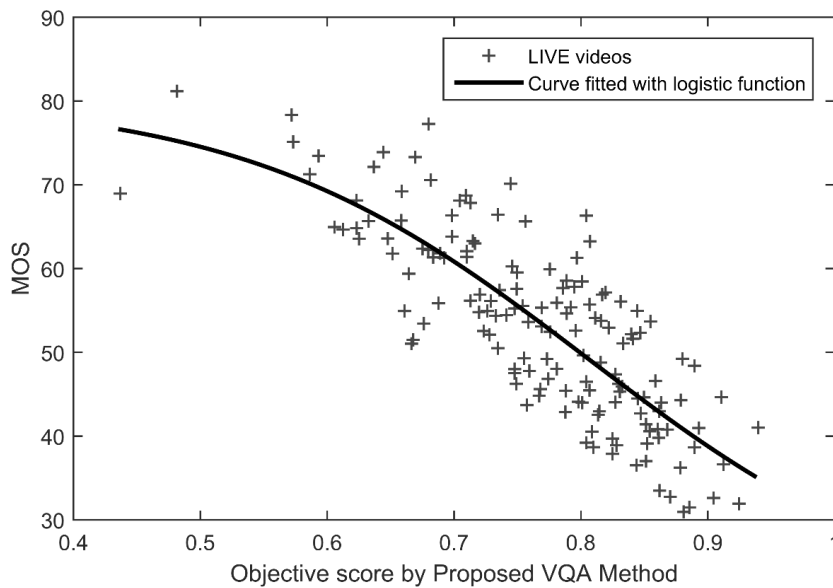


Figure 4.3: Scatter plot of Index2 for LIVE videos.

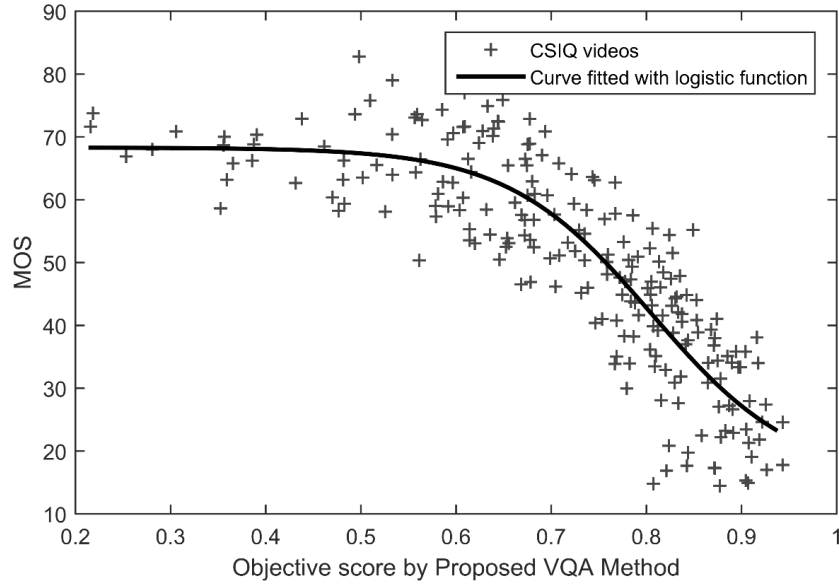


Figure 4.4: Scatter plot of Index2 for CSIQ videos.

Table 4.7: RMSE of Index2 for LIVE and CSIQ databases.

Database	RMSE
LIVE	6.3934
CSIQ	8.1878

4.2.2 Statistical performance of Index1

The results of Index1 are also compared with its predecessor, Mod-MSE, and Index1 without applying JND. This is because Index1 is based on the improvement from Mod-MSE and the effects of applying JND can be examined. Since only LIVE database is used for Mod-MSE, only the results from LIVE database is used. Detailed results of Index1, Mod-MSE, and Index1 without JND are shown in Table 4.8 and Table 4.9. The best performing method for each type of distortion is highlighted.

Table 4.8: SROCC of Index1 related VQA methods for LIVE videos.

Distortions	Mod-MSE	Index1 without JND	Index1
Wireless	0.6694	0.7795	0.8261
IP	0.5871	0.6796	0.7206
H.264	0.4638	0.8161	0.8026
MPEG-2	0.4170	0.7300	0.7480
Overall	0.5931	0.7880	0.7967

Table 4.9: CC of Index1 related VQA methods for LIVE videos.

Distortions	Mod-MSE	Index1 without JND	Index1
Wireless	0.6974	0.8056	0.8522
IP	0.6033	0.7165	0.7613
H.264	0.4988	0.8408	0.8418
MPEG-2	0.4582	0.7616	0.7783
Overall	0.6016	0.7961	0.8156

In terms of SROCC, Index1 has the best performance for all videos in LIVE database except H.264 compressed videos. Index1 without JND has the best results for evaluating the quality of videos distorted by H.264 compression errors, which is 0.8161. Index1 only achieves SROCC of 0.8026. For the remaining videos, the second best performing method is Index1 without JND. Mod-MSE has the worst correlation results for all videos in LIVE database.

For CC, Index1 has the best performance for all videos in LIVE database. The condition is reversed for H.264 compressed videos. Index1 attains the best performance while Index1 without JND falls behind Index1. The results for other videos remain the same as SROCC results. Mod-MSE still has the poorest performance for all videos in LIVE database. Generally, Index1 method has better performance compared to the Mod-MSE and Index1 without JND.

4.2.3 Statistical performance of Index2

Similar to Index1, the results of Index2 are also compared with its predecessor, MD-SSIM, and Index2 without weight. For Index2 without weight, the SSIM map in equations (3.29) and (3.36) are averaged instead of weighted average. The results are compared in both LIVE and CSIQ databases. For each database, the results of these methods are compared according to the distortion types and in overall. SROCC and CC results are shown in Table 4.10 and Table 4.11. The best performing method for each type of distortion is highlighted.

Table 4.10: SROCC of Index2 related VQA methods for LIVE videos.

Distortions	MD-SSIM	Index2 without weight	Index2
Wireless	0.7364	0.4520	0.8066
IP	0.6743	0.3580	0.7976
H.264	0.7812	0.6850	0.7769
MPEG-2	0.7686	0.5892	0.7720
Overall	0.7787	0.4972	0.8013

Table 4.11: CC of Index2 related VQA methods for LIVE videos.

Distortions	MD-SSIM	Index2 without weight	Index2
Wireless	0.7438	0.4850	0.8127
IP	0.7260	0.4863	0.8260
H.264	0.7903	0.6947	0.7907
MPEG-2	0.7901	0.6167	0.8030
Overall	0.7862	0.5131	0.8129

In terms of SROCC, Index2 has the best performance for all videos in LIVE database except H.264 compressed videos. MD-SSIM has the best results for evaluating the quality of videos distorted by H.264 compression errors, which is 0.7812. Index2 achieves a very close SROCC of 0.7769. For the remaining videos, the second best

performing method is MD-SSIM. Index2 without weight has the worst correlation results for all videos in LIVE database.

For CC, Index2 has the best performance for all videos in LIVE database. The condition is reversed for H.264 compressed videos. Index2 attains the best performance while MD-SSIM falls behind Index1. The performance in terms of SROCC results is in the same ranking. Index2 without weight still has the poorest performance for all videos in LIVE database. Generally, Index2 method has better performance compared to the MD-SSIM and Index2 without weight.

