Perceptual Video Quality Assessment and Enhancement

by

Kai Zeng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

With the rapid development of network visual communication technologies, digital video has become ubiquitous and indispensable in our everyday lives. Video acquisition, communication, and processing systems introduce various types of distortions, which may have major impact on perceived video quality by human observers. Effective and efficient objective video quality assessment (VQA) methods that can predict perceptual video quality are highly desirable in modern visual communication systems for performance evaluation, quality control and resource allocation purposes. Moreover, perceptual VQA measures may also be employed to optimize a wide variety of video processing algorithms and systems for best perceptual quality.

This thesis exploits several novel ideas in the areas of video quality assessment and enhancement. Firstly, by considering a video signal as a 3D volume image, we propose a 3D structural similarity (SSIM) based full-reference (FR) VQA approach, which also incorporates local information content and local distortion-based pooling methods. Secondly, a reduced-reference (RR) VQA scheme is developed by tracing the evolvement of local phase structures over time in the complex wavelet domain. Furthermore, we propose a quality-aware video system which combines spatial and temporal quality measures with a robust video watermarking technique, such that RR-VQA can be performed without transmitting RR features via an ancillary lossless channel. Finally, a novel strategy for enhancing video denoising algorithms, namely poly-view fusion, is developed by examining a video sequence as a 3D volume image from multiple (front, side, top) views. This leads to significant and consistent gain in terms of both peak signal-to-noise ratio (PSNR) and SSIM performance, especially at high noise levels.

Acknowledgements

During the completion of this doctoral work, I have been accompanied and supported by many people. It is my great pleasure to take this opportunity to thank all of them.

First of all, I would like to express my deepest gratitude to my senior supervisor, Professor Zhou Wang, who helped me most and guided me during the studies toward my doctoral degree at University of Waterloo. Dr. Wang is not only a great professor with knowledgable and deep vision in academia research but also most importantly a very nice and kind person. He always provided me great opportunities, incredible encouragements, and generous supports during my research career. It is impossible for my to finish any of my research work in this thesis without his supervision.

I would also like to express my sincere gratitude to Dr. Dake He at Research In Motion (now BlackBerry), who supervised and supported me during my eight months of internship. His invaluable ideas and guidance introduced me to the real industry work. I am very grateful to Professor Sherman Shen, Professor Pin-Han Ho, and Professor Alex Wong for serving as the examiner of my thesis. I also gratefully thank the external examiner, Professor Ivan V. Baji'c at Simon Fraser University, and defense chair Professor Brent Doberstein, for their precious time to attend the defense in the midst of their busy activities.

I am also thankful to all amazing members at Image Visual Computing (IVC) laboratory, for making my life during the past four years a fun, challenging and memorable one. Especially, I would like to thank Abdul Rehman, Hojatollah Yeganeh, Rania Hassen, Jiheng Wang, Tiesong Zhao, Yuming Fang, Kede Ma, and Nima Nikvand to name but a few, for their insightful comments and helpful discussions.

Finally, I would like to dedicate this thesis to my parents, sister, and my wife, Lin Xie, for their endless love, limitless encouragement, and unselfish sacrifice throughout my doctoral education. Especially thanks for my wife's deepest love and her effort to taking care of me. Really appreciated!

Contents

List of Tables

List of Figures

List of Abbreviations

3D-DCT	3 Dimensional Discrete Cosine Transform
3D-SSIM	3 Dimensional Structure SIMilarity
ACR	Absolute Categorical Rating
ALM	Additive Log-logistic Model
AQIM	Angle Quantization Index Modulation
BLS-GSM	Bayes Least Square estimation based on Gaussian Scale Mixture
BM3D	Block Matching and 3D transform shrinkage
CfP	Call for Proposals
CIM	Curvature Interpolation Method
CRC	Cyclic Redundancy Check
CSF	Contrast Sensitivity Function
CV	Circular Variance
DCT	Discrete Cosine Transform
DMOS	Difference of Mean Opinion Score
FR-VQA	Full Reference Video Quality Assessment
GOPs	Groups of Pictures
HD	High Definition
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IP	Internet Protocol
IQA	Image Quality Assessment
ITS	Institute for Telecommunication Sciences
JCT-VC	Joint Collaborative Team on Video Coding
JND	Just Noticeable Difference

KLD	Kullback-Leibler Distance
KRCC	Kendall's Rank Correlation Coefficient
LDPC	Low-Density Parity-Check
LHS	Local Harmonic Strength
MAE	Mean Absolute Error
MOS	Mean Opinion Score
MOVIE	MOtion-based Video Integrity Evaluation
MPEG	Moving Picture Experts Group
MSE	Mean Squared Error
MS-SSIM	Multi-Scale Structure SIMilarity
NORM	NO-Reference video quality Monitoring
NR-VQA	No Reference Video Quality Assessment
NTIA	National Telecommunications and Information Administration
PLCC	Pearson Linear Correlation Coefficient
PQI	Perceptual Quality Index
PSNR	Peak Signal-to-Noise Ratio
PTQM	Perceptual Temporal Quality Metric
PVF	Poly-View Fusion
PVQM	Perceptual Video Quality Measure
QAI	Quality-Aware Image
QAV	Quality-Aware Video
QoE	Quality of Experience
RA-HE	Random-Access High-Efficiency
RD	Rate Distortion
RDO	Rate Distortion Optimization
RMS	Root Mean Square
ROI	Region Of Interest
RR-VQA	Reduced Reference Video Quality Assessment
SRCC	Spearman Rank-order Correlation Coefficient
SSIM	Structure SIMilarity
SURE-LET	Stein's Unbiased Risk Estimator Linear Expansion of Threshold
SVR	Support Vector Regression

TMS	Temporal Motion Smoothness
VCEG	Video Coding Experts Group
VIF	Visual Information Fidelity
VPVF	Variance weighted Poly-View Fusion
VQA	Video Quality Assessment
VQEG	Video Quality Expert Group
VQM	Video Quality Model

Chapter 1

Introduction

1.1 Motivation

Digital video has become ubiquitous and indispensable in our everyday lives. With the rapid development of communication technologies, video has played a significant role in multimedia communication systems. Therefore, it is crucial to maintain the quality of video at an acceptable level in diverse application environments such as network visual communications.

The first problem is how to define and measure *video quality*. The definition of video quality in Wikipedia is as follows:

> *Video quality is a characteristic of a video passed through a video transmission/processing system, a formal or informal measure of perceived video degradation (typically, compared to the original video). Video processing systems may introduce some amounts of distortion or artifacts in the video signal, so video quality evaluation is an important problem.*

Because the video quality measure is a fundamental problem related to the majority of video processing applications, it has attracted a large amount of effort from academia and industry during the past decades. In recent years, extensive studies have identified the drawbacks of traditional video quality measurements: mean squared error (MSE) or peak signal-to-noise ratio (PSNR). They have been criticized for their low correlation with the perceptual quality measurement of the human visual system (HVS). Specifically, the MSE/PSNR do not take into account the correlation among adjacent pixels, not to mention that among neighboring frames for video. They also do not consider the properties of HVS, such as multi-scale and multi-channel characteristics.

Steady progress has been made recently in still image quality assessment (IQA). Successful IQA approaches include structural similarity (SSIM) [1] and its derivatives (multi-scale SSIM [2], information-weighted SSIM [3], complex-wavelet SSIM [4], feature SSIM [5]), and visual information fidelity (VIF) [6]. However, for VQA, theoretically cohesive and practically effective methods are still lacking. Although IQA approaches can be easily extended to VQA scenarios on a frame-by-frame basis, some significant aspects of video, in particular, the temporal correlation or motion information among adjacent frames, are ignored.

According to the availability of a reference video, there is a general agreement [7] that objective VQA metrics can be divided into three categories: full-reference (FR), no-reference (NR), and reduced-reference (RR) methods. In order to evaluate the quality of a distorted video, FR-VQA always assumes full access to the original video. Thus, FR methods usually provide the most-precise evaluation results in comparison with NR and RR methods. However, it is hard or expensive to satisfy this assumption in practical applications. NR-VQA methods are designed to support quality measurement without the corresponding reference, but existing methods perform reasonably well only when distortions are known and modeled precisely. To provide a compromise between FR and NR, RR-VQA approaches have been proposed that employ partial information (quality features) of the reference.

This thesis mainly focuses on FR and RR VQA research.

Firstly, by considering video as a 3D volume image, it proposes a 3D structural similarity (3D-SSIM) based FR-VQA approach. Instead of evaluating the video quality frame-by-frame, 3D-SSIM is able to capture spatial and temporal distortion simultaneously. In addition to the quality estimation, a local information content and local distortion based weighting method is developed to pool the quality map into a single quality measure.

Secondly, the thesis proposes an RR-VQA method in which the evolvement of local phase structures is traced over time in the complex wavelet domain. Temporal motion smoothness, a novel descriptor of motion, is developed for the evaluation of perceptual video quality. The proposed measure is capable of detecting a variety of common distortions, including noise contamination, blurring, line or frame jittering, and frame dropping. Moreover, it does not require a costly motion estimation process and has a very low RR data rate, both of which make its adoption in visual communication applications much easier.

Thirdly, a quality-aware video (QAV) system is constructed for the deployment of RR-VQA method. In RR-VQA literature, the quality features are extracted from the reference video at the sender side and used to measure the quality of distorted video at the receiver side. Those features are assumed to be transmitted through an ancillary error-free channel. However, it is generally impossible or very costly to provide such an additional channel in practical scenarios. To resolve this problem, a digital watermarking technique is employed to embed features into the original video invisibly and extract them when needed. The error-control coding scheme is also integrated to enhance the robustness of the QAV system.

The last problem of interest is how to improve video quality based on the idea behind the effective VQA method. One of the most useful applications is *video de-*

noising, because video signals are subject to noise contamination during acquisition, compression, transmission, and reproduction. Therefore, an effective video denoising algorithm that can remove or reduce the noise is highly desirable. Such as algorithm improves not only video signals' perceptual quality, but also the performance of subsequent processes such as compression, segmentation, resizing, de-interlacing, and object detection, recognition, and tracking [7].

1.2 Contribution

The focuses of this thesis are to develop effective and efficient VQA methods for multimedia communication systems and improve perceptual video quality based on novel VQA measures. The major contributions of this thesis are as follows.

- For FR-VQA, a video signal is considered to be a 3D volume image, and a "region" in the image is defined as a localized 3D block. A 3D quality map can then be generated by applying a block-wise quality measure within local regions. This step is followed by a pooling stage that merges the quality map into an overall quality score. Based on the assumption that a video region that contains more information (computed based on statistical image models) or more-severe distortion is more likely to attract visual attention, local information content and local distortion-weighted pooling for VQA is developed. The combination of quality measurement and pooling strategies leads to consistent gain when tested using several independent databases.

