USC-SIPI REPORT #424

Experimental Design and Evaluation Methodology for Human-Centric Visual Quality Assessment

by

Yu-Chieh Lin

December 2015

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

In dedication to everyone has supported me

and especially to my wife, Amy

Acknowledgments

I am deeply appreciative of Prof. Kuo who has supported my work and continually encouraged me. Without his time, attention, encouragement, thoughtful feedback, and patience, I would not have been able to see it through. I am grateful to Dr. Wesley Szu-Wei Lee and Dr. Ioannis Katsavounidis for their valuable suggestions in various stages of this research.

Contents

List of Tables

List of Figures

Abstract

The problem of human-centric visual quality assessment (VQA) is extensively studied in this thesis. Our study includes three major topics: 1) design of a dataset for streaming video quality assessment, 2) development of a new and effective video quality assessment index, 3) exploration of a new methodology for human visual quality assessment based on the notion of just-noticeable-differences (JND).

For the first topic, we present a high-definition VQA dataset that captures two typical video distortion types in streaming video services in Chapter 3. The VQA dataset, called MCL-V, contains 12 source video clips and 96 distorted video clips with subjective assessment scores. The source video clips are selected from a large pool of public-domain video sequences with representative and diversified contents. Both distortion types are perceptually adjusted to distinguishable distortion levels. An improved pairwise comparison method is adopted for subjective evaluation to save evaluation time. Several VQA algorithms are evaluated against the MCL-V dataset.

For the second topic, we propose two objective assessment indices to predict subjective video quality in Chapter 4. They are a fusion-based video quality assessment (FVQA) index and an ensemble-learning video quality assessment (EVQA) index. The FVQA index first classifies video sequences according to their content complexity so as to reduce content diversity within each group. Then, it fuses several VQA methods to provide the final video quality score, where fusion coefficients are learned from

training samples in the same group. Being motivated by ensemble learning, we propose another video quality assessment index to extend FVQA furthermore, and call it the EVQA index. The basic idea is to fuse multiple VQA methods with diverse and complementary merits so that the fused outcome outperforms that of any single method. The superior performance of EVQA is demonstrated by comparing it with other video quality assessment indices with several benchmarking video quality datasets.

For the third topic, we propose a new human-centric methodology for visual quality assessment based on the JND notion in Chapter 5. JND is characterized by the detectable minimum amount of two visual stimuli, and has been used to enhance perceptual visual quality in the context of image/video compression. We first argue that the perceived quality of coded image/video is a stairwise function with several discrete jump points defined by JND. Then, we present a novel bisection method in performing the JND test on JPEG-coded images. Finally, we construct a JND dataset called MCL-JCI that contains 50 source images and analyze the relationship between the source content and the number of its distinguishable quality levels. The impact of JND-based quality assessment on image/video coding is also discussed.

Chapter 1

Introduction

1.1 Significance of the Research

Video streaming service grows and evolves in a incredible speed. Thousands of titles are monthly added to major service providers, such as Netflix, YouTube, and Amazon. Consumers enjoy such on-demand video services from service provides, and watching high-definition (HD) programs becomes the mainstream for video content consumption. According to the report in [110], more than half of US population watches on-line movies or dramas. Specifically, the viewers have increased from 37% in 2010 to 51% in 2013. The watched video programs vary in bit rates and resolutions due to the available bandwidth of their networks. The main reason to the blooming of streaming video is because abundant video genres. Thousands of movies and TV shows are provided by streaming video service such that consumers have tremendous choices of video contents. Therefore, *the diversity of video contents* is an significant issue in assessing video quality. When using the streaming service, the delivered titles are compressed and scaled in various bit rates and resolutions to match clients' bandwidth and end-terminal. Different sizes of video are transmitted at lower bit rates and up-scaled for display on HDTV (e.g., playing a 720p movie on the 1080p screen). The streaming bandwidths range from 500 Kbits/sec to 12 Mbits/sec and the resolutions vary from CIF to UHD. Thus, the distortion comes not only from compression but also resizing, where a video of a smaller size is scaled up to a larger resolution to match the dimension of the display device. Such artifacts usually appear after the production process. Therefore, service providers

are looking for an automatic to detect these distortions in the video clips of their huge libraries.

To control video quality well, there are two key factors, 1) diversity of content and 2) presence of distortions. These two factors make the nature of streaming video is so complicated that assessing the video quality becomes a significant issue. In the field of video compression, the mean squared error (MSE) is widely applied to assess the quality. However, [33] indicates several psychovisual phenomena affects the human visual system, and MSE does not reflect these phenomena. These phenomena affects are so called spatial masking effects, which are have been studied in [5, 6, 8, 31, 43, 44]. For video, temporal masking effect has a significant impact on human perception as indicated in [19, 75, 78, 94]. The masking effects are created by the characteristics of video contents. Thus, researchers in the VQA field strongly intend to model these phenomena and apply to VQA metrics. Li *et al.* [62] and Brandao and Queluz [16] both adopted Daly's contrast sensitivity function (CSF) [23] in their work. Their results show that modeling masking effects well can improve the accuracy of the VQA indices. However, only a few masking effects are modeled as well as CSF. Since it is difficult to find the closed-form functions to model all the related masking effects, the research results from the fields of vision and psychology are rarely to be combined to VQA related researches. Therefore, machine-learning based VQA indices[73, 81, 83] are introduced to find the relationship from data-oriented approach, rather than digging out the interrelationship between human brain and vision systems.

In order to take the data-oriented approach of assessing video quality, we need accurate and representative ground truth to develop our algorithm. However, the existing video quality datasets have several issues such that the development would be limited. First, some prior datasets [82, 103] contain scenes that are not representative in video applications. For example, there are video clips with a close view on the water surface

or the blue sky in the LIVE database [103]. These sequences were used for video coding performance test since they contain specific contents which are difficult to encode. However, they are not common scenes in movies or dramas. We prefer more representative scenes since they can better reveal human visual experience. Second, the dataset should have sufficient diversity in terms of several characteristics. Since the streaming service may have higher bit rates to maintain video quality [3, 45], the blocking effect is not as strong as that in existing video quality datasets. Furthermore, video quality is blurred due to video resizing. Spatial and temporal masking effects appear in various forms due to content diversity. For instance, visual artifacts are likely to be seen in still scenes than fast-motion scenes. These properties have not yet been covered by existing datasets such that we decide to build a new dataset, called MCL-V to address the shortcomings of existing datasets.

Our goal is to develop an automatic VQA method that is scalable to diversified video contents and highly correlated to human perceived quality. By surveying the existing datasets, we have not found any suitable one which can be used for our purpose. Thus, we decide to build two new datasets, MCL-V and MCL-JCI, and develop our method based on it.

1.2 Review of Previous Work

There are quite a few video quality assessment datasets available to the public [9, 12, 13, 17, 24, 25, 30, 35, 54, 56, 57, 72, 82, 85, 86, 87, 88, 89, 90, 91, 92, 93, 103, 109, 114, 116, 128, 129]. They were however limited in the following areas [26] and [122]. First, the source video set is not representative or diversified enough. We check the existing datasets by the video characteristics listed in 1.1.

Table 1.1: Video characteristics used in diversity check

Video Genres	**Video Semantics**	**Video Features**
• Cartoon • Sports • Indoor	• Face • People • Water • Number of objects • Salience	• Brightness • Contrast • Texture • Motion • Color Variance • Color Richness • Sharpness • Film Grain • Camera motion • Scene change

None of them is able to cover all the properties. The lack of these contents will not provide an extensive evaluation of viewers' experience. Second, the video resolution is low. The resolution of sequences in all VQA datasets except five [12, 35, 91, 116, 129] are lower than 1920×1080. Third, the distortion is not complete for the target application. For example, all above-mentioned VQA datasets except [56, 57, 91] do not cover video up-scaling, which is encountered frequently in our daily life. Although the work in [91] includes practical distortion types, it has only three video sources.

In the last decade, the state-of-the-art VQA metrics follow two main approaches: formula-based and learning-based. The traditional formula-based approach creates a close form expression of perceptual quality. The famous ones include SSIM[118], VIF[105], FSIM[130]. This type of metrics are good to handle specific distortions. The other approach, the learning-based approach, integrates several features and predict the perceptual quality based on the machine learning algorithms. [73] is an full-reference index that fuses existing quality metrics and predicts image quality with high accuracy. [81] combines a large number of computational statistical features and predict perceptual quality without reference. In this work, we focus on the learning-based approach,

because it is more scalable. In addition, to let the model learn from the opinion scores, we adopt supervised learning algorithms in this work.

One significant point of learning-based methods is adopting ensemble learning to fuse existing methods rather than low-level features. Ensemble learning is a classic divide-and-conquer strategy that has been widely adopted in solving classification and regression problems[7, 27, 28, 37, 49, 52, 79]. Generally, in the framework of ensemble learning, multiple methods with diverse and complementary skills are fused to tackle a task such that the joint outcome outperforms any single method. Liu *et al.*[73] proposed a multi-method fusion (MMF) method for image quality assessment. A regression approach is used to combine scores of multiple IQA methods in the MMF. The MMF score is obtained by a non-linear fusion of scores computed by multiple methods with suitable weights obtained by a training process. So far, MMF offers one of the best IQA results in several popular datasets such as LIVE[106], CSIQ[55], and TID2008[97].

Unlike successful IQA research, there are several challenges in designing accurate learning-based VQA methods. First, Limited number of data The total number of images in image quality datasets (e.g., [95, 96, 97, 106]) are larger than the number of sequences in video quality datasets (e.g., [25, 82, 103]). Hence, supervised learning operates well with abundant samples for training and develop accurate models. However, learning-based VQA method suffers from the limited samples, and the limited training set would cause over-fit and inaccurate model. Second, no ground truth to learn for temporal variation of video sequence. Temporal pooling has been studied by [100, 101, 104] over a decade. These researches show that using different methods does not generally provide significant improvement, and no solid conclusion is made in this topic.

All of above methods provide a quality measure of continuous-scale. However, it is well known that the HVS cannot perceive small changes in pixel differences. In reality,

humans cannot perceive continuous-scale but discrete-scale quality changes over a range of coding bitrates. This phenomenon is well-known as just-noticeable difference (JND) [50]. JND is a statistical quantity that accounts for the maximum difference unnoticeable to a human being. It has been extensively studied to understand human visual sensitivity [66]. In the context of image/video compression, Watson [121] proposed a way to use JND for video quality measurement. His work adopts "pair comparison" or "two-alternative forced-choice". That is, subjects are asked to determine which of two videos (*i.e.* the original source and the compressed one) is more distorted. Then, the distortion level of compressed video can be derived from the JND test result. Although Watson's pioneering work offers a statistical relationship between visual quality assessment and JND, his result is difficult to apply to a real-world video coding system. Furthermore, the test duration required for each subject is long. Recently, several JND estimators based on HVS properties were investigated in [51, 123]. It was also shown in [67, 77, 131] that JND-guided coding schemes can achieve perceptually similar quality with lower bitrates.

1.3 Contributions of the Research

To address the shortcomings of existing VQA datasets, we build a new VQA dataset called MCL-V in Chapter 3. The specific contributions are given below.

- The MCL-V dataset contains 12 source video clips and 96 distorted video clips with subjective assessment scores. The source video clips are selected from a large pool of public-domain high-definition (HD) video sequences with representative and diversified contents.

- The MCL-V dataset captures two typical video distortion types; namely, "compression" and "compression followed by scaling". The distortion types are

designed to simulate streaming video services. Both distortion types are perceptually adjusted to yield distinguishable distortion levels. An improved pairwise comparison method is adopted for subjective evaluation to save evaluation time. We use an improved pairwise comparison method to make the final MOS more stable and meaningful.

- Several image and video quality assessment (IQA and VQA) algorithms are evaluated against the MCL-V dataset. We show that the MCL-V dataset is one of the most challenging video quality assessment datasets to today's IQA and VQA indices.

Being inspired by [73], we propose two VQA methods in Chapter 4. They are the Fusion-based Video Quality Assessment (FVQA) Index and the Ensemble Learning Video Quality Assessment (EVQA) Index. It has the following specific contributions.

- The proposed FVQA index first classifies the whole MCL-V dataset into groups depending on the characteristics of video contents and then fuse several EVQA algorithms to predict the perceptual quality within each group. It has two distinctive features. First, video content grouping reduces content diversity and increases the fusion performance via machine learning. Second, different quality assessment methods adopted by FVQA can compensate each other with respect to different quality levels.

- The proposed EVQA method adopts frame-level training, and its solution is more scalable as compared to FVQA. It uses recursive grouping and the machine learning technique to reduce content diversity and improve the fusion performance. Furthermore, it uses an ensemble learning approach by taking different quality assessment methods and a wide variety of video content characteristics into account for better video assessment performance.

Finally, we argue that humans cannot perceive continuous-scale but discrete-scale quality changes over a range of coding bitrates, quantify this phenomenon based on JND, and then propose a new methodology to characterize the human visual experience on coded image/video content in Chapter 5. Specific contributions include the following.

- We study the problem of coded image/video quality assessment using a brand new framework based on JND. It is demonstrated by a small-scale subjective test that human perceived quality of coded images can be characterized by a piecewise constant function of the QF/QP with discontinuities at JND locations. Although these locations are content-dependent and statistically distributed, they do provide consistent and useful information in understanding the human visual experience.

- Given coded image/video content with densely sampled QF or QP values, we develop a new methodology to measure the number of JND points and locations, and analyze the inter-person variance of measured quantities.

- We build a JND dataset called MCL-JCI that contains 50 source images. Furthermore, we analyze the relationship between the source content and the number of its distinguishable quality levels.

1.4 Organization of the Dissertation

The rest of this dissertation is organized as follows. A brief review of previous related work is described in Chapter 2. Next, a newly-built MCL-V video quality dataset is described in detail in Chapter 3. The proposed FVQA and EVA indices are presented in Chapter 4. The JND-based visual quality assessment methodology is described, and a JND-based JPEG-coded image dataset is presented in Chapter 5. Finally, concluding remarks and future research directions are given in Chapter 6.

Chapter 2

Background Review

2.1 Video Quality Assessment Indices

Full-reference (FR) video quality indices take both the test video and the reference video as inputs. Because the information is retrieved from the reference video, the FR approach is more reliable than the no-reference approach, which only considers the test video. In this paper, we extend the image quality (IQA) metrics as VQA indices by averaging frame-level quality scores. The PSNR value is the most common FR VQA. It is calculated from the mean squared error (MSE), which can discriminate slight change between reference and distorted videos. However, it is highly content dependent due to perceptual effects[33], and its values are not comparable between different video contents. For example, video with grain noise is heavily penalized in PSNR although its perceptual quality is high as shown in Fig. 4.1. Even for identical content coded by different bit rates, the difference in PSNR is not a good indicator of subjective quality difference. Since PSNR is not well correlated with subjective human visual experience, other VQA indices are proposed to assess video quality[68, 74].

The state-of-the-art VQA metrics follow two main approaches: formula-based and learning-based. The formula-based approach creates a close form expression of perceptual quality. The famous ones include SSIM[118], VIF[105], FSIM[130]. This type of metrics are good to handle specific distortions. The other approach, the learning-based approach, integrates several features and predict the perceptual quality based on

the machine learning algorithms. [73] is an FR index that fuses existing quality metrics and predicts image quality with high accuracy. [81] combines a large number of computational statistical features and predict perceptual quality without reference.

Formula-based VQA indices were designed to measure "similarity" or "difference" with close forms. The former approach usually retrieves features from reference and distorted videos and computes the similarity index from the ratios of specific features. Examples include: SSIM[118], MSSIM[119], VIF[105] and FSIM[130]. The difference-based approach adopts the human visual system (HVS) model to predict how human perceives the difference between reference and distorted videos. WSNR[80], ADM[62], and VADM[61] belong to the difference-based category. These indices work reasonably well against simpler video quality databases developed before MCL-V[64]. The correlation coefficient of the best metric, VADM[61] is less than 0.75. However, these metric takes various approaches to justify the quality degradation between reference and distorted videos. It would be interesting to see the strength of each metric.

2.2 Formula-based Quality Indices: SSIM and FSIM

SSIM and FSIM are two commonly used IQA indices. They can be applied to the entire video sequence frame by frame followed by an averaging. We briefly review both of them below.

The computation of the SSIM index consists of two steps. First, it measures the local change in luminance, contrast and structure with the following formula:

$$SSIM(\mathbf{x}) = \frac{(2\mu_d(\mathbf{x})\mu_r(\mathbf{x}) + c_1)(2\sigma_{d,r}(\mathbf{x}) + c_2)}{(\mu_d^2(\mathbf{x}) + \mu_r^2(\mathbf{x}) + c_1)(\sigma_d^2(\mathbf{x}) + \sigma_r^2(\mathbf{x}) + c_2)}, \tag{2.1}$$

where **x** denotes the same block in distorted and reference images. The block size in SSIM is 11×11. A block-wise SSIM map is generated accordingly. Second, the SSIM index is the average of the SSIM map in form of

$$SSIM = \overline{SSIM_{\mathbf{x}\in\Omega}(\mathbf{x})}, \tag{2.2}$$

where Ω denotes the whole image domain.

The FSIM and FSIM Color (FSIMc) indices are computed based on the following two formulas:

$$FSIM = \frac{\Sigma_{\mathbf{x}\in\Omega}\left[S_{pc}(\mathbf{x})\right]\left[S_G(\mathbf{x})\right]\left[PC_m(\mathbf{x})\right]}{\Sigma_{\mathbf{x}\in\Omega}\left[PC_m(\mathbf{x})\right]}, \tag{2.3}$$

$$FSIMc = \frac{\Sigma_{\mathbf{x}\in\Omega}\left[S_{pc}(\mathbf{x})\right]\left[S_G(\mathbf{x})\right]\left[S_I(\mathbf{x})\cdot S_Q(\mathbf{x})\right]^{\lambda}\left[PC_m(\mathbf{x})\right]}{\Sigma_{\mathbf{x}\in\Omega}\left[PC_m(\mathbf{x})\right]}, \tag{2.4}$$

where S_{pc}, S_G, S_I, S_Q, and PC_m represent the phase congruency (PC), the gradient magnitude (GM), the I color compoent and the Q color component in the YIQ color space, and the maximum PC between distorted and reference images, respectively. FSIM uses PC and GM only while FSIMc includes all of them. Being similar to that the SSIM index is averaged by all local SSIM values, the resulting FSIM and FSIMc indices are the normalized mean values of all blocks. The these formula-based indices are difficult to be generalized to adapt to diversified contents. This is their main weakness.

2.3 Comparison of VQA Indices

Generally speaking, compressed video contains two types of visual artifacts, blurriness and blockiness, as illustrated in Fig. 2.1. High quality video has slight blurring but no obvious blocking. It may appear sharp to ordinary viewers, yet blurred edges can be

found by experts. The artifact of medium quality video is visible to common assessors, yet the overall quality is still acceptable. Low quality video has the strongest blurring and blocking artifacts, which is unacceptable for broadcasting. The appearance of visual artifacts changes along the scale from high to low quality.

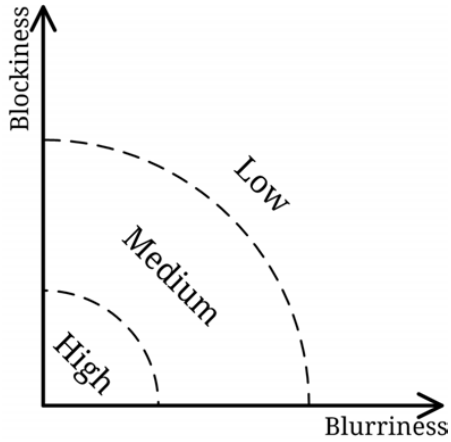


Figure 2.1: Illustration of video quality scale.

To understand the strength of different VQA indices in assessing streaming video quality, we test them with videos of different quality levels given in Fig. 4.1. Specifically, we split the 48 compression-only videos in MCL-V into three classes according the their mean opinion scores (MOS):

- Class-High: It is the high quality class whose video is slightly blurred with little blocking artifacts;
- Class-Medium: It is the medium quality class whose video has medium level blurring and blocking artifacts;
- Class-Low: It is the low quality class whose video has the strongest blurring and blocking among the three.

If a method performs well in Class-High, we claim that it performs well in detecting perceptual blurriness. We conduct experiments for videos in each class and compute the

root mean squared error (RMSE) of predicted quality scores. It is worthwhile to point out that correlation coefficients may not provide an accurate result on their performance due to the small number of sequence numbers (i.e. 16) in each quality class. The ranking of several VQA methods (namely, SSIM, VIF, FSIM, and VADM) for each video quality class is shown in Table 4.1. We see that the best VQA indices with respect to Class-High, Class-Medium and Class-Low are SSIM, FSIM and VIF, respectively. Thus, if we can fuse them into a single index via learning, it may offer the best solution among all three individual indices.

Table 2.1: The ranking of quality methods with respect to different quality classes.

Quality Class	Comparison of the Methods in RMSE		
High	SSIM	VIF	VADM
	0.565	0.599	0.617
Medium	FSIM	VIF	VADM
	0.683	0.684	0.738
Low	VIF	VADM	FSIM
	0.283	0.466	0.506

2.4 Learning-Based VQA Methodology

In this work, we focus on the learning-based approach, because it is more scalable. In addition, to let the model learn from the opinion scores, we adopt supervised learning algorithms in this work. Supervised learning entails learning a model between input data and labeled data and applying the model to predict unseen test data. In order to learn the subjective opinions, supervised learning is generally adopted in quality assessment research. The input data are calculated from image or video features and the labeled data are the mean of opinion scores (MOS) of quality databases. Since modeling HVS is so complicated that learning-based image quality assessment methods adopt the strategy to boost accuracy of the model. Narwaria and Lin[83] extracts the major structural

information in images by singular value decomposition and use support vector machine to map the feature to MOS. Moorthy and Bovik[81] developed a two-stage framework for learning-based IQA metric. In the first stage, the algorithm classifies the input image into one of the five distortions by SVM. In the second stage, image features are calculated and mapped to MOS by SVM.

One approach of learning-based methods is combing low-level features to final scores as introduced above. The other approach adopts ensemble learning to fuse existing methods rather than low-level features. Ensemble learning is a classic divide-and-conquer strategy that has been widely adopted in solving classification and regression problems[49, 52, 79, 7, 37, 27, 28]. Generally, in the framework of ensemble learning, multiple methods with diverse and complementary skills are fused to tackle a task such that the joint outcome outperforms any single method. Liu *et al.*[73] proposed a multi-method fusion (MMF) method for image quality assessment. A regression approach is used to combine scores of multiple IQA methods in the MMF. The MMF score is obtained by a non-linear fusion of scores computed by multiple methods with suitable weights obtained by a training process. So far, MMF offers one of the best IQA results in several popular databases such as LIVE[106], CSIQ[55], and TID2008[97].

Unlike successful IQA research work, there are several challenges in designing accurate learning-based VQA methods. First, Limited number of data The total number of images in image quality databases (e.g., [106, 97, 95, 96]) are larger than the number of sequences in video quality databases (e.g., [25, 103, 82, 64]). Hence, supervised learning operates well with abundant samples for training and develop accurate models. However, learning-based VQA method suffers from the limited samples, and the limited training set would cause over-fit and inaccurate model. Second, no ground truth to learn for temporal variation of video sequence. Temporal pooling has been studied by [101, 100, 104] over a decade. These researches show that using different methods does

not generally provide significant improvement, and no solid conclusion is made in this topic.

As compared with previous work[65], the proposed EVQA method adopts frame-level training to solve the problems of limited samples and temporal pooling. When we take one frame as one sample, we use the sequence MOS of the whole sequence as ground truth. Since the MOS is represented for the whole sequence, there exists mismatch between a single frame and the whole sequence. In order to compensate the mismatch, we assume each frame has close perceptual quality and the frame scores should be adjusted by video content. Therefore, we should use certain spatial and temporal indices to denote the spatial and temporal context of the current frame, such that the learning process can take account of the spatial and temporal masking effects.

Chapter 3

MCL-V: A streaming video quality assessment database

3.1 Introduction

The high-definition video broadcasting and streaming services are blooming nowadays. Consumers can enjoy on-demand video services from Netflix, Hulu or Amazon, and watching high-definition (HD) programs becomes the mainstream for video content consumption. According to the report in [110], more than half of US population watches on-line movies or dramas. Specifically, the viewers have increased from 37% in 2010 to 51% in 2013. The watched video programs vary in bit rates and resolutions due to the available bandwidth of their networks. Different sizes of video are transmitted at lower bit rates and up-scaled for display on HDTV (e.g., playing a 720p movie on the 1080p screen). This is common in people's daily life [11], yet users' video quality of experience on HD video has not yet been extensively studied in the past.