Index2, MD-SSIM, and Index2 without weight are also compared in CSIQ database. The results of SROCC and CC are shown in Table 4.12 and Table 4.13. In terms of SROCC, Index2 has the best performance for three distortion types, they are MJPEG, white noise, and HEVC. For H.264 and wavelet distorted videos, MD-SSIM achieves the best results, which are 0.9483 and 0.9001 respectively. Yet, Index 2 has close performances to MD-SSIM, which are 0.9117 and 0.8952 respectively. Index2 without weight attains the best results for plr distorted videos. It achieves 0.9202 while Index 2 achieves moderate results of 0.8690. In overall, Index2 has the best performance compared to MD-SSIM and Index2 without weight. The results of Index2 are far better than Index2 without weight.

Table 4.12: SROCC of Index2 related VQA methods for CSIQ videos.

Distortion	MD-SSIM	Index2 without weight	Index2
H.264	0.9483	0.9037	0.9117
plr	0.7601	0.9202	0.8690
MJPEG	0.8517	0.7156	0.8870
Wavelet	0.9001	0.7897	0.8952
White noise	0.9143	0.9359	0.9375
HEVC	0.9449	0.8551	0.9542
Overall	0.7934	0.5908	0.8676

Table 4.13: CC of Index2 related VQA methods for CSIQ videos.

Distortion	MD-SSIM	Index2 without weight	Index2
H.264	0.9390	0.9156	0.9197
plr	0.7323	0.8589	0.8858
MJPEG	0.8570	0.7318	0.8911
Wavelet	0.9188	0.8429	0.9144
White noise	0.9372	0.9687	0.9685
HEVC	0.9645	0.8636	0.9635
Overall	0.8000	0.6447	0.8704

For the results in terms of CC, Index2 still achieves the best performance for MJPEG distorted videos. The condition reversed for the case of white noise and HEVC. MD-SSIM achieves the best results of 0.9645 for HEVC distorted videos while Index2 without weight has the best performance on white noise distorted videos with CC of 0.9687. Again, Index2 has close performances to the best performing methods on these two types of distortions. MD-SSIM still attains the best CC for H.264 and wavelet distorted videos with 0.9390 and 0.9188 respectively. Meanwhile, Index2 still has very close performances of 0.9197 and 0.9144. For plr distorted videos, Index2 is able to achieve the best results. This is the reversed condition in the case of SROCC where Index2 without weight achieves the best results. On the average, Index2 still has the best performances compare to MD-SSIM and Index2 without weight.

In conclusion, Index2 performs better than MD-SSIM and Index2 without weight. This shows that the weights calculated in equations (3.28) and (3.25) play an important role in improving the SSIM map. Especially for the case of Index2 without average, its performance differences are large.

4.2.4 Statistical performance for videos

From Section 4.2.1 to Section 4.2.2, performances of VQAs are compared according to databases. In this section, the performance of all VQA methods is measured for different types of videos from the two databases. For instance, the performance of the VQA methods for H.264 compressed videos from the LIVE database is combined with the CSIQ database. This analysis can give further insight into the performance of each VQA method towards each type of distortion.

The first type of videos is H.264 compressed videos. The results from LIVE and CSIQ databases are utilized. The results are shown in Table 4.14 and Figure 4.5. It is observed that Index1 and Index2 have good performance for H.264 compressed videos. Index1 has the best performance. It achieves 0.8616 and 0.8855 for SROCC and CC respectively. Index2 has the second best result. It attains 0.8443 for SROCC and 0.8552 for CC. MOVIE has the poorest performance in terms of SROCC. It achieves 0.8317 while ViS₃ attains 0.8385. Thus, ViS₃ has better performance than MOVIE in terms of SROCC. The performance in terms of CC is the opposite of SROCC results. ViS₃ has poorer result than MOVIE. ViS₃ attains 0.8455 while MOVIE achieves 0.8471.

Table 4.14: SROCC and CC for H.264 compressed videos.

Metric	SROCC	CC
MOVIE	0.8317	0.8471
ViS ₃	0.8385	0.8455
Index1	0.8616	0.8855
Index2	0.8443	0.8552

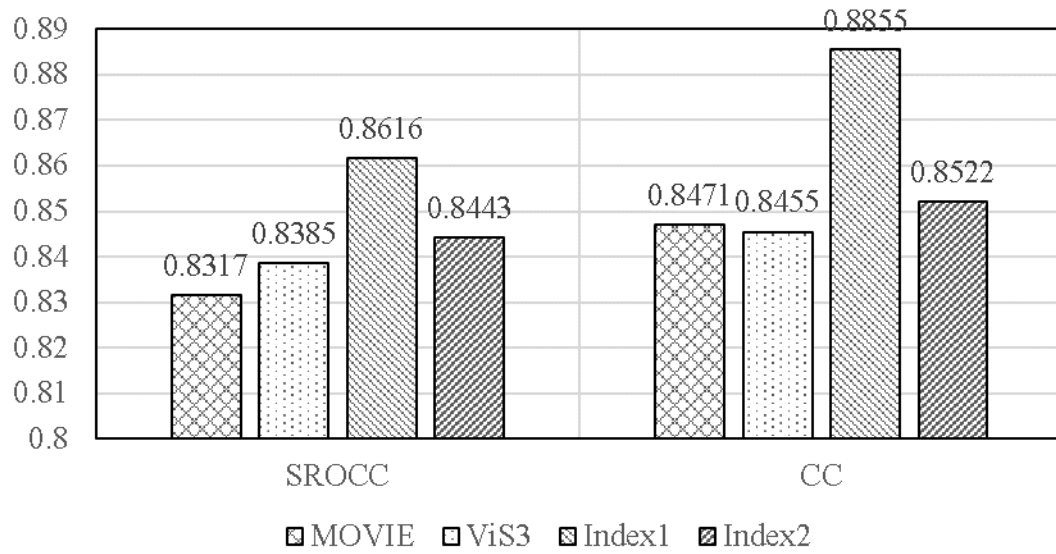


Figure 4.5: SROCC and CC for H.264 compressed videos.

The scatter plots of the two proposed VQA methods are shown in Figure 4.6 and Figure 4.7. The RMSE results are shown in Table 4.15. The scatter plot in Figure 4.6 shows that the points are close to the curve fitted by the logistic function. Index1 only has RMSE of 6.6349 while Index2 attains 8.1561. This is also evident by comparing the two scatter plots. The points in the bottom right of Figure 4.7 have larger distances to the curve. Thus, Index1 performs better than Index2 for H.264 compressed videos. Nevertheless, both of the proposed VQA methods have better performance than MOVIE and ViS₃ in evaluating H.264 compressed videos

Table 4.15: RMSE of Index1 and Index2 for H.264 compressed videos.

VQA Methods	RMSE
Index1	6.6349
Index2	8.1561

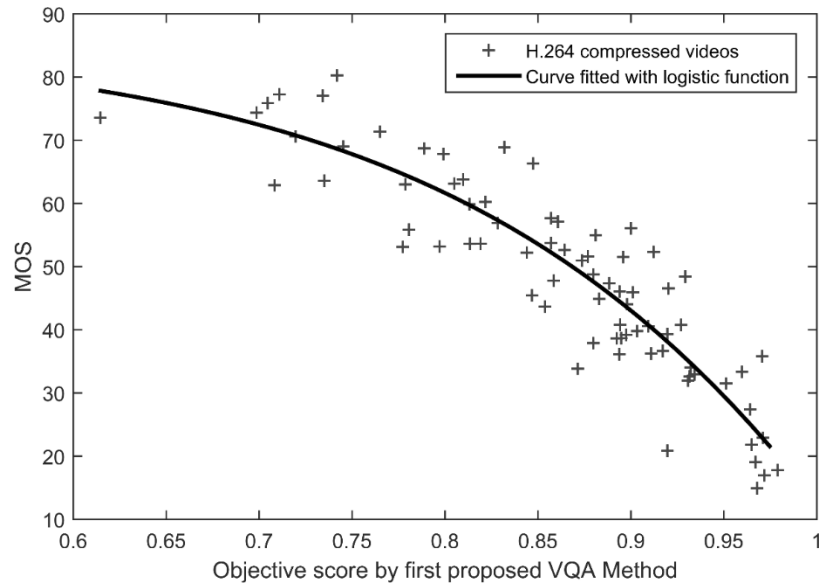


Figure 4.6: Scatter plots of Index1 for H.264 compressed videos.