- For RR-VQA, discovering quality features that can capture video quality degradation is crucial, especially those related to motion (because the capability of representing motion is probably the most critical feature that distinguishes video from still images). Thus, this thesis presents a novel method for quantifing the temporal motion smoothness [8] of video sequences, which is affected by many types of distortions commonly encountered in real-world video acquisition, communication and processing systems.

- A novel QAV system is constructed based on spatial and temporal RR-VQA and a robust video watermarking technique. At the sender side, two quality features are extracted: (1) intra-frame features based on a statistical model of the marginal distribution of wavelet coefficients [9] and (2) inter-frame features calculated by temporal motion smoothness measurement in the complex wavelet transform domain [10]. An error-control encoding scheme is employed to improve the robustness in the subsequent transmission of the quality features. This is followed by embedding the encoded features into the original video invisibly using a robust video watermarking approach. The angle quantization index modulation (AQIM) [11] is employed to hide those features in the video after a 3D discrete cosine transform (3D-DCT). The resulting video is called a QAV, which is transmitted to the receiver through a lossy communication channel. At the receiver side, the same feature-extraction process as at the sender side is applied to the distorted video. Meanwhile, the hidden messages are extracted, followed by error-control decoding to recover the quality features. The recovered features, together with the corresponding features extracted from the distorted video, are employed by an RR-VQA algorithm, that evaluates the perceptual quality degradation of the distorted QAV.

- For video denoising, a novel strategy called polyview fusion (PVF) is proposed to boost existing video denoising approaches. In particular, the same noisy video volume is denoised using 2D approaches but from three different views, i.e., front-, top-, and side-views. An optimal fusion scheme is then employed to combine the three denoised versions of the video. By doing so, each pixel is denoised by its neighboring pixels from all three dimensions. Moreover, a variance-weighted PVF (VPVF) scheme is proposed. After three denoised videos from three views are obtained, a normalization procedure inspired by the SSIM measure [1] and a fusion process based on local variance are employed to produce better denoised video. It is shown that those two strategies lead to significant gain of video denoising performance over different base-denoising algorithms, especially at high noise levels.

1.3 Organization

Following the above introductory section, the remainder of this thesis is organized as follows. Chapter 2 gives a detailed literature review of the VQA problem and video denoising. State-of-the-art algorithms are briefly introduced and summarized in chronological order. Chapter 3 focuses on a description of the proposed FR-VQA method using 3D-SSIM. A novel RR-VQA approach and the QAV system based on temporal motion smoothness are introduced in Chapter 4. The proposed methods for video denoising enhancement are depicted in Chapter 5. Finally, the conclusion of the research work in this thesis and potential future work topics are discussed in Chapter 6. In the Appendix, the performance of existing VQA methods is investigated under the context of video compression, with the aim of raising new problems regarding both objective VQA and video coding schemes.

Chapter 2

Literature Review

This chapter provides an overview of video quality assessment (VQA) techniques and video denoising approaches. Generally, existing VQA approaches can be divided into two classes:

- Subjective methods, which seek opinions from the observer about the perceptual quality [12].
- Objective methods, which provide a computational model for automatic quality estimation [13].

In the most of network visual communication systems, subjective VQA experiments offer the most reliable quality measure because human eyes are the ultimate receivers. Groups of trained or untrained subjects are recruited to watch videos and rate the quality. In addition, the setup of a subjective test environment needs to be carefully designed (e.g., following the ITU-T recommendations [14]) in terms of viewing distance, room illumination, test duration, subject selection, and quality rating strategy. The results of subjective VQA in terms of mean opinion score (MOS) are generally considered to be the benchmark for performance evaluation of objective VQA methods. The MOSs are provided by all of the existing VQA databases, including VQEG FR-TV Phase I Database, VQEG HDTV Database [15], LIVE Video Quality Database [16], LIVE Mobile Video Quality Database [17], EPFL-PoliMl VQA Database [18], and IRCCyN/IVC Databases [19]. A detailed introduction and analysis of publicly available image and video databases for quality assessment can be found in [20]. However, a major drawback of the subjective VQA is the high costs in terms of time, labor, and money. Therefore, the subjective VQA approach is infeasible or extremely difficult to deploy in practical communication systems.

In contrast, objective VQA metrics can be employed for video quality evaluation fully automatically. They can play an essential role in network visual communication systems for the evaluation, control, and improvement of the perceptual quality of video. The benefits to video service providers of measuring video quality are manifold:

- Video service providers can choose the best equipment or technique among available products based on an effective VQA metric as benchmark.
- The parameters of existing equipments can be adjusted to maximize output video quality.
- It allows video service providers to control the visual quality of videos that are produced/processed/encoded/transcoded/bought/sold.

As the traditional video quality measurements, PSNR and MSE are widely accepted because they are computationally simple and tractable for algorithm optimization. However, their major drawback is that they are inconsistent with subjective opinion. Therefore, an effective and efficient objective VQA approach is urgently needed. Based on the availability of reference videos, objective VQA approaches can be further categorized into full-reference (FR), reduced-reference (RR), and no-reference

(NR) methods. Because of the broad scope of VQA research, major contributions in each category are reviewed in chronological order.

2.1 Full-reference Video Quality Assessment

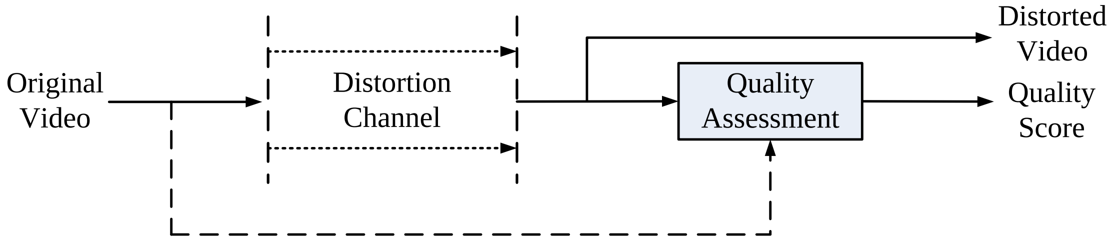


Figure 2.1: Framework of full-reference video quality assessment

For FR-VQA, the design principle is to measure the similarity or distance between reference and distorted videos. The basic framework of FR-VQA is depicted in Fig. 2.1. A straightforward method is to study the characteristics of HVS and simulate them using carefully designed algorithms to measure the similarity quantitatively. Because HVS is an extremely complex system for which we have only limited knowledge, FR-VQA is still a difficult task. Recent FR-VQA has achieved notable success in predicting perceived video quality [13].

The simplest method for FR-VQA is applying the mature FR-IQA to video on a frame-by-frame basis. During the past several decades, the key FR-IQA methods have included Sarnoff's just noticeable difference (JND) metrics [21], picture quality scale (PQS) [22], noise quality measure (NQM) [23], structural similarity (SSIM) [1] and its derivatives (multi-scale SSIM [2], information weighted SSIM [3], complex wavelet SSIM [4], feature SSIM [5]), and visual information fidelity (VIF) [6]. They were used to measure the quality of a video without taking into account the correlation among adjacent frames. The advanced FR-VQA approaches integrated the motion information by either a searching procedure or calculating the optical flow between neighboring frames.

- In 1993, Webster *et al.* [24] from the institute for telecommunication sciences (ITS) described several objective video quality measures that are able to predict the subjective ratings. They claimed that the color distortions were insignificant relative to the spatial and temporal artifacts.

- In 1996, Christian and Olivier [25] proposed a FR-VQA model based on a multi-channel model of human spatio-temporal vision. A spatio-temporal filter bank was adopted to simulate the mechanisms of vision. The contrast sensitivity and masking effect were also taken care of by the decomposition.

- In 1997, Olsson *et al.* [26] introduced several perceptual objective models for quality assessment and compared them with PSNR/MSE.

- In 1998, Tan *et al.* [27] proposed a two-stage objective quality model for MPEG-coded video. After detection of several coding artifacts, a cognitive emulator was employed to simulate human high-level processing of visual information. This technique is suitable for evaluating the temporal quality variations in long sequences. Based on the DCT transform, Watson [28] developed a FR-VQA metric that incorporated human spatial, temporal, chromatic sensitivity, light adaptation, and contrast masking.

- In 1999, for MPEG-coded color videos, Winkler [29] proposed a distortion metric, which took into account the spatial and temporal aspects of vision, as well as the color perception. Wolf and Pinson [30] developed a spatial-temporal distortion metric for quality monitoring over a wide range of quality levels. Tong *et al.* [31] introduced a spatial-temporal quality model for MPEG-coded videos in the CIE-LAB color domain.

- In 2000, Tan and Ghanbari [32] designed a multi-metric quality model for MPEG video, which comprises a perceptual quality model and a blockiness detector. After a review on HVS based VQA, Yu and Wu [33] introduced how to incorporate the characteristics of HVS into quality metrics. Rohaly *et al.* [34] described the progress of VQEG at that time.

- In 2001, Kwon and Lee [35] developed a FR-VQA system based on a recursive biorthogonal wavelet transform. It took reference and distorted videos as input and applied different weights to different spatial frequencies according to the sensitivity of HVS.

- In 2002, Hekstra *et al.* [36] proposed a perceptual video quality measure (PVQM), in which three quality indicators ("edginess" of the luminance, normalized color error, and temporal de-correlation) are linearly combined. This method aims to predict the degree of distortion generated by video coding systems. Tested by the video quality expert group (VQEG), it was recognized as the best-quality model at that time.

- In 2004, Wang *et al.* [37] introduced an effective and efficient FR-VQA approach based on the design philosophy of Structural SIMilarity (SSIM). The masking effect of HVS was integrated for weighted pooling in the spatial-temporal domain. Based on the multi-channel properties of HVS, Guo *et al.* [38] employed Gabor filtering to imitate the psycho-perceptual properties of HVS for quality measurement.

- In 2007, Wang and Li [39] proposed a statistical model of human visual speed perception for VQA under the information theory framework. It is able to estimate the motion information content and perceptual uncertainty of video so as to facilitate the weighting process. Yang *et al.* [40] proposed a perceptual temporal quality metric (PTQM) that focuses on the temporal quality degradation caused by both regular and irregular frame loss. The PTQM is capable of estimating perceived visual discomfort induced by temporal distortion under various combinations of scenes and motion activities.

- In 2008, an extensive review was conducted by Winkler and Mohandas in [41], which described the evolution of VQA techniques. They analyzed the merits and drawbacks of a wide range of VQA models, from the traditional PSNR to state-of-the-art models. The potential research directions in this area were also discussed.

- In 2009, Liu *et al.* [42] studied the effects of packet losses in low bit-rate wireless network and lossy compression of H.264/AVC coding. They proposed a FR-VQA scheme by considering five distortion factors: error length, loss severity, loss location, number of losses, and loss patterns. Focusing on temporal evolutions of spatial distortions, Ninassi *et al.* [43] proposed dividing a sequence into short-term spatio-temporal segments to calculate a quality map. A long-term temporal pooling strategy was adopted to compute the overall score. The temporal characteristics of video have also been studied and employed by Barkowsky *et al.* [44] in their temporal trajectory aware video quality measurement system.