There are quite a few video quality assessment databases available to the public [9, 12, 13, 17, 24, 25, 30, 35, 54, 56, 57, 72, 82, 85, 86, 87, 88, 89, 90, 91, 92, 93, 103, 109, 114, 116, 128, 129]. They were however limited in the following areas [26], [122]. First, the source video set is not representative or diversified enough. For example, they do not contain dark scenes, sports scenes, traditional cartoon, and computer animation. The lack of these contents will not provide an extensive evaluation of viewers' experience. Second, the video resolution is low. The resolution of sequences in all VQA databases

except five [12, 35, 91, 116, 129] are lower than 1920×1080. Third, the distortion is not complete for the target application. For example, all above-mentioned VQA databases except [91, 56, 57] do not cover video up-scaling, which is encountered frequently in our daily life. Although the work in [91] includes practical distortion types, it has only three video sources. Being motivated by these observations, we build a new VQA database called MCL-V to address the shortcomings of existing VQA databases. The MCL-V database provides 12 source video clips, 96 distorted video clips and their associated mean opinion score (MOS). In this paper, we will elaborate on the methodology of building MCL-V such as collecting suitable video sources, generating distortions and conducting subjective evaluation.

One key issue in our design is to choose an appropriate subjective test procedure to collect opinion scores. Several subjective test methodologies have been recommended in VQEG [114, 115] and ITU [47, 48] as shown in Table 3.1. Since the precision of the final MOS is not improved by adopting the continuous scale [18, 111], the discrete scale is adopted in this work for user friendliness. Furthermore, we use an improved pairwise comparison method to make the final MOS more stable and meaningful.

Table 3.1: Classification of subjective testing methods.

	Discrete Scale	Continuous Scale
Single Stimulus	Absolute Category Rating (ACR) [47]	Single Stimulus Continuous Quality Evaluation (SSCQE) [48]
Double Stimulus	Degradation Category Rating (DCR) [47] Comparison Category Rating (CCR) [48]	Double Stimulus Continuous Quality Scale (DSCQS) [48]

The rest of this paper is organized as follows. Section 2 describes ways to choose representative and diversified reference sequences, to generate practical distortion types

and to determine the reasonable distortion levels. Section 3 presents an improved pairwise comparison method for subjective evaluation and elaborates on the process of collecting and normalizing opinion scores in the subjective test. We study the MOS values and analyze the performances of several existing IQA and VQA metrics against the MCL-V database in Sections 4 and 5, respectively. Finally, concluding remarks are given in Section 6. The whole database is publicly available on the USC Media Communication Lab website http://mcl.usc.edu/mcl-v-database/.

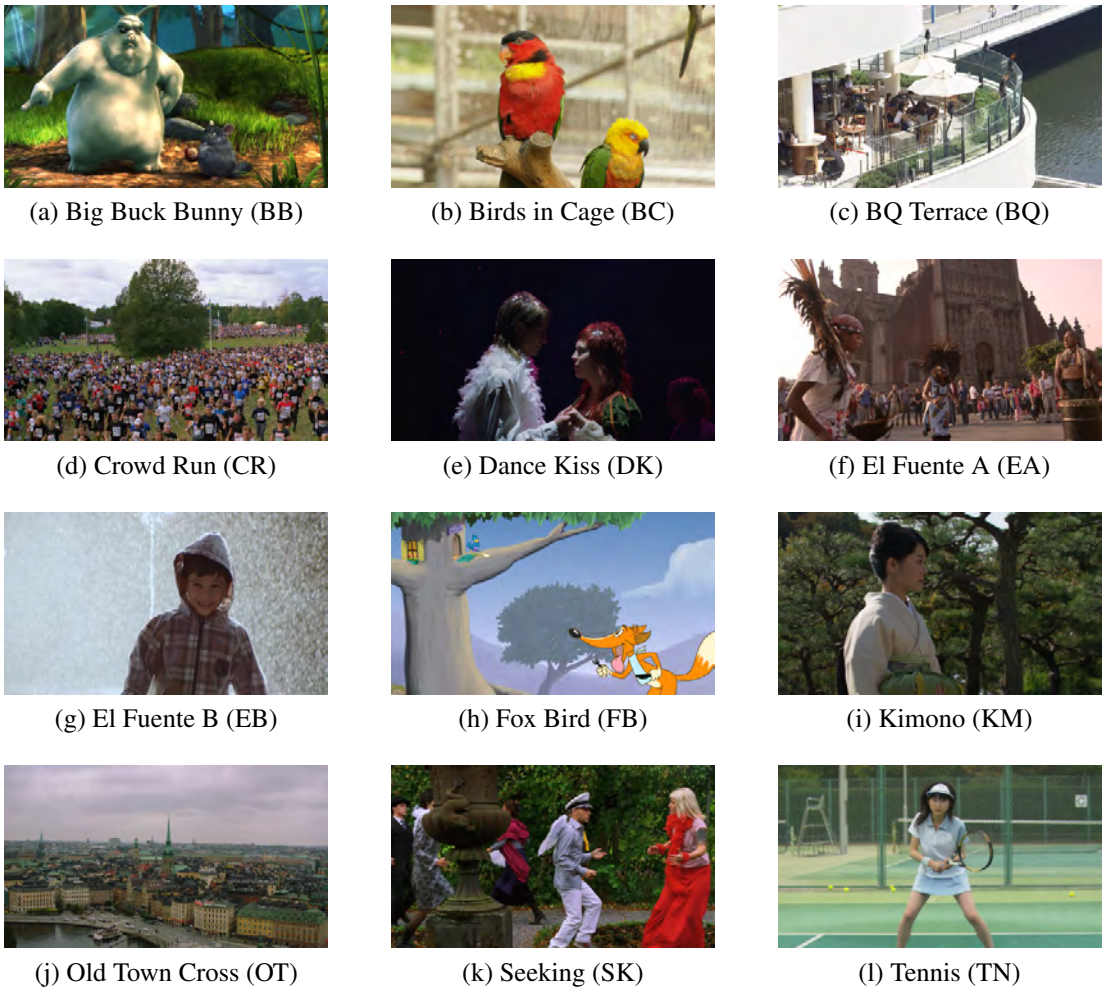
3.2 Construction of MCL-V Database

3.2.1 Source Video Selection

We selected 12 uncompressed HD video clips as the source sequences. Some sequences are originally in YUV444p or YUV422p, and we converted them into YUV420p using [10] to make all videos included in the MCL-V database be YUV420p at a fixed resolution of progressive 1920×1080. The frame rates of the sequences range from 24 fps to 30 fps, and the length of each video is 6 seconds. Figure 3.1 shows all reference videos with a single frame.

The selected sequences are freely available from several sources, including HEVC test sequences [84], TUM dataset [53], CDVL [20], and others [29, 34, 40]. They were professionally acquired and recorded in digital form. We select some of them to construct the MCL-V database based on the following two criteria.

First, some prior databases [82, 103] contain scenes that are not representative in video applications. For example, there are video clips with a close view on the water surface or the blue sky in the LIVE databse [103]. These sequences were used for video coding performance test since they contain specific contents which are difficult to



(a) Big Buck Bunny (BB) (b) Birds in Cage (BC) (c) BQ Terrace (BQ)

(d) Crowd Run (CR) (e) Dance Kiss (DK) (f) El Fuente A (EA)

(g) El Fuente B (EB) (h) Fox Bird (FB) (i) Kimono (KM)

(j) Old Town Cross (OT) (k) Seeking (SK) (l) Tennis (TN)

Figure 3.1: Selected Source Video Sequences

encode. However, they are not common scenes in movies or dramas. We prefer more representative scenes since they can better reveal human visual experience.

Second, the database should have sufficient diversity in terms of several characteristics. We list various characteristics for diversity consideration in Table 3.2. They are categorized into three groups: 1) high-level video genres, 2) mid-level video semantics and 3) low-level video features. We aim to make the database cover a wide range of characteristics given in the table.

Table 3.2: Video characteristics used in diversity check

Video Genres	**Video Semantics**	**Video Features**
• Cartoon • Sports • Indoor	• Face • People • Water • Number of objects • Salience	• Brightness • Contrast • Texture • Motion • Color Variance • Color Richness • Sharpness • Film Grain • Camera motion • Scene change

For video genres, we take several new genre types such as animation and sports into account. These video genres have different characteristics from others. For instance, cartoons scenes contain clear edges and simple color components while sports scenes contain fast moving objects with simple background. These videos are commonly seen in applications and should be included in the MCL-V database.

For video semantics, we consider factors that will have a great impact on human visual perception. For example, while other databases usually do not include video scenes with a close-up face, we take this feature into consideration since it is typical in many dramas. In addition, the human face is typically a region of visual salience which attracts human attention.

For video features, we examine brightness, contrast, motion, texture and color since these features are related to the level of the video compression distortion. These features also have influence on the visual masking effect. For example, there is no obviously dark scene or fast-motion scene in existing video quality databases [25, 82, 103]. As a result, they do not contain representative video clips for horror movies or action films. The diversity of video features can be captured by the Spatial Information (SI) versus the

Temporal Information (TI) plot as defined in the ITU-T Recommendation [47]. Eqs. 4.4 and 4.5 of SI and TI are shown as follows:

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\}\,, \tag{3.1}$$

$$TI = max_{time}\{std_{space}[M_n(i,j)]\}\,, \tag{3.2}$$

SI is calculated based on the Sobel filter. The n^{th} video frame, F_n, is first filtered with the Sobel filter and taken the standard deviation over space domain. Then, the maximum value along the time is chosen to present SI. TI is based on motion difference. $M_n(i,j)$ is the difference in pixel at the i^{th} row and j^{th} column between F_n and F_{n-1}. TI is computed as the time maximum of the space standard deviation of $M_n(i,j)$. These two indices correspond to the texture and the motion features in Table 3.2, respectively. As shown in Fig. 3.2, the 12 video sequences in the MCL-V database are well scattered in the feature space spanned by SI and TI, which demonstrates the diversity of the MCL-V database.

Not all characteristics can be quantitatively measured. We conducted subjective evaluation on the characteristics of video clips to illustrate the diversity of the MCL-V database and show the results in Table 3.3. The main characteristics are listed from high-to-low levels in the first column while the 12 source sequences are listed in the top row in this table. Each column in the table represents the characteristics of the corresponding source sequence. The subjective evaluation was conducted by a group of professionals. Since there are only a few levels defined for each property, the results can be easily verified and they are quite consistent among viewers. This table shows that the selected source video clips in the MCL-V database well span all characteristics with excellent diversity.

The contents of the 12 source video clips are described below.

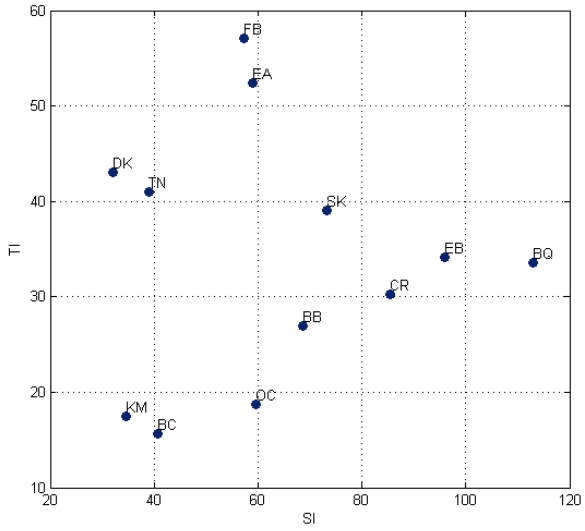


Figure 3.2: Plot of the Spatial Information (SI) and the Temporal Information (TI) indices for selected video sequences.

- Big Buck Bunny (BB) in Fig. 3.1a: An animated sequence, where there are two animals in the video, with clear textures and rich backgrounds.

- Birds in Cage (BC) in Fig. 3.1b: Two colorful birds standing in front of a clean background in a still scene.

- BQ Terrace (BQ) in Fig. 3.1c: Plenty of vehicles moving on a bridge, and below the bridge are the water. The camera pans in a diagonal direction.

- Crowd Run (CR) in Fig. 3.1d: A crowd of people running together, with big trees and the blue sky as the background.

- Dance Kiss (DK) in Fig. 3.1e: People dancing in a dark room. There are scene changes, and the motions are fast. People will focus on two main characters that kiss in the middle of the scene.

Table 3.3: MCL-V source video diversity

	BB	BC	BQ	CR	DK	EA	EB	FB	KM	OT	SK	TN
Cartoon								√				
CG Animation	√											
Sports												√
Indoor					√							
Scene change					√	√	√	√	√			√
Camera motion (*)	P	S	P	M	Z	P	P	SPZ	P	P	P	P
Face close-up					√		√		√			
People					√	√	√		√		√	√
Water Surface			√									
Salience	√	√			√		√	√	√			√
Film grain noise										√		
Flat, low gradient area		√				√						
Object number (**)	1	1	2	3	1	2	1	2	1	0	2	1
Brightness	2	3	2	2	1	2	3	3	2	2	3	2
Contrast	3	3	2	3	1	2	3	2	2	1	2	2
Texture (spatial variance)	2	1	2	3	2	2	3	2	3	2	2	1
Motion (temporal variance)	2	1	1	3	3	2	2	3	2	1	2	3
Color variance	1	3	1	3	1	1	1	3	2	1	2	1
Color richness	2	3	1	2	1	1	1	3	2	1	3	2
Sharpness	2	3	2	1	2	2	1	3	3	2	2	1

For high-level video genres, √ indicates the video contains this features, and vice versa.
For low-level video features, the number represents the level of the feature, where (1 , 2 , 3) means (low , median , high), respectively.
(*) Camera motion types: S for Still, P for Pan, Z for Zoom, M for irregular movements.
(**) Object number: (0 , 1 , 2 , 3) means (no main object , one , a few , many)

- El Fuente A (EA) in Fig. 3.1f: Several people in the tribe dancing around a man who is drumming. In addition to fast motions, the scene also contains large portions of ground and sky that are with low gradient.

- El Fuente B (EB) in Fig. 3.1g: A boy walking in front of a fountain. In another scene, we have a close view to the frontal face of the boy. The water drops in the background make it very difficult for video coding.

- Fox Bird (FB) in Fig. 3.1h: A cartoon sequence with a fox running rapidly. There are scene changes, and several camera motions are involved.

- Kimono (KM) in Fig. 3.1i: A woman walking slowly toward the camera in front of the woods. The woman is close to the camera and the face of the women can be seen clearly.

- Old Town Cross (OT) in Fig. 3.1j: A bird's eye view of an old town with slow camera movements. Except the sky and the buildings, there are no other objects in the scene. Film grain noise can be observed in this video sequence.

- Seeking (SK) in Fig. 3.1k: Several people in different colors moving around.

- Tennis (TN) in Fig. 3.1l: Girls playing tennis, and running very fast to chase the ball. There is also a scene change in this sequence.

3.2.2 Distorted Video Generation

We consider two typical distortion types in video applications.

- H.264/AVC compression

  H.264/AVC is the most popular video format used in IP-based video streaming. The compression artifact due to lower coding bit rates is one main distortion source.

- compression followed by scaling (or simply called scaling below)

  The image size has to be scaled when a video clip of a lower resolution is displayed in a display panel of higher resolution. This effect can be simulated via a

cascade of operations: *down-sampling, encoding, and then resizing to the original resolution*.

We adopt four distortion levels for each distortion type. Since there are 12 source reference sequences, we have $12 \times 2 \times 4 = 96$ distorted sequences in total.

We used x264 [4] as the encoder to generate compressed video files. Rate control was enabled with a variable bit rate, and a two-pass encoding scheme was used to ensure consistent perceptual quality frame by frame so that viewers can determine the opinion reasonably. At most two B frames were allowed between an I and a P frames. Both the input and the output video resolutions are kept at 1080p as shown in Fig. 3.3. The distortion levels are controlled by the target bit rates. Since we select a wide variety of video sequences, the bit rate range is from 0.2 Mbps to 10 Mbps.

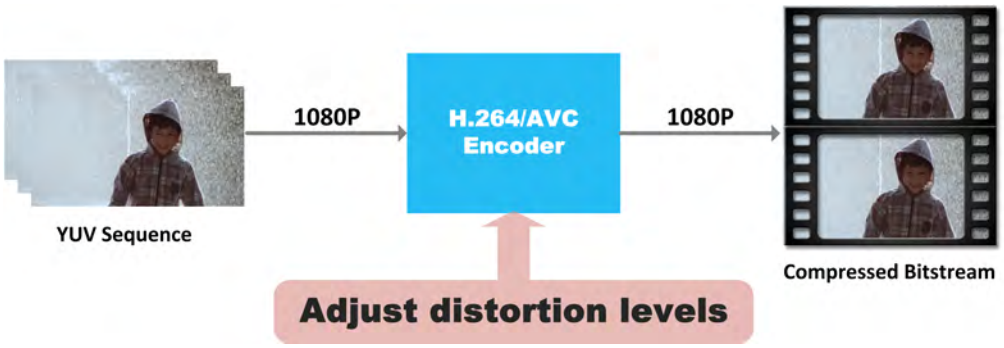


Figure 3.3: The process of generating distorted video contents with an H.264/AVC codec.

Since the bit rates depend on video contents, we used the following method to subjectively select distinguishable levels. First, we generated 300 compressed sequences with different bit rates in the above range and drew a plot of "the PSNR value versus the bit rate" as shown in Fig. 3.4. Although the PSNR value could be used as an auxiliary tool, We do not rely on PSNR to determine perceptual quality. In this bit rate range, there is a region where coded video quality is no distinguishable any longer as the bit rate increases. We also set up a lower bound in the sense that the quality of video clip

will not be acceptable if the bit rate is lower than this bound. The perceptual upper and lower bounds are plotted as two solid horizontal lines in Fig. 3.4. We generated 300 clips for selection within the interval, and divided them into four regions with respect to the PSNR value - A, B, C and D. Finally, we choose four suitable distortion levels (namely, one from each region) based on subjective visual experience.

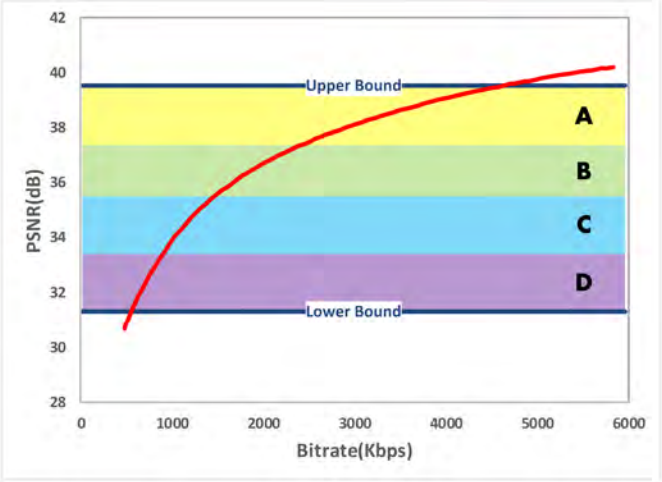


Figure 3.4: The process of selecting compression-distorted video clips with four distortion levels.

To generate scaling-distorted video files, we follow the process as depicted in Fig. 3.5. First, all video sequences are converted to 720p before compression. The down-sampling process is achieved by using the Lanczos algorithm so as to preserve as many details as possible. Different video players may have different settings in video resizing. To make a controllable environment, we choose the bilinear interpolation as the up-sampling algorithm. The format conversion is done by FFmpeg [10]. In the subjective test, we play up-sampled YUV sequences. The distortion levels are adjusted in the compression step, which is the same methodology as before.

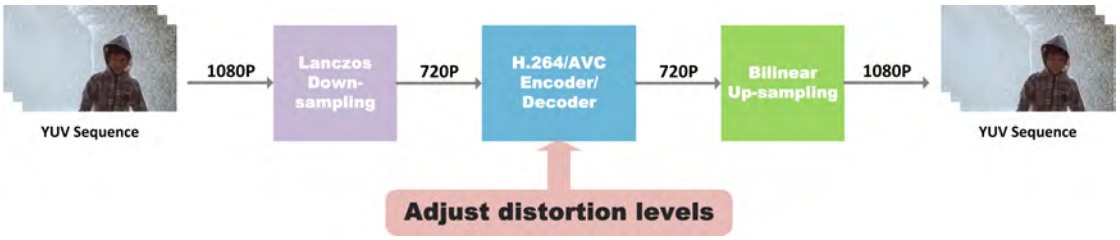


Figure 3.5: The process of generating scaling-distorted video clips.

3.3 Subjective Video Quality Assessment

3.3.1 Subjective Assessment Methodology

Quite a few subjective test methods for multimedia applications have been recommended by VQEG [114, 115] and ITU [47, 48]. There are various discrete scoring methods, for example, five score levels in DCR [47] and seven score levels in CCR [48]. When the number of choices increases, it becomes more difficult to get consistent and stable scores across multiple assessors. That is, the same choice made by a different person may have a different meaning. Sometimes, the decision of the same person may also vary along the test time. To mitigate these problems, we adopt the pairwise comparison method in the subjective test.

Video clips of the same source but with a different distortion level were selected to form a pair for comparison. An assessor was only asked to decide which video has better quality out of the pair. The objective of a sequence of pairwise comparisons by the same assessor is to create an ordered list of multiple distorted video sequences according to the perceptual quality. The shortcoming of a straightforward pairwise comparison method is its long assessment time. For example, if one attempts to compare the quality of N samples, the total number of an exhaustive pairwise comparison is C_2^N. Several methods were proposed to lower the complexity of the pairwise comparison method, *e.g.*, [107, 108, 58, 98, 60, 59, 124]. Here, we propose another simplification method as

illustrated in Fig. 3.6, where each circle represents one distorted sequence. The basic idea is sketched below.

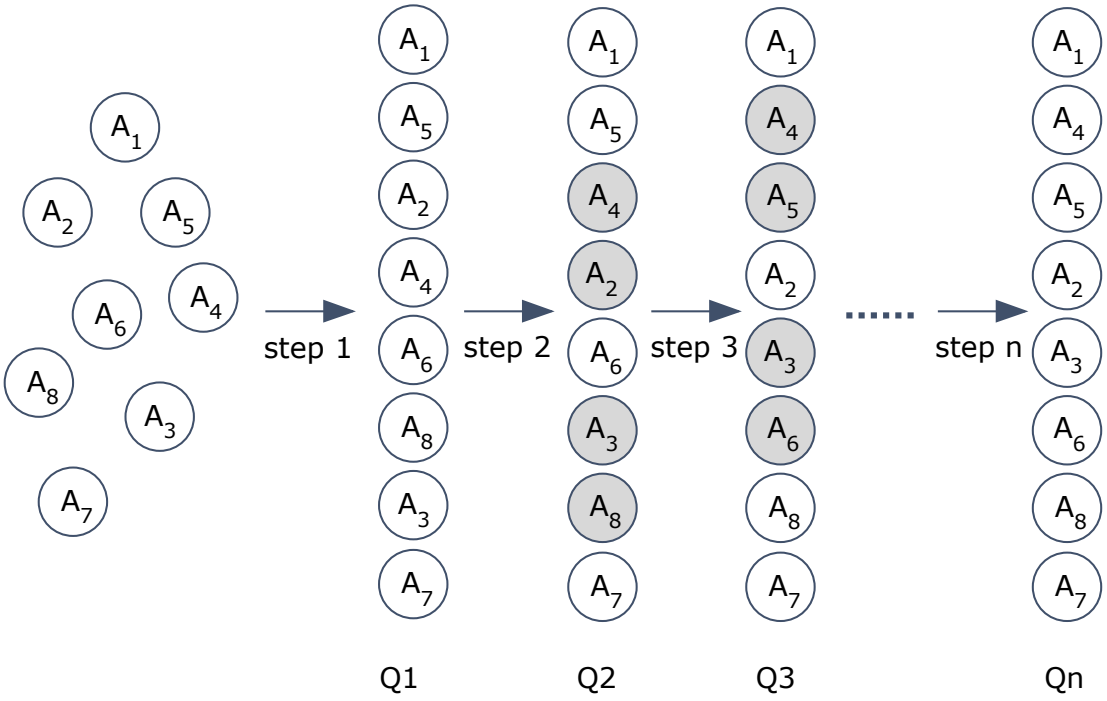


Figure 3.6: Illustration of a simplified pairwise comparison process.

It is desirable to get a good initial list for pairwise comparison. The distorted sequences were first sorted by visual inspection. When the two sequences are far from each other in the queue, it means the visual quality gap between them is obvious. This initialization process is illustrated in Step 1, which is used to generate a rough sorted list of all distorted video sequences for the initialization purpose at a low complexity. Specifically, we ask a small number of professionals to participate in the subjective evaluation with the ACR [47] to achieve this goal. The sorted list result is shown in Q_1, where A_1 and A_7 denote sequences of the best and worst quality, respectively.