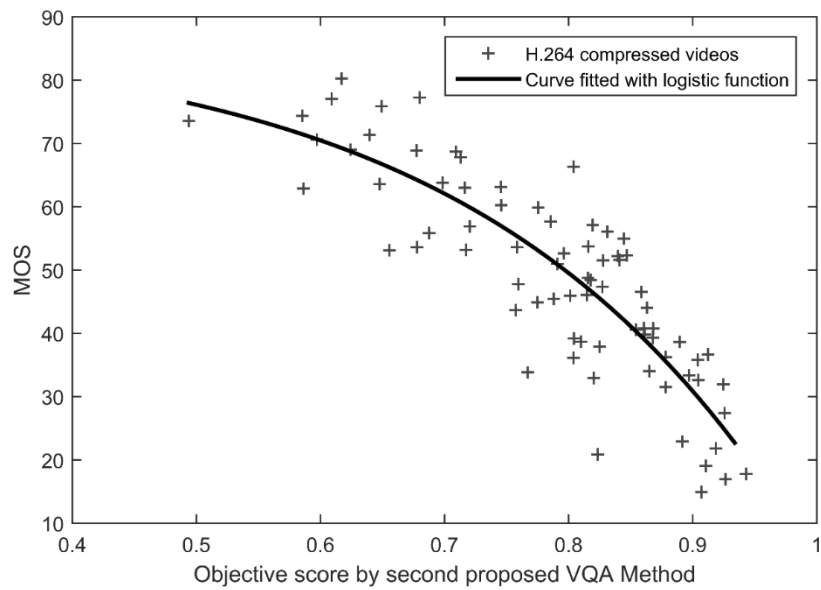


Figure 4.7: Scatter plots of Index2 for H.264 compressed videos.

Next, the performance for compressed videos is analysed for all VQA methods. In LIVE video database, there are two types of compressed videos, namely H.264 and MPEG-2 compressed videos. Meanwhile, there are four types of compressed videos in CSIQ database. They are videos distorted by H.264, MJPEG, wavelet, and HEVC compression errors. The performances of VQA methods for these compressed videos are

determined. The results are shown in Table 4.16 and Figure 4.8. Index1 has the worst performance in terms of SROCC and CC. It only achieves 0.7934 and 0.7541 respectively. Index2 has the best performance. It attains 0.8661 and 0.8804 for SROCC and CC respectively. MOVIE, in turns, has the second best results. It achieves 0.8595 for SROCC and 0.8618 for CC. ViS₃ has the correlation results that is slightly lower than MOVIE. It attains 0.8368 for SROCC. For CC, it achieves 0.8463. In overall, Index2 has the best performance in assessing the quality of videos distorted by compression errors. This makes sense as it is based on MD-SSIM, which is specially designed for compression errors.

Table 4.16: SROCC and CC for compressed videos.

Metric	SROCC	CC
MOVIE	0.8595	0.8618
ViS ₃	0.8368	0.8463
Index1	0.7934	0.7541
Index2	0.8661	0.8804

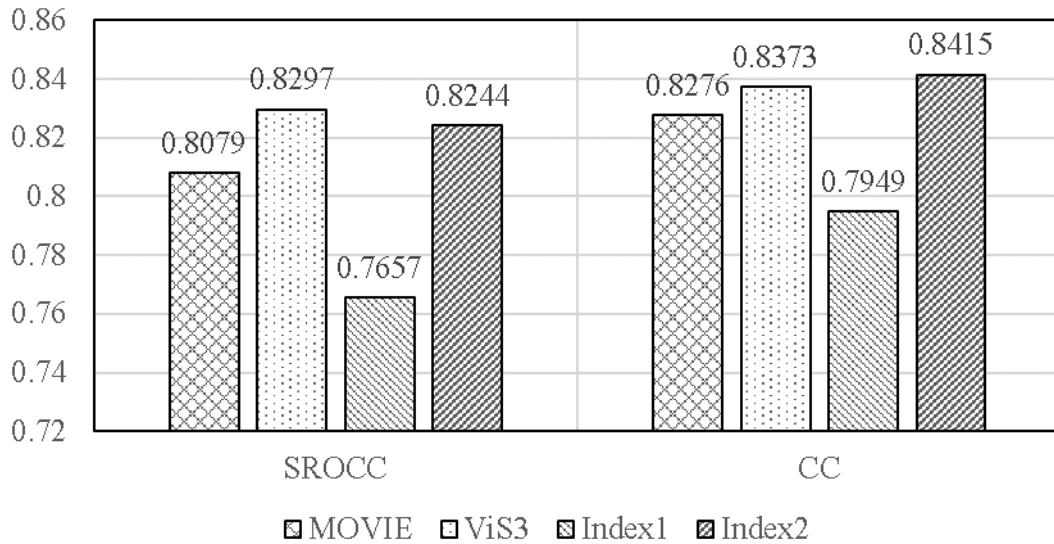


Figure 4.8: SROCC and CC for compressed video.

The results for compressed videos are also shown in scatter plots and RMSE. The results are shown in Figure 4.9, Figure 4.10, and Table 4.17. From Figure 4.9, the points are scattered around the whole area of the figure. Thus, the distances of the points to the fitted curve are also larger. Index1 has RMSE of 11.9706, which is considered high. Index2 only has RMSE of 8.5599. Thus, the points in Figure 4.10 are closer to the curve. There are some points that are quite far at the bottom right of the figure. Yet, Index2 still performs better than the Index1 for compressed videos.

Table 4.17: RMSE of Index1 and Index2 for compressed videos.

VQA Methods	RMSE
Index1	11.9706
Index2	8.5599

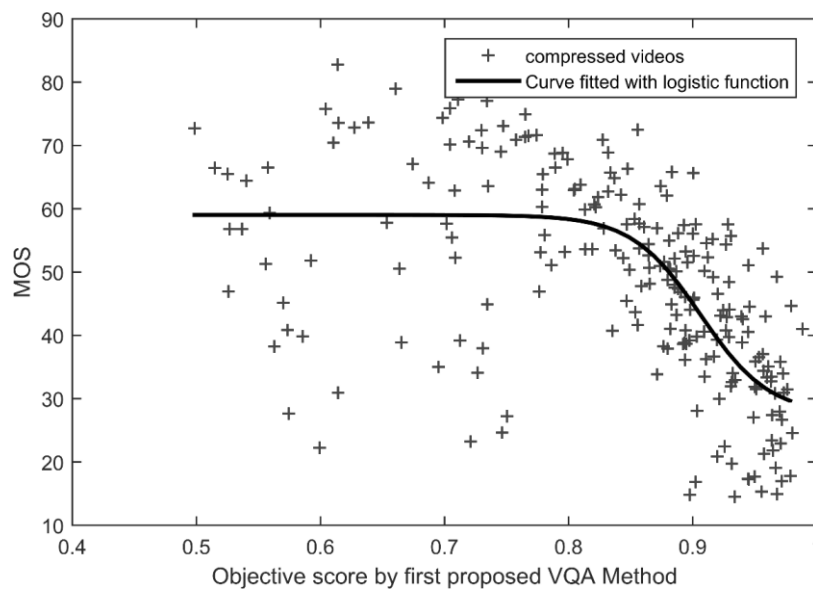


Figure 4.9: Scatter plots of Index1 for compressed videos.