- In 2010, the motion-based video integrity evaluation (MOVIE) index was proposed by Seshadrinathan and Bovik [45, 16]. This general-purpose spatio-spectrally localized multiscale framework employs Gabor decomposition to integrate both spatial and temporal (and spatio-temporal) aspects of distortion evaluation. Moorthy *et al.* [46] made an in-depth study of the subjective and objective quality assessment of H.264 compressed videos transmitted over a wireless channel. Moorthy and Bovik [47] also developed an efficient VQA algorithm based on the motion compensated structural similarity index.

Huynh-Thu and Ghanbari [48] reported that the impact of spatial quality on overall video quality is dependent on the temporal quality and vice-versa. The spatial quality contributes more than temporal quality to the overall quality.

- In 2011, You *et al.* [49] proposed a visual-attention-driven FR-VQA framework under the motivations that attention mechanism plays an important role in HVS and that unattended stimuli can still contribute to the perception of visual content. Based on the advanced attention selection theory, the overall quality score was computed by combing global and local quality features using an adaptive fusion technique. Zhao *et al.* [50] introduced the perceptual quality index (PQI) by incorporating a series of fundamental HVS characteristics. After examining the influence of temporal video quality variation, Yim and Bovik [51] proposed a VQA algorithm that combines a simple frame-based VQA method with a temporal quality variance factor. Ćulibrk *et al.* [52] explored the effect of bottom-up motion saliency features for the problem of MPEG-2 coded VQA and proposed a video quality estimator by employing the selected best features. Narwaria and Lin [53] combined MOVIE, multiscale SSIM, and motion vector similarity into one metric based on adaptive basis function regression-based machine learning.

- In 2012, Li *et al.* [54] incorporated the motion information and temporal HVS characteristics with two types of spatial distortions (detail losses and additive impairments) for objective VQA. Narwaria *et al.* [55] developed a low-complexity VQA approach that combines the temporal quality fluctuations, worst case pooling strategy, and machine learning scheme. Leszczuk *et al.* [56] employed both SSIM and temporal pooling techniques to derive a quality of experience (QoE) model for high definition video with different patterns of packet losses artifact. Wang *et al.* [57] proposed to deal with the motion information by structural features in the localized spatio-temporal regions. Three dimensional structure tensors was employed to extract two descriptors for structural information representation.

- In 2013, Park *et al.* [58] proposed a content adaptive spatial and temporal pooling strategy based on the distribution of spatio-temporal local quality scores to account for the effect of severe distortion on overall perceived video quality. The "worst" local scores along the spatial and temporal dimensions of a video were emphasized implicitly.

2.2 No-reference Video Quality Assessment

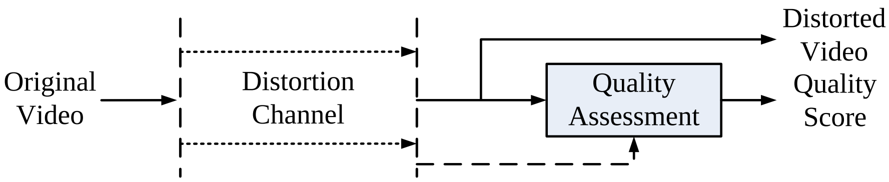


Figure 2.2: Framework of no-reference video quality assessment.

No-reference (NR) VQA, whose framework is shown in Fig. 2.2, aims to estimate the quality of a received video without any access to reference video. In this case, the natural video statistics and distortion model become much more important for quality evaluation. Most of existing NR-VQA approaches are application-specific and aim for one or several specific distortions.

- In 2005, Farias and Mitra [59] introduced a real-time NR-VQA measure based on the detection of three artifacts (blockiness, blurriness, noisiness) and their combination. Yang *et al.* [60] took into account the temporal dependency among neighboring frames to create a general-purpose NR-VQA method.

- In 2008, Kawayoke and Horita [61] suggested a continuous scored NR-VQA system which used the information from both video content and motion for frame-level quality evaluation. Tao *et al.* [62] developed a loss-distortion model for real-time video quality monitoring in IP networks. It accounts for the impact of various network-dependent and application-specific factors on the quality of decoded video. Based on this model, a relative metric was also defined to evaluated the video quality without parsing or decoding the transmitted video bitstream.

- In 2009, Naccari *et al.* [63] proposed a NORM (no-reference video quality monitoring) system to evaluate the quality change of H.264/AVC compressed video with transmission errors. Keimel *et al.* [64] combined blocking and blurring measurement with temporal pooling scheme for high-definition (HD) NR-VQA. Saad and Bovik [65] studied the potential of using natural motion statistics for NR-VQA. They mainly cared about Internet Protocol (IP) transmission distortion. Ćulibrk *et al.* [66] focused on the MPEG-2 coded video sequences and developed a multi-layer neural network based feature selection scheme for NR-VQA. Huynh-Thu and Ghanbari [67] developed a no-reference temporal quality metric based on a proposed *freeze event* detector to model the impact of frame freezing artifacts on perceived video quality.

- In 2010, Brandão and Queluz [68] proposed a NR-VQA metric for H.264/AVC encoded videos by combining the coding error estimation and corresponding perceptual weighting. Discrete cosine transform (DCT) was used to measure the quantization noise, and spatio-temporal contrast sensitivity function (CSF) was applied to pool the error map into a single score. Hemami and Reibman [69] conducted a survey about the NR-VQA problem and related applications. The proposed three-stage framework provides a potential scheme to facilitate NR-VQA system design. Yang *et al.* [70] proposed a NR-VQA method using transmitted bitstream only. They considered the quantization distortion, packet loss and error propagation, as well as the temporal effects of HVS. Based on the estimation of the spatio-temporal complexity of video content, Liao and Chen [71] developed a packet-layer model for quality monitoring. They also studied the interaction between content features and the influence of error concealment and propagation. Kawano *et al.* [72] proposed a media-layer model through blockiness and blurring detection for compressed video. Specifically for mobile devices, Liu *et al.* [73] presented a real-time quality monitoring system, which provided valuable information for network diagnosis and quality-scalable service planning.

- In 2011, Argyropoulos *et al.* [74] proposed a NR-VQA model for the quality evaluation of SD and HD H.264/AVC sequences distorted by packet loss. Based on continuous estimation of packet loss visibility, support vector regression (SVR) was employed to build the relationship between subjective quality ratings and a set of spatiotemporal features from bitstream. Boujut *et al.* [75] combined spatio-temporal saliency maps with macro-block error detection for HDTV quality estimation. The compressed bitstream doesn't need to be fully decoded in this system. Liu *et al.* [76] studied the effects of video compression on perceived quality and proposed a NR-VQA model considering three artifacts (blurring, blocking, jittering) for luminance and chrominance, separately. Shi and Jiang [77] studied the video quality degradation effected by lossy compression and proposed a fully-decoding-free VQA model based on three factors (average quantization parameters (QP), number of skipped macroblock, average motion vectors).

- In 2012, Valenzise *et al.* [78] developed a NR video quality monitoring approach for packet loss distorted videos which are transmitted through error-prone network. With decoded pixel value only, the described system was able to provide an accurate estimation of mean-square-error distortion introduced by channel errors. Wang *et al.* [79] claimed that the subjective quality of MPEG-2 encoded video is best correlated with three features, including quantiser-scale factor, bit rate, and statistics of intra macroblock. Lin *et al.* [80] proposed to estimate the video quality by measuring the effect of blockiness and blur distortions. The distortion measure is incorporated with region of interest (ROI) which is identified by bitstream information and HVS characteristics. Boujut *et al.* [81] considered the semantics of the visual scene for a bottom-up spatio-temporal saliency map enhancement and developed a NR-VQA model for broadcasted HD video over IP networks. Yao *et al.* [82] proposed to measure the spatial distortion for individual frame using statistics of wavelet coefficients and temporal distortion using a motion-compensated approach based on block and motion vector. Bailey *et al.* [83] presented a full analytic NR-VQA model for pause intensity, which is based on the video playout buffer behavior at the receiver side.

- In 2013, Zhang *et al.* [84] tried to solve the NR-VQA problem using additive log-logistic model (ALM) which adds the distortions due to each type of impairment in a log-logistic transformed space of subjective opinions. A large amount of features, which may effect the perceived quality, were investigated. The final selected key features include average QP, visible error rate, freezing duration, spatial-temporal complexity of the video clip, as well as the mean of motion vectors. Staelens *et al.* [85] proposed a bitstream-based NR-VQA model that is constructed by genetic programming-based symbolic regression. They studied 42 different parameters extracted from bitstream to characterizing the encoding settings, type of distortions, and video content characteristics. It was reported that only 20% of those parameters significantly contributes to the final VQA, including temporal duration of distortion, percentage of slices lost of the picture where the loss originates, loss originates from I- and P-picture, slices per picture, number of B-pictures, number of consecutive slice drops. For HEVC encoded video, Lee and Kim [86] proposed a reference-free PSNR estimation approach based on Laplacian mixture probability density function, which characterized the distribution of transformed residual coefficients in different quadtree depths.

2.3 Reduced-reference Video Quality Assessment

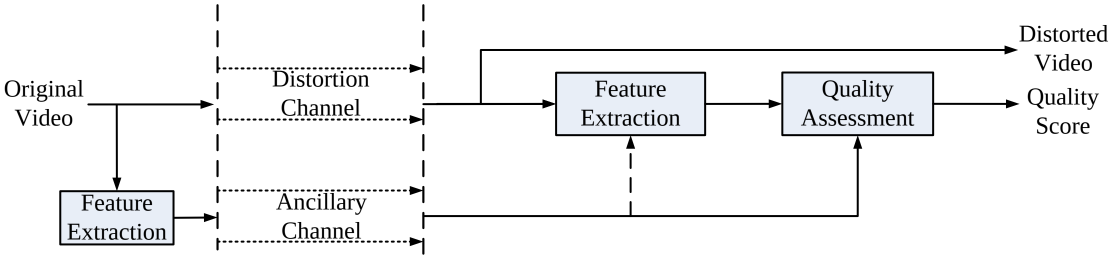


Figure 2.3: Framework of reduced-reference video quality assessment.

The FR-VQA approaches may not applicable in visual communication scenarios, because full access to the original video is expensive or not available. Meanwhile, NR-VQA, especially general-purpose NR-VQA, is extremely difficult to design due to our limited knowledge about HVS and video signal statistics. Reduced-reference (RR) VQA measure provides a compromised solution, which evaluates video quality with only partial information about the original video. One or more RR features are extracted from the original video at the sender side [13] and transmitted to the receiver side through an extra ancillary channel. The general framework of RR-VQA is showed in Fig. 2.3. The most challenging task in the design of RR-VQA is to find appropriate RR features that 1) provide an efficient summary of the reference video; 2) are sensitive to the targeted types of video distortions; 3) are relevant to the perceptual characteristics of the HVS; and 4) have relatively low data rate (so that they do not add too much burden to the visual communication systems that need to transmit the RR features) [9]. Most existing RR-VQA models are developed and trained for specific applications such as lossy compression [13]. This makes the design task easier because the distortion types are known and fixed. However, it also significantly limits their application scope at the same time.