After the initialization, all assessors are invited to participate in the subjective test. When the distance of two distorted sequences in the ordered list is longer, their quality

difference is more obvious. Thus, each assessor is asked to conduct pairwise comparison of adjacent nodes only. In the given example, if the assessor prefer A_4 to A_2, then A_4 and A_2 are swapped. Furthermore, A_8 and A_3 are swapped similarly. After this round, the assessor is led to a new ordered list denoted by Q_2. With Q_2, the four new adjacent pairs $(A_4, A_5), (A_2, A_6), (A_6, A_3)$, and (A_8, A_7) will be compared by the assessor, and the assessors decision will create Q3. The process is repeated for the same assessor until no further swap is needed. A comparison record matrix is used to record whether any pair of nodes has been compared or not. If two adjacent nodes have been compared by this assessor once, no further comparison will be conducted. All adjacent nodes in the final ordered list, Q_n, will be compared by the same assessor once, and the sequence in the list reflects the preference of this assessor. A preference matrix for the nth assessor, denoted by P_n, can be created accordingly.

By aggregating the preference matrices of multiple assessors, we get the group preference matrix, M. Here, we use the Bradley-Terry model [14, 15, 113] to derive the final absolute scale score from the group preference matrix. Note that the Bradley-Terry (BT) model and the Thurstone-Mosteller (T-M) model are two well-known models to convert pair comparison data to psychophysical scale values for all stimuli. To verify its accuracy, we compute the point score, as defined in [97], for the nth assessor based on his/her ordered list Q_n and compare them in Fig. 3.7, where the horizontal axis is the point score and the vertical axis is the absolute scale number obtained by using the Bradley-Terry model. We see that the two results are very consistent. The Pearson Correlation Coefficient (PCC) between them is 0.9961. The absolute scale score can also be derived by using the Morrisey Gulliksen incomplete matrix solution [36, 42].

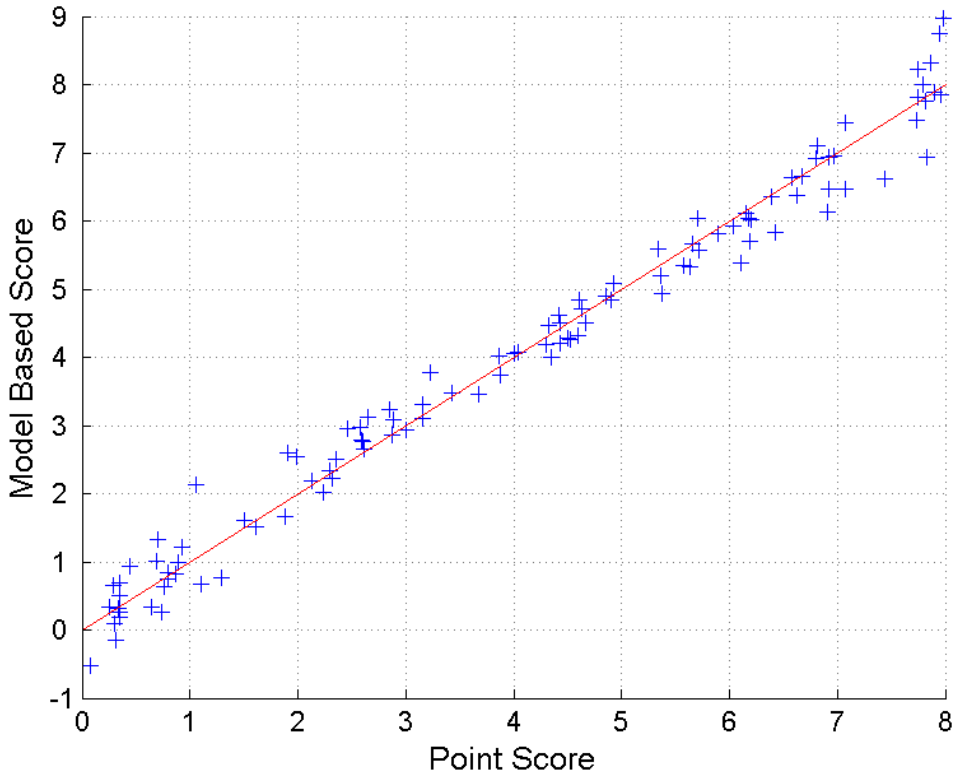


Figure 3.7: Correlation between the point score and the absolute scale number calculated based on the Bradley-Terry Model.

3.3.2 Test Setting and Procedure

The assessors are seated in a controlled environment to assess the quality of video. The view distance is strictly kept in 3 meters (3.5 times of the picture height), from the center of the monitor to the seat. The videos are displayed on the HDTV, LG 47LW5600, with native 1920x1080 resolution, thorough this work.

The total number of assessors is 45 consisting of 13 females and 32 males. Their age is distributed from 20 to 40. Some of the assessors are PhD students in image processing field. Others are naive and inexperienced with the topic of video quality assessment. The assessors are confirmed verbally with sound or corrected vision.

Before each test session, a lesson is offered to assessors on how to provide their opinion scores. The training session consists of two parts. For the first part, a 5-minute video is played with various video quality. In the second part, assessors learn to see the difference in video quality and the way to operate the software. After the training lesson, assessors will see the notification on the screen and start their test session.

The subjective test is conducted based on the modified pairwise comparison. The software is written in Python. Video clips of the same source but with a different distortion level were selected to form a pair for comparison. They were randomly ordered and played one by one with a 3-second break in-between. An assessor was given three choices: "the first one is better", "the second one is better", or "no difference". Eight video sets were tested at each session and 45 sessions were conducted. One video set includes all distorted video clips from the same video content. We collected 32 opinion scores for each video set. Most assessors have no prior experience in video coding. The test time and each decision made by every assessor were recorded for outlier detection and score conversion. The duration of a test session ranges from 20 to 30 minutes.

3.4 Analysis of Subjective Opinion Scores

The collected opinion scores are processed according to the ITU recommendation [48]. The screening of possible outlier subjects is done by following that in [97]. That is, the highest 10% and lowest 10% of the point scores are discarded. The final MOS values with the 95% confidence interval sorted along the decreasing preset level are shown in Figure 3.8. We see that the MOS values range from 0 to 8. The mean and the standard deviation of assessors' scores for each distorted video file are provided in the MCL-V database. In this section, we discuss how video properties affect these scores in the following two scenarios.

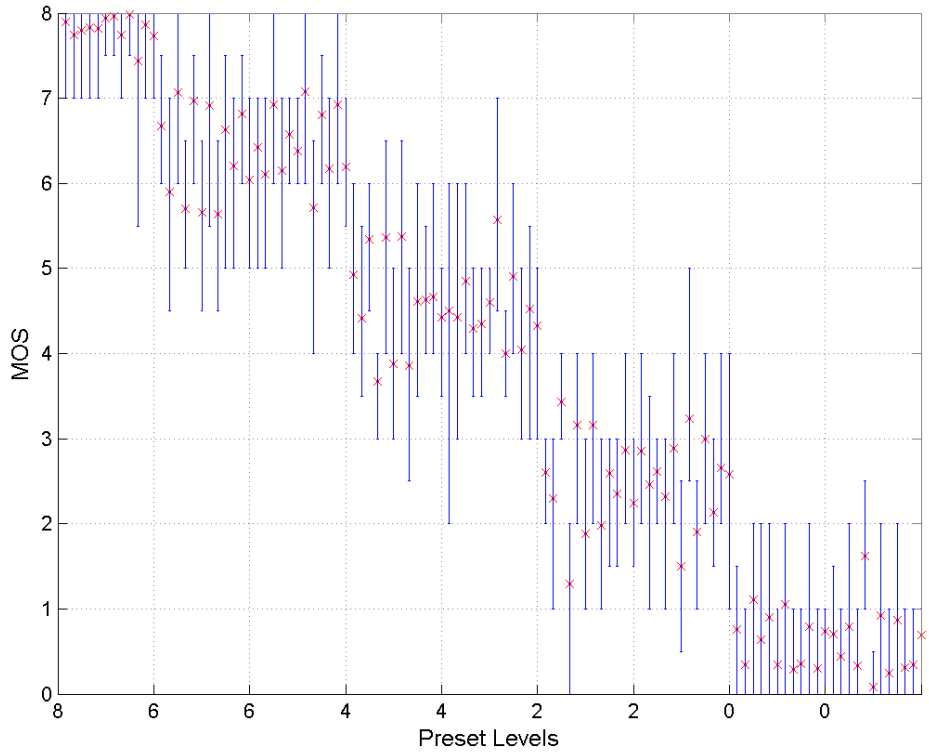


Figure 3.8: Sorted Mean Opinion Scores with the 95% confidence interval, where the red cross is the mean and the blue line indicates the stand deviation range between -1 and 1.

First, we compare the mean and the variance of opinion scores for two compressed sequences in Table 3.4: Crowd Run and Kimono. As shown in the table, the variance of Kimono is significantly lower than that of Crowd Run. This can be explained as follows. Human has a clear visual attention region in Kimono, which is the Japanese lady in the scene. In contrast, there is no clear visual attention region in Crowd Run. As a result, viewers' opinions are more diverse for Crowd Run.

Next, we compare the MOS values for two compression-distorted and scaling-distorted video clips, Dance Kiss and Fox Bird, in Table 3.5. Since the variances are close in each level between two distortion types, we only list the MOS here. For Dance

Table 3.4: Comparison of the mean and the variance of opinion scores for compression-distorted Crowd Run and Kimono sequences in MCL-V.

Level	Mean		Variance	
	Crowd Run	Kimono	Crowd Run	Kimono
Good	6.91	6.92	0.25	0.22
Fair	5.38	4.85	0.51	0.38
Poor	3.16	2.61	0.41	0.16
Bad	1.05	0.80	0.39	0.35

Table 3.5: Comparison of the MOS values of the compression- and scaling-distorted Dance Kiss and Fox Bird sequences.

Level	Dance Kiss		Fox Bird	
	Compression	Scaling	Compression	Scaling
Good	6.63	6.20	6.58	6.38
Fair	4.61	4.62	4.34	4.59
Poor	2.59	2.35	2.88	1.50
Bad	0.35	0.79	1.61	0.07

Kiss, the MOS of scaling distortion is close to that of compression distortion. For Fox Bird, we observe a significant MOS drop in scaling distortion when the bit rate becomes low. Fox Bird is a bright video clip that contains stronger edges as a result of the cartoon content. The scaling distortion is more visible in a bright scene with strong edges. In contrast, Dance Kiss is a dark video clip in our selection. It has smoother textures. The scaling distortion is reduced by the dark scene and the smooth texture.

Since the MCL-V database contains a wide range of video contents, it can capture the characteristics of the human visual system better and allow researchers to develop better objective video quality assessment algorithms.

3.4.1 Performance Comparison of Objective VQA Methods

Several full-reference (FR) IQA and VQA algorithms [118, 119, 105, 130, 71, 126, 61, 117] are evaluated against the collected MOS of the MCL-V databased and reported in

this subsection. IQA methods can be extended to VQA methods by averaging frame-level quality scores. The IQA source codes of [118, 119, 105] are downloaded from [32]. Others are downloaded from respective authors' websites. The state-of-the-art VQA methods take both spatial and temporal artifacts into account. For example, the VADM method decouples two spatial distortion types; i.e. detail losses and additive impairments, and evaluate them separately. Furthermore, the motion information is adopted by VADM to measure the temporal masking effect. The ST-MAD method employs the spatio-temporal images to model the interaction between these two artifacts.

Three performance measure for these IQA and VQA methods are calculated and compared. They are: 1) the Pearson correlation coefficient (PCC) [114, 115], 2) the Spearman rank-order correlation coefficient (SROCC) [114, 115], and 3) the root mean squared error (RMSE) [114, 115]. The PCC and SROCC are computed after nonlinear regression on the quality scores using the logistic function as recommended in [22]. Mathematically, we have

$$y = \beta_1 \cdot (0.5 - \frac{1}{1 + e^{\beta_2(x-\beta_3)}}) + \beta_4 \cdot x + \beta_5, \tag{3.3}$$

where x is an objective quality score and β_i, $i = 1...5$, are fitting parameters.

First, the performance of these quality metrics with respect to the compression distortion is shown in Table 3.6. We see that FSIM and VADM give the best performance among the test group for the compression distortion due to their good distortion models. They are close in PCC and RMSE while VADM provides a better SROCC measure. However, their PCC and RMSE values are still lower than 0.75, which allows room for further improvement.

Next, the performance of these quality metrics with respect to the scaling distortion is given in Table 3.7. First, we see that these metrics perform worse for the scaling

Table 3.6: Performance comparison of objective quality metrics with respect to the compression distortion in MCL-V.

	PCC	SROCC	RMSE
PSNR	0.471	0.422	1.994
MSSIM[119]	0.617	0.609	1.779
SSIM[118]	0.650	0.633	1.718
VIF[105]	0.667	0.637	1.685
GMSD[126]	0.653	0.644	1.712
GSM[71]	0.715	0.713	1.580
FSIM[130]	0.770	0.775	1.441
S-MAD[117]	0.702	0.701	1.609
T-MAD[117]	0.625	0.623	1.763
ST-MAD[117]	0.657	0.663	1.702
VADM[61]	0.747	0.735	1.515

Table 3.7: Performance comparison of objective quality indices with respect to the scaling distortion in MCL-V.

	PCC	SROCC	RMSE
PSNR	0.463	0.493	1.881
MSSIM[119]	0.609	0.630	1.683
SSIM[118]	0.635	0.649	1.639
VIF[105]	0.636	0.661	1.637
GMSD[126]	0.634	0.662	1.642
GSM[71]	0.692	0.707	1.531
FSIM[130]	0.722	0.702	1.468
S-MAD[117]	0.659	0.624	1.594
T-MAD[117]	0.580	0.548	1.728
ST-MAD[117]	0.617	0.585	1.669
VADM[61]	0.728	0.741	1.469

distortion than the compression distortion. Second, VADM and FSIM are still the top two performers among the test group while VADM outperforms FSIM in all three scores.

Finally, we list the performance of all methods against the entire MCL-V database that contains both compression and scaling distortion types in Table 3.8. Furthermore,

Table 3.8: Performance comparison of objective quality indices with respect to both compression and scaling distortions in MCL-V.

	PCC		SROCC		RMSE	
Database	MCL-V	LIVE[103]	MCL-V	LIVE[103]	MCL-V	LIVE[103]
PSNR	0.472	0.549	0.464	0.523	1.956	9.176
MSSIM[119]	0.621	0.739	0.623	0.732	1.740	7.398
SSIM[118]	0.650	0.542	0.648	0.525	1.687	9.223
VIF[105]	0.660	0.570	0.655	0.557	1.666	9.019
GMSD[126]	0.650	0.737	0.661	0.726	1.686	8.414
GSM[71]	0.709	0.650	0.711	0.684	1.565	8.341
FSIM[130]	0.750	0.690	0.755	0.689	1.466	8.240
S-MAD[117]	0.681	0.737	0.670	0.721	1.624	7.669
T-MAD[117]	0.600	0.818	0.584	0.815	1.774	6.562
ST-MAD[117]	0.634	0.830	0.623	0.824	1.714	6.133
VADM[61]	0.742	0.844	0.752	0.835	1.489	5.945

we list their performance against the LIVE database [103] for side-by-side comparison. The top three performers for the LIVE database are VADM, ST-MAD and T-MAD. Their PCC and SROCC scores are all above 0.80. In contrast, their performance degrades substantially in the MCL-V database, which indicates that MCL-V is a more challenging video quality database. This can be explained by that the source video in MCL-V is more diversified, and it is not easy to find an ideal metric to cover all of them.

3.5 Conclusion and Future Work

The construction of a new HD video quality assessment database, called MCL-V, was described in this work. MCL-V contains 12 source video clips and 96 distorted video clips with subjective assessment scores. The source video clips were selected from a large pool of public-domain HD video sequences with representative and diversified contents. Several existing IQA and VQA algorithms were evaluated against the MCL-V database. The database is publicly available at http://mcl.usc.edu/mcl-v-database/ for future research and development.

We attempted to analyze the relationship between video properties and the MOS values using 4 video sequences as examples in Section 3.4. A thorough analysis of the acquired MOS involves visual salience detection/tracking and a good understanding of the spatial/temporal masking effects. Although this is beyond the scope of our work, it is an interesting topic for further study. Furthermore, as shown in Section 3.4.1, there is no objective quality metric that has a PCC (or SROCC) value higher than 0.75 against the MCL-V database. The development of a better VQA method is also in need.

Chapter 4

Objective Assessment Methods

4.1 Introduction

Video streaming service grows and evolves in a incredible speed. Thousands of titles are monthly added to major service providers, such as Netflix, Hulu, and Amazon. The delivered titles are compressed and scaled in various bit rates and resolutions to match clients' bandwidth and end-terminal. The streaming bandwidths range from 500 Kbits/sec to 12 Mbits/sec and the resolutions vary from CIF to UHD. Thus, the distortion comes not only from compression but also resizing, where a video of a smaller size is scaled up to a larger resolution to match the dimension of the display device. The nature of streaming video is so complicated that assessing the video quality becomes a significant issue.

Video quality assessment (VQA) is essential to video coding and processing. Since the streaming service may have higher bit rates to maintain video quality [3, 45], the blocking effect is not as strong as that in existing video quality databases. Furthermore, video quality is blurred due to video resizing. Spatial and temporal masking effects appear in various forms due to content diversity. For instance, visual artifacts are likely to be seen in still scenes than fast-motion scenes. A new video quality database, called MCL-V[64], was recently constructed for streaming video quality evaluation. MCL-V provides richer diversity in source video contents with practical distortions.

To measure the distortion, the mean squared error (MSE) is widely applied to assess the quality. However, [33] indicates several psychovisual phenomena affects the human

visual system, and MSE does not reflect these phenomena. These phenomena affects are so called spatial masking effects, which are have been studied in [31, 6, 8, 44, 5, 43]. For video, temporal masking effect has significant impact on human perception as indicated in [78, 19, 75, 94]. Thus, researchers in the VQA field strongly intend to model these phenomena and apply to VQA metrics. Li *et al.* [62] and Brandao and Queluz [16] both adopted Daly's contrast sensitivity function (CSF) [23] in their work. Their results show that modeling masking effects well can improve the accuracy of the VQA indices. However, only a few masking effects are modeled as well as CSF. Since it is difficult to model other masking effects by experiments, the research results from Vision and Psychology are rarely to be combined to VQA related researches. Therefore, machine-learning based VQA indices[83, 81, 73] are introduced to find the relationship from data-oriented approach, rather than digging out the interrelationship between human brain and vision systems.

Several successful learning-based IQA methods[83, 73, 81] are proposed and demonstrate high-correlated prediction to the subjective opinions. Unlike successful IQA research work, there are many challenges in designing accurate learning-based VQA methods. First, Limited number of data restricts the learning performance. Second, no ground truth is available to learn for temporal variation of video sequence. Temporal pooling has been studied by [101, 100, 104] over a decade, but using different methods does not generally show significant improvement. As compared with previous work[65], we consider frame-level training to solve the problems of limited samples and temporal pooling. When we take one frame as one sample, we use the mean opinion score (MOS) of the whole sequence as ground truth. In order to compensate the mismatch of between frame data and sequence MOS, we use video content indices to indicate the spatial and temporal context of the current frame, such that the learning process is capable to take account of the spatial and temporal masking effects.

We propose an ensemble learning-based Video quality Assessment (EVQA) index in this work. First, The frame samples are applied with multiple IQA methods and the proposed video content indices. Then, we justify and reposition the IQA methods as essential metric, group classifier, and fusion candidate. In the second step, the group classifier and essential metric are used to recursively partition the whole sample space into several groups. Finally, several VQA algorithms are selected and fused to predict the perceptual quality within each group. The contributions of this novel EVQA index are 1) a new approach to justify the existing IQA metrics and reposition their roles in a scalable VQA framework, 2) two proposed video content indices for addressing video property as input features, 3) recursive grouping by experts to reduce content diversity and improve the fusion performance via machine learning, and 4) optimized ensemble learning to adopt different quality assessment methods and video content indices to compensate each other with respect to different quality levels. The rest of this paper is organized as follows. A brief review of previous related work is described in Section II. The details of the proposed EVQA index is presented in Section III. Experimental results are shown in Section IV. Finally, concluding remarks are given in Section V.

4.2 Background Review

Full-reference (FR) video quality indices take both the test video and the reference video as inputs. Because the information is retrieved from the reference video, the FR approach is more reliable than the no-reference approach, which only considers the test video. In this paper, we extend the image quality (IQA) metrics as VQA indices by averaging frame-level quality scores. The PSNR value is the most common FR VQA. It is calculated from the mean squared error (MSE), which can discriminate slight change between reference and distorted videos. However, it is highly content dependent due to

perceptual effects[33], and its values are not comparable between different video contents. For example, video with grain noise is heavily penalized in PSNR although its perceptual quality is high as shown in Fig. 4.1. Even for identical content coded by different bit rates, the difference in PSNR is not a good indicator of subjective quality difference. Since PSNR is not well correlated with subjective human visual experience, other VQA indices are proposed to assess video quality[68, 74].

The state-of-the-art VQA metrics follow two main approaches: formula-based and learning-based. The formula-based approach creates a close form expression of perceptual quality. The famous ones include SSIM[118], VIF[105], FSIM[130]. This type of metrics are good to handle specific distortions. The other approach, the learning-based approach, integrates several features and predict the perceptual quality based on the machine learning algorithms. [73] is an FR index that fuses existing quality metrics and predicts image quality with high accuracy. [81] combines a large number of computational statistical features and predict perceptual quality without reference.

4.2.1 Evaluation of Formula Based Quality Indices

Formula-based VQA indices were designed to measure "similarity" or "difference" with close forms. The former approach usually retrieves features from reference and distorted videos and computes the similarity index from the ratios of specific features. Examples include: SSIM[118], MSSIM[119], VIF[105] and FSIM[130]. The difference-based approach adopts the human visual system (HVS) model to predict how human perceives the difference between reference and distorted videos. WSNR[80], ADM[62], and VADM[61] belong to the difference-based category. These indices work reasonably well against simpler video quality databases developed before MCL-V[64]. The correlation coefficient of the best metric, VADM[61] is less than 0.75. However, these

metric takes various approaches to justify the quality degradation between reference and distorted videos. It would be interesting to see the strength of each metric.

Generally speaking, compressed video contains two types of visual artifacts, blurriness and blockiness, as illustrated in Fig. 4.1. High quality video has slight blurring but no obvious blocking. It may appear sharp to ordinary viewers, yet blurred edges can be found by experts. The artifact of medium quality video is visible to common assessors, yet the overall quality is still acceptable. Low quality video has the strongest blurring and blocking artifacts, which is unacceptable for broadcasting. The appearance of visual artifacts changes along the scale from high to low quality.

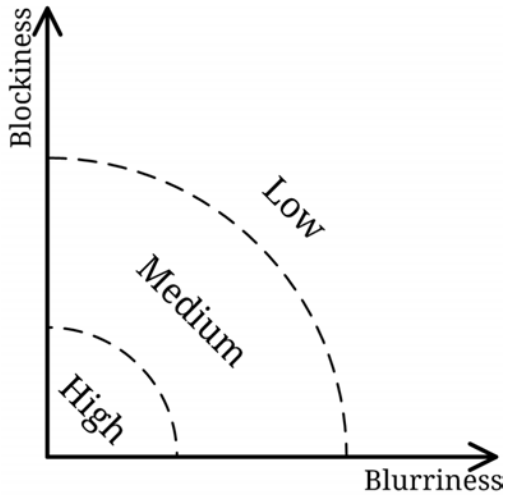


Figure 4.1: Illustration of video quality scale.

To understand the strength of different VQA indices in assessing streaming video quality, we test them with videos of different quality levels given in Fig. 4.1. Specifically, we split the 48 compression-only videos in MCL-V into three classes according the their mean opinion scores (MOS):

- Class-High: It is the high quality class whose video is slightly blurred with little blocking artifacts;

- Class-Medium: It is the medium quality class whose video has medium level blurring and blocking artifacts;

- Class-Low: It is the low quality class whose video has the strongest blurring and blocking among the three.

If a method performs well in Class-High, we claim that it performs well in detecting perceptual blurriness. We conduct experiments for videos in each class and compute the root mean squared error (RMSE) of predicted quality scores. It is worthwhile to point out that correlation coefficients may not provide an accurate result on their performance due to the small number of sequence numbers (i.e. 16) in each quality class. The ranking of several VQA methods (namely, SSIM, VIF, FSIM, and VADM) for each video quality class is shown in Table 4.1. We see that the best VQA indices with respect to Class-High, Class-Medium and Class-Low are SSIM, FSIM and VIF, respectively. Thus, if we can fuse them into a single index via learning, it may offer the best solution among all three individual indices.

Table 4.1: The ranking of quality methods regarding the quality classes.

Quality Class	Comparison of the Methods in RMSE		
High	SSIM	VIF	VADM
	0.565	0.599	0.617
Medium	FSIM	VIF	VADM
	0.683	0.684	0.738
Low	VIF	VADM	FSIM
	0.283	0.466	0.506

4.2.2 Learning-based Visual Quality Assessment Methods

There are two approaches to the design of an IQA/VQA index; namely, formula-based and learning-based. They are reviewed below. In the formula-based approach, a closed

form mathematical model is derived to predict perceptual quality, such as frame-based structural similarity (SSIM) [118], visual information fidelity (VIF) [105], feature similarity (FSIM) [130], and video additive impairments and detail losses measure (VADM) [61]. However, it is extremely difficult to provide a good mathematical HVS model to cover a wide range of video collections. To give an example, Li *et al.* [61] used Daly's contrast sensitivity function (CSF) [23] to improve the performance of VQA indices. However, there are more visual properties than contrast in the HVS, including luminance adaptation, visual saliency among others, the applicability of which is rather limited.