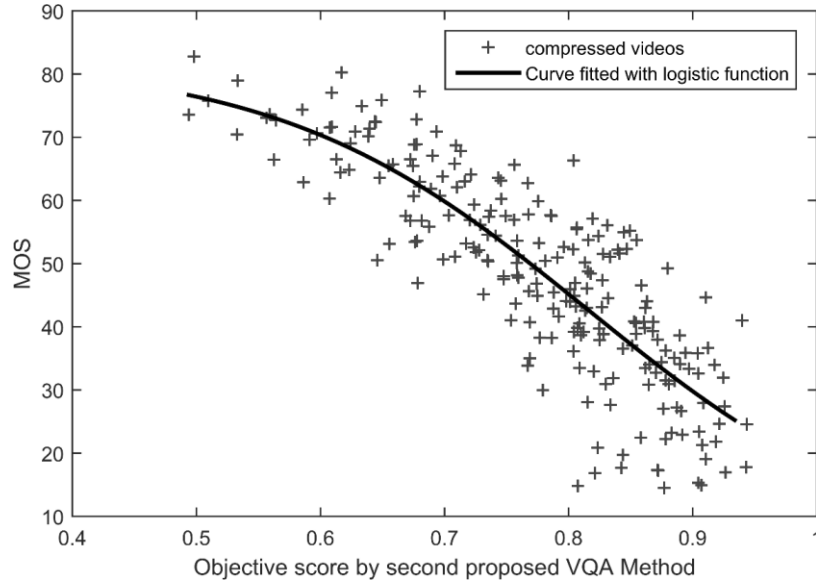
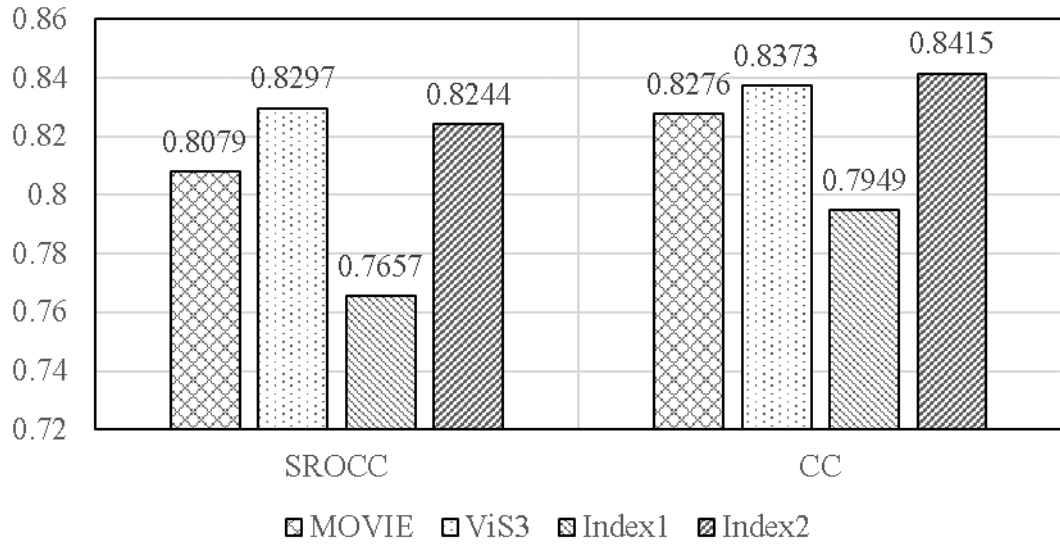


Figure 4.10: Scatter plots of Index2 for compressed videos.

The last type of videos is the transmission medium distorted videos. In LIVE video database, there are wireless and IP distorted videos. These two types of videos have transmission-related distortions. There are only plr distorted videos in CSIQ database that suffer from transmission medium distortion. As stated in Section 2.1, plr is the most common type of distortion during transmission. The results of the proposed VQA methods are shown in Table 4.18 and Figure 4.11. In terms of SROCC, ViS₃ has the best performance, which is 0.8297. This is followed by Index2 at 0.8244. Index1 has the worst performance that is 0.7657. MOVIE, on the other hand, achieves results better than Index1 but poorer than Index2 at 0.8079. The performance for Index1 and MOVIE for CC is similar to the SROCC results. They attain 0.7949 and 0.8276 respectively. Index2 has the highest correlation results in terms of CC that attains 0.8415 while ViS₃ achieves 0.8373. This is a reversed condition as compared to the results of SROCC. In overall, Index1 has poor performance than Index2 in transmission medium distorted videos.

Table 4.18: SROCC and CC for medium distorted videos.

Metric	SROCC	CC
MOVIE	0.8079	0.8276
ViS ₃	0.8297	0.8373
Index1	0.7657	0.7949
Index2	0.8244	0.8415

**Figure 4.11:** SROCC and CC for medium distorted videos.

The scatter plots of the two proposed VQA methods are shown in Figure 4.12 and Figure 4.13. The RMSE results are shown in Table 4.19. From the figures, the points of the two proposed VQA methods have similar distances to their respective fitted curve. This is also shown in the RMSE results. Index1 has RMSE of 6.4017 while Index2 achieves 6.7655. The results in scatter plots and RMSE are different to the correlation results. This may due to the limited types of distortion in the two databases.

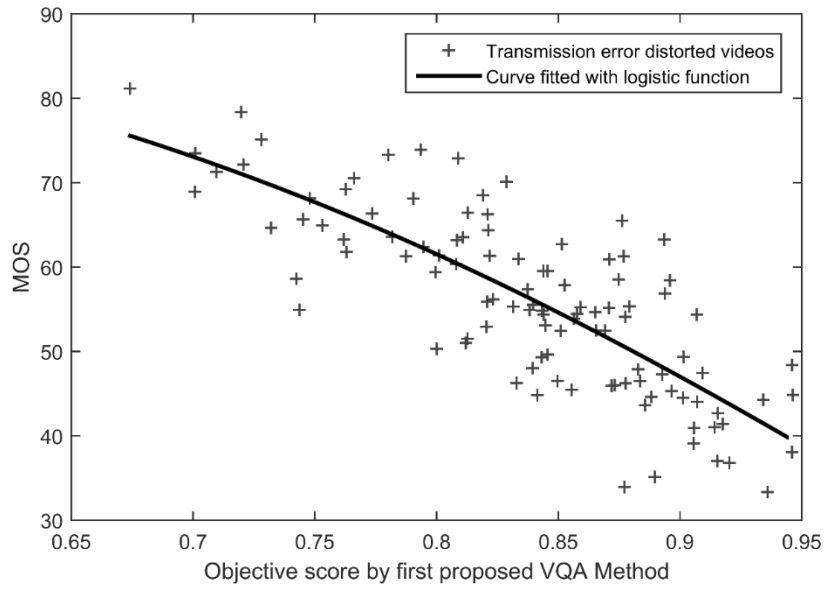


Figure 4.12: Scatter plots of Index1 for transmission medium distorted videos.