- In 2004, Pinson and Wolf [87] proposed a highly complex VQA scheme, known as national telecommunications and information administration (NTIA) General Model, which, along with its associated calibration techniques, has been accepted as a North America Standard for objective video quality estimation. Its competitive performance has been demonstrated on the VQEG FR-TV Phase II data set. However, the major drawbacks of this approach include that 1) a large number of parameters need to be trained beforehand, and 2) a lossless ancillary channel is required, which is hard to satisfy in practical scenarios.

- In 2006, LeCallet *et al.* [88] employed a convolutional neural network to explore the nonlinear relationship between subjective quality scores and RR features. The MOS were obtained from a subjective test with single stimulus continuous quality evaluation protocol. The considered RR features include power of frame difference, blockiness measure, blurring detection, and tiling evaluation.

- In 2007, Oelbaum and Diepold [89] employed multivariate data analysis to combine a set of detected features (blurring, blocking) for RR-VQA of H.264/AVC encoded sequences.

- In 2008, Gunawan and Ghanbari [90] used a discriminative analysis of harmonic strength, combined with motion information for weighting, for RR-VQA. Later, they [91] also presented an efficient RR-VQA approach for continuous quality monitoring based on local harmonic strength (LHS), in which harmonics gain and loss corresponding to blockiness and blurriness.

- In 2010, Lu *et al.* [92] employed three-dimensional wavelet transform, combined with spatio-temporal CSF and perceptual threshold, to mimic the characteristics of HVS, including multichannel structure, nonlinearity processing, and distortion tolerance. Temporal perceptual mechanism, namely short-term memory, is also considered for temporal pooling of quality scores. Garcia and Raake [93] described a parametric packet-layer RR-VQA model for HD and SD sequences. The proposed model took into account variable factors, including bit-rate, packet-loss-rate, burstiness factor, as well as the information about the codec configurations.

- In 2011, Wang *et al.* [94] investigated the perceptual video quality score effected by the content features, including AC energy of DCT coefficients for picture activity, spectral entropy for randomness of DCT coefficients, the percentage of intra coded macroblock and skipped macroblock, bit-rate, as well as the mean and standard deviation of quantizer-scale factors over each frame and the whole video. They reported that a statistic of proportion of skipped block possesses the best correlation with subjective quality. Based whether or not the video quality was effected by the contents, Yang *et al.* [95] categorized quality-related features into two classes: (1) semantically dependent and (2) semantically independent. They reported that the semantic-independent features showed more promising performance in terms of subjective quality estimation. The investigated semantic-independent features include motion space, hand-shaking, color harmonic, and composition. Meanwhile, the semantic-dependent features include motion direction entropy, color saturation and value, and lightness.

- In 2012, Niu and Liu [96] tackled the problem of what makes a professional video and proposed a computational aesthetics approach for RR-VQA. A variety of features were selected to distinguish professional videos from amateur ones, including noise, focus control, exposure control, color palette, camera motion, shot length, and visual continuity. Ma *et al.* [97] proposed a RR-VQA approach by exploiting the spatial information loss and temporal statistical characteristics of the inter-frame histogram. Energy variation descriptor is employed to evaluate the individual frame quality and simulate the texture masking property of HVS. The temporal quality degradation was captured by the city-block distance between generalized Gaussian density distribution of reference and distorted video. Yang *et al.* [98] designed a content-adaptive packet-layer model for RR-VQA of networked video services. Only packet headers are employed for real-time and non-intrusive video quality monitoring. Aiming for the compression artifacts and packet losses, the proposed model was composed by information extracted from packet headers, frame type detection, temporal complexity estimation, as well as a two-level temporal pooling strategy. Atzori *et al.* [99] proposed an efficient visual quality estimator based on probability of starvation. It was embedded into a wireless channel based video streaming system for source rate control. This approach allows the video streaming provider adjust the system settings automatically to optimize the user-perceived video quality. Karacali and Krishnakumar [100] focused on the video conference-type applications and proposed a real-time RR-VQA scheme based on the face detection and discrepancies of face location between sent and received video frames. Ou *et al.* [101] thoroughly studied the impact of spatial, temporal, and amplitude resolution on perceived video quality and proposed an effective RR-VQA model, Q-STAR, with high performance on several databases. Q-STAR is consisted by three three different models to account for the relationship between normalized subjective quality with spatial resolution, quantization, and temporal resolution.

2.4 Video Denoising

Existing video denoising algorithms can be roughly classified into three categories. In the first category, the video signal is denoised on a frame-by-frame basis, where all that is needed is a 2D still image denoising algorithm applied to each frame of the video sequence independently. Well-known and state-of-the-art still image denoising algorithms include spatially adaptive Wiener filtering [102], Bayes least square estimation based on Gaussian scale mixture model (BLS-GSM) [103], non-local means denoising (NLM) [104], K-SVD method [105], Stein's unbiased risk estimator-linear expansion of threshold algorithm (SURE-LET) [106], and block matching and 3D transform shrinkage method (BM3D) [107]. For the purpose of video denoising, the major advantage of these approaches is memory efficiency, as no storage of previous frames are necessary in order to denoise the current frame. However, since the correlation between neighboring frames is completely ignored, the denoising process does not make use of all available information and thus cannot achieve the best denoising performance.

In natural video signals, there exists strong correlation between adjacent frames. The second category of video denoising approaches exploited such correlation by incorporating both intra- and inter-frame information. It was found that motion estimation and compensation could further enhance denoising performance [108, 109, 110]. In [108], a motion estimation algorithm was employed for recursive temporal denoising along estimated motion trajectory. Motion compensation processes had also been incorporated into BLS-GSM and SURE-LET methods, leading to the ST-GSM [109] and video SURE-LET algorithms [110]. In [111], it was claimed that finding single motion trajectory may not be the best choice for video denoising. Instead, multiple similar patches in neighboring frames are found that may not reside along a single trajectory. This is followed by transform and shrinkage based denoising procedures. One of the most successful video denoising methods in recent years is the extension of BM3D method for video, namely VBM3D [112], which searches similar patches in both intra- and inter-frames and uses 3D bilateral filtering for noise removal after aggregating the similar patches together.

The third category of denoising algorithms treat video sequences as 3D volumes. The algorithms can operate in the space-time domain by adaptive weighted local averaging [113], 3D order-statistic filtering [114], 3D Kalman filtering [115], or 3D Markov model based filtering [116]. They may also be applied in 3D transform domain, where soft/hard thresholding or Bayesian estimation are employed to eliminate noise, followed by an inverse 3D transform that brings the signal back to the space-time domain. The method in [117] is one such example, where 3D dual-tree complex wavelet transform was employed that demonstrates some interesting and desired properties. Recently, several authors investigated 3D-patch based methods and achieved highly competitive denoising performance [118, 119].

Ideally, to make the best use of all available information, the best video denoising algorithms would need to operate in 3D (Category 3). However, when there exists significant motion in the video, direct space-time 3D filtering or 3D transform based approaches are difficult to effectively cover all motion-related video content within local region. Meanwhile, 3D-patch based methods are expensive in finding similar 3D-patches in the 3D volume. By contrast, 2D denoising algorithms that use intra- and/or inter-frame information (Categories 1 and 2) can be made much more efficient, but their performance is restricted by not fully making use of the neighboring pixels in all three dimensions simultaneously.

Chapter 3

Full-reference Video Quality Assessment

In this chapter, we mainly focus on full-reference video quality assessment (FR-VQA) approach. The design of FR-VQA algorithms depends on how a video signal is interpreted. If we consider it as a stack of still images, a natural approach is to apply still image quality assessment (IQA) algorithms on a frame-by-frame basis, followed by pooling the frame level quality measures into a single quality score. However, this approach missed the temporal correlation between adjacent frames. Specifically, it disregards the motion information, which is the most critical characteristic that distinguishes a video sequence from a stack of independent still image frames. As a result, advanced FR-VQA algorithms take into account the temporal correlation or motion information. This can be done by combining multichannel spatio-temporal filtering and spatio-temporal just noticeable difference (JND) models [120, 50]. It can also be implemented by block- or optical flow-based motion estimation, followed by weighted pooling based on models of human visual motion perception [37, 43]. More sophisticated method combines both spatio-temporal filtering and motion estimation, and then incorporates both spatial and temporal distortion measures [45].

In this study, we consider a video signal as a 3D volume image and define a "region" in the image as a localized 3D block. We can then generate a 3D quality map by applying a block-wise quality measure within local regions. This is followed by a pooling stage that merges the quality map into an overall quality score. Recently, pooling has become an active research topic in IQA/VQA research. Most existing methods are based on the hypothesis that the regions that are more likely to attract visual attention should be assigned larger weights. The critical issue here is how visual attention is predicted, which may include a spectrum of approaches, ranging from saliency-based low-level vision models [3] to motion detection and object tracking based high-level cognitive methods [39, 45, 49, 121]. In [3], a number of different pooling strategies were compared in the context of IQA. It was found that the approaches that lead to the most significant performance gain are local information content and local distortion weighted pooling, which are based on the assumptions that the image regions that contain more information (computed based on statistical image models) or more severe distortions are more likely to attract visual attention. Moreover, these methods can be implemented with low computational cost, which is often an important factor in real world deployment of VQA techniques. In this research, we extend these pooling strategies to FR-VQA and find that they lead to consistent gain when tested using several independent video quality databases.