In learning-based approaches, a statistical model is built to model the relation between features of training image/video data and their mean opinion scores (MOS). Then, it is used to predict the quality of unseen test video. This approach has been used by researchers to design IQA indices in recent years. Narwaria and Lin [83] extracts the structural information in images with the singular value decomposition and then use the support vector regression (SVR) to map the feature to MOS. Liu *et al.* [73] proposed a multi-method fusion (MMF) IQA index, where a regression approach is used to combine scores of multiple IQA indices. The MMF score is obtained by a nonlinear fusion of scores computed by multiple methods with suitable weights obtained by a training process. The MMF index offers the state-of-the-art IQA results in several popular databases, including LIVE [106] and TID2008 [97].

The learning-based approaches have also been applied to the design of VQA indices. The fusion-based VQA (FVQA) technique was proposed in [65]. The FVQA method treats each video clip as a single data sample. In the training stage, it first classifies video clips based on their spatial and temporal complexities into several groups to reduce intra-group content diversity. Then, it learns the fusion rule of multiple VQA indices in each

group. In the testing stage, the FVQA index first classifies a test video clip into a group and then applies the fusion rule in that group for VQA score prediction.

Since the development of formula-based VQA indices is hindered by video content diversity and HVS complexity, we adopt a learning-based approach in this work. It is however important to emphasize that the number of training images in image quality databases [106, 97] is significantly larger than that in video quality databases. The accuracy of a statistical VQA model can be severely affected by the small size of the training data. This is the main issue to be addressed in our current work. Two exemplary video quality databases are used in our experiments. They are the LIVE database [103] and the MCL-V database [64]. The LIVE database contains 80 video clips of resolution 768×432 and with a duration of 10 seconds. They were coded by H.264 and MPEG-2. The MCL-V database contains 12 source video clips of resolution 1080p and with a duration of 6 seconds. There are 96 distorted video clips due to compression and scaling.

4.3 Proposed Fusion-based VQA (FVQA) Index

The block-diagram of the proposed fusion-based VQA (FVQA) method is illustrated in Fig. 4.2. It consists of two stages. In the first stage, reference videos are grouped according to their properties. In the second stage, several FR VQA methods are applied to the reference and distorted videos and their scores are fused to generate the final quality score. The performance of the FVQA index is evaluated by cross-validation. The details are described below.

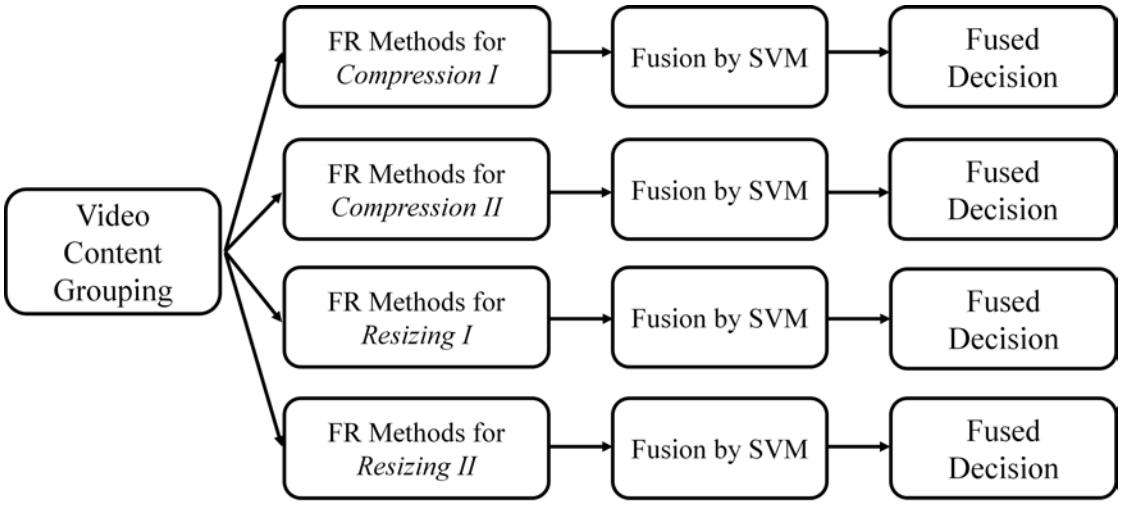


Figure 4.2: The block-diagram of the FVQA method.

4.3.1 Video Grouping

By classifying videos of similar content into a group, we can build a more accurate quality prediction model within each group. On the other hand, we want a sufficient number of samples in each group to allow the machine learning approach. Several features are chosen in [122] to characterize source images and videos along the color, space and time dimensions. For the VQA purpose, we choose the Spatial Information (SI) and the Temporal Information (TI) defined by the ITU-T Recommendation [47] to represent the spatio-temporal characteristics of source videos and utilize them to group video contents. Mathematically, they are expressed as

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\}\,, \tag{4.1}$$

where SI is calculated by applying the Sobel filter to the n^{th} video frame, F_n, and then taking the standard deviation over the space domain. Then, the maximum value along the time is chosen to present SI, and

$$TI = max_{time}\{std_{space}[M_n(i,j)]\}\,. \tag{4.2}$$

where $M_n(i,j)$ is the pixel difference between frames F_n and F_{n-1} located in the i^{th} row and j^{th} column. TI is computed as the maximum of the space standard deviation of $M_n(i,j)$ along the time axis.

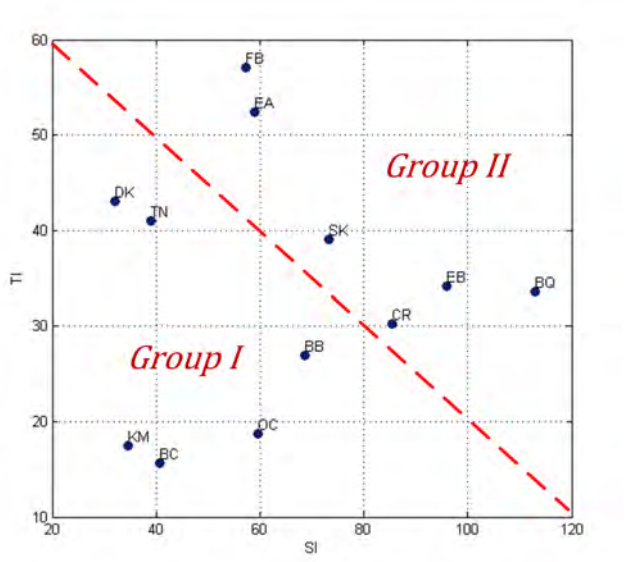


Figure 4.3: Grouping video contents based on SI and TI.

By plotting the SI and TI values of each MCL-V video sample in Fig. 4.3, we can divide all MCL-V source videos into two groups: Group I and Group II with the following separating line:

$$G_X = sign(a \cdot SI_X + b \cdot TI_X + c), \tag{4.3}$$

where X denotes a particular video sample, SI_X and TI_X are its SI and TI values. We choose $a = 1$, $b = 2$, and $c = -140$ in this work. Video X belongs to Group I and Group II if $G_X = -1$ and $G_X = 1$, respectively.

Moreover, there are two distortion types, compression and resizing, in MCL-V. Thus, we classify video samples into four groups: Group I/Compression, Group I/Resizing, Group II/Compression, Group II/Resizing eventually. In each group, the spatial-temporal characteristics of video samples are closer. Then, we will adopt a suitable machine learning technique for each group.

4.3.2 Learning Algorithm for FVQA

In this section, we consider the FVQA method that fuses scores from five VQA indices; namely, ADM, VIF, FSIM, PSNR and SNR. The supervised learning algorithm is adopted by FVQA in the determination of their weight coefficients. We compare two popular algorithms; namely, the Support Vector Machines [21] (SVM) and the Artificial Neural Network[38] (ANN), and choose the better one as the target training algorithm. The 4-fold cross-validation is applied to the MCL-V database for performance comparison.

For the SVM implementation, we adopted the software developed by Lin *et al.*[21]. We chose the Nu-SVR[102] for regression with two kernel functions - the Sigmoid kernel and the Radial Basis Function (RBF) kernel. The grid search [63] method was used to find the optimal pair of (C, γ), where C is the penalty cost in the training process and γ is the parameter in the kernel functions. For the ANN implementation, we considered two back-propagation training algorithms. They are the Resilient Propagation (RPROP)[99] and the Levenberg-Marquardt (LM) algorithm [39]. We tested a wide range of hidden node parameters to find the best ANN configuration.

Table 4.2: Comparison of Learning methods

	PCC	SROCC	RMSE
SVM (Sigmoid)	0.732	0.724	1.512
SVM (RBF)	0.901	0.891	0.992
ANN (RPROP)[99]	0.780	0.774	1.389
ANN (LM)[39]	0.877	0.877	1.067

The performance of four FVQA methods is given in Table 4.2 for comparison. As shown in this table, the SVM method with the RBF kernel offers the best performance among the four. Hence, it is adopted in the framework given in Fig. 4.2.

4.3.3 Selection of Contributing VQA Indices

We examined the FVQA method with five contributing VQA indices. However, a poor VQA index may contribute to the final fusion performance in a negative way. In this case, we want to remove it from the contributing set. Besides, by reducing the number of participating VQA methods, we can reduce the complexity of FVQA. To achieve this goal, we adopt the Sequential Forward Method Selection (SFMS) method as proposed in [73]. It is summarized in Algorithm 1. For each video group, we use SFMS to select two to three contributing VQA indices. The results are shown in Table 4.3.

Table 4.3: Selected VQA indices for 4 video groups.

Group	Selected VQA Indices
Group I/Compression	ADM and PSNR
Group II/Compression	ADM, VIF, and SNR
Group I/Resizing	FSIM and PSNR
Group II/Resizing	ADM, VIF, and SNR

Algorithm 1 Sequential Forward Method Selection (SFMS)

Require: M: The set of available quality metrics; M^*: The set of the selected metrics;
Ensure: Optimal M^* by minimizing RMSE;
1: Initial $M^* = \{\phi\}$; $J(\{\phi\}) = \infty$;
2: Sort M in ascending RMSE;
3: **while** $M \neq \{\phi\}$ **do**
4: Pick the next best method;
5: m = $\arg\min_{m \in M - M^*} J(M^* + m)$;
6: **if** $J(M^* + m) < J(M^*)$ **then**
7: Update Selection; $M^* = M^* + m$; $M = M - m$;
8: **else**
9: Break the while loop;
10: **end if**
11: **end while**
12: **return** M^*;

4.4 Proposed EVQA Index

Motivation and Overview. A video stream is composed by image frames, where frame-to-frame variation is usually small except for scene change. It is a commonly believed fact that perceptual quality remains stable within a short period of time. This property was exploited in [61, 117], where a spatial (or frame-level) quality index is first computed for each frame independently and the index scores across multiple consecutive frames can be weighted by a temporal pooling method. In this way, a VQA index can be constructed from frame-level IQA indices. To tackle the problem of limited training data, our proposed method adopts a frame-based learning mechanism, inspired by the same principle. Each source video clip in the MCL-V database lasts for 6 seconds and the frame rate is 30 frames per second (fps). The total number of frames for one sequence is 180 frames. In other words, each training video clip can offer 180 data samples, instead of just one. There is however a missing link in the aforementioned strategy; namely, the frame-level MOS score is not available during the training process. Since all video quality databases contain short and homogeneous video clips, it is assumed

that the MOS of the whole sequence can be used as an approximate ground truth of its frames. This assumption will be verified in Section 4.

The EVQA method consists of three steps in the training phase. Step 1: Feature Extraction. Several IQA methods are applied to each individual frame and their scores are stored as its feature vector. The raw scores of IQA indices are properly normalized to match the MOS value. Step 2: Frame Space Partitioning. The frame space is partitioned into several subspaces to enhance the learning performance. Step 3: IQA index Fusion. The fusion rule of combining multiple IQA indices into one single IQA score for a frame is learned in each partitioned frame subspace.

In the testing process, the EVQA method predicts the quality index of each frame by following the above steps. After that, all predicted frame scores are integrated to generate one MOS value for a short test video clip via *temporal pooling*. Since feature extraction is straightforward, we will elaborate on the following three topics below: frame space partitioning, IQA index fusion and temporal pooling.

Frame Space Partitioning. The main purpose of frame space partitioning is to allow more efficient learning rule in a smaller subspace, where frames share properties of higher similarity. This can be done based on spatial, temporal and quality/distortion properties of frames. The spatial and temporal complexities are related to the spatial and temporal masking effects of HVS. For the quality/distortion property, the predicted performance of an IQA index can be exploited. That is, each IQA index has its own strength in assessing some distortion types [73] and, if two frames can be well predicted by a common set of IQA indices, they must share certain similarity in their quality/distortion property.

Spatial and temporal complexities are computed based on the undistorted reference frames. The spatial information (SI) and the temporal information (TI) introduced in [47] are two well-known parameters for video sequences. However, they are not suitable

for our purpose since we are concerned with the properties of a single frame. Some modifications are needed, and we call extended SI (ESI) and extended TI (ETI) the modified metrics.

For the ESI, we first obtain the edge magnitude map G_M of frame F_n using the 3×3 Sobel filter. Then, the ESI for this frame is defined as

$$ESI_n = \frac{std[G_M(F_n)]}{mean[G_M(F_n)]}\,. \tag{4.4}$$

Complex scenes with a large amount of texture have a larger ESI value. For the ETI, the basic idea is to compute the pixel-based luminance difference of two adjacent frames. Sequences with large motion have large ETI values. Since the fine structure of the frame data, such as film noise, will have a negative impact on the accuracy of ETI, we apply the 5×5 Gaussian filter to their pixel difference, which is equivalent to taking the difference after we filter out each frame by the same Gaussian filter. Then, the ETI is defined as

$$ETI_n = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H} D_n(x,y)\,. \tag{4.5}$$

where W and H are the width and the height, correspondingly of the nth frame and

$$D_n(x,y) = |G * (F_n(x,y) - F_{n-1}(x,y))| \tag{4.6}$$

is the absolute value of the Gaussian-smoothed frame difference, and where G is the Gaussian filter.

Besides spatial and temporal complexities, it is desired to classify image frames based on their distortion type. However, it is difficult to obtain this information directly, yet it is possible to be obtained indirectly by analyzing its IQA scores. This analysis is conducted with respect to frames in the training set. Suppose that there are T training

frames. For a given IQA index, we can divide all training frames into C clusters of equal size $N = T/C$, based on its score distribution. We map raw IQA scores in one cluster, denoted by x, to a normalized score, Q, using a logistic function [130] as follows:

$$Q = \beta_1 \cdot (0.5 - \frac{1}{1 + e^{\beta_2(x-\beta_3)}}) + \beta_4 \cdot x + \beta_5, \tag{4.7}$$

where $\beta_i, i = 1...5$, are the fitting parameters determined by known IQA/MOS score pairs. Eq. (4.7) is used to convert an IQA score of an arbitrary range to a suitable range which is compatible with measured MOS values. After the score conversion, we can compute the root-mean-squared-error (RMSE), denoted by E, between Q and MOS in that cluster via

$$E(Q, MOS) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(Q_n - MOS_n)^2}, \tag{4.8}$$

where n is a frame index, MOS_n is its MOS value and Q_n is its transformed IQA index value. Furthermore, we can choose a threshold value to determine if an IQA method performs well in a cluster. For example, the RMSE values of 8 clusters are shown in Fig. 4.4. By setting the threshold value to $E = 1$, we see that the IQA index performs well for frames in Cluster Nos. 6-8 but poorly for frames in Cluster Nos. 1-5.

If an IQA index performs equally well (or poorly) for all clusters as shown in Fig. 4.5 (a), it cannot be used to partition the frame space. On the other hand, if it performs well for some clusters but poorly for other clusters as shown in Fig. 4.5 (b), we can use it to partition the frame space into two subspaces according to its preference - favored and unfavored subspaces.

Furthermore, we can use a sequence of IQA indices with preference to partition a frame space into multiple subspaces as illustrated in Fig. 4.6, where each split is defined by one IQA index. In this figure, the favored and unfavored subspaces of the first IQA index is denoted by A and A^c, respectively. Similarly, the frame space can be partitioned

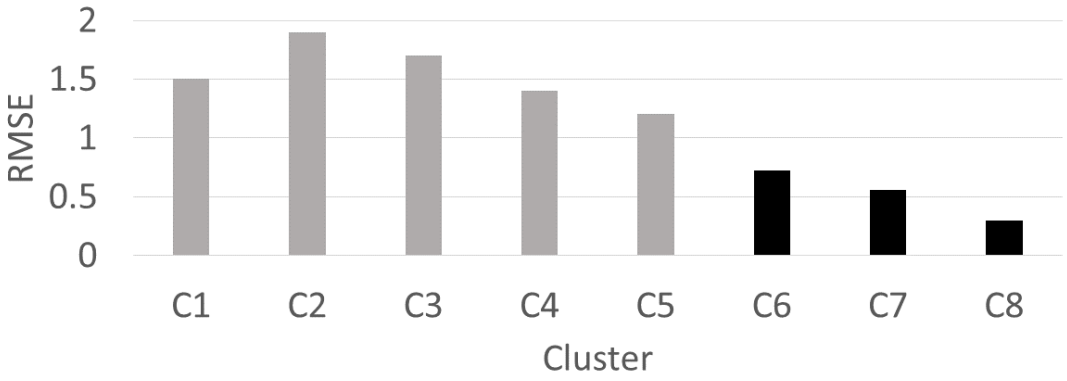
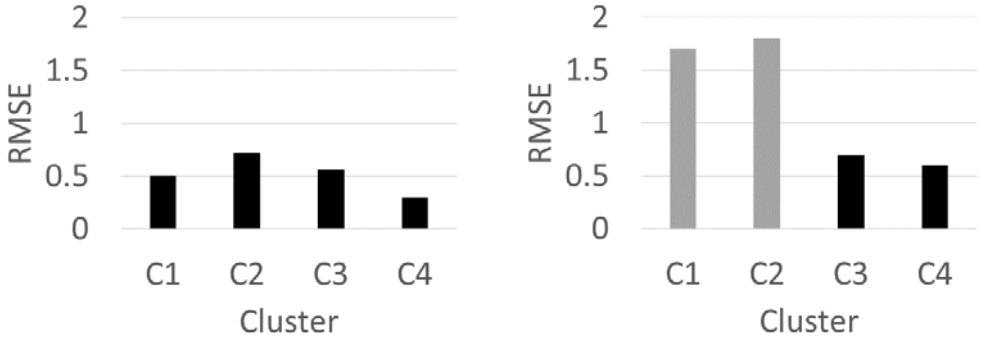


Figure 4.4: The plot of the RMSE of the predicted MOS values against the actual ones in 8 clusters for an exemplary IQA index, where a black and a gray bars indicate that its RMSE is lower and higher than a pre-selected threshold value, respectively.



(a) Index without preference (b) Index with preference

Figure 4.5: Illustration of an IQA index (a) without and (b) with preference.

by another IQA index into the favored and unfavored subspaces denoted by B and B^c, respectively. Then, the frame space can be decomposed into four subspaces as shown in the third stage of Fig. 4.6.

The frame space partitioning process can be organized as a binary tree as shown in Fig. 4.7. Each partition creates two children, and grows the tree to the next level. The partition should stop if the number of frames in a node is too small since each group should have a sufficient number frames for the learning purpose. On the other hand, for

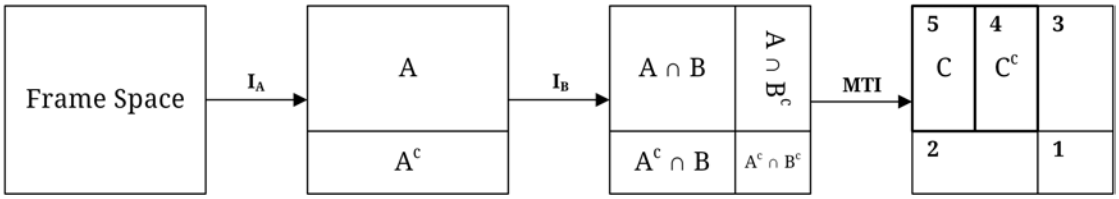


Figure 4.6: Frame space partitioning using multiple IQA indices with preference.

nodes that have a large number of frames, after we exhaust all IQA indices, we can use the frame′s ETI and ESI value to further partition them. Then, we can use frame's ETI and ESI to partition them.

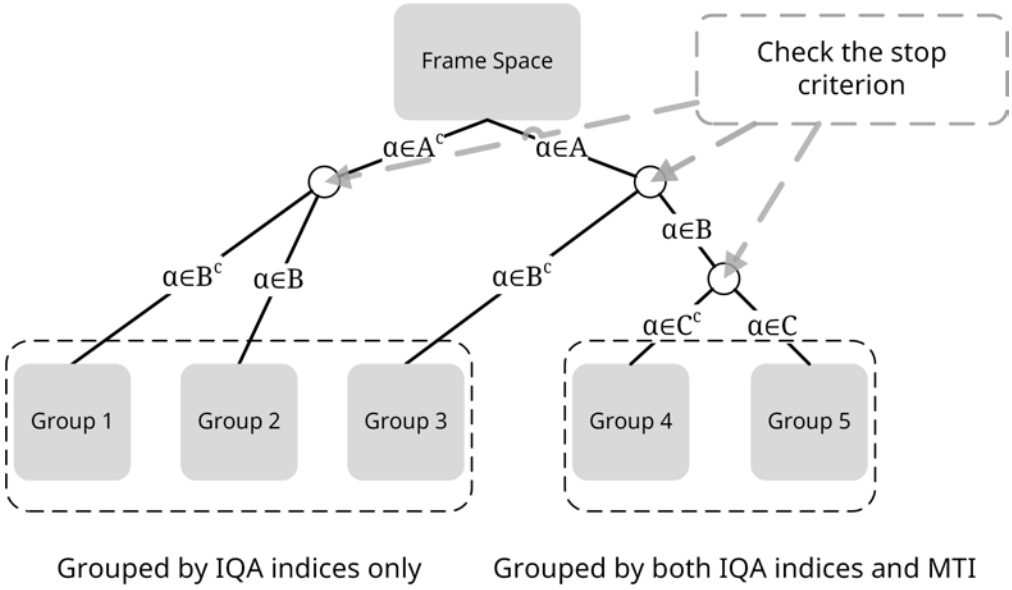


Figure 4.7: Illustration of frame space partitioning using a binary tree structure, where the stop criterion is checked at each node.

IQA index Fusion. In our experiment, six state-of-the-art IQA indices [62, 118, 105, 130, 119, 71] are included in the IQA candidate pool. For each partitioned frame subspace, we use the sequential forward method selection (SFMS) scheme [73, 65] to select a set of IQA indices to fuse so as to optimize an objective function such as the Pearson linear correction coefficient (PCC) value. The SFMS scheme is a greedy search

algorithm that selects the optimal IQA index in the candidate pool to yield a better prediction at each iteration. The iteration will terminate if the improvement becomes negligible. The SFMS scheme is determined by training data while the fusion rule is also learned through SVR from training data in each frame subspace.

Temporal Pooling. Temporal pooling is necessary to generate the final MOS value for the entire test video based on the predicted MOS value of each individual frame. Several temporal pooling methods such as the mean, median, Minkowski, percentile was studied and compared in [100, 104]. There is however no universal method that offers the best performance for all video contents. We adopt a simple average scheme here, which is justified by experimental results in Section 4.

4.5 Experimental Results

We present experimental results in two parts in this section. In the first part, we study the relationship between the frame-level and the sequence-level quality indices to justify two items: 1) the assumption that the sequence-level MOS can be used to approximate the frame-level MOS, and 2) the adoption of simple averaging as the temporal pooling method in EVQA.

Relationship between Frame-Level and Sequence-Level MOS Values. The frame-to-frame quality level is assumed to be stable for a short period if no scene change occurs. To verify this assumption, we plot the predicted frame-level MOS as a function of the frame index for the BC (Birds in Cage) sequence coded under ”good” quality in the MCL-V quality database in Fig. 4.8. We see that the predicted frame-level MOS is nearly constant.