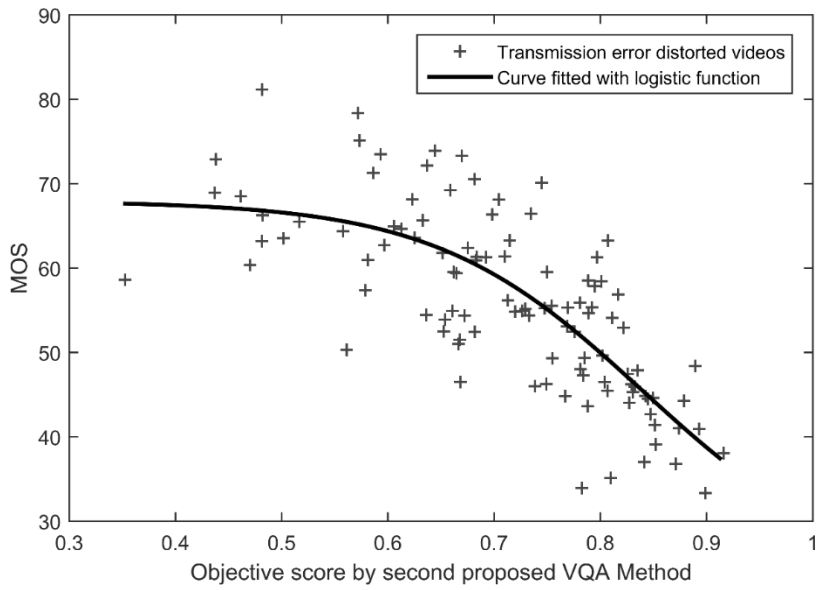


Figure 4.13: Scatter plots of Index2 for transmission medium distorted videos.

Table 4.19: RMSE of Index1 and Index2 for medium distorted videos.

VQA Methods	RMSE
Index1	6.4107
Index2	6.7655

4.2.5 Computational complexity

Other than the performance of the VQA methods, the computational complexities of every VQA methods are also compared. The computational complexity is measured in terms of computational time. Table 4.20 shows the results of the computational time. The test is done for a video (bs2_25fps.yuv). Each method is being computed for ten times and the average time is considered the final computational time. When compared to existing methods, Index2 has the lowest computational time. It only requires 17.79s to assess the video quality. This is followed by Index1. It requires 18.62s to rate the video quality. Both of the proposed methods have almost the same computational complexity. MOVIE has the longest computational time, which consumes 2,923,261.93s for assessing the video quality. ViS₃ has moderate computational time that is 314.32s. The results show that both of the proposed VQA methods have much lower computational time compare to the state-of-the-art VQA methods.

Table 4.20: Computation complexity for all VQA Methods.

Metric	Computational Time (s)
Mod-MSE	7.42
MD-SSIM	9.53
Index2	17.79
Index1	18.62
ViS ₃	314.32
MOVIE	2,923,261.93

When compared to predecessors, Index1 and Index2 need more computational time. Yet this shows the advantages of Index1 and Index2 as they perform better than their predecessors with only a slight increase in computational time. The only drawback of Index1 is its ability in assessing the quality of videos distorted by MJPEG compression. In overall, Index1 and Index2 have low computational time while having acceptable performances in both databases.

4.3 Summary

In this chapter, the experiment results, analysis and discussion are shown. The experiment results of Index1 and Index2 of the databases, in terms of SROCC and CC, are compared to MOVIE and ViS₃. They are also compared to their predecessors to show the improvements. Other than comparing the performances according to databases, Index1 and Index2 are also compared among different types of videos. Their respective computational complexities are also compared. All of the experiment results and computational complexity are shown in Section 4.2.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Overview

In this thesis, two new FR VQA method, Index1 and Index2, were proposed. In Chapter 2, the literature review related to the research topic is presented. Some common distortions related to video are also discussed. Two types of VQA method, i.e. objective and subjective VQA methods, are discussed briefly. Some state-of-the-art VQA methods are also elaborated. The algorithms for the two proposed VQA methods are discussed in details in Chapter 3. The experiment results, analysis and discussion are presented in Chapter 4. The proposed VQA methods perform well in comparison with state-of-the-art methods. In addition, computational complexities in terms of computational time of the methods are compared with various methods.

5.2 Summary of the proposed methods

The summary of the two proposed methods is presented in this section. The first method, Index1, is the JND-based VQA method while the second method, Index2, is based on MD-SSIM. An overview of Index1 is presented in Section 5.2.1 while the summary of Index2 is discussed in Section 5.2.2.

5.2.1 Index1

Index1 is based on Mod-MSE. Different from MSE, SSIM is involved for both spatial and temporal part. JND is also applied to spatial and temporal part of Index1 through threshold. The distortions above the threshold are assumed to be not perceivable by HVS and thus eliminated. For the temporal part, three frames are involved at an instant. This is different to most of the proposed VQA methods by other researchers that only involve two frames at a time. The overall video quality index is the average of spatial and temporal indices. In short, Index1 is summarized next paragraph.

First of all, video frames from reference and distorted videos assessed by using SSIM. JND is then applied to the SSIM indices. Spatial index is calculated from the averaged SSIM indices. For the temporal part, D_1 and D_2 are generated from frames of reference and distorted videos respectively. Next, they are evaluated by SSIM index to form SSIM map. JND is then applied to the as in spatial part. Temporal index is the average of JND-based SSIM maps. Overall video quality index is the average of spatial and temporal indices.

5.2.2 Index2

Index2 is a modified and improved version of MD-SSIM method. In both method, SSIM map is being weighted. Yet the weight is difference. For MD-SSIM, MSE map is used as the weight. Direct frame differences with surrounding pixels consideration are used as the weights for Index2. From the results in Section 4.2.3, the weight of Index2 performs better than MD-SSIM and direct average of SSIM map. For the temporal part, variation of temporal information is computed. This is very different from the existing methods. To the best of my knowledge, no other works involve variation of temporal information. This

variation can be computed by direct difference of the temporal information for three particular frames. Lastly, the overall video quality index is the average of the spatial and temporal indices. The process is summarized in the next paragraph.

First, SSIM map and error map are generated from reference and distorted frames. Error map is modified by considering surrounding errors. Next, modified error map is utilized to weight the SSIM map. Spatial index is calculated from the weighting process. For the temporal part, two difference maps are obtained by computing the variation in TI of the reference and distorted frames. Weight computation from spatial part are repeated using the two difference maps to calculate the temporal index. Overall video quality index is the arithmetic mean of spatial and temporal indices.

5.3 Contribution of the research

Two new VQA methods, Index1 and Index2, were proposed in detail in Chapter 3 of this thesis. This corresponds to the first objective of the research. In Chapter 4, the statistical performances of the VQA methods were compared and analysed with existing methods. This fulfills the second objective of the research. The computational complexities of the two methods were also presented, compared, and discussed in Chapter 4. The computational complexity of each VQA method was measured in terms of computational time to meet the third objective. This proved the feasibility of Index1 and Index2.

Index1 performed comparatively well for LIVE database. For CSIQ database, Index1 also performed quite well except for MJPEG compressed videos. It has the lowest performance among all the VQA methods. Index1 also has poor performance in videos

with plr distortion. Yet, the performance is still acceptable when compared to MJPEG compressed videos.

For Index2, it was proposed to combat the drawback of Index1. It has good performance for all databases. For CSIQ database, it has the best performance when compared to state-of-the-art VQA methods. For LIVE database, it has the best performance in assessing IP, MPEG-2, white noise, and HEVC distorted videos. It has slightly poorer performance for wireless and H.264 distorted videos. However, the performance is only slightly lower than the best performing VQA methods for each type of distortions.