3.1 3D Structural Similarity for RR-VQA

The diagram of the proposed method, namely three-dimensional structural similarity (3D-SSIM) algorithm, is shown in Fig. 3.1. The input reference and distorted videos are first divided into non-overlapping 3D blocks. Within each block, a local

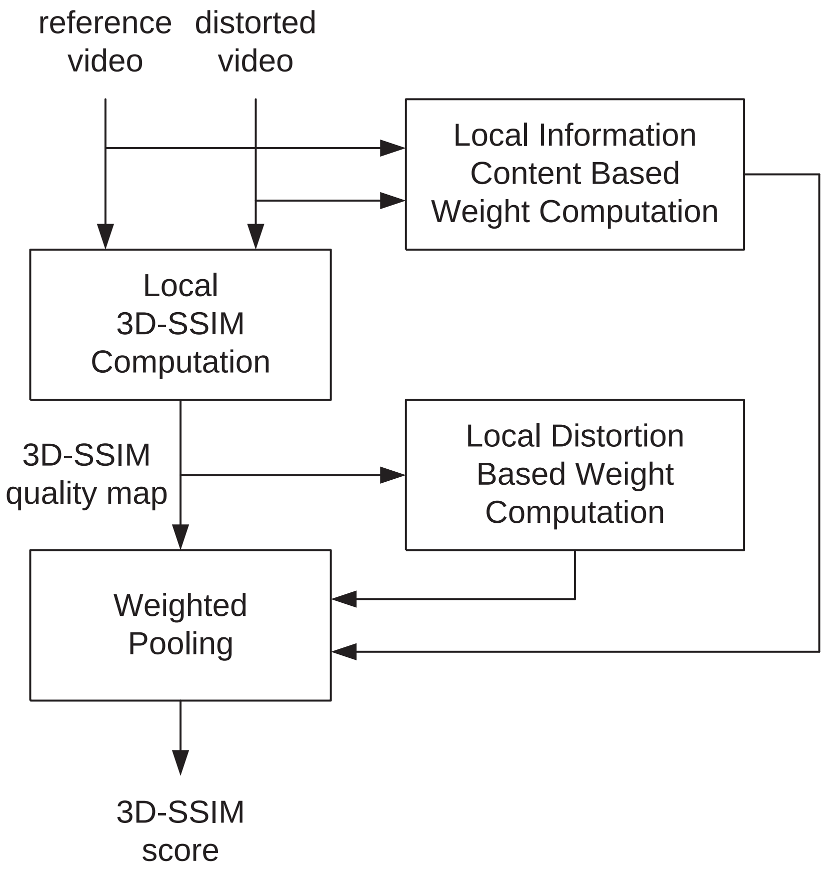

reference video
distorted video
Local Information Content Based Weight Computation
Local 3D-SSIM Computation
3D-SSIM quality map
Local Distortion Based Weight Computation
Weighted Pooling
3D-SSIM score

Figure 3.1: Framework of 3D-SSIM algorithm.

3D-SSIM measure and a local information content measure are computed. The local 3D-SSIM values collected from all blocks form a 3D quality map of the video, which are used to compute a local distortion-based weight map. Both the local information content and local distortion based weights are involved in the weighted pooling stage of the 3D-SSIM map, resulting in an overall 3D-SSIM score.

Let $\mathbf{x} = \{x_i | i = 1, \cdots, N\}$ and $\mathbf{y} = \{y_i | i = 1, \cdots, N\}$ be two sets of pixel values collected from corresponding 3D blocks from the reference and distorted videos, respectively. As in the spatial domain SSIM method [1], the local 3D-SSIM between the 3D blocks is computed as

$$\mathrm{S}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \, , \tag{3.1}$$

where where μ_x, σ_x^2 and σ_{xy} represent the mean, variance and covariance of the image blocks, respectively, and C_1 and C_2 are small positive constants to avoid instability when the means and variances are close to zero.

Effective estimation of perceptual information content relies on good statistical models of both natural images and perceptual distortion channels [3]. While sophisticated models such as the Gaussian scale mixtures [3] are available for still images, they often lead to substantially increased complexity, which becomes a major barrier to overcome when applied to large volume video data. To achieve a good compromise between accuracy and simplicity, here we assume a simple model, where Gaussian distributed image source passes through an additive Gaussian channel and the mutual information between the source and received signals is employed to quantify the perceived information content. When this model is applied to local 3D image blocks of both the reference and distorted video signals, a simple computational model of the overall perceptual information content is given by [122]

$$w_{ic}(\mathbf{x}, \mathbf{y}) = \frac{1}{2} \log \left[\left(1 + \frac{\sigma_x^2}{\sigma_0^2}\right) \left(1 + \frac{\sigma_y^2}{\sigma_0^2}\right) \right] , \tag{3.2}$$

where, as in [123], σ_0^2 is a constant that accounts for the noise power of the additive Gaussian channel. This measure is computationally efficient because the values of σ_x^2 and σ_y^2 are readily available in the local 3D-SSIM computation.

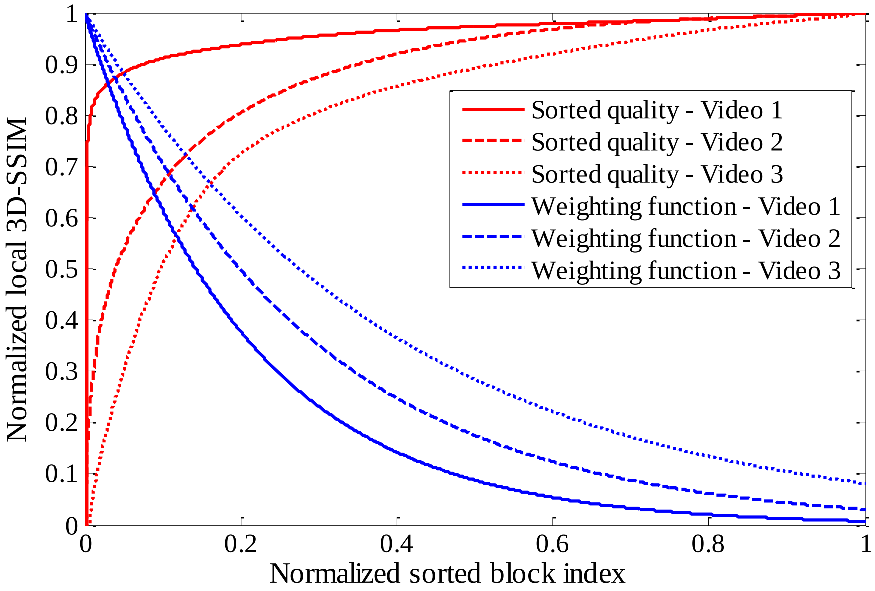


Figure 3.2: Samples of sorted local 3D-SSIM curves and local distortion based weighting functions.

Previous studies had shown that assigning larger weights to higher distortion regions generally has positive effect on the performance of IQA/VQA algorithms [122, 3, 121]. In Fig. 3.2, the local 3D-SSIM measures computed from different regions are sorted in ascending order for three different distorted video sequences. It can be observed that the shapes of the ascending curves vary for different video sequences, which may depend on the nature of the videos as well as the type and level of the distortions. It was demonstrated in [121] the usefulness of adapting the weight assignment strategy based on the shape. In this work, we propose to use a width-adapted exponential weighting function applied upon sorted block index.

Assume that there are totally K 3D blocks extracted from the video, and let $\mathbf{y}_k$ be the block with the k-th lowest local 3D-SSIM value. The local distortion-based weighting function is defined upon the normalized index $\alpha_k = k/K$ by

$$w_d(\mathbf{y}_k) = e^{-\frac{|\alpha_k|}{\alpha_0}}\,, \tag{3.3}$$

where α_0 is a width parameter that controls the speed of falloff of the exponential function. As shown in Fig. 3.2, the ascending speeds of the sorted local 3D-SSIM curves vary for different video sequences. This motivates us to adapt the weighting function accordingly which can be readily implemented by adjusting α_0. Specifically, we preset an S^* parameter on the normalized 3D-SSIM value and find the corresponding block index α^* value on the sorted 3D-SSIM curve. We then compute the α_0 parameter by

$$\alpha_0 = \beta\alpha^*\,, \tag{3.4}$$

where β is a scaling parameter to control the relative widths of the sorted 3D-SSIM curve and the weighting function. Examples of the weighting functions computed based on the sorted 3D-SSIM curves are shown in Fig. 3.2.

Finally, the local 3D-SSIM map is pooled based on both local information content and local distortion based weighting and the overall 3D-SSIM measure of the entire video sequence is given by

$$\text{3D-SSIM} = \frac{\sum_{k=1}^{K}[w_{ic}(\mathbf{x}_k,\mathbf{y}_k)]^{\mu}[w_d(\mathbf{y}_k)]^{\nu}S(\mathbf{x}_k,\mathbf{y}_k)}{\sum_{k=1}^{K}[w_{ic}(\mathbf{x}_k,\mathbf{y}_k)]^{\mu}[w_d(\mathbf{y}_k)]^{\nu}}. \tag{3.5}$$

where μ and ν are two parameters used to control the relative importance of the two weighting functions.

Table 3.1: Specifications about the tested VQA databases. SRC denotes the number of source reference videos and HRC denotes the number of distorted videos created from each source video.

Database	# of video	SRC	HRC	Resolution
VQEG FR-TV I	320	20	16	480i, 576i
IRCCyN/IVC	192	24	7	720×576
EPFL-PoliMI	156	16	9	CIF, 4CIF
LIVE	150	10	15	768×432p

3.2 Implementation and Experiment

The implementation details of the proposed 3D-SSIM algorithm are as follows. As in the default SSIM implementation [124], the input reference and distorted video signals first go through an automatic downsampling (or auto-scale) process on a frame-by-frame basis. This is followed by dividing the 3D volume image into non-overlapping $7 \times 7 \times 7$ blocks, within which the local 3D-SSIM measure (3.1), the local information content weighting function (3.2), and the local distortion weighting function (3.3) are calculated. The parameters C_1, C_2 and σ_0^2 are the same as in the default SSIM [124] and VIF [123] implementations. The other parameters are obtained empirically to optimize the performance on the EPFL-PoliMI VQA database and are given by $S^* = 0.95$, $\beta = 0.4$, $\mu = 4.5$ and $\nu = 1$, respectively. Our simulation shows that the values of those parameters are stable across different VQA databases and will have slightly change if trained on other databases. The information content weights go through another normalization step so that its value is between 0 and 1 before being plugged into the final computation of the overall 3D-SSIM measure.

The proposed approach was tested on four publicly available VQA databases, as described in Table 3.1, where the main distortion types include standard video compression (MPEG and H.264) at different bit rates and simulated transmission errors. The subjective scores are in the form of either mean opinion score (MOS) or difference of mean opinion score (DMOS) (difference between the MOS values of the reference and distorted videos). The following two evaluation metrics are adopted to compare the performance of different VQA measures[125, 126, 127].

- Pearson Linear correlation coefficient (PLCC) after a nonlinear mapping between the subjective and objective scores. For the i-th image in an image database of size N, given its subjective score o_i (MOS or DMOS between reference and distorted images) and its raw objective score r_i, we first apply a nonlinear function to r_i given by [126]

  $$q(r) = a_1 \left\{ \frac{1}{2} - \frac{1}{1 + \exp[a_2(r - a_3)]} \right\} + a_4 r + a_5 \,, \tag{3.6}$$

  where a_1 to a_5 are model parameters found numerically using a nonlinear regression process in MATLAB optimization toolbox to maximize the correlations between subjective and objective scores. The PLCC value can then be computed as

  $$\text{PLCC} = \frac{\sum_i (q_i - \bar{q}) * (o_i - \bar{o})}{\sqrt{\sum_i (q_i - \bar{q})^2 * \sum_i (o_i - \bar{o})^2}} \,. \tag{3.7}$$

- Spearman's rank correlation coefficient (SRCC) is defined as:

  $$\text{SRCC} = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \,, \tag{3.8}$$

  where d_i is the difference between the i-th image's ranks in subjective and objective evaluations. SRCC is a non-parametric rank-based correlation metric, independent of any monotonic nonlinear mapping between subjective and objective scores.