The MCL-V video quality database consists of 12 sequences with five quality levels caused by different coding bitrates. We show the mean (μ) and the standard deviation

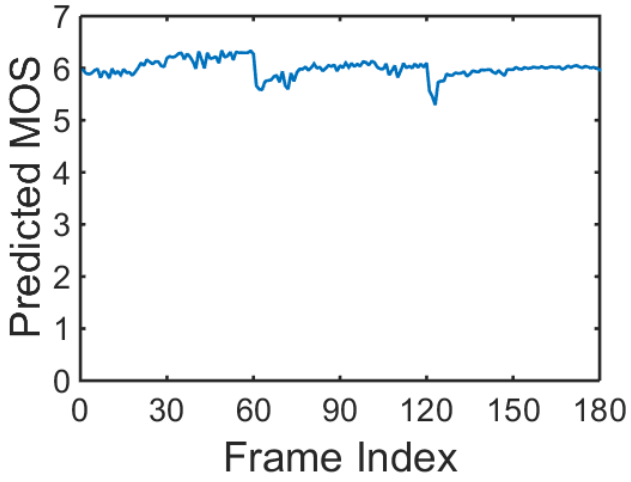


Figure 4.8: The predicted frame-level MOS value is plotted as a function of the frame index for the BC sequence coded under ”good” quality, where predicted sequence-level MOS is 6.81 by simple averaging while the true MOS value is 7.06.

(σ) of the predictive frame-level MOS values for all of them, in four quality levels (good, fair, poor and bad) in Table 4.4. The first column is the acronym for the title of each sequence. When the standard deviation value is low, it means that the frame-level MOS is nearly a constant. There are several sequences with larger standard deviation values such as DK (Dance Kiss), EA (El Fuente A), EB (El Fuente B), FB (Fox Bird) and TN (Tennis). These sequences were shot with more complex camera-object relative motion; thus they deviate slightly from the homogeneous frame-level MOS assumption.

To examine such a deviation in detail, we plot the predicted frame-level MOS as a function of the frame index for the DK sequence coded under ”good” quality in the MCL-V quality database in Fig. 4.9. We do observe the fluctuation of the predicted frame-level MOS values between frames 45-90 caused by camera-object relative motion. However, the predicted sequence-level MOS (6.48) is still close to its ground truth (6.63).

Performance Comparison of VQA Indices. The performance of the proposed EVQA method is evaluated on the MCL-V database [64] and the coding distortion of

Table 4.4: Mean and standard deviation of frame scores with compression distortion in MCL-V

Qual. Level	Good		Fair		Poor		Bad	
Seq. Title	μ	σ	μ	σ	μ	σ	μ	σ
BB	6.17	0.43	4.92	0.58	2.97	0.67	1.82	1.53
BC	6.81	0.09	5.79	0.59	2.69	0.59	0.71	0.26
BQ	6.88	0.19	6.42	0.52	3.58	0.55	1.80	0.64
CR	6.76	0.50	4.50	0.41	3.22	0.42	1.29	0.64
DK	6.48	1.05	5.42	1.88	2.74	1.65	0.42	1.72
EA	6.64	0.82	4.37	0.92	1.72	0.70	0.29	0.06
EB	4.92	0.92	3.48	1.18	1.52	0.68	1.76	0.79
KM	6.24	0.50	5.00	0.90	2.81	0.71	1.05	0.52
FB	6.00	0.72	3.57	1.08	1.37	0.89	1.43	1.26
OT	6.55	0.24	6.34	0.19	2.80	0.49	0.73	0.28
SK	6.75	0.08	5.34	0.52	3.35	0.42	0.75	0.36
TN	6.32	0.47	4.82	0.58	4.12	0.69	1.29	0.84

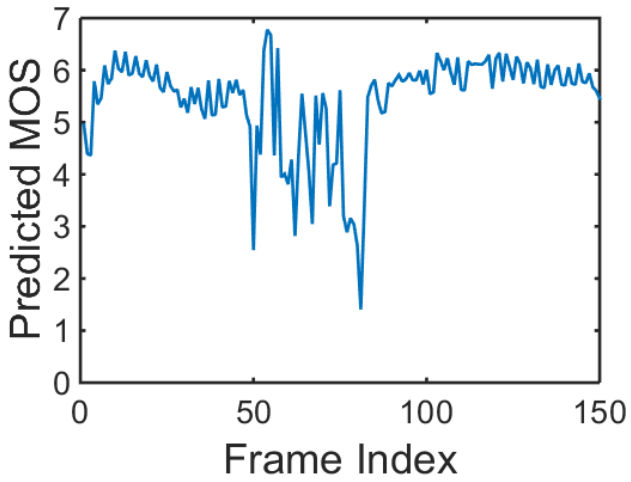


Figure 4.9: The predicted frame-level MOS value is plotted as a function of the frame index for the DK sequence coded under ”good” quality, where predicted sequence-level MOS is 6.48 by simple averaging while the true MOS value is 6.63.

the LIVE video database [103]. We follow the validation process proposed by VQEG in [115]. First, IQA index scores [62, 118, 105, 130, 80, 119, 126, 71] are mapped by the logistic function given in Eq. 4.7. Then, we consider three commonly used performance measures: 1) the Pearson correlation coefficient (PCC), 2) the Spearman rank-order correlation coefficient (SROCC), and 3) the root mean squared error (RMSE). PCC computes the correlation between the true and predicted MOS values. SROCC measures

prediction monotonicity. RMSE calculates the error between the true and predicted MOS values.

Table 4.5: Performance comparison of video quality indices for video clips in the LIVE video quality database with compression distortion (H.264 and MPEG-2).

	PCC	SROCC	RMSE
PSNR	0.478	0.449	9.034
VIF [105]	0.600	0.607	8.236
MSSIM [119]	0.591	0.692	8.294
FSIM [130]	0.634	0.698	7.955
GSM [71]	0.614	0.658	8.117
ST-MAD [117]	0.838	0.825	5.607
VADM [61]	0.847	0.850	5.469
EVQA	0.934	0.926	3.664

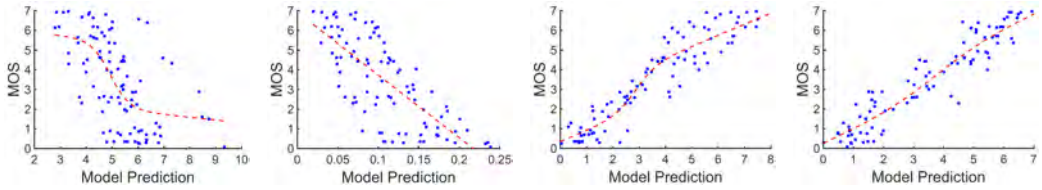
We adopt the 10-fold cross-validation strategy to select training and testing sets in the experiments. The performance of EVQA is compared with several benchmarking IQA and VQA indices. If an IQA index is used, its simple averaging is adopted to yield the final sequence-level MOS value. PCC, SROCC and RMSE results against the LIVE and the MCL-V databases are shown in Tables 4.5 and 4.6, respectively. Clearly, EVQA outperforms all other indices in every performance measure in both databases.

Table 4.6: Performance comparison of video quality indices for video clips in the MCL-V video quality database.

	PCC	SROCC	RMSE
PSNR	0.476	0.426	1.984
VIF [105]	0.660	0.655	1.666
MSSIM [119]	0.621	0.623	1.740
FSIM [130]	0.755	0.747	1.455
GSM [71]	0.709	0.711	1.565
ST-MAD [117]	0.634	0.623	1.714
VADM [61]	0.742	0.752	1.489
FVQA [65]	0.945	0.932	0.727
EVQA	0.956	0.947	0.652

Fig. 4.10 shows the scatter plots of four leading methods in Table 4.6, where each dot gives the predicted MOS value and the actual MOS value in its x-coordinate and y-coordinate, respectively, for each test sequence in the MCL-V database. The red dash

line indicates the optimal regression curve for these points. The ideal case is a straight line starting from zero along either the positive or the negative 45-degree direction with little deviation. The ST-MAD [117] regression curve is not straight while its data points are too spread out. The VADM [61] has a more straight regression line, yet its data points are still quite spread out. In contrast, data points in FVQA [65] and EVQA are much closer to their regression lines. Furthermore, the regression line of EVQA is more straight than that of FVQA.



(a) STMAD [117] (b) VADM [61] (c) FVQA [65] (d) EVQA

Figure 4.10: Scatter plots and their regression curves for all sequences in the MCL-V database using (a) ST-MAD, (b) VADM, (c) FVQA and (d) EVQA indices.

4.6 Conclusion

A novel FR VQA index, called EVA, was proposed to assess compressed and resized video quality in the work. EVA classifies video contents into groups and adopts a machine learning technique to fuse several known VQA indices to predict the MOS value within each group. The effectiveness of EVA was demonstrated using the MCL-V database. We would like to extend the current work in two directions. First, it is interesting to improve the proposed grouping technique by considering other HVS attributes such as luminance and contrasts. Second, it is desired to develop no-reference VQA based on the same fusion idea.

Chapter 5

JND-based Visual Quality Assessment

5.1 Introduction

The mean-squared-error (MSE) has been widely used in measuring the distortion of compressed image and video content. It offers a continuous-scale distortion measure as a function of coding parameters, such as the bitrate, the quality parameter (QP) in video coding or the quantization factor (QF) in image coding, since it is computed based on pixel differences. Recent efforts have been made to develop new objective quality/distortion metrics that are better correlated with the human visual system (HVS) [69]. All of them provide a quality measure of continuous-scale. However, it is well known that the HVS cannot perceive small changes in pixel differences. In reality, humans cannot perceive continuous-scale but discrete-scale quality changes over a range of coding bitrates. In this work, we quantify this phenomenon based on the notion of just-noticeable difference (JND) [50], and propose a new methodology to characterize the human visual experience on coded visual content.

JND is a statistical quantity that accounts for the maximum difference unnoticeable to a human being. It has been extensively studied to understand human visual sensitivity [66]. In the context of image/video compression, Watson [121] proposed a way to use JND for video quality measurement. His work adopts "pair comparison" or "two-alternative forced-choice". That is, subjects are asked to determine which of two videos (*i.e.* the original source and the compressed one) is more distorted. Then, the distortion level of compressed video can be derived from the JND test result. Although Watson's

pioneering work offers a statistical relationship between visual quality assessment and JND, his result is difficult to apply to a real-world video coding system. Furthermore, the test duration required for each subject is long. Recently, several JND estimators based on HVS properties were investigated in [51, 123]. It was also shown in [67, 77, 131] that JND-guided coding schemes can achieve perceptually similar quality with lower bitrates.

In this chapter, we study the problem of coded image/video quality assessment from a new angle based on JND. Given coded image/video content with densely sampled QF or QP values, we would like to measure the number of JND points and locations, and analyze the inter-person variance of measured quantities. One major contribution of this work is to develop a new methodology to achieve these tasks. It is confirmed by a small-scale subjective test that human perceived quality of coded images can be characterized by a piecewise constant function of the QF/QP with discontinuities at JND locations. Although these locations are content-dependent and statistically distributed, they do provide consistent and useful information in understanding the human visual experience.

The rest of this paper is organized as follows. The JND-based quality assessment problem is defined and its solution methodology is described in Section 5.3. A brief review of JND related work is given in Section 5.2. We start from a pilot study of JND test. The process of JND data collection is presented and its statistics are given in Section 5.3. The JND data post-processing technique and the final output quality plot as a function of the QF/QP are detailed in Section 5.3.4. Next, we extend the techniques to develop a JND dataset, called MCL-JCI. The number reference images is increased from 5 (in the pilot study) to 50 (in the MCL-JCI) such that the MCL-JCI enables us to study how content affects JND properties. The details of MCL-JCI is given in Section 5.4. Finally, concluding remarks are given in Section 5.5.

5.2 Background Review

JND is essentially the visibility threshold of perceptual change above which can be perceived by the HVS[50]. The earliest study can be traced to Weber-Fechner law in 1800s, which states that the JND of the test and background signal is proportional to the intensity of background signal. Lin [66] represents the general relation of the test signal x_t, the original or background signal x_o and the visual stimulus t as

$$x_t = x_o + t. \tag{5.1}$$

According to Weber-Fechner law, if x_o is a uniform image and t is a luminance variation, the difference of x_t and x_o can be perceived only when t is greater than the JND. $CSF1999$ studied the contrast JND of image quality. When x_o is a uniform image but t is with varying contrast the JND depends on contrast sensitivity function of t.

Determining JND is a challenging task in general because of the complex nature of the HVS. The process of investigating JND is related to the HVS characteristics and underlying human brain activities such as sensation, perception and recognition. Moreover, personal experience and preference play an important role in the process. Experienced video experts, for example, can find compression artifacts more easily than ordinary ones. In this paper, we focus on the image compression system. The subjective test [121] has been studied to determine the JND in a compressed image x_d against the original image x_o, and the test image x_t for such experiments can be expressed as

$$x_t = x_o + h(x_d - x_o), \tag{5.2}$$

where h is a scaling factor ($0 < h < 1$) to be adjusted in the tests. x_d is a decompressed image of x_o. In the tests, observers compare x_o and x_t with an increasing h, and a JND

can be determined with an h value when 75% of observers are able to distinguish x_t from x_o. If the distorted image x_t corresponding to a JND is denoted as x_1, a 2-JND difference can be decided when x_o is substituted with x_1 in Eq. 5.2, i.e., x_1 is regarded as the original image. The differences with 3-JND, 5-JND, and so on, can be determined with a similar process. In real-world coding system, x_d is controlled by bitrates rather than scaling $(x_d - x_o)$. Therefore, the results of [121] is difficult to apply to coding systems. The tests have not been primarily designed to facilitate mathematical JND modeling since the JND determined in this way is highly contextual to the contents of the image under test. However, such JND maps can be utilized to calibrate a perceptual distortion metric that has been developed [76].

Knowledge on JND is useful to evaluate the HVS tolerance for the compression distortion. For visual quality/distortion prediction, a metric can be defined or fine-tuned according to JND [120, 51, 125, 123] for better matching the HVS perception. It was also shown in [67, 131, 77] that JND-guided coding schemes can achieve perceptually equivalent quality with lower bitrates. JND has been used to determine not only the noticeable visual distortion but also the possibly noticeable visual quality enhancement[70, 51]. According to the operating domains, we can divide JND models into two basic categories: subband-based models and pixel-based models. The former category has been relatively well investigated with DCT decomposition [120, 112, 41, 66, 77], because of the popularity of DCT-based coders for compression. However, the latter category [127, 67, 131] is more convenient to be used in some situations (e.g., in motion estimation and residue manipulations[127] or quality/distortion evaluation with decoded signal[70, 51]). The work of both categories are developed based on psychological model of low-level HVS characteristics such as edge, contrast or luminance. The subjective tests in these research are used to examine the methods.

The high- level or complex HVS features such as saliency and masking effects are not well-studied due to the lack of JND ground-truth.

5.3 Problem Statement and Solution Methodology

5.3.1 Problem Formulation

To conduct the JND test on coded images, the first step is to collect a proper set of diverse, high-resolution images. Diversity can be can be characterized by their spatial complexity index (called the "Spatial Information" index and denoted by SI), colorfulness index (denoted by CF) [122], and semantic properties (e.g. a human object, etc.). We select five source images in our preliminary test as shown in Fig. 5.1. The SI index increases (or become more complext) from DB, RP, FT, CC to HS while the CF index increases (or become more colorful) from HS, RP, DB, CC to FT.

(a) CC (b) DB (c) FT (d) HS (e) RP

Figure 5.1: Five images selected for the JND test: (a) Color Checker (CC) [2], (b) Dark Building (DB) [2], (c) Food Truck (FT) [1], (d) Houses (HS) [2], and (e) Railway Platform (RP) [2].

Each source image is encoded by the JPEG encoder [46] with densely sampled bitrates. In the experiment, we encode each image 100 times with $QF = 1, 2, \cdots, 99, 100$, where QF controls the scaling factor (denoted by S) via

$$S = \begin{cases} \frac{5000}{QF}, & \text{if } QF > 50 \\ 200 - 2QF. & \text{otherwise} \end{cases} \tag{5.3}$$

The default quantization table (denoted by T) and the matrix of quantization step (denoted by Q) is related by the scaling factor S as

$$Q(i,j) = \frac{T(i,j) \times S + 50}{100}, \tag{5.4}$$

where $Q(i,j)$ and $T(i,j)$ are the $(i,j)^{th}$ entry of Q and T, respectively. Therefore, a larger QF value results in a smaller quantization step size and, hence, a higher coding bitrate. Since the quality of a coded image is a monotonically increasing function of its coding bitrate, a higher QF will not yield a lower quality image. There are 101 images in total for each source by including the original one. It is expected that humans are not able to differentiate all 101 levels. We would like to ask the following three questions.

- Q #1: How many quality levels can a person discern?
- Q #2: What are the JND locations in the perceived quality level versus QF plot?
- Q #3: Are the above two results stable among multiple subjects?

In the following section, we attempt to develop a methodology to measure the quantities raised in Questions #1 and #2. Then, we would like to analyze the obtained results so as to answer Question #3.

5.3.2 Solution Methodology

We represent N coded images with N nodes and arrange them in a bitrate ascending order from the left to the right. Under the assumption that quality is a monotonically increasing piecewise constant function of the bitrate, we would like to find JND locations. This can be done efficiently using the bisection search algorithm by starting from the two ends of the entire interval (denoted by nodes a and b) as shown in the top case of Fig. 5.2a. If images at the two end nodes exhibit noticeable difference (ND), we

expect that at least one JND point exists in that interval and will check the middle position (or the closest rounded integer) of the interval. In this example, the quality at node $(a+b)/2$ is not distinguishable from that at node b. Then, the search terminates for the interval $[(a+b)/2, b]$. However, there still exists notice difference between nodes a and $(a+b)/2$, so that we will continue the bisection search in the corresponding interval. This process is repeated until the finest resolution is reached. Then, we can find the JND location between two adjacent nodes in the finest level.

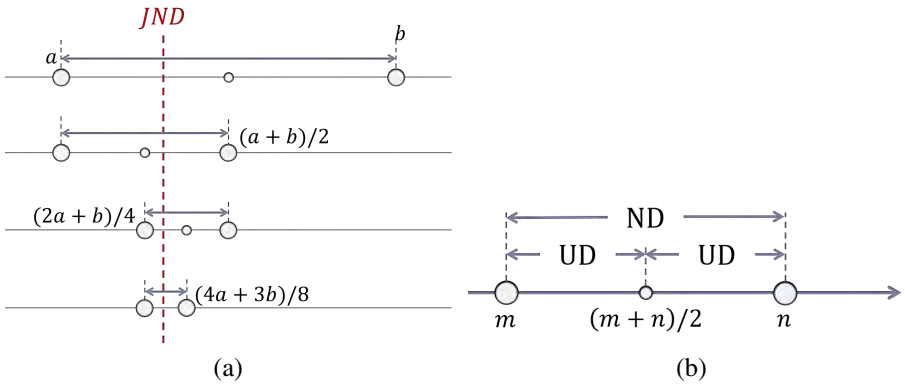


Figure 5.2: Illustrations of (a) the bisection search process to determine the JND location in the finest level and (b) the early termination condition of a bisection search process.

For some cases, the process is terminated earlier before hitting the finest level. That is, the JND location may appear between nodes at a coarser level. An example is shown in Fig. 5.2b. We observe noticeable differences (ND) in image quality at nodes m and n. However, there is unnoticeable difference (UD) in visual quality between nodes m and $(m+n)/2$ and between nodes $(m+n)/2$ and n. This happens when the quality levels at nodes m and n are perceivable but very close. Then, one cannot tell the differences furthermore by going to the next finer level.

5.3.3 Subjective Test and Data Analysis

In the preliminary study, we invited 20 subjects to attend the subjective test with five source images of resolution 1920×1080 and coded by a JPEG encoder with integer QF values of $1, 2, \cdots, 100$. Among all subjects, only eight of them had previous experience in image coding. They were seated in a controlled environment. The viewing distance was 2 meters (1.6 times of the picture height) from the center of the monitor to the seat. The image pair was displayed on a 65" TV with native resolution of 3840×2160. A subject compared two images displayed side by side and determined whether these two images are noticeably different (ND) or unnoticeably different (UD). The JND location can be identified by tracking the transitional position from ND to UD. The bisection search methodology described in Section 5.3 was adopted in the test procedure for speed-up. Due to the efficiency of this search algorithm, the test duration of each subject for five images was about 15 minutes.

The JND points in the domain of QF values are collected for each subject. First, we want to find the total number of JND points for a given source, which in turn determines the number of perceived JND levels. That is, the number of levels equals the number of points plus one. The mean and the standard deviation of JND points for each source are given in Table 5.1. Furthermore, the corresponding box plot is shown in Fig. 5.3 (a), where the bottom and the top of each box indicate the 25th and 75th percentiles of the samples, respectively, and the middle line is the mean value. We see from the table and the figure that the number of JND points does not vary significantly among subjects. This is especially true by focusing on the interquartile range (i.e. between the tops and bottoms).

Furthermore, we show the box plots of the highest and lowest JND locations in Figs. 5.3 (b) and (c), respectively. By comparing Figs. 5.3 (b) and (c), we see that there is correlation between the number of JND points and the highest JND location. That is,

Table 5.1: Statistics of the number of JND points for five test images.

	Mean	Stdev
CC	5.20	1.47
DB	3.75	1.41
FT	7.05	1.63
HS	6.00	1.62
RP	5.00	1.41

if the highest JND point occurs at a higher QF, the subjects will see more JND levels. For the lowest JND locations, since the images are heavily distorted and can be easily detected, the opinions are close among subjects. In contrast, the highest JND locations depend on the subject's experience. Trained subjects can find differences at higher QF more easily than inexperienced ones. Thus, as compared with Fig. 5.3 (b), the ranges of the highest JND points from experienced subjects are significantly narrower as shown in Fig. 5.3 (d).

5.3.4 JND-based Quality Level Plot

Since some subjects may have more JND points than others, we need to normalize the JND values across different subjects. The normalization process is straightforward If a subject has K JND points, we give a weight of K^{-1} to each of his/her selected JND points. After normalization, we can aggregate the JND data from all subjects and plot the corresponding histogram. The result for the Color Checker (CC) image is shown in Fig. 5.4(a). We see that there is rarely any JND point for QF greater than 50. The quality of these images is high enough that they are not differentiable from the original. The same observation applies to other test images. Thus, we will focus on the interval where QF ranges from 1 to 50.

Next, we describe a procedure to derive the JND level plot from the histogram plot. Since JND is a random variable, we need to account for its statistical variation. We

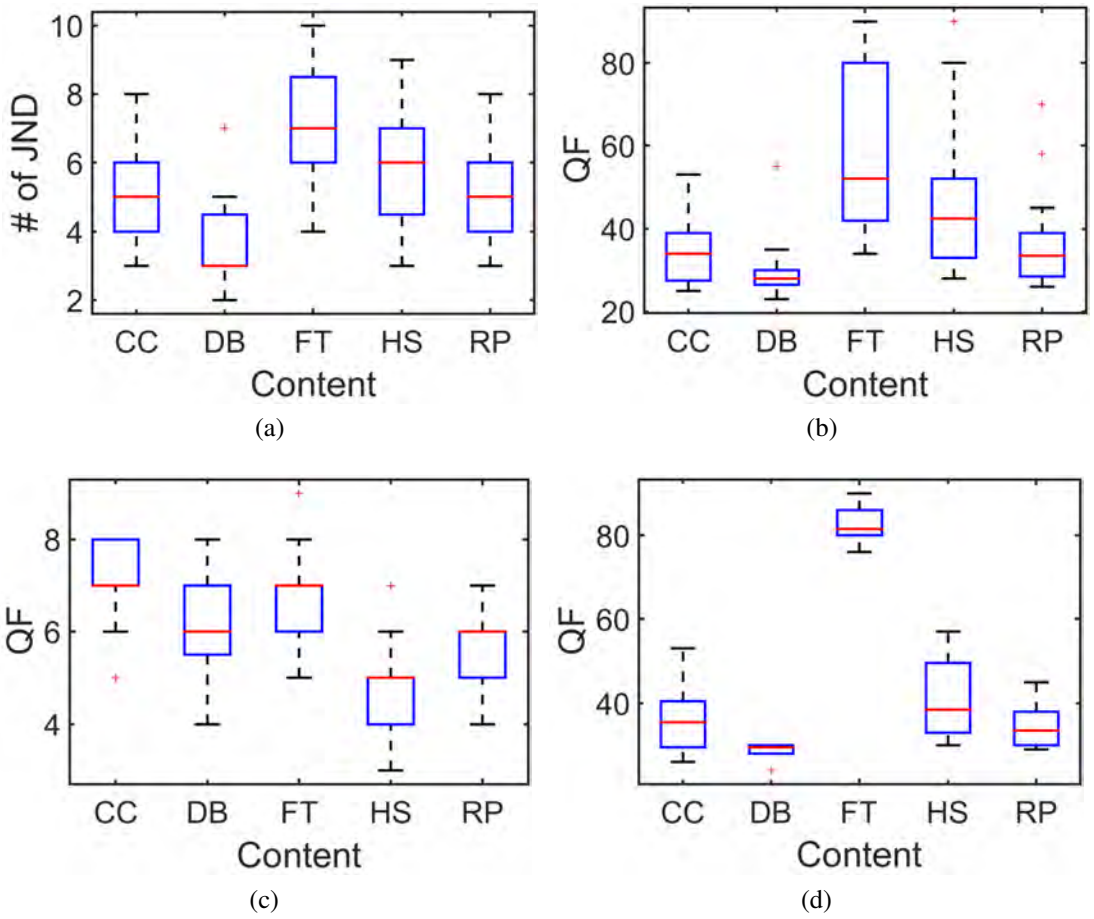


Figure 5.3: The box plots of (a) JND numbers, (b) the highest JND locations of all subjects, (c) the lowest JND locations of all subjects, and (d) the highest JND locations from experienced subjects only.

use the rounded mean number in Table 5.1 as the target number of JND points. For example, we choose 5 JND points for the CC image. To cluster the JND histogram into 5 sub-intervals, we run the k-means algorithm with $k = 5$ on the JND histogram and partition the whole QF interval into 5 sub-intervals. The JND location bars in these 5 sub-intervals are shown in 5 different colors in Fig. 5.4(a). After that, we select a high peak that is close to the mean of each sub-interval as the desired JND point. The selected JND location bar is labeled by a circle on its top. The reason of selecting the high peak

rather than the mean value as the JND location is that we would like to align it with the JND experimental data.