By comparing both of the proposed VQA methods, Index2 outperformed Index1. This is evident from the correlation results and scatter plots as elaborated in Chapter 4. Index2 succeeded in overcoming the drawback of the Index1. It can evaluate the quality of MJPEG and plr distorted videos well. The performance of Index2 is even better than Index1 for LIVE database.

For the case of different videos, Index2 can assess compressed video the best. However, Index1 performs better than Index2 for H.264 compressed videos. For transmission medium distorted videos, Index2 has similar performance to Index1. Both of them have moderate performance in assessing the quality of videos distorted by transmission medium distortions.

In terms of computational complexity, both methods have lower computational complexity than the state-of-the-art VQA methods. Index2 even has computational time which is slightly lower than Index1. This shows that the proposed methods are more applicable in real-life scenario than existing state-of-the-art VQA methods.

Due to the good performance and low computational complexity, Index1 and Index2 are both suitable to be applied in the application that involves hardware with limited processing capability. Some examples of these devices are smartphones, tablets, smart watches, and smart glasses. Assessment of the video qualities of these products is crucial as a quality feedback tool for device manufacturers and content service providers. The proposed methods also can be modified according to specific task requirements such as the applications in telementoring in open surgery (Hashimoto *et al.*, 2016), underwater scientific applications (Moreno-Roldán *et al.*, 2017), and inter-vehicular streaming (Roy *et al.*, 2017). Better performance and low computational complexity can increase the efficiency and productivity.

5.4 Future Works

New VQA methods can be modelled to increase the performance while maintaining a low computational complexity. HVS concept can be applied to the proposed method to further enhance the performance. For example, the concept of eye foveation can be applied to mimic the human perception. The more advanced concept of JND can also be applied to further improve Index1. Since the proposed methods are spatial-based, frequency-based methods or concepts can be explored with the proposed method to include more aspects that can affect video quality.

REFERENCES

- Aggarwal, N. (2014). A review on video quality assessment. In *2014 Recent Advances in Engineering and Computational Sciences (RAECS)* (pp. 1-6). Chandigarh: IEEE.
- Akramullah, S. (2014). *Digital video concepts, methods, and metrics*. New York: Springer Science.
- Aramvith, S, & Sun, M. T. (2009). Chapter 9 - MPEG-1 and MPEG-2 video standards. In Bovik, A. C. (Ed.), *The essential guide to video processing* (2nd ed.) (pp. 267-293). Boston: Academic Press.
- Biemond, J., Lagendijk, R., & Mersereau, R. (1990). Iterative methods for image deblurring. *Proceedings of the IEEE*, 78(5), 856-883.
- Bong, D. B. L., & Khoo, B. E. (2014a). An efficient and training-free blind image blur assessment in the spatial domain. *IEICE Transactions on Information and Systems*, 97(7), 1864-1871.
- Bong, D. B. L., & Khoo, B. E. (2014b). Blind image blur assessment by using valid reblur range and histogram shape difference. *Signal Processing: Image Communication*, 29(6), 699-710.
- Bong, D. B. L., & Khoo, B. E. (2015). Objective blur assessment based on contraction errors of local contrast maps. *Multimedia Tools and Applications*, 74(17), 7355-7378.
- Borer, S. (2010). A model of jerkiness for temporal impairments in video transmission. In *2010 Second IEEE International Workshop on Quality of Multimedia Experience (QoMEX)* (pp. 218-223). Trondheim: IEEE.
- Bowker, D. (2013, June 13). *What is contrast in photography?* Retrieved March 07, 2018, from <http://www.photocritic.org/articles/what-is-contrast>

- Bridgewater, K. E., Lockridge, T. W., & Horlander, T. E. (2009). *U.S. Patent No. 7,512,158*. Washington, DC: U.S. Patent and Trademark Office.
- Britanak, V., Yip, P. C., & Rao, K. R. (2010). *Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations*. UK: Academic Press.
- Cardoso, J. V. M., Alencar, M. S., Regis, C. D. M., & Oliveira, I. P. (2014). Temporal analysis and perceptual weighting for objective video quality measurement. In *2014 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)* (pp. 57-60). San Diego, CA, USA: IEEE.
- Channappayya, S. S. (2007). *Image communication system design based on the structural similarity index*. USA: ProQuest.
- Chen, M. J., & Bovik, A. C. (2011). Fast structural similarity index algorithm. *Journal of Real-Time Image Processing*, 6(4), 281-287.
- Chen, Z. & Liu, H. (2014). JND modeling: approaches and applications. In *2014 19th International Conference on Digital Signal Processing (DSP)* (pp. 827-830). Hong Kong: IEEE.
- Cisco. (2015). *Cisco visual networking index: Global mobile data traffic forecast update, 2015–2020 White Paper*. Retrieved May 12, 2016, from <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html>
- Damera-Venkata, N., Kite, T. D., Geisler, W. S., Evans, B. L., & Bovik, A. C. (2000). Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4), 636-650.
- Erol, B., Kossentini, F., Joch, A., Sullivan, G. J., & Winger, L. (2009). Chapter 10 - MPEG-4 visual and H.264/AVC: Standards for modern digital video. In Bovik, A. C (Ed.), *The essential guide to video processing* (2nd ed.) (pp. 295-330). Boston: Academic Press.

- Fang, B., Shen, G., Li, S., & Chen, H. (2005). Techniques for efficient DCT/IDCT implementation on generic GPU. In *2005 IEEE International Symposium on Circuits and Systems (ISCAS)* (pp. 1126-1129). Kobe, Japan: IEEE.
- Fleet, D. J., & Jepson, A. D. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1), 77-104.
- Fraser, K., Wang, Z., & Liu, X. (2010). *Microarray image analysis: An algorithmic approach*. Boca Raton, FL: Chapman & Hall/CRC.
- Galton, F. (1879). The geometric mean, in vital and social statistics. *Proceedings of the Royal Society of London*, 29(196-199), 365-367.
- Geisler, W. S. & Perry, J. S. (1998). Real-time foveated multiresolution system for low-bandwidth video communication. In *Photonics West'98 Electronic Imaging* (pp. 294-305). San Jose, CA: International Society for Optics and Photonics.
- Gomez, J. M. (2002). *Satellite broadcast systems engineering*. Norwood, Massachusetts: Artech House.
- Harvey, L. O., Rentschler, I., & Weiss, C. (1985). Sensitivity to phase distortions in central and peripheral vision. *Perception & Psychophysics*, 38(5), 392-396.
- Hashimoto, D. A., Phitayakorn, R., Fernandez-del Castillo, C., & Meireles, O. (2016). A blinded assessment of video quality in wearable technology for telementoring in open surgery: the Google Glass experience. *Surgical Endoscopy*, 30(1), 372-378.
- Hilbert, M. & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025), 60-65.
- Hosam, O. (2012). Motion compensation for video codec based on disparity estimation. In *2012 IEEE Symposium on Computers and Communications (ISCC)* (pp. 000559-000565). Cappadocia: IEEE.