PLCC is adopted to evaluate *prediction accuracy* [125], and SRCC is employed to assess *prediction monotonicity* [125]. A better objective VQA measure should have higher PLCC and SRCC values.

Table 3.2: PLCC performance comparison of VQA algorithms

Database	VQEG	IRCCyN	EPFL-PoliMI	LIVE
PSNR	0.7683	0.4160	0.7351	0.5621
2D-SSIM [124] (auto-scale)	0.8113	0.6139	0.6770	0.7177
VQM [87]	0.8170	0.4850	0.8434	0.7236
MOVIE [45]	0.8210	0.4850	0.9210	0.8116
You *et al.* [49]	0.8170	0.7680	0.9470	**0.8450**
2D-SSIM [1] (no weighting)	0.8215	0.5012	0.6781	0.5444
2D-SSIM (w_{ic} only)	0.8301	0.5206	0.7685	0.5985
2D-SSIM (w_d only)	0.8297	0.5827	0.8716	0.7062
2D-SSIM (with both weighting)	0.8311	0.6612	0.9092	0.7621
3D-SSIM (no weighting)	0.8079	0.6212	0.7591	0.7026
3D-SSIM (w_{ic} only)	0.8203	0.7357	0.8136	0.7497
3D-SSIM (w_d only)	0.8295	0.7209	0.9091	0.7832
3D-SSIM (with both weighting)	**0.8403**	**0.8194**	**0.9621**	0.8353

The evaluation results in terms of PLCC and SRCC are given in Tables 3.2 and 3.3, respectively. First, the proposed 3D-SSIM approach in (3.5) is compared with other pooling options (that are based on the same local 3D-SSIM map), where no weighting or only one of the weighting approaches (w_{ic} in (3.2) or w_d in (3.3) only) is applied. Apparently, either information content or distortion based weighting scheme significantly improves upon the no-weighting case and the best results are obtained when both of them are applied. The proposed 3D-SSIM algorithm is

Table 3.3: SRCC performance comparison of VQA algorithms

Database	VQEG	IRCCyN	EPFL-PoliMI	LIVE
PSNR	0.7714	0.4510	0.7440	0.5398
2D-SSIM [124] (auto-scale)	0.7919	0.6058	0.6949	0.6947
VQM [87]	0.7760	0.4820	0.8383	0.7026
MOVIE [45]	0.8330	0.5930	0.9200	0.7890
Yu *et al.* [49]	0.8030	0.7910	0.9450	0.8180
2D-SSIM [1] (no weighting)	0.7880	0.5126	0.6770	0.5257
2D-SSIM (w_{ic} only)	0.7941	0.5382	0.7655	0.5752
2D-SSIM (w_d only)	0.7917	0.5971	0.8612	0.6878
2D-SSIM (with both weighting)	0.8085	0.6301	0.9034	0.7490
3D-SSIM (no weighting)	0.7804	0.6147	0.7483	0.6810
3D-SSIM (w_{ic} only)	0.8147	0.7143	0.8003	0.7397
3D-SSIM (w_d only)	0.8208	0.7012	0.9016	0.7712
3D-SSIM	**0.8396**	**0.7916**	**0.9608**	**0.8244**

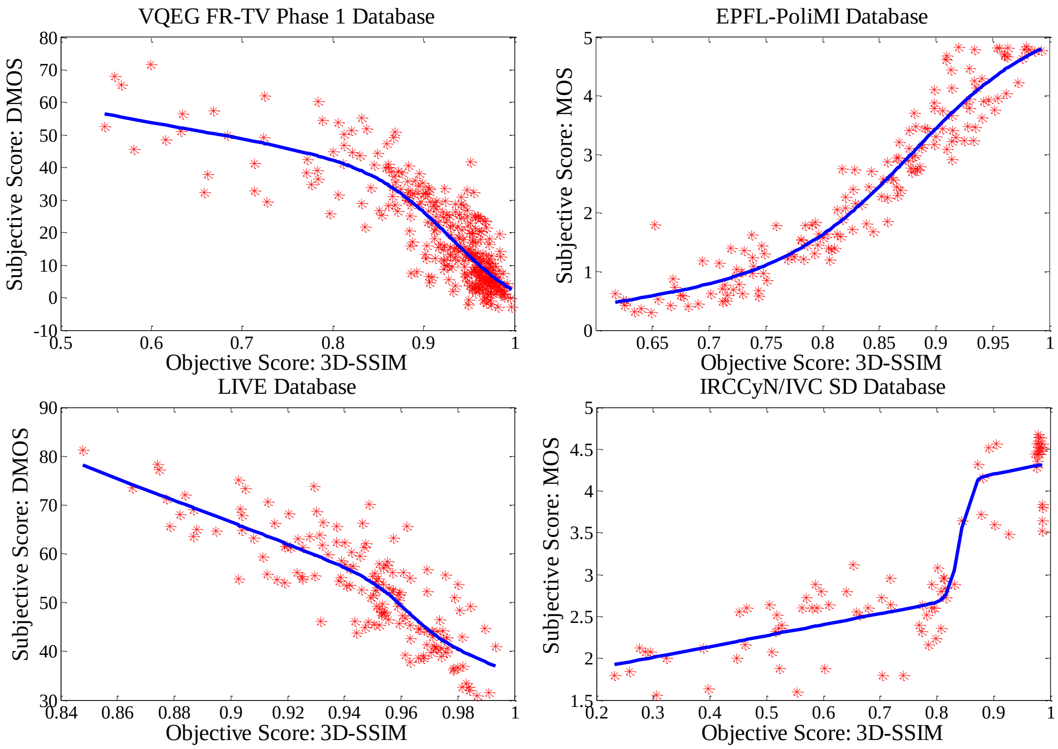


Figure 3.3: Scatter plots of 3D-SSIM versus subjective score for four VQA databases.

also compared with six other VQA approaches: peak signal-to-noise-ratio (PSNR), direct 2D-SSIM [1], 2D-SSIM with auto-scaling [124], video quality model (VQM) [87], motion-based video integrity evaluation index (MOVIE) [45], and a most recent method proposed by Yu *et. al* [49]. In addition, the results of applying 2D version of two weighting approaches to direct 2D-SSIM are also included in two tables, so that the effect of weighted pooling can be better examined. The best results obtained for each database are highlighted in bold. It can be observed that 3D-SSIM appears to be the most reliable measure across all four databases and achieves the best performance in most cases. The scatter plots of 3D-SSIM values versus subjective quality scores over the four databases, together with the nonlinear fitting functions, are shown in Fig. 3.3.

It is worth emphasizing that the highly competitive performance of 3D-SSIM is obtained with vastly reduced computational complexity. Our Matlab implementation of the 3D-SSIM algorithm takes around 4.64 seconds (excluding data loading time) to evaluate a video sequence of 768×432 in spatial resolution and 217 frames in length on a computer with Intel Core2 Duo CPU E8600 processor at 3.33GHz. This is estimated to be only less than 1% and 0.1% of the well known VQM [87] and MOVIE [45] algorithms, respectively. This could be a critical advantage in many real world applications.

3.3 Summary

In this chapter, a novel FR-VQA algorithm, namely 3D-SSIM, is proposed. It regards a video signal as a 3D volume image, which can be further divided into multiple local regions. Localized SSIM is employed to create a 3D quality map, which is further merged into a single quality score using two pooling strategies. The first one is local information content weighted pooling, which is calculated based on the assumption of Gaussian distribution of source signal and channel noise. The second one is local distortion based pooling method, which is based on the philosophy that more severe distortion would attract more attention. The resulting 3D-SSIM measure is computationally efficient and achieves highly competitive performance when compared with state-of-the-art VQA approaches. One potential drawback of the proposed approach is the memory requirement to store 3D volume data. This problem may be alleviated by dividing the video sequence into segments based on the size of the 3D block involved in the computation. Additionally, the proposed FR-VQA approach is directly useful in video codec and video processing system development, but may require excessive computational power to be applied in real-time communication systems, such as video streaming and conferencing, due to its complexity. In the future, the proposed method may be improved by incorporating more accurate statistical models in the estimation of local information content and investigating more advanced adaptive strategies for local distortion based pooling. In addition, because all local quality measure and weights calculation are conducted in 3D block, the size of 3D block should be adaptive to the spatial and temporal resolution of target videos, so that the frame size and rate can be better accounted for.

Chapter 4

Reduced-reference Video Quality Assessment

Objective FR-VQA models typically require the full access to the reference video that is assumed to have perfect quality. In practical visual communication applications, such methods may not be applicable because the reference video are unavailable [13]. On the other hand, NR-VQA is extremely difficult, especially when the types of distortions between senders and receivers are unknown [13]. Reduced-reference video quality assessment (RR-VQA) methods provide solutions that lies between FR and NR models. They are designed to evaluate the visual quality of the distorted video with only partial information about the reference video. One difficulty in the deployment RR-VQA approaches is that they require the RR features to be transmitted to the receiver through a lossless ancillary channel [13], which is often hard to provide in real-world application environment. This motivated the ideas of quality-aware image (QAI) [9] and quality-aware video (QAV) [128], where the extracted RR features are embedded into the original image/video signal as invisible messages and transmitted to the receiver together with the image/video content. In this chapter, we develop a novel RR-VQA measurement and further construct a more effective QAV system.

4.1 Temporal Motion Smoothness Measurement for RR-VQA

Because of the high correlation between adjacent frames in natural video signals, we explore temporal statistics using phase difference in complex wavelet domain. The simulation demonstrates that it is effective to be employed in an RR-VQA framework as a feature to capture several kinds of common quality degradations.

4.1.1 Temporal Motion Smoothness

Let $f(x)$ be a given real static signal, where x is the index of spatial position. When $f(x)$ represents an image, x is a 2-D vector. For simplicity, in the derivations below, we assume x to be one dimensional. However, the results can be easily generalized to two and higher dimensions. A time varying image sequence can be created from the static image $f(x)$ with rigid motion and constant variations of average intensity:

$$h(x,t) = f(x + u(t)) + b(t)\,. \tag{4.1}$$

Here $b(t)$ is real and accounts for the time-varying background luminance changes, and $u(t)$ indicates how the image positions move spatially as a function of time. We call the motion N-th order smooth if the $(N+1)$-th and higher order derivatives of $u(t)$ with respect to t are all zeros [8]. Because the video is transformed into complex wavelet domain, this assumption is valid for the local regions covered by a wavelet envelop.