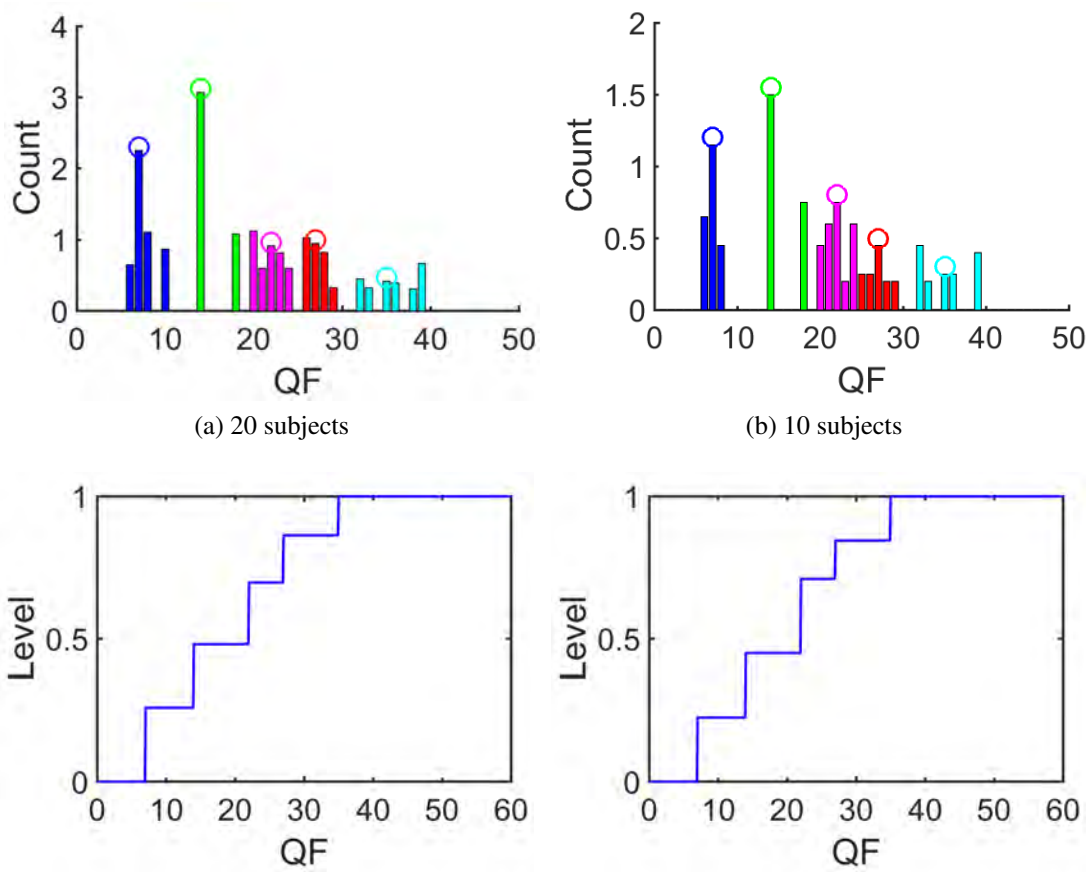
To demonstrate the robustness of our experimental design and the post-processing technique described above, we aggregate the JND data from those subjects whose JND numbers are in the interquartile range (with 10 subjects in total) to obtain the corresponding JND histogram plot and the JND level plot as shown in Figs. 5.4(c) and (d), respectively. The JND level plots are very similar for the two cases with 10 and 20 subjects. Finally, the JND level plot for the other four test images with all 20 subjects are shown in Fig. 5.5.

5.4 JND-based Coded Image Quality Dataset

5.4.1 Data Collection and Processing

From the preliminary study, we developed the method of collecting and generating statistical JND data from subjective evaluations. Next, we aim to develop a dataset with rich content features. Therefore, we selected 50 images (as shown in Fig. 5.6) with the resolution 1920×1080.

The diversity of selection can be characterized by their spatial complexity index (called the "Spatial Information" index and denoted by SI), colorfulness index (denoted by CF) [122], and semantic properties (e.g. a human object, etc.). SI is calculated by applying the Sobel filter to image, taking the standard deviation over the space domain, and then the maximum value is chosen. Fig. 5.7 shows SI and CF to all selected images. From Fig. 5.7 , we see that our selected images cover a wide range of diversity. Moreover, the source images can be also further classified into different categories according to its content as summarized in Table 5.2.



(a) 20 subjects

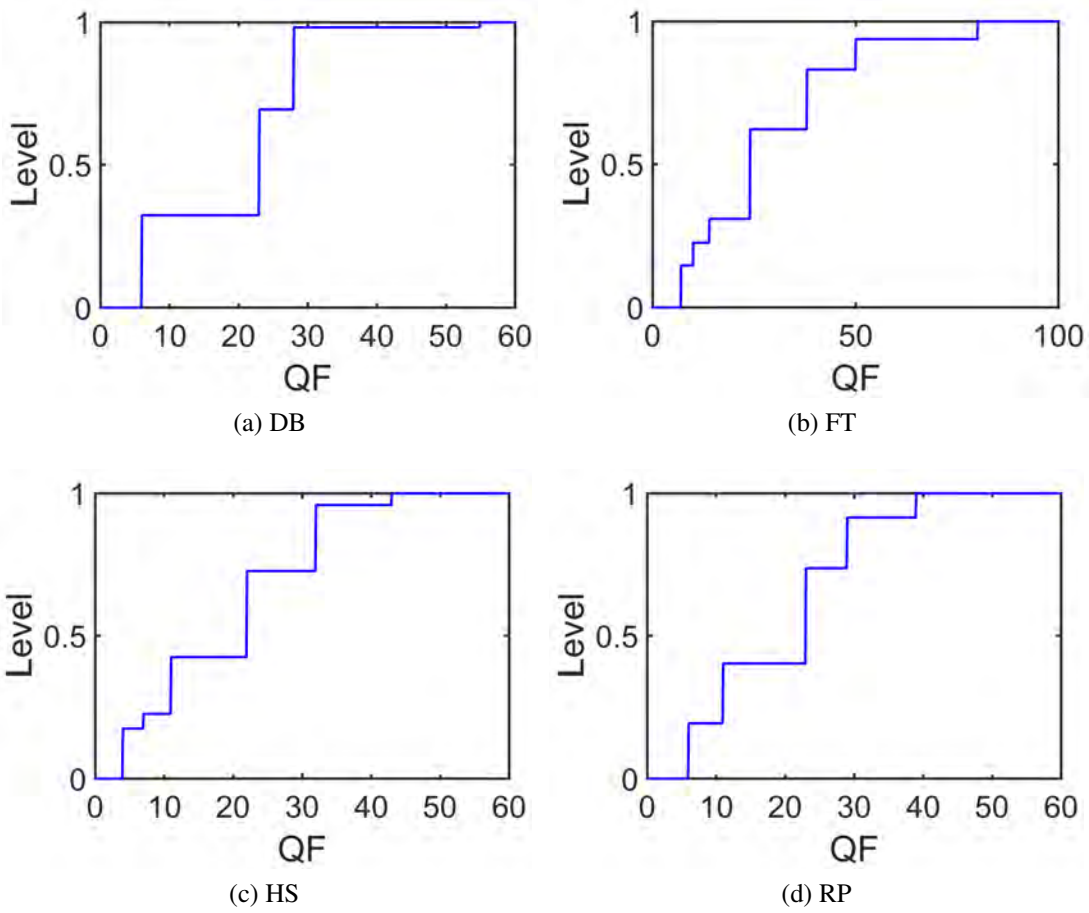
(b) 10 subjects

(c) 20 subjects

(d) 10 subjects

Figure 5.4: Comparison of the JND location histogram plots in (a) and (b) and the JND level plots in (c) and (d), where (a) and (c) are based on all 20 subjects and (b) and (d) are based on the 10 subjects whose JND numbers are in the interquartile range.

More than 50 volunteers participated in the subjective test, with equally stratified by gender and by age between 20 and 40 years old. 10 out of the subjects are experts in technical implementation in quality assessment or image compression. The rest of them have little or no prior experience of quality evaluation experiments. They were seated in a controlled environment. The viewing distance was 2 meters (1.6 times of the picture height) from the center of the monitor to the seat. The image pair was displayed on a 65" TV with native resolution of 3840×2160. A subject compared two images displayed side by side and determined whether these two images are noticeably different



(a) DB (b) FT

(c) HS (d) RP

Figure 5.5: The JND level versus the QF plot for (a) DB, (b) FT, (c) HS and (d) RP.

(ND) or unnoticeably different (UD). The JND location can be identified by tracking the transitional position from ND to UD. The bisection search methodology described in Section 5.3 was adopted in the test procedure for speed-up. Due to the efficiency of this search algorithm, the test duration of each subject was about 45 minutes.

The JND points in the domain of QF values are collected for each subject. First, we want to find the total number of JND points for a given source image, which in turn determines the number of perceived JND levels. That is, the number of levels equals the number of points plus one. The mean and the standard deviation of JND points for each source are given in Fig. 5.8. Furthermore, the corresponding box plot is shown in Fig.

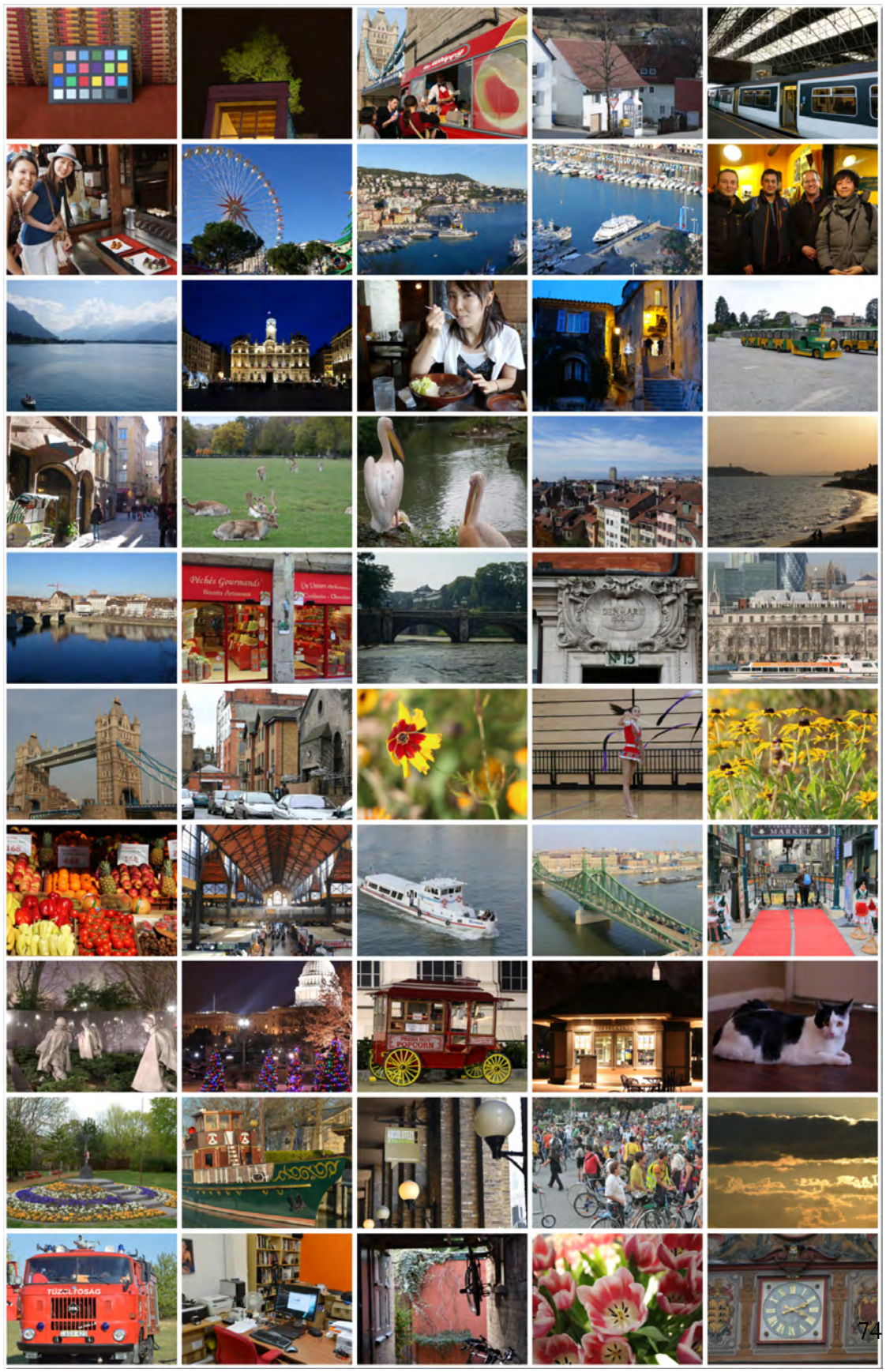


Figure 5.6: 50 reference images in the MCL-JCI Dataset.

Table 5.2: Count of Reference Images in MCL-JCI.

	number
People	5
Dark Scene	6
Animals	3
Plants	4
Building	8
Water or Lake	5
Sky	3
Bridge	3
Boat or Cars	5
Indoor	8

5.9, where the bottom and the top of each box indicate the 25th and 75th percentiles of the samples, respectively, and the middle line is the mean value. We see from the table and the figure that the number of JND points does not vary significantly among subjects. This is especially true by focusing on the interquartile range (i.e. between the tops and bottoms). Most of images have 4-6 JND points . We also show the average JNDs of the highest and lowest JND locations in Figs. 5.10, which are marked in red and blue,

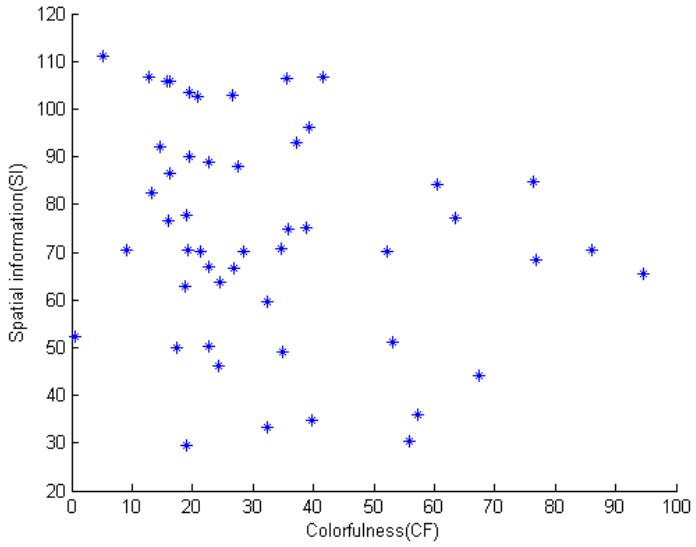


Figure 5.7: Reference image spatial information and colorfulness.

respectively. We see that there is rarely any JND point for QF greater than 60. The quality of these images is high enough that they are not differentiable from the original. The same observation applies to other test images. Thus, we will focus on the interval where QF ranges from 1 to 60.

By comparing average number of JND points in Fig. 5.8 and the highest and lowest JNDs in Figs. 5.10, we see that there is correlation between the number of JND points and the highest JND location. That is, if the highest JND point occurs at a higher QF, the subjects will see more JND levels. For the lowest JND locations, since the images are heavily distorted and can be easily detected, the opinions are close among subjects.

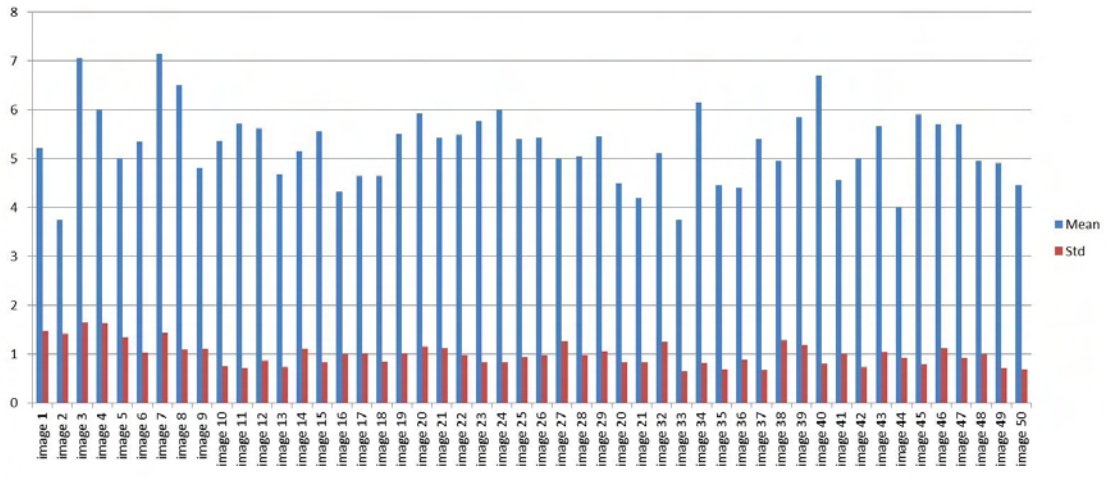


Figure 5.8: Statistics of the number of JND points for all images in MCL-JCI.

The data-processing techniques introduced in Section 5.3 are applied to the dataset. The JND level plot for all 50 reference images in 5.6 with all 20 subjects are shown in Fig. 5.11.

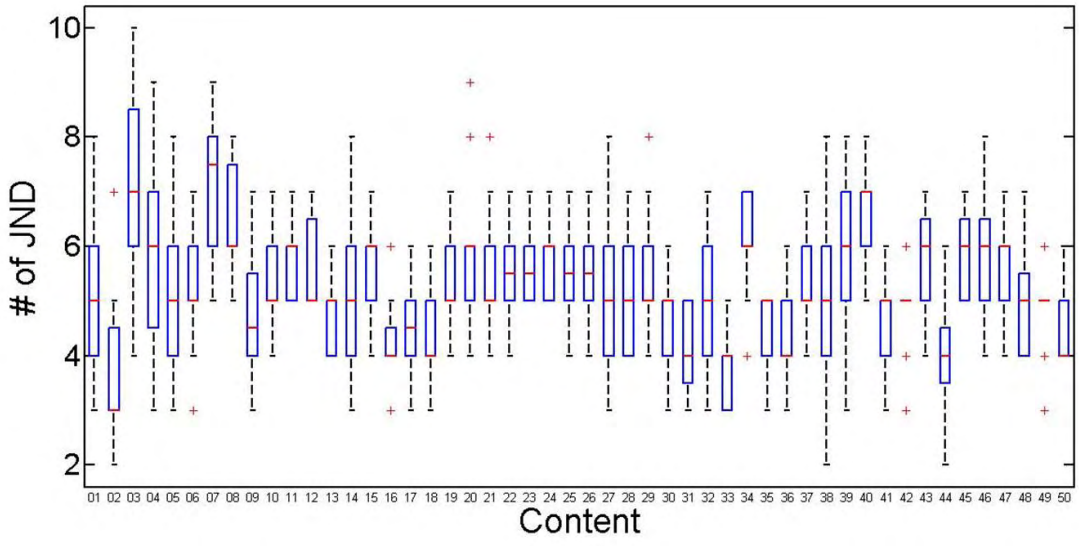


Figure 5.9: The box plot of the number of JND points for all images in MCL-JCI.

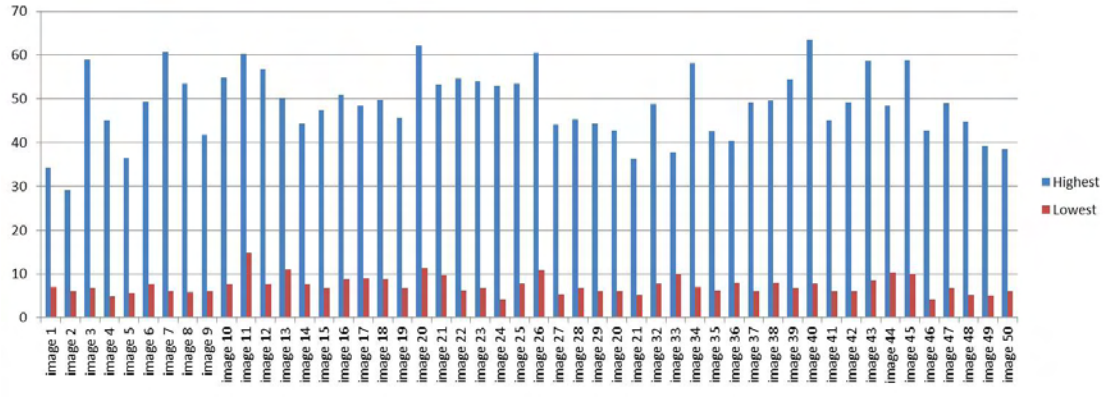


Figure 5.10: Statistics of JND locations of the highest and lowest acceptable quality for all images in MCL-JCI.

5.4.2 Relationship between Contents and JND-based Quality Levels

With the JND data, it is a desire to know which and how image feature influence to the visual quality. We compare the images from different JND levels and figure out which part is change and what is the trending.

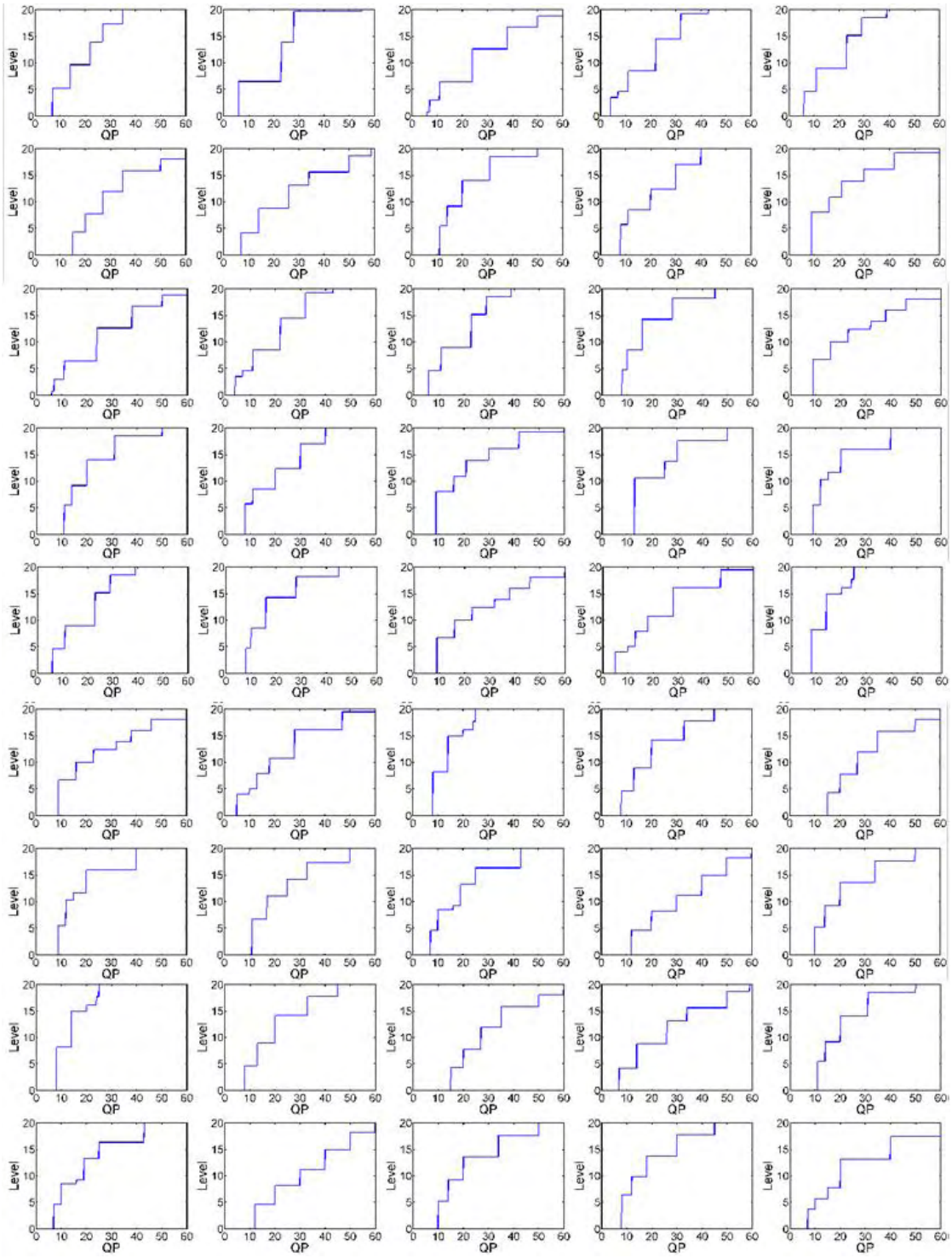


Figure 5.11: Final output JND points and location to all reference image corresponding in Fig.5.6.