- Hou, H. (1987). A fast recursive algorithm for computing the discrete cosine transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10), 1455-1461.
- Internet Users. (n.d.). Retrieved May 12, 2016, from <http://www.internetlivestats.com/internet-users/>
- ITU-T. (1999). Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation*, P.910. Switzerland: ITU.
- ITU-R. (2007). Methodology for the subjective assessment of video quality in multimedia applications. *ITU-R BT.1788*. Switzerland: ITU.
- ITU-R. (2003). Methodology for the subjective assessment of the quality of television pictures: *ITU-R BT.500-10*. Switzerland: ITU.
- Joseph, V., & Chapman, B. (2009). *Deploying QoS for Cisco IP and next generation networks: The definitive guide*. USA: Morgan Kaufmann.
- Karam, L. J., Ebrahimi, T., Hemami, S. S., Pappas, T. N., Safranek, R. J., Wang, Z., & Watson, A. B. (2009). Introduction to the issue on visual media quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 3(2), 189-192.
- Kaspersky Lab. (2015). *Consumer security risks survey, from scared to aware: Digital lives in 2015*. Russian Federation: Kaspersky Lab.
- Lafferty, P. A. (2008). *Obfuscation and the steganographic active warden model*. USA: ProQuest.
- Larson, E. C. & Chandler, D. M. (2010). Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1), 011006-011006.
- Le Callet, P., Viard-Gaudin, C., & Barba, D. (2006). A convolutional neural network approach for objective video quality assessment. *IEEE Transactions on Neural Networks*, 17(5), 1316-1327.

- Lee, S., Pattichis, M. S., & Bovik, A. C. (2002). Foveated video quality assessment. *IEEE Transactions on Multimedia*, 4(1), 129-132.
- Liu, K. C. (2009). Just noticeable distortion model and its application in color image watermarking. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 92(2), 563-576.
- Ma, L., Zhang, F., Li, S., & Ngan, K. N. (2010). Video quality assessment based on adaptive block-size transform just-noticeable difference model. In *2010 17th IEEE International Conference on Image Processing (ICIP)* (pp. 2501-2504). Hong Kong: IEEE.
- Monma, A., & Oishi, K. (2007). *U.S. Patent No. 7,276,950*. Washington, DC: U.S. Patent and Trademark Office.
- Moreno-Roldán, J. M., Luque-Nieto, M. Á., Poncela, J., & Otero, P. (2017). Objective Video Quality Assessment Based on Machine Learning for Underwater Scientific Applications. *Sensors*, 17(4), 664.
- Nayar, S. K., & Ben-Ezra, M. (2004). Motion-based motion deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6), 689-698.
- Narwaria, M., Lin, W., & Liu, A. (2012). Low-complexity video quality assessment using temporal quality variations. *IEEE Transactions on Multimedia*, 14(3), 525-535.
- Ng, M. K., Chan, R. H., & Tang, W. C. (1999). A fast algorithm for deblurring models with Neumann boundary conditions. *SIAM Journal on Scientific Computing*, 21(3), 851-866.
- Oliveira, Í, Cardoso, J. V. M., Regis, C. D., & Alencar, M. S. (2013). Spatial and temporal analysis considering relevant regions applied to video quality assessment. In *XXXI Brazilian Symposium on Telecommunications, SBrT 2013* (pp. 1-5). Fortaleza - CE, Brazil: Brazilian Telecommunications Society.

- Oosa, K. (1998). A new deblocking filter for digital image compression. *Nippon steel technical report*. (77-78), 105-108.
- Ooyala, Q4 2015 global video index. (2016). Retrieved May 12, 2016, from <http://go.ooyala.com/wf-video-index-q4-2015.html>.
- Oxford Dictionaries. (n.d.). *Definition of video in English*. Retrieved May 12, 2016, from <http://www.oxforddictionaries.com/definition/english/video>.
- Pal, A. (2011). *Data communications and computer networks*. New Delhi: PHI Learning Private Limited.
- Park, H. W., & Lee, Y. L. (1999). A postprocessing method for reducing quantization effects in low bit-rate moving picture coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(1), 161-171.
- Peng, P., Cannons, K., & Li, Z. N. (2013). Efficient video quality assessment based on spacetime texture representation. In *Proceedings of the 21st ACM international Conference on Multimedia* (pp. 641-644). Barcelona, Spain: ACM.
- Peterson, H. A., Ahumada Jr, A. J., & Watson, A. B. (1993). Improved detection model for DCT coefficient quantization. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology* (pp. 191-201). San Jose, CA: International Society for Optics and Photonics.
- Phan, T., Sohoni, S., Chandler, D. M., & Larson, E. C. (2012, April). Performance-analysis-based acceleration of image quality assessment. In *2012 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)* (pp. 81-84). Santa Fe, NM, USA: IEEE.
- Pinson, M. H., & Wolf, S. (2003). Comparing subjective video quality testing methodologies. In *Visual Communications and Image Processing 2003* (pp. 573-582). Lugano, Switzerland: International Society for Optics and Photonics.

- Pinson, M. H., & Wolf, S. (2004). A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3), 312-322.
- Rimac-Drlje, S., Vranješ, M., & Žagar, D. (2010). Foveated mean squared error—a novel video quality metric. *Multimedia Tools and Applications*, 49(3), 425-445.
- Rodríguez, D. Z., Rosa, R. L., Alfaia, E. C., Abrahão, J. I., & Bressan, G. (2016). Video quality metric for streaming service using DASH standard. *IEEE Transactions on Broadcasting*, 62(3), 628-639.
- Samsung Electronics America. (n.d.). *The new Galaxy S7*. Retrieved August 18, 2016, from <http://www.samsung.com/us/explore/galaxy-s7-features-and-specs/>
- Seshadrinathan, K., & Bovik, A. C. (2009). Chapter 14 - video quality assessment. In Bovik, A. C. (Ed.), *The essential guide to video processing* (2nd ed.) (pp. 417-436). Boston: Academic Press.
- Seshadrinathan, K., & Bovik, A. C. (2010). Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2), 335-350.
- Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. (2010a). A subjective study to evaluate video quality assessment algorithms. In *Proc. SPIE 7527, Human Vision and Electronic Imaging XV* (pp. 75270H-75270H). San Jose, California: International Society for Optics and Photonics.
- Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. (2010b). Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6), 1427-1441.
- Shang, X., Wang, Y., Luo, L., & Zhang, Z. (2013). Perceptual multiview video coding based on foveated just noticeable distortion profile in DCT domain. In *2013 20th IEEE International Conference on Image Processing (ICIP)* (pp. 1914-1917). Melbourne: IEEE.

- Sheikh, H. R., & Bovik, A. C. (2005). A visual information fidelity approach to video quality assessment. In *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics* (pp. 23-25). Scottsdale, Arizona: Springer.
- Sheikh, H. R., Bovik, A. C., & De Veciana, Gustavo. (2005a). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12), 2117-2128.
- Sheikh H. R., Wang Z., Cormack L., & Bovik A. C. (2005b). *LIVE image quality assessment database release 2*. Retrieved August 15, 2014, from <http://live.ece.utexas.edu/research/quality>.
- Simpson, W., & Greenfield, H. (2012). *IPTV and internet video: Expanding the reach of television broadcasting*: USA: CRC Press.
- Singh, R., & Aggarwal, N. (2014). State of the art and research issues in video quality assessment. In *2014 Recent Advances in Engineering and Computational Sciences (RAECS)* (pp. 1-6). Chandigarh: IEEE.
- Stern, M. K., & Johnson, J. H. (2010). Just noticeable difference. *Corsini Encyclopedia of Psychology*. Hoboken, NJ, USA: John Wiley & Sons.
- Tokdemir, S., & Belkasim, S. (2011). Parallel processing of DCT on GPU. In *2011 Data Compression Conference (DCC)* (pp. 479-479). Snowbird, Utah, USA: IEEE.
- Triantafyllidis, G. A., Tzovaras, D., & Strintzis, M. G. (2002). Blocking artifact detection and reduction in compressed data. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(10), 877-890.
- Unser, M. (1994). Fast Gabor-like windowed Fourier and continuous wavelet transforms. *IEEE Signal Processing Letters*, 1(5), 76-79.