Now consider a family of symmetric complex wavelets whose "mother wavelets" can be written as a modulation of a low-pass filter $w(x)= g(x)\, e^{j\omega_c x}$, where ω_c is the center frequency of the modulated band-pass filter, and $g(x)$ is a slowly varying and symmetric function. The family of wavelets are dilated/contracted and translated versions of the mother wavelet: $w_{s,p}(x) = \frac{1}{\sqrt{s}}\, w\left(\frac{x-p}{s}\right)$, where $s \in R^+$ is the scale factor, and $p \in R$ is the translation factor. Using the convolution theorem and the scaling and modulation properties of the Fourier transform, we can compute the complex wavelet transform of $f(x)$ as

$$\begin{aligned} F(s,p) &= \int_{-\infty}^{\infty} f(x)\, w^*_{s,p}(x)\, dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega p}\, d\omega\,, \end{aligned} \tag{4.2}$$

where $F(\omega)$ and $G(\omega)$ are the Fourier transforms of $f(x)$ and $g(x)$, respectively. Applying such a complex wavelet transform to both sides of Eq. (4.1) at time instance t, we have

$$\begin{aligned} H(s,p,t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\sqrt{s}\, G(s\,\omega - \omega_c)\, e^{j\omega(p+u(t))}\, d\omega \\ &\approx F(s,p)\, e^{j(\omega_c/s)u(t)}\,. \end{aligned} \tag{4.3}$$

Here $b(t)$ is eliminated because of the bandpass nature of the wavelet filters. The approximation is valid when the envelope $g(t)$ is slowly varying and the motion $u(t)$ is small. A more convenient way to understand Eq. (4.3) is to take a logarithm on both sides, which gives

$$\log H(s,p,t) \approx \log F(s,p) + j(\omega_c/s)u(t)\,. \tag{4.4}$$

The key point here is that at a given scale s and a given spatial position p, the first term is a constant and the imaginary part of the logarithm of the complex wavelet coefficient changes linearly with $u(t)$. In other words, the local phase structures over time can be fully characterized by the movement function $u(t)$.

In order to relate temporal motion smoothness with the time-varying complex wavelet transform relationship, we must examine the complex wavelet coefficients at multiple time instances. A convenient choice is to start from a time instance t_0 and sample the sequence at consecutive time steps $t_0 + n\Delta t$ for $n = 0, 1, ..., N$. We define the N-th order temporal correlation function as [8]

$$L_N(s,p) = \sum_{n=0}^{N} (-1)^{n+N} \binom{N}{n} \log H(s,p,t_0 + n\Delta t)\,. \tag{4.5}$$

When the motion is (N-1)-th order smooth, i.e., $u^{(N)}(t_0) = 0$, then it can be derived that $L_N(s,p) \approx 0$ [8]. It needs to be kept in mind that this approximation is achieved based on the ideal formulation of Eq. (4.1) and the ideal assumption of (N-1)-th order temporal motion smoothness. Real natural image sequences are expected to deviate from these assumptions. However, by looking at the statistics of the imaginary part of $L_N(s,p)$, one may be able to quantify such deviation and use it as an indicator of temporal motion smoothness.

As a counterpart of the temporal correlation function $L_N(s,p)$, we can also define a temporal energy function

$$M_N(s,p) = \sum_{n=0}^{N} \binom{N}{n} \log H(s,p,t_0 + n\Delta t)\,, \tag{4.6}$$

which is useful for us to observe the strength of temporal motion smoothness as a function of local energy. An example of the imaginary part of $L_N(s,p)$ conditioned on the real part of $M_N(s,p)$ is shown in Figure 4.1(a), where each column in the 2-D histogram is normalized to one. The conditional histogram shows strong temporal motion smoothness (when the values of $imag\{L_2(s,p)\}$ are close to zero), and such a statistical regularity becomes stronger with the increase of local signal strength (as the width of the column in the 2D histogram becomes narrower). This is not surprising because small magnitude coefficients typically come from the smooth background regions in an image and are easily disturbed by background noise.

4.1.2 RR Video Quality Assessment

A full RR-VQA system consists of three modules: 1) RR feature extraction at the sender side; 2) Transmission of RR features from the sender to the receiver (maybe through an ancillary channel [13] or through the same channel as video transmission [9, 128]); 3) Feature extraction and quality evaluation of the distorted video at the receiver side. This section focuses on the first and the third modules.

At the sender side, the given reference video sequence is first divided into groups of pictures (GOPs), each containing three consecutive frames. For each GOP, all three frames were decomposed using the complex version [129] of the steerable pyramid [130], an overcomplete wavelet transform that avoids aliasing in subbands. The second order temporal correlation and temporal energy functions $L_2(s, p)$ and $M_2(s, p)$ are then computed for each subband. Instead of using the marginal histogram of $imag\{L_2(s, p)\}$ to quantify temporal motion smoothness (as in [8]), here we extract RR features based on the conditional histogram of $imag\{L_2(s, p)\}$ versus $real\{M_2(s, p)\}$. The reason behind this choice is that temporal motion smoothness is much stronger at high energy coefficients (as can be seen in Figure 4.1(a)), but marginal histogram of $imag\{L_2(s, p)\}$ cannot distinguish such differences and takes all coefficients into equal account. Furthermore, the trend of how temporal motion smoothness varies with the increase of local signal energy provides additional information that can help characterize the reference video. Specifically, we use the circular variance (CV) [131] of each column in the conditional histogram to quantify the spread of the angle variables. For each column, the circular variance is computed as

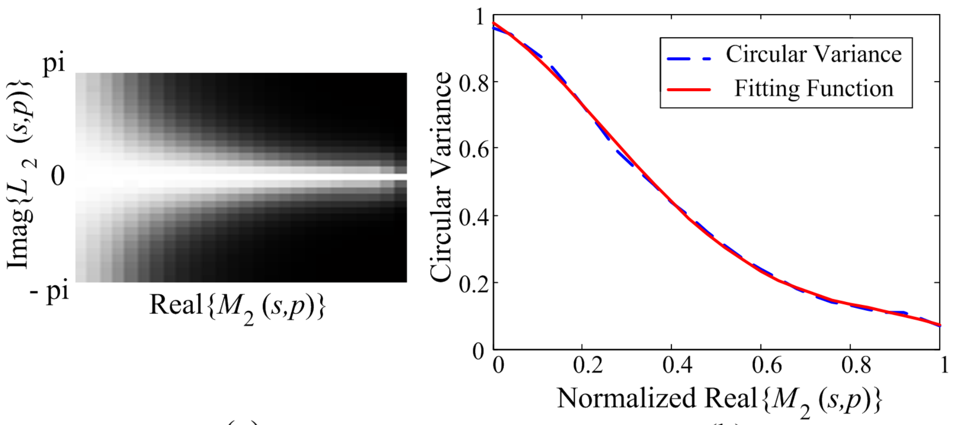
$$CV = 1 - \frac{\left|\sum_{i=1}^{M} h_i e^{j\theta_i}\right|}{\sum_{i=1}^{M} h_i}\,, \tag{4.7}$$

where M is the total number of histogram bins, and h_i and θ_i are the height and center angle of the i-th histogram bin, respectively. The column CV values computed based on the conditional histogram of Figure 4.1(a) are shown in Figure 4.1(b) as a dashed curve, which provides an adequate description about the variation trend of temporal motion smoothness. Depending on the application environment, transmitting the CV curve as the RR features to the receiver may not be a realistic solution because it requires a fairly large RR data rate. To overcome this problem, we use a parametric model to describe the CV curve and only send the model parameters to the receiver. In particular, we find that a fourth order polynomial can very well approximate a typical CV curve, as demonstrated by the solid fitting curve in Figure 4.1(b). Consequently, only 5 parameters (that uniquely define the fourth order polynomial) are employed for every three consecutive frames as RR features and are transmitted to the receiver. They have been further quantized into integer numbers to reduce the necessary transmission data rate.

At the receiver side, the distorted video sequence is processed the same way as at the sender side, i. e., GOP division and complex wavelet signal decomposition, followed by the computation of the conditional histogram and the CV curve. Meanwhile, the received RR features (polynomial parameters) are used to reconstruct the model CV curve. Finally, we quantify the overall video quality distortion as

$$D = \left\{ \frac{1}{K} \sum_{k=1}^{K} [CV(k) - CV_{model}(k)]^2 \right\}^{1/2}\,, \tag{4.8}$$

where K is the total number of columns in the conditional histogram, and $CV(k)$ and $CV_{model}(k)$ are the CV values of the k-th column of the distorted CV curve and



(a) (b)

Figure 4.1: Typical conditional histogram and variation of circular variance. (a) Conditional histogram of $imag\{L_2(s,p)\}$ versus $real\{M_2(s,p)\}$ of a natural video sequence; (b) Variation of circular variance and the best fourth order polynomial fitting.

the model CV curve, respectively. Because the CV values are bounded between 0 and 1, this distortion measure is also bounded by the same range.

4.2 Quality Aware Video based on Intra- and Inter-Frame Features

One of the most significant differences of video from image is the temporal redundant information between frames. Therefore, both intra- and inter- frame knowledge has been considered in the proposed quality-aware video (QAV) system, whose framework is shown in Figure 4.2. Basically, the system consists of three parts: (1) feature extraction for VQA; (2) error control coding and decoding; (3) information hiding by watermarking technique.

4.2.1 RR-VQA Method

In order to capture the video degradation more effectively, both intra- and inter-frame RR features are considered in the complex wavelet transform domain. The marginal distribution of the amplitude of complex wavelet coefficients in each subband can be employed as quality indicator within a frame [132]. Meanwhile, the temporal motion smoothness can be calculated using local phase coherence of consecutive frames as quality indicator along temporal direction [10].

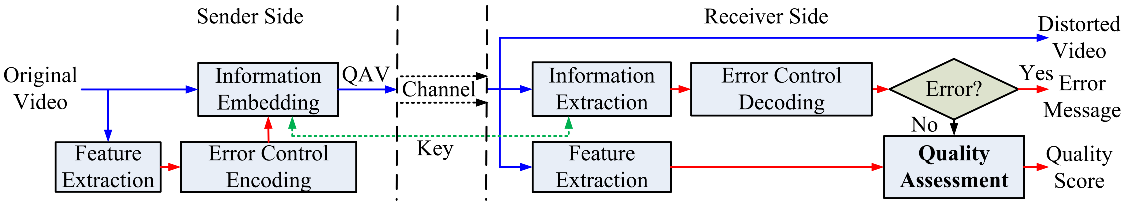


Figure 4.2: Framework of the proposed QAV system.