Luminance. The experiment results show that the 6 dark sources in Fig. 5.12 have less JND points are than other bright images. However, if the dark image contains homogeneous information with chrome change, people could also see more different quality levels. The phenomenon can be explained by Weber-Fechner law. Since the image is in low luminance, the background creates higher bar for the HVS to trigger stimuli of the difference. However, quantifying the luminance is still an open problem.

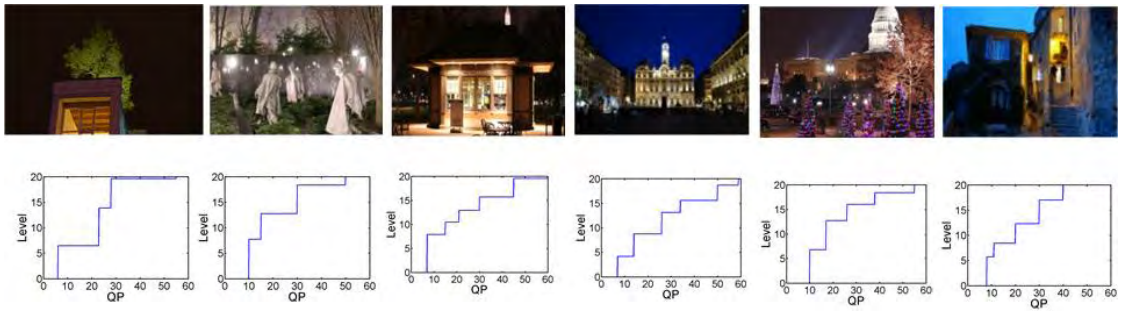


Figure 5.12: Statistics of the JND location to 6 dark sources in MCL-JCI.

Texture. The JPEG compression adopts block-based quantization of the transform coefficients. Therefore, the high frequency components are reduced after JPEG compression. For textureless or flat area, if the color is gradually changed in the block, the compression distortion is significant. In contrast, if the block is with rich texture, the distortion is not obvious. Fig. 5.13 shows the compressed image that is one JND to the original one. We can see that most details are preserved after compression. Fig. 5.14 shows the compressed image with six JND to the original one. We select two patches, "Sky" and "Tree", as our example to show how texture affect JND. We calculate the spectrums and show them next to the patches. In the Sky region, the spectrum looks less degradation but the visual quality is much worse than the "Tree" region.

When we look into the distortion, the fake contours in the "Sky" draw our attention. In contrast, the fuzziness on the edges in the "Tree" is not obvious. Therefore, the texture on the background influences the thresholds of JND. If the texture of certain regions is

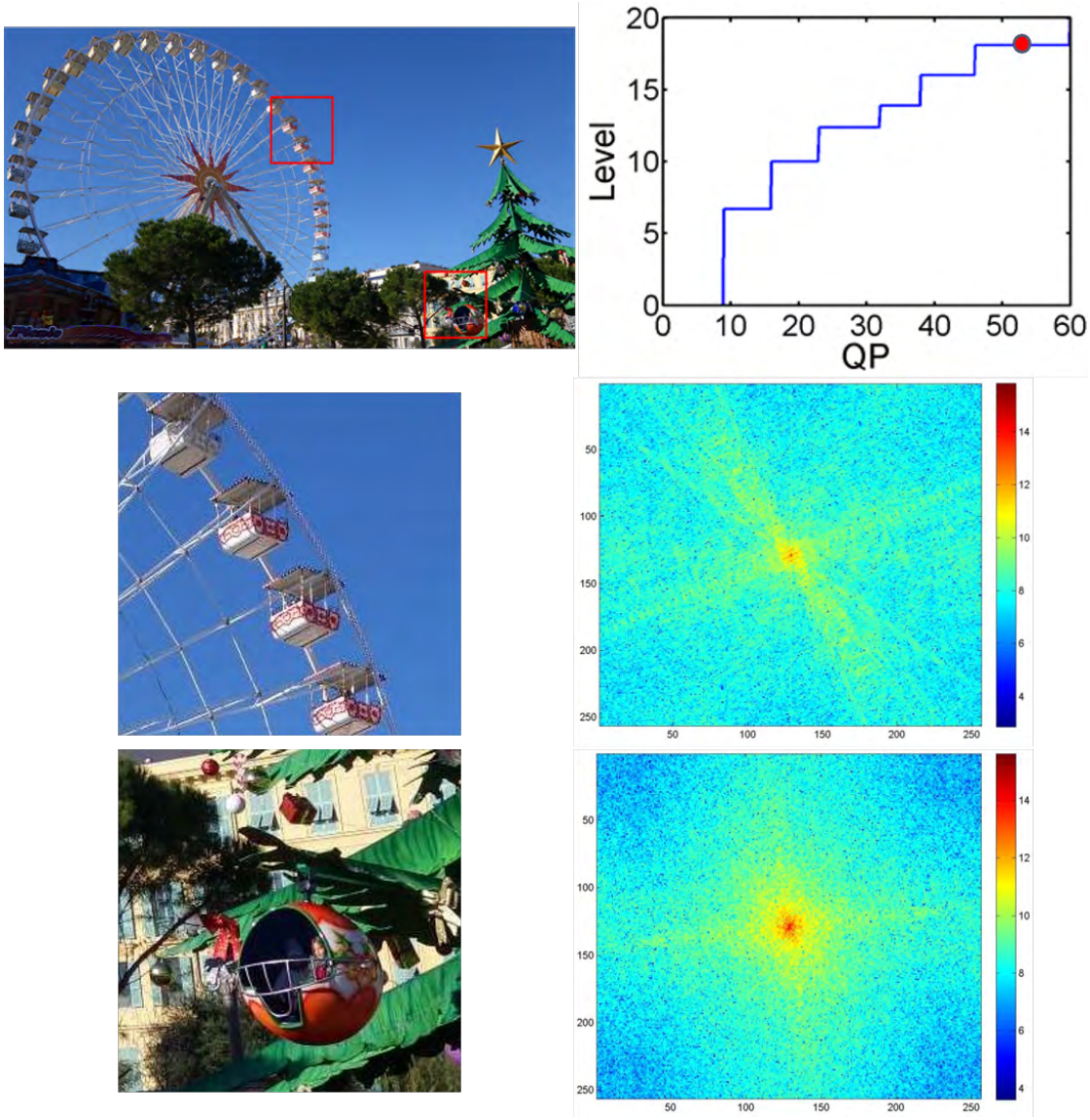


Figure 5.13: Compressed image with one JND level to the original Source 8.

homogeneous, the threshold is lower than the ones with rich texture. If the whole image is filled with rich texture, the viewers may perceive less JND levels due to the higher thresholds.

In MCL-JCI, Source 44 is one of the source images with rich textures. In Fig. 5.15, the middle row of the three images show the details of the ”People” region, and the bottom row shows the details of the ”Ground” region. We can see that differences

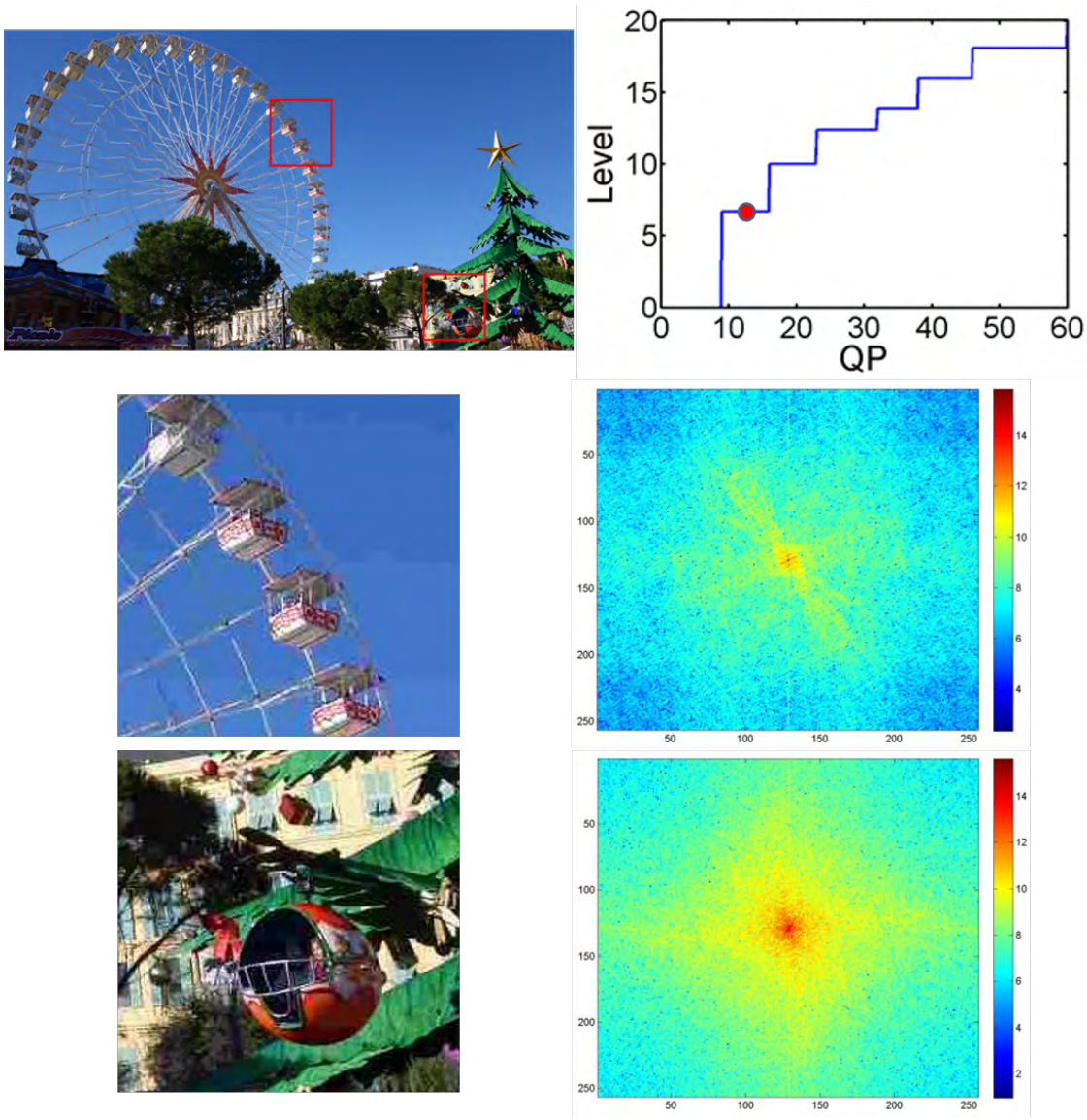
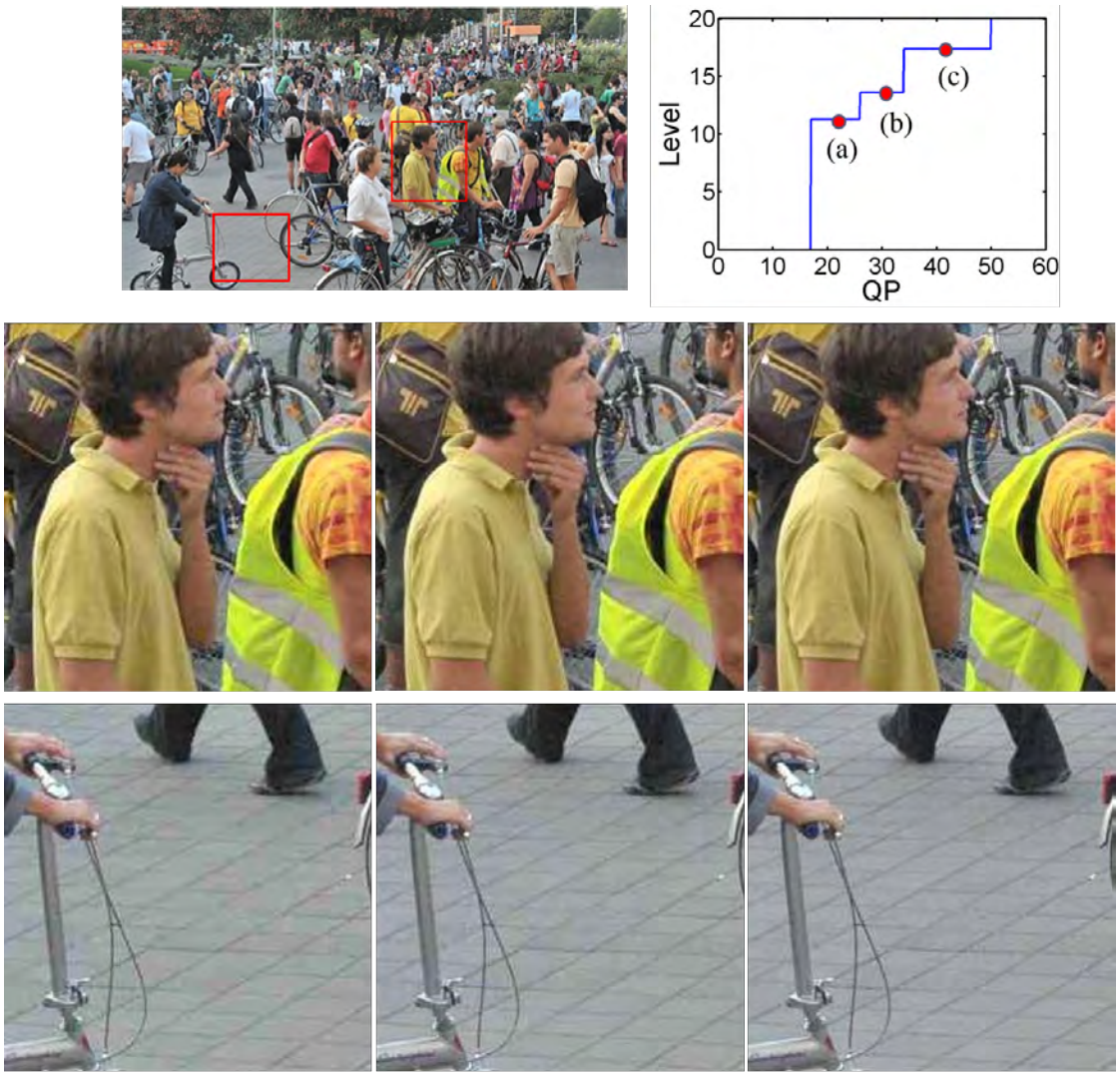


Figure 5.14: Compressed image with six JND levels to the original Source 8.

among JND levels are not significant. Therefore, the number of JND points is much less than other source images with flat regions.

Semantic. Some images have stronger semantic meanings on certain regions such that the viewers intend to focus on these regions of the images. When the distortion appears on these semantic regions, the HVS is more sensitive to these distortions. Fig. 5.16 shows that the JND points are more related to the quality change in the face region.



(a) (b) (c)

Figure 5.15: ”People” and ”Ground” regions of the Source 44.

Most of Source 13 is background and the quality are close for each level as shown in the bottom three images of Fig. 5.16. Even though the area of the face is much smaller than the background, the viewers tend to evaluate the quality based on the face region.

Image JND are influenced by both low-level (like luminance and texture) and high-level features (like face, sky, building and bridge). From the collected JND data, the

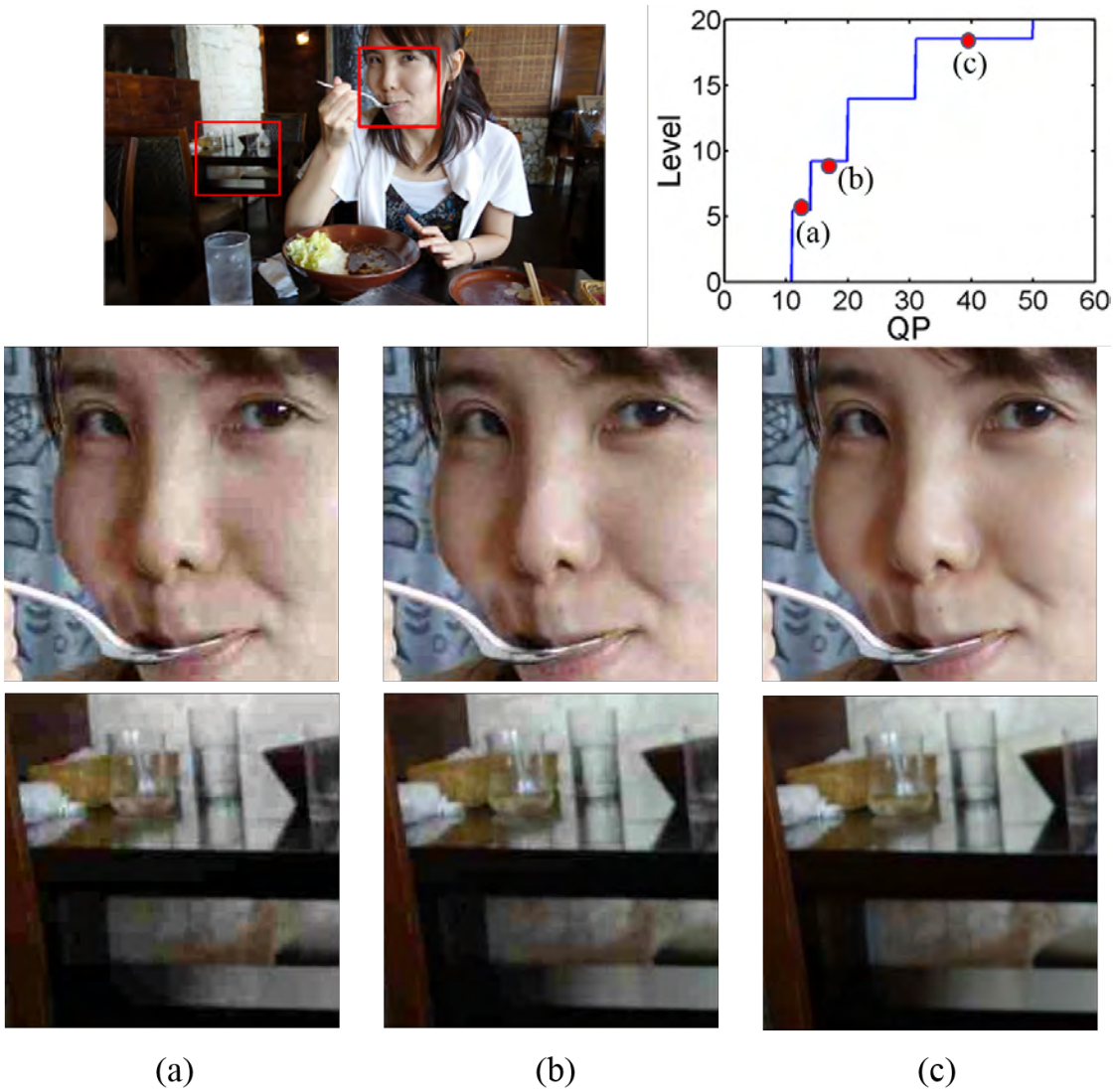


Figure 5.16: Quality change on the face and background regions of Source 13.

relationships of these image features and JND properties can be analyzed. This leads us a better understanding of the underlying scheme in the HVS and provides a new angle to efficient compression.

5.5 Conclusion

A new methodology for human visual experience measurement was proposed in this work. A preliminary subjective test was carried out to collect the JND data to demonstrate the feasibility of this idea. The collected raw JND data was analyzed and post-processed to derive the JND-based quality level plot. It was shown that this quality level plot is robust by shrinking the number of test subjects from 20 to 10. Currently, we are working on large-scale JND-based compressed image quality assessment datasets to gain a deeper understanding of the relationship between the JND-based quality level plot and the underlying image content. Then, for a given image content, we would like to be able to predict its JND numbers and locations based on extracted features.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In Chapter 3, the construction of a new video quality assessment database, called MCL-V, was described in this work. MCL-V contains 12 source video clips and 96 distorted video clips with subjective assessment scores. The source video clips were selected from a large pool of public-domain video sequences with representative and diversified contents. Several existing IQA and VQA algorithms were evaluated against the MCL-V database.

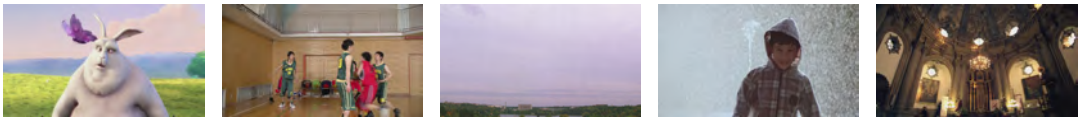
In Chapter 4, two novel VQA indices, FVQA and EVQA, were proposed to assess compressed and resized video quality in the work. FVQA classifies video contents into groups according to their content complexity to reduce content diversity within each group. The proposed EVQA extended the framework of FVQA by take frame-level prediction into account. The frame samples were classified by the distortion significance measured by selected quality methods. In the first step, multiple IQA methods and two proposed video content indices are applied to video frames. According to the outcomes of the IQA methods, they are repositioned as essential metric, grouping classifier, and fusion candidate. In the second stage, the group classifier and essential metric are used to recursively partition the whole sample space into several groups. Finally, several VQA algorithms are selected and fused to predict the perceptual quality within each group. We demonstrate the superior performance of EVQA as compared with other video quality assessment methods using the MCL-V video quality dataset.

In Chapter 5, a thorough discussion about JND was described. A new methodology for human visual experience measurement was proposed in this work. A pilot subjective test was carried out to collect the JND data to demonstrate the feasibility of this idea. The collected raw JND data were analyzed and post-processed to derive the JND-based quality level plot. It was shown that this quality level plot is robust by shrinking the number of test subjects from 20 to 10. With the experience from the pilot study, we developed the first JND image dataset, called MCL-JCI. The MCL-JCI dataset contained 50 image sources that cover a wide range of visual features (like people, bridge, building and indoor scenes). The analysis of JND and content showed image content affects JND.

6.2 Future Work

6.2.1 Preliminary video JND

We built the MCL-JCI dataset in Chapter 5 and wanted to extend the research to video JND. For the JND test on coded video, we collected five high-resolution and diverse video sequences. Again, their diversity can be characterized by their spatial and temporal information [47] and semantic properties [64]. The five video sequence selected in the preliminary test are shown in Fig. 6.1.



(a) BB (b) BD (c) CS (d) FB (e) IC

Figure 6.1: Five video sequences selected for the JND test: (a) Bunny and Butterfly (BB) [34], (b) Basketball Drive (BD), (c) City Sky (CS) [20], (d) Fountain Boy (FB), and (e) Inside Church (IC).

Each source video is encoded by x264 [4]. For video coding, the quality is controlled by quality parameter (QP). Since the QP value can be dynamically adjusted by

a rate control algorithm so that rate control methods may affect visual quality. In the experiment, we compare the following two rate control methods for the x264 encoder.

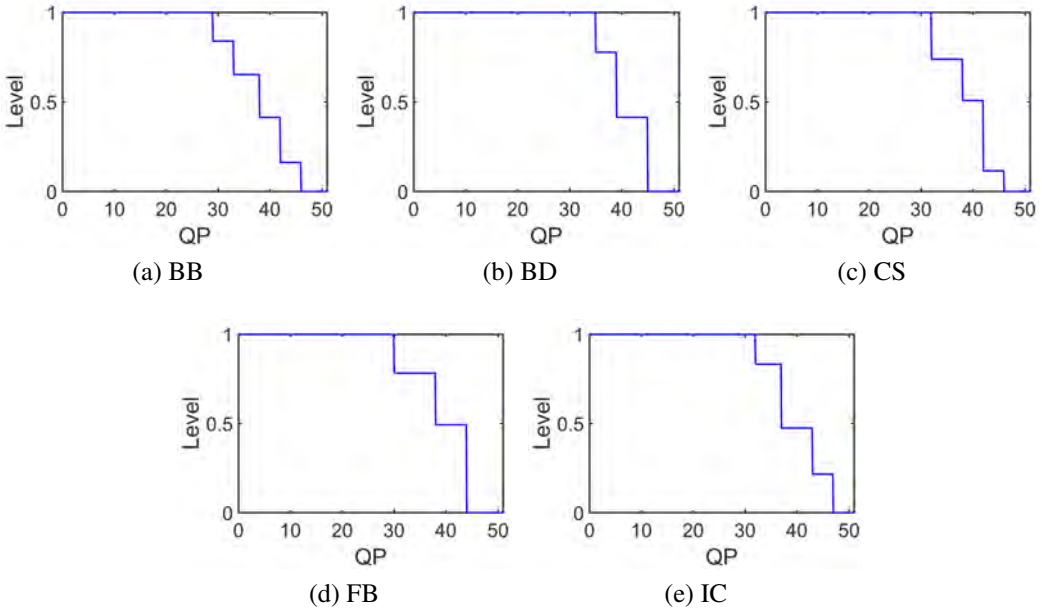
- Fixed QP (FQP). Each video is encoded with $QP = 1, 2, \cdots, 51$.
- Variable Bitrate (VBR). We enable the default variable bitrate (VBR) of the x264 encoder. To get the same number of encoded video as that in the FQP setting, we set the target bitrates to those obtained from FQP. For example, if the bit rate is 3000 kbps with $QP = 20$, we set the target bitrate to 3000 kbps to get a corresponding data point under the VBR setting.