- Unterweger, A. (2012). Compression artifacts in modern video coding and state-of-the-art means of compensation. In Farrugia, R. A, & Debono C. J. (Eds.) *Multimedia Networking and Coding* (pp. 28-49). USA: IGI Global.
- USD Internet Sensation & Perception Laboratory. (n.d.). *Weber's Law of just noticeable difference*. Retrieved August 22, 2016, from <http://apps.usd.edu/coglab/WebersLaw.html>
- Verma, K., Kumar, A., Gupta, S ., & Venkatesh, K. S. (2014). Fast and robust real time digital video stabilization with smear removal using integral projection curve warping technique. In *2014 Twentieth National Conference on Communications (NCC)* (pp. 1-6). Kanpur: IEEE.
- Video Quality Experts Group (2003). *Final report from the video quality experts group on the validation of objective models of video quality assessment. Phase II (FR_TV2)*. Retrieved September 13, 2014, from ftp://ftp.its.bldrdoc.gov/dist/ituvidq/Boulder_VQEG_jan_04/VQEG_PhaseII_FRTV_Final_Report_SG9060E.doc, 2003.
- Vu, C., & Deshpande, S. (2012). ViMSSIM: from image to video quality assessment. In *Proceedings of the 4th Workshop on Mobile Video* (pp. 1-6). New York: ACM.
- Vu, P. V., & Chandler, D. M. (2014). ViS₃: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1), 013016(1)-013016(24).
- Wang, H. S., & Mersereau, R. M. (1999). Fast algorithms for the estimation of motion vectors. *IEEE Transactions on Image Processing*, 8(3), 435-438.
- Wang, Y. (2006). *Survey of objective video quality measurements*. Retrieved April 25, 2016, from <http://digitalcommons.wpi.edu/cgi/viewcontent.cgi?article=1043&context=computerscienc>e-pubs.

- Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1), 98-117.
- Wang, Z., Lu, L., & Bovik, A. C. (2004). Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2), 121-132.
- Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *2004 Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers* (pp. 1398-1402). Pacific Grove, CA, United States: IEEE.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.
- Wang, Z., Simon, S., Baroud, Y., & Najmabadi, S. M. (2015). Visually lossless image compression extension for JPEG based on just-noticeable distortion evaluation. In *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 237-240). London: IEEE.
- Wang, Z., Wang, J., Wang, F., Li, C., Fei, Z., & Rahim, T. (2017). A Video Quality Assessment Method for VoIP Applications Based on User Experience. *Sensing and Imaging*, 18(1), 12.
- Watson, A. B., Hu, J., & McGowan, J. F. (2001). Digital video quality metric based on human vision. *Journal of Electronic Imaging*, 10(1), 20-29.
- Wiegand, T., Sullivan, G. J., Bjøntegaard, G., & Luthra, A. (2003). Overview of the H. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7), 560-576.
- Winkler, S. (2005). *Digital video quality: Vision models and metrics*. England: John Wiley & Sons.
- Winkler, S., & Mohandas, P. (2008). The evolution of video quality measurement: from PSNR to hybrid metrics. *IEEE Transactions on Broadcasting*, 54(3), 660-668.

- Wolf, S. (2001). Objective video quality measurement using a peak-signal-to-noise-ratio (PSNR) full reference technique, *Technical Report T1.TR.74*. Washington: Alliance for Telecommunications Industry Solutions.
- Wolf, S. & Pinson, M. H. (1998). In-service performance metrics for MPEG-2 video systems. In *Proc. Made to Measure 98-Measurement Techniques of the Digital Age Technical Seminar* (pp. 12-13). Montreux, Switzerland: International Academy of Broadcasting (IAB), ITU and Technical University of Braunschweig.
- Wong, T. T., Leung, C. S., Heng, P. A., & Wang, J. (2007). Discrete wavelet transform on consumer-level graphics hardware. *IEEE Transactions on Multimedia*, 9(3), 668-673.
- Wu, H. R., & Rao, K. R. (2005). *Digital video image quality and perceptual coding*. Boca Raton: USA: CRC/Taylor & Francis.
- Xue, Y., Erkin, B., & Wang, Y. (2015). A novel no-reference video quality metric for evaluating temporal jerkiness due to frame freezing. *IEEE Transactions on Multimedia*, 17(1), 134-139.
- Yuan, W., Hao, P., & Xu, C. (2006). Matrix factorization for fast DCT algorithms. In *2006 Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 948-951). Toulouse, France: IEEE.
- Zakhor, A. (1992). Iterative procedures for reduction of blocking effects in transform image coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2(1), 91-95.
- Zhao, H., Yang H., and Zheng, S. (2009). An efficient method for detecting ghosts and left objects in intelligent video surveillance. In *2nd International Congress on Image and Signal Processing, 2009 (CISP'09)* (pp. 1-6). Tianjin, China: IEEE.
- Zhang, D., Zhou, D., & Jin, X. (2017). A content-adaptive video quality assessment method for online media service. *Multimedia Tools and Applications*, 76(5), 7175-7195.

- Zhang, K., Wang, S., & Zhang, X. (2002). New metric for quality assessment of digital images based on weighted mean square error. In *Second International Conference on Image and Graphics* (pp. 491-497). Hefei, China: International Society for Optics and Photonics.
- Zhang, X., Lin, W., & Xue, P. (2008). Just-noticeable difference estimation with pixels in images. *Journal of Visual Communication and Image Representation*, 19(1), 30-41.
- Zheng, M., Su, K., Wang, W., Lan, C., & Yang, X. (2013). Enhanced subband JND model with textural image. In *2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)* (pp. 1-4). Kunming, China: IEEE.

APPENDIX

List of Publication

- Loh, W. T., & Bong, D. B. L. New Spatiotemporal Method for Assessing Video Quality. *International Journal of Computational Vision and Robotics*. (Accepted on 2017).
- Loh, W. T., & Bong, D. B. L. (2016). Temporal video quality assessment method involving structural similarity index. In *2016 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)* (pp. 1-2). Nantou County, Taiwan: IEEE.
- Loh, W. T., Voo, K. H. B., & Bong, D. B. L. (2015). Video quality assessment based on a modified mean squared error. In *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (pp. 500-504). IEEE. Kuala Lumpur, Malaysia: IEEE.
- Loh, W. T., & Bong, D. B. L. (2015). Video quality assessment method: MD-SSIM. In *2015 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)* (pp.290-291). Taipei City, Taiwan: IEEE.
- Bong, D. B. L., Khoo, B. E., Voo, K. H. B., & Loh, W. T. (2015a). Perceptual quality metrics for image blurriness. *UNMAS R&D Expo 2015 – Strengthening Research and Innovation for Communities Sustainability*. Kota Samarahan: UNIMAS. (Silver Award) (Poster)
- Bong, D. B. L., Khoo, B. E., Voo, K. H. B., & Loh, W. T. (2015b). Perceptual quality metrics for image blurriness. *14th International Conference and Exposition on Inventions by Institutions of Higher Learning (PECIPTA 2015)*. Kuala Lumpur: Ministry of Higher Education Malaysia. (Bronze Award) (Poster)
- Bong, D. B. L., Khoo, B. E., Voo, K. H. B., & Loh, W. T. (Filing Date: 9 September 2015). A Method for Measuring Blurriness of Digital Image. *Application No. PI 2015704041*. Kuala Lumpur, Malaysia. Malaysia Patents Registration Office. (Filed)