As a result, each source video is encoded 102 times, and there are 103 video contents in total for each source by including the source. A larger QP value leads to a larger quantization step size and a lower coding bitrate in video coding. Again, humans are not able to differentiate all levels. The mean and the standard deviation of JND points for each video source are given in Table 6.1. From Table 6.1, we observe fast-moving

Table 6.1: Statistics of the number of JND points for x264 coded video.

	FQP		VBR	
	Mean	Stdev	Mean	Stdev
BB	4.50	1.32	4.60	1.39
BD	3.15	0.59	3.20	0.77
CS	3.50	0.95	3.65	0.93
FB	3.15	0.99	2.80	1.01
IC	3.50	0.69	3.30	0.86

scene, BD and FB, tend to have less JND points than other 3 scenes. Since the temporal masking effect has significant impact on human perception as indicated in [78, 19, 75, 94], the HVS is unable to differentiate as many quality levels of fast-moving scenes as static scenes. In video compression, useful QP range depends on content. When choosing video bitrates, we can use the QP in the range to achieve reasonable quality with lower bitrates. The JND results give us a rough QP range for acceptable visual



(a) BB (b) BD (c) CS

(d) FB (e) IC

Figure 6.2: The JND-based quality level plot for x264 with FQP: (a) BB (b) BD, (c) CS, (d) FB and (e) IC.

quality. Table 6.2 shows the useful QP for video coding are between 26 and 45. It is obvious that the range is so wide due to the diversity of video content. In video JND, content temporality plays an important role. It can affect the number and location of JND points. If we want to achieve better compression efficiency, predicting JND enables us how to choose bitrates. However, the prediction is difficult since the diversity of content.

6.2.2 JND prediction and its application

We adopted the absolute scale for the representation of video quality in the current research. However, we observe that the visual quality is not continuous and only several levels can be perceived. This created the mismatch between the test principle and score representation. In other words, if we believe pairwise comparison is reliable, we should not convert the results to the typical way. Thus, we want to examine this problem from

JND. The JND is in general defined to be the smallest detectable difference between the starting and the secondary level of a particular sensory stimulus. In the context of video quality assessment, we define the JND is the least quality difference that a trained person can tell from two videos of same content. According the analysis in Chapter 5, image content has a impact to JND. We assume that the JND is strongly affected by video content characteristics. Consider the coding of two movies with the same bit rate, where one movie is grainy and the other is a cartoon. To achieve the same visual quality, the former demands a higher bit rate than the latter. The JND values are expected to be different. To achieve better coding efficiency, we should consider a coding scheme that is content-adaptive. Since we compare two coded video programs to determine the JND, we can record the results of subjective evaluation as JND pairs. It is natural to assume that the JND metric is transitive from pair to pair. By collecting multiple JND pairs with an anchor sequence whose bit rate is known, we can plot the JND vs. bitrate curve, which is simply called the JND curves. The JND curve offers a relationship between video contents and bit rates. Given a target JND, we can use a model, based on the video content, to predict the corresponding bitrate. A good understanding of the JND model will help a video encoder choose a set of suitable coding bit rates according to its content. To develop the JND model, one need to conduct video complexity analysis. With the experiences in MCL-JCI, we want to develop a video dataset, called MCL-JCV. The MCL-JCV dataset can offer a new training set for the proposed EVQA quality metric under our current investigation. Along this line, we set the following three objectives for the proposed research.

- We will conduct the subject test to construct a JND video dataset, called MCL-JCV. This dataset should contain diverse video sources for developing related applications.

- Based on the current MCL-JCI and future MCL-JCV datasets, we will build the image/video analysis model to predict the JND levels of various operating points for a test image/video.

- We will develop a bit rate control scheme based on the image/video analysis model. We expect that this model indicates anchor points according to image/video contents and these points will be adopted as target bitrates.

Bibliography

[1] PhotographyBLOG. `http://www.photographyblog.com`. Accessed: 2015-03-16.

[2] RAW-Samples. `http://www.rawsamples.ch`. Accessed: 2015-01-04.

[3] V. K. Adhikari, Y. Guo, F. Hao, M. Varvello, V. Hilt, M. Steiner, and Z.-L. Zhang. Unreeling netflix: Understanding and improving multi-cdn movie delivery. In *INFOCOM, 2012 Proceedings IEEE*, pages 1620–1628. IEEE, 2012.

[4] L. Aimar, L. Merritt, E. Petit, M. Chen, J. Clay, M. Rullgrd, C. Heine, and A. Izvorski. x264-a free h264/avc encoder. 2005. http://www.videolan.org/developers/x264.html.

[5] U. Ansorge, G. Francis, M. H. Herzog, and H. Öğmen. Visual masking and the dynamics of human perception, cognition, and consciousness a century of progress, a contemporary synthesis, and future directions. *Advances in Cognitive Psychology*, 3(1-2):1, 2007.

[6] P. Atchley and L. Hoffman. Aging and visual masking: sensory and attentional factors. *Psychology and Aging*, 19(1):57, 2004.

[7] R. Avnimelech and N. Intrator. Boosted mixture of experts: an ensemble learning scheme. *Neural computation*, 11(2):483–497, 1999.

[8] N. Bacon-Macé, M. J.-M. Macé, M. Fabre-Thorpe, and S. J. Thorpe. The time course of visual processing: Backward masking and natural scene categorisation. *Vision research*, 45(11):1459–1469, 2005.

[9] M. Barkowsky, N. Staelens, L. Janowski, Y. Koudota, M. Leszczuk, M. Urvoy, P. Hummelbrunner, I. Sedano, K. Brunnström, et al. Subjective experiment dataset for joint development of hybrid video quality measurement algorithms. In *QoEMCS 2012-Third Workshop on Quality of Experience for Multimedia Content Sharing*, pages 1–4, 2012.

[10] F. Bellard, M. Niedermayer, et al. FFmpeg. 2012. http://ffmpeg.org.

[11] S. E. Bird. *The audience in everyday life: Living in a media world*. Routledge, 2013.

[12] H. Boujut, J. Benois-Pineau, O. Hadar, T. Ahmed, and P. Bonnet. Weighted-mse based on saliency map for assessing video quality of h. 264 video streams. In *IS&T/SPIE Electronic Imaging*, pages 78670X–78670X. International Society for Optics and Photonics, 2011.

[13] F. Boulos, W. Chen, B. Parrein, and P. Le Callet. Region-of-interest intra prediction for h. 264/avc error resilience. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3109–3112. IEEE, 2009.

[14] R. A. Bradley. Rank Analysis of Incomplete Block Designs: II. Additional Tables for the Method of Paired Comparisons. *Biometrika*, 41(3):502–537, 1954.

[15] R. A. Bradley and M. E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3):324–345, 1952.

[16] T. Brandão and M. P. Queluz. No-reference quality assessment of h. 264/avc encoded video. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(11):1437–1447, 2010.

[17] T. Brandao, L. Roque, and M. P. Queluz. Quality assessment of h. 264/avc encoded video. *Proc of Confernce on Telecommunications-ConfTele, Sta. Maria da Feira, Portugal*, 2009.

[18] M. D. Brotherton, Q. Huynh-thu, D. S. Hands, and K. Brunnstrom. Subjective Multimedia Quality Assessment. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E89-A(11):2920–2932, Nov. 2006.

[19] M. Carrasco, A. Marie Giordano, and B. McElree. Temporal performance fields: Visual and attentional factors. *Vision research*, 44(12):1351–1365, 2004.

[20] CDVL. The Consumer Digital Video Library. 2010. http://www.cdvl.org/.

[21] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[22] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *Broadcasting, IEEE Transactions on*, 57(2):165–182, 2011.

[23] S. J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pages 2–15. International Society for Optics and Photonics, 1992.

[24] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi. Subjective assessment of h. 264/avc video sequences transmitted over a noisy channel. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pages 204–209. IEEE, 2009.

[25] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi. A h.264/avc video database for the evaluation of quality metrics. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2430–2433, 2010.

[26] T. Ebrahimi. Quality of multimedia experience: Past, present and future. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 3–4, New York, NY, USA, 2009. ACM.

[27] R. Ebrahimpour, E. Kabir, H. Esteky, and M. R. Yousefi. View-independent face recognition with mixture of experts. *Neurocomputing*, 71(4):1103–1107, 2008.

[28] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *Image Processing, IEEE Transactions on*, 20(10):2967–2979, 2011.

[29] European Broadcasting Union (EBU). EBU HDTV Test Sequences. 2006. http://tech.ebu.ch/.

[30] X. Feng, T. Liu, D. Yang, and Y. Wang. Saliency based objective quality assessment of decoded video affected by packet losses. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 2560–2563. IEEE, 2008.

[31] J. M. Foley. Human luminance pattern-vision mechanisms: masking experiments require a new model. *JOSA A*, 11(6):1710–1719, 1994.

[32] M. Gaubatz and S. Hemami. MeTriX MuX visual quality assessment package. 2011. http://foulard.ece.cornell.edu/gaubatz/metrix_mux.

[33] B. Girod. What's wrong with mean-squared error? In *Digital images and human vision*, pages 207–220. MIT press, 1993.

[34] S. Goedegebure, A. Goralczyk, E. Valenza, N. Vegdahl, W. Reynish, B. V. Lommel, C. Barton, J. Morgenstern, and T. Roosendaal. Big Buck Bunny. 2008. http://www.bigbuckbunny.org/.

[35] L. Goldmann, F. De Simone, and T. Ebrahimi. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. In *IS&T/SPIE Electronic Imaging*, pages 75260S–75260S. International Society for Optics and Photonics, 2010.

[36] H. Gulliksen. A least squares solution for paired comparisons with incomplete data. *Psychometrika*, 21(2):125–134, June 1956.

[37] S. Gutta, J. R. Huang, P. Jonathon, and H. Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *Neural Networks, IEEE Transactions on*, 11(4):948–960, 2000.

[38] M. T. Hagan, H. B. Demuth, M. H. Beale, et al. *Neural network design*. Pws Pub. Boston, 1996.

[39] M. T. Hagan and M. B. Menhaj. Training feedforward networks with the marquardt algorithm. *Neural Networks, IEEE Transactions on*, 5(6):989–993, 1994.

[40] L. Haglund. The SVT high definition multi format test set. *Swedish Television Stockholm*, 2006.

[41] P. J. Hahn and V. J. Mathews. An analytical model of the perceptual threshold function for multichannel image compression. In *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, pages 404–408. IEEE, 1998.

[42] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, June 2005.

[43] F. Hermens, G. Luksys, W. Gerstner, M. H. Herzog, and U. Ernst. Modeling spatial and temporal aspects of visual backward masking. *Psychological review*, 115(1):83, 2008.

[44] M. H. Herzog. Spatial processing and visual backward masking. *Advances in Cognitive Psychology*, 3(1-2):85, 2007.

[45] T.-Y. Huang, N. Handigol, B. Heller, N. McKeown, and R. Johari. Confused, timid, and unstable: picking a video streaming rate is hard. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 225–238. ACM, 2012.

[46] Independent JPEG Group. JPEG image compression software. `http://www.ijg.org`. Accessed: 2015-03-20.

[47] ITU. Recommendation ITU-T P.910, Subjective video quality assessment methods for multimedia applications. *International Telecommunication Union, Geneva, Switzerland*, 910, 1999.

[48] ITU. Recommendation ITU-R BT.500-11, Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva, Switzerland*, 2002.

[49] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[50] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proceedings of the IEEE*, 81(10):1385–1422, Oct 1993.

[51] Y. Jia, W. Lin, and A. Kassim. Estimating just-noticeable distortion for video. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(7):820–829, July 2006.

[52] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[53] C. Keimel, A. Redl, and K. Diepold. The TUM high definition video datasets. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 97–102, July 2012.

[54] A. Khan, L. Sun, E. Ifeachor, J. O. Fajardo, and F. Liberal. Impact of rlc losses on quality prediction for h. 264 video over umts networks. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 702–707. IEEE, 2010.

[55] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010.

[56] J.-S. Lee, F. De Simone, and T. Ebrahimi. Subjective quality evaluation via paired comparison: application to scalable video coding. *Multimedia, IEEE Transactions on*, 13(5):882–893, 2011.

[57] J.-S. Lee, F. De Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, and T. Ebrahimi. Subjective evaluation of scalable video coding for content distribution. In *Proceedings of the international conference on Multimedia*, pages 65–72. ACM, 2010.

[58] J.-S. Lee, L. Goldmann, and T. Ebrahimi. Paired comparison-based subjective quality assessment of stereoscopic images. *Multimedia tools and applications*, 67(1):31–48, 2013.

[59] J. Li, M. Barkowsky, and P. Le Callet. Analysis and improvement of a paired comparison method in the application of 3dtv subjective experiment. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 629–632, Sept 2012.

[60] J. Li, M. Barkowsky, and P. Le Callet. Subjective assessment methodology for preference of experience in 3dtv. In *IVMSP Workshop, 2013 IEEE 11th*, pages 1–4, June 2013.

[61] S. Li, L. Ma, and K. N. Ngan. Full-reference video quality assessment by decoupling detail losses and additive impairments. *Circuits and Systems for Video Technology, IEEE Transactions on*, 22(7):1100–1112, 2012.

[62] S. Li, F. Zhang, L. Ma, and K. N. Ngan. Image quality assessment by separately evaluating detail losses and additive impairments. *Multimedia, IEEE Transactions on*, 13(5):935–949, 2011.

[63] C. J. Lin, C. Hsu, and C. Chang. A practical guide to support vector classification. Technical report, 2003.

[64] J. Y. Lin, R. Song, C.-H. Wu, H. W. TsungJung Liu, and C.-C. J. Kuo. MCL Video Quality Database. 2014. http://mcl.usc.edu/mcl-v-database/.

[65] J. Y. L. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo. Fvqa. In *APSIPA 2014*. IEEE, 2014.

[66] W. Lin. Computational models for just-noticeable difference. *Digital video image quality and perceptual coding*, 2005.

[67] W. Lin, L. Dong, and P. Xue. Visual distortion gauge based on discrimination of noticeable contrast changes. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(7):900–909, July 2005.

[68] W. Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.

[69] W. Lin and C.-C. J. Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.

[70] W. Lin, D. Li, and P. Xue. Discriminative analysis of pixel difference towards picture quality prediction. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 3, pages III–193. IEEE, 2003.

[71] A. Liu, W. Lin, and M. Narwaria. Image quality assessment based on gradient similarity. *Image Processing, IEEE Transactions on*, 21(4):1500–1512, April 2012.

[72] T. Liu, Y. Wang, J. M. Boyce, H. Yang, and Z. Wu. A novel video quality metric for low bit-rate video considering both coding and packet-loss artifacts. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):280–293, 2009.

[73] T.-J. Liu, W. Lin, and C.-C. Kuo. Image quality assessment using multi-method fusion. *Image Processing, IEEE Transactions on*, 22(5):1793–1807, 2013.

[74] T.-J. Liu, Y.-C. Lin, W. Lin, and C.-C. J. Kuo. Visual quality assessment: recent developments, coding applications and future trends. *APSIPA Transactions on Signal and Information Processing*, 2:e4, 2013.

[75] Z.-L. Lu, S.-T. Jeon, and B. A. Dosher. Temporal tuning characteristics of the perceptual template and endogenous cuing of spatial attention. *Vision research*, 44(12):1333–1350, 2004.

[76] J. Lubin and D. Fibush. Sarnoff JND vision model, 1997.

[77] Z. Luo, L. Song, S. Zheng, and N. Ling. H.264/Advanced Video Control Perceptual Optimization Coding Based on JND-Directed Coefficient Suppression. *Circuits and Systems for Video Technology, IEEE Transactions on*, 23(6):935–948, June 2013.

[78] D. McKeefry, M. Burton, and C. Vakrou. Speed selectivity in visual short term memory for motion. *Vision research*, 47(18):2418–2425, 2007.

[79] D. J. Miller and H. S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in neural information processing systems*, pages 571–577, 1997.

[80] T. Mitsa and K. L. Varkur. Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 5, pages 301–304. IEEE, 1993.

[81] A. Moorthy and A. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *Image Processing, IEEE Transactions on*, 20(12):3350–3364, Dec 2011.

[82] A. Moorthy, L. K. Choi, A. Bovik, and G. de Veciana. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *Selected Topics in Signal Processing, IEEE Journal of*, 6(6):652–671, 2012.

[83] M. Narwaria and W. Lin. Objective image quality assessment based on support vector regression. *Neural Networks, IEEE Transactions on*, 21(3):515–519, 2010.

[84] J. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand. Comparison of the coding efficiency of video coding standards-including high efficiency video coding (hevc). *Circuits and Systems for Video Technology, IEEE Transactions on*, 22(12):1669–1684, 2012.

[85] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang. Modeling the impact of frame rate on perceptual quality of video. *City*, 70(80):90, 2008.

[86] Y.-F. Ou, Z. Ma, and Y. Wang. A novel quality metric for compressed video considering both frame rate and quantization artifacts. *City*, 80:100, 2009.

[87] Y.-F. Ou, Y. Zhou, and Y. Wang. Perceptual quality of video with frame rate variation: A subjective study. In *ICASSP*, pages 2446–2449, 2010.

[88] S. Péchard, R. Pépion, P. Le Callet, et al. Suitable methodology in subjective video quality assessment: a resolution dependent paradigm. In *Proceedings of the Third International Workshop on Image Media Quality and its Applications, IMQA2008*, 2008.

[89] Y. Pitrey, M. Barkowsky, P. Le Callet, and R. Pépion. Evaluation of mpeg4-svc for qoe protection in the context of transmission errors. In *SPIE Optical Engineering+ Applications*, pages 77981C–77981C. International Society for Optics and Photonics, 2010.

[90] Y. Pitrey, M. Barkowsky, P. Le Callet, and R. Pepion. Subjective quality assessment of mpeg-4 scalable video coding in a mobile scenario. In *Visual Information Processing (EUVIP), 2010 2nd European Workshop on*, pages 86–91. IEEE, 2010.

[91] Y. Pitrey, M. Barkowsky, P. Le Callet, R. Pepion, et al. Subjective quality evaluation of h. 264 high-definition video coding versus spatial up-scaling and interlacing. *QoE for Multimedia Content Sharing*, 2010.

[92] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet. Subjective quality of svc-coded videos with different error-patterns concealed using spatial scalability. In *Visual Information Processing (EUVIP), 2011 3rd European Workshop on*, pages 180–185. IEEE, 2011.

[93] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, P. Le Callet, et al. Aligning subjective tests using a low cost common set. *QoE for Multimedia Content Sharing*, 2011.

[94] U. Polat, A. Sterkin, and O. Yehezkel. Spatio-temporal low-level neural networks account for visual masking. *Advances in Cognitive Psychology*, 3(1-2):153, 2007.

[95] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Kuo. Color image database tid2013: Peculiarities and preliminary results. In *Visual Information Processing (EUVIP), 2013 4th European Workshop on*, pages 106–111, June 2013.

[96] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. A new color image database tid2013: Innovations and results. In *Advanced Concepts for Intelligent Vision Systems*, pages 402–413. Springer, 2013.

[97] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[98] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[99] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Neural Networks, 1993., IEEE International Conference on*, pages 586–591. IEEE, 1993.

[100] S. Rimac-Drlje, M. Vranjes, and D. Zagar. Influence of temporal pooling method on the objective video quality evaluation. In *Broadband Multimedia Systems and Broadcasting, 2009. BMSB'09. IEEE International Symposium on*, pages 1–5. IEEE, 2009.

[101] A. M. Rohaly, J. Lu, N. R. Franzen, and M. K. Ravel. Comparison of temporal pooling methods for estimating the quality of complex video sequences. In *Electronic Imaging'99*, pages 218–225. International Society for Optics and Photonics, 1999.

[102] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

[103] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack. Study of subjective and objective quality assessment of video. *Image Processing, IEEE Transactions on*, 19(6):1427–1441, 2010.

[104] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp. to pool or not to pool: A comparison of temporal pooling methods for http adaptive video streaming. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 52–57. IEEE, 2013.

[105] H. Sheikh and A. Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430–444, 2006.

[106] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing, IEEE Transactions on*, 15(11):3440–3451, 2006.

[107] D. A. Silverstein and J. E. Farrell. Quantifying perceptual image quality. In *PICS*, volume 98, pages 242–246, 1998.

[108] D. A. Silverstein and J. E. Farrell. Efficient method for paired comparison. *Journal of Electronic Imaging*, 10(2):394–398, 2001.

[109] N. Staelens, G. Van Wallendael, R. Van de Walle, F. De Turck, and P. Demeester. High definition h. 264/avc subjective video database for evaluating the influence of slice losses on quality perception. In *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pages 130–135. IEEE, 2013.

[110] D. Tice. How People Use Media: Over-the-Top TV 2013. Technical report, "GfK Media", Aug. 2013.

[111] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi. Performance comparisons of subjective quality assessment methods for mobile video. In *2010 2nd International Workshop on Quality of Multimedia Experience, QoMEX 2010*, pages 82–87, Trondheim, Norway, June 2010. IEEE.

[112] T. D. Tran. A locally adaptive perceptual masking threshold model for image coding. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 4, pages 1882–1885. IEEE, 1996.

[113] K. Tsukida and M. R. Gupta. How to analyze paired comparison data. Technical report, DTIC Document, 2011.

[114] Video Quality Experts Group (VQEG). Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I. 2000. ftp://ftp.crc.ca/crc/vqeg/TestSequences/.

[115] Video Quality Experts Group (VQEG). Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II. 2003.

[116] Video Quality Experts Group (VQEG). Report on the validation of video quality models for high definition video content. 2010.

[117] P. Vu, C. Vu, and D. Chandler. A spatiotemporal most-apparent-distortion model for video quality assessment. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2505–2508, 2011.

[118] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[119] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. IEEE, 2003.

[120] A. B. Watson. DCTune: A technique for visual optimization of DCT quantization matrices for individual images. In *Sid International Symposium Digest of Technical Papers*, volume 24, pages 946–946. SOCIETY FOR INFORMATION DISPLAY, 1993.

[121] A. B. Watson. Proposal: Measurement of a JND scale for video quality. *IEEE G-2.1. 6 Subcommittee on Video Compression Measurements*, 2000.

[122] S. Winkler. Analysis of public image and video databases for quality assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 6(6):616–625, Oct 2012.

[123] J. Wu, G. Shi, W. Lin, A. Liu, and F. Qi. Just noticeable difference estimation for images with free-energy principle. *Multimedia, IEEE Transactions on*, 15(7):1705–1710, Nov 2013.

[124] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin. Random partial paired comparison for subjective video quality assessment via hodgerank. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 393–402, New York, NY, USA, 2011. ACM.

[125] J. Xue and C. W. Chen. Mobile JND: Environment Adapted Perceptual Model and Mobile Video Quality Enhancement. In *Proceedings of the 3rd Multimedia Systems Conference*, MMSys '12, pages 173–183, New York, NY, USA, 2012. ACM.

[126] W. Xue, L. Zhang, X. Mou, and A. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *Image Processing, IEEE Transactions on*, 23(2):684–695, Feb 2014.

[127] X. Yang, W. Lin, Z. Lu, E. P. Ong, and S. Yao. Perceptually adaptive hybrid video encoding based on just-noticeable-distortion profile. In *Visual Communications and Image Processing 2003*, pages 1448–1459. International Society for Optics and Photonics, 2003.

[128] J. You, T. Ebrahimi, and A. Perkis. Modeling motion visual perception for video quality assessment. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 1293–1296, New York, NY, USA, 2011. ACM.

[129] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan. IVP subjective quality video database. 2011. http://ivp.ee.cuhk.edu.hk/research/database/subjective/.

[130] L. Zhang, D. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *Image Processing, IEEE Transactions on*, 20(8):2378–2386, 2011.

[131] X. Zhang, W. Lin, and P. Xue. Just-noticeable difference estimation with pixels in images. *Journal of Visual Communication and Image Representation*, 19(1):30–41, 2008.