4.2.1.1 Feature extraction and distortion measure

In [9], it is demonstrated that the distance between the wavelet coefficient distributions of a reference and a distorted image can be used to characterize perceptual degradations. Let $p(x)$ and $q(x)$ denote the probability density functions of the wavelet coefficients in the same subband of the same frame in the reference and distorted images, respectively. The Kullback-Leibler distance (KLD) between them is

$$d(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx \, . \tag{4.9}$$

$q(x)$ can be easily calculated from the distorted frame at the receiver. $p(x)$ needs to be transmitted from the sender. To do that efficiently, it is useful to summarize it using a 2-parameter generalized Gaussian density model that provides a good approximation [9]

$$p_m(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(|x|/\alpha)^{\beta}} \,, \tag{4.10}$$

where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ (for $a > 0$) is the Gamma function. The model approximation error is computed as the KLD between $p_m(x)$ and $p(x)$:

$$d(p_m||p) = \int p_m(x) \log \frac{p_m(x)}{p(x)} dx \,. \tag{4.11}$$

In the end, only three RR parameters, α, β and $d(p_m||q)$, are extracted from each subband. At the receiver side, the intra-frame distortion is computed as an estimate of $d(p||q)$ given by

$$D_{\text{intra}} = \hat{d}(p||q) = d(p_m||q) - d(p_m||p) \,. \tag{4.12}$$

For inter-frame case, we adopted those features introduced in Section 3.1, as equation (4.8) which is rewritten as follows

$$D_{\text{inter}} = \left\{ \frac{1}{N} \sum_{n=1}^{N} [\text{CV}(n) - \text{CV}_{\text{model}}(n)]^2 \right\}^{1/2} \,, \tag{4.13}$$

where N is the number of samples in CV curve, and $\text{CV}(n)$ and $\text{CV}_{\text{model}}(n)$ are the n-th sample computed from the distorted video and the model CV curve, respectively. Finally, the overall distortion is computed as the average of intra- and inter-frame distortions:

$$D = \frac{1}{2}(D_{\text{intra}} + D_{\text{inter}}) \,. \tag{4.14}$$

4.2.2 Robust Information Embedding

Robustness of information embedding is a critical issue to the success of QAV systems. To achieve it, the scalar RR features are first quantized to 7-bit representations, resulting in a binary RR bitstream. The bitstream is then expanded by a 16-bit CRC code for error detection, and then encoded using a binary LDPC code for error correction [133]. The column number of the sparse parity-check matrix of LDPC encoder was designed to be twice of the row number, so that it can correct up to 1 bit of error out of every 2 bits.

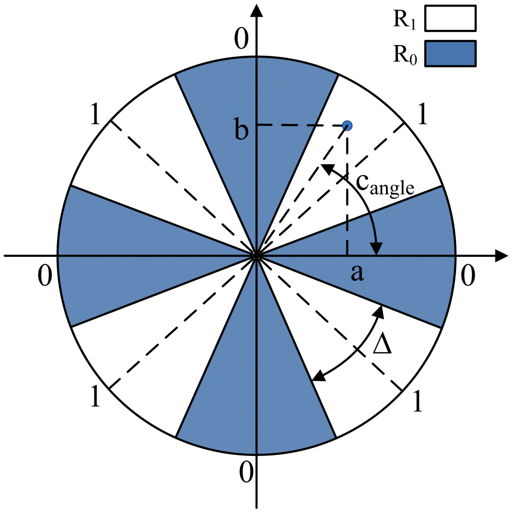


Figure 4.3: Illustration of AQIM for $\Delta = \pi/4$.

The error control coded bitstream is embedded invisibly into the original video using a watermarking scheme. Our method is based on an AQIM approach, which was shown to be highly robust to contrast scaling attacks [11]. The novelty of our scheme is to apply it to pairs of coefficients in 3D-DCT domain, so that it is not only robust to scaling, but also to blur and other types of attacks. An example is illustrated in Figure 4.3, where one bit of information is embedded into the plane composed of two 3D-DCT coefficients. The plane is divided into R_0 and R_1 regions, corresponding to 0 and 1, respectively. The division is based on angular values and the angular quantization step is $\Delta = \pi/4$. Let a and b be the values of a pair coefficients, and $\angle c$ be the angle of the complex number $c = a + jb$. Then the

AQIM embedding scheme is given by an angular quantization operation

$$\begin{aligned} \angle c_{\text{new}} &= Q(\angle c + d(m)) - d(m) \equiv Q^m(\angle c)\,, \\ c_{\text{new}} &= |c| \exp(j \angle c_{\text{new}})\,, \end{aligned} \tag{4.15}$$

where m is the bit being embedded, Q is an angular quantization operator as exemplified by Figure 4.3, c_{new} is the complex coefficient pair after embedding, and $d(m)$ is a dithering operator defined as

$$d(m) = \begin{cases} \quad -\Delta/4, & \text{if } m = 0 \\ \Delta/4, & \quad \text{if } m = 1\,. \end{cases} \tag{4.16}$$

At the receiver side, after a distorted version (denoted as c_d) of the embedded complex coefficient pair c_{new} is received, the embedded bit can be estimated using a minimum angular distance criterion:

$$\hat{m}(\angle c_d) = \underset{m \in \{0,1\}}{\operatorname{argmin}} \| \angle c_d - Q^m(\angle c_d) \|\,. \tag{4.17}$$

3D-DCT often leads to strong energy concentration when applied to natural video signals. As a result, the coefficients corresponding to low spatial and temporal frequencies have much higher energy than that of the high frequency ones. To maximize robustness, we choose the low frequency coefficients for AQIM embedding that are much less sensitive to typical distortions such as compression and noise contamination. Since both 3D-DCT and contrast scaling are linear operators, 3D-DCT domain AQIM is automatically robust to contrast scaling attack because the angular value in Figure 4.3 is invariant to scaling. In addition, the coefficients selected for embedding are paired so that two coefficients that form a pair correspond to the same spatial and temporal frequencies (though may be different in orientation). This is critical to make the AQIM scheme robust to blur attack, because blur causes the two coefficients to scale down by the same ratio, such that the angular value in Figure 4.3 remains unchanged. The value of Δ is tuned to achieve a compromise between robustness and imperceptibility of information embedding. The locations of the selected 3D-DCT coefficients are shared between the sender and receiver as the embedding key, as illustrated in Figure 4.2.

4.3 Experimental Results

4.3.1 Temporal Motion Smoothness for RR VQA

The proposed RR video distortion measure is tested using simulated five distortion types at different distortion levels. These include 1) Gaussian noise contamination, where the distortion level is defined as the standard deviation of the noise; 2) Gaussian blur, where the standard deviation of the Gaussian filter size defines the distortion level; 3) Line jittering, where each line in a frame is shifted horizontally by a random number uniformly distributed between $[-S, S]$, and S defines the jittering level; 4) frame jittering, where the whole frame is shifted together by a random number uniformly distributed between $[-S, S]$; and 5) frame dropping, which is simulated by discarding every 1 of N frames and repeating the previous frame to fill the empty frame, and $12-N$ defines the distortion level. All distortion types are associated with certain real-world scenarios. For example, line jittering occurs when two fields of interlaced video signals are not synchronized; frame jittering is often caused by irregular camera movement such as hand shaking; and frame dropping usually happens when the bandwidth of a real-time communication channel drops and some video frames have to be discarded to reduce the bit rate of the video signal being transmitted.

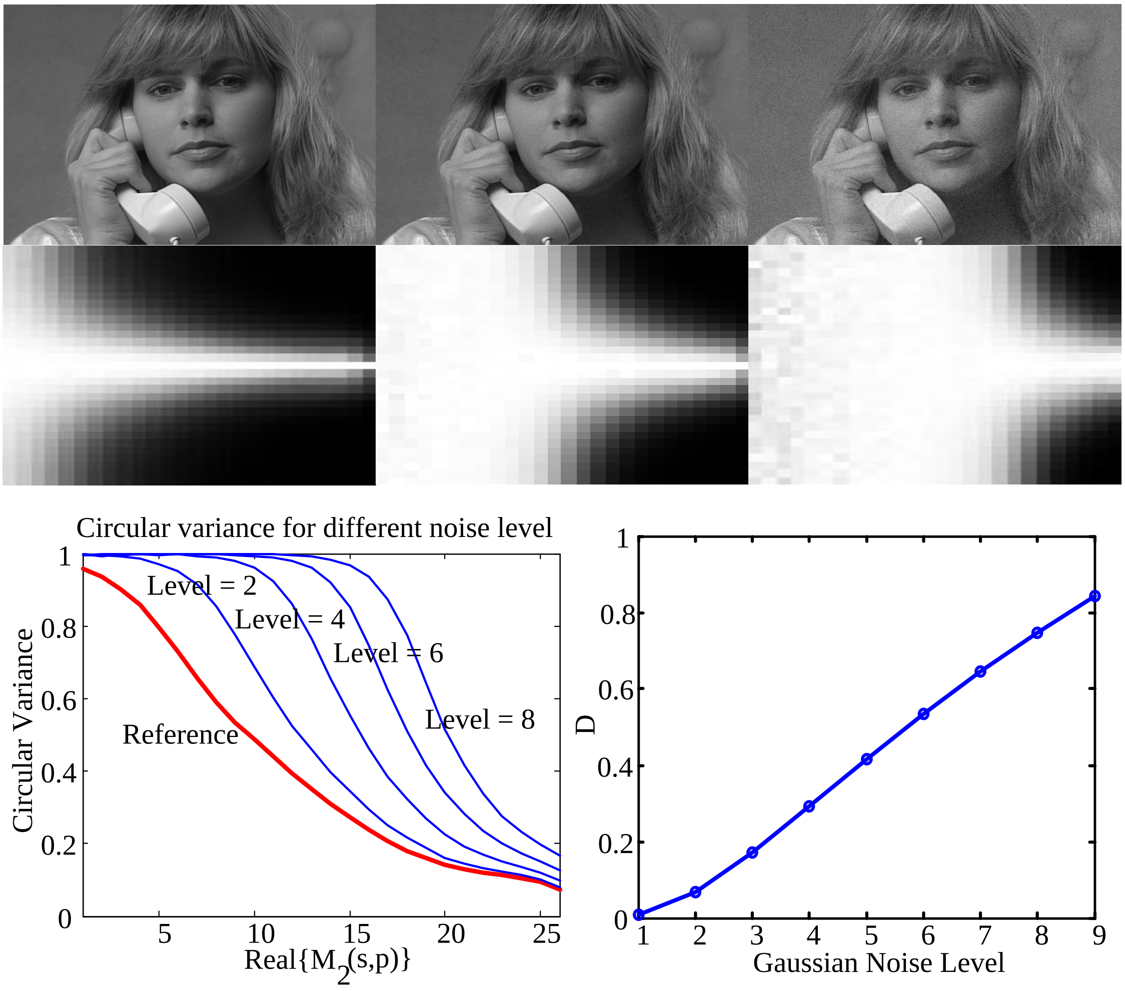


Figure 4.4: Experimental results for proposed RR-VQA approach on Gaussian noise contamination. Top row: three images contaminated by Gaussian noise at low, middle and high levels. Middle row: conditional histograms of $Imag\{L_2(s,p)\}$ versus $Real\{M_2(s,p)\}$ of Gaussian noise contamination at low, middle and high distortion levels; Bottom left: circular variance as a function of $Real\{M_2(s,p)\}$ for the reference video sequence and distorted sequences at different distortion levels; Bottom right: proposed distortion measure as a function of Gaussian noise contamination level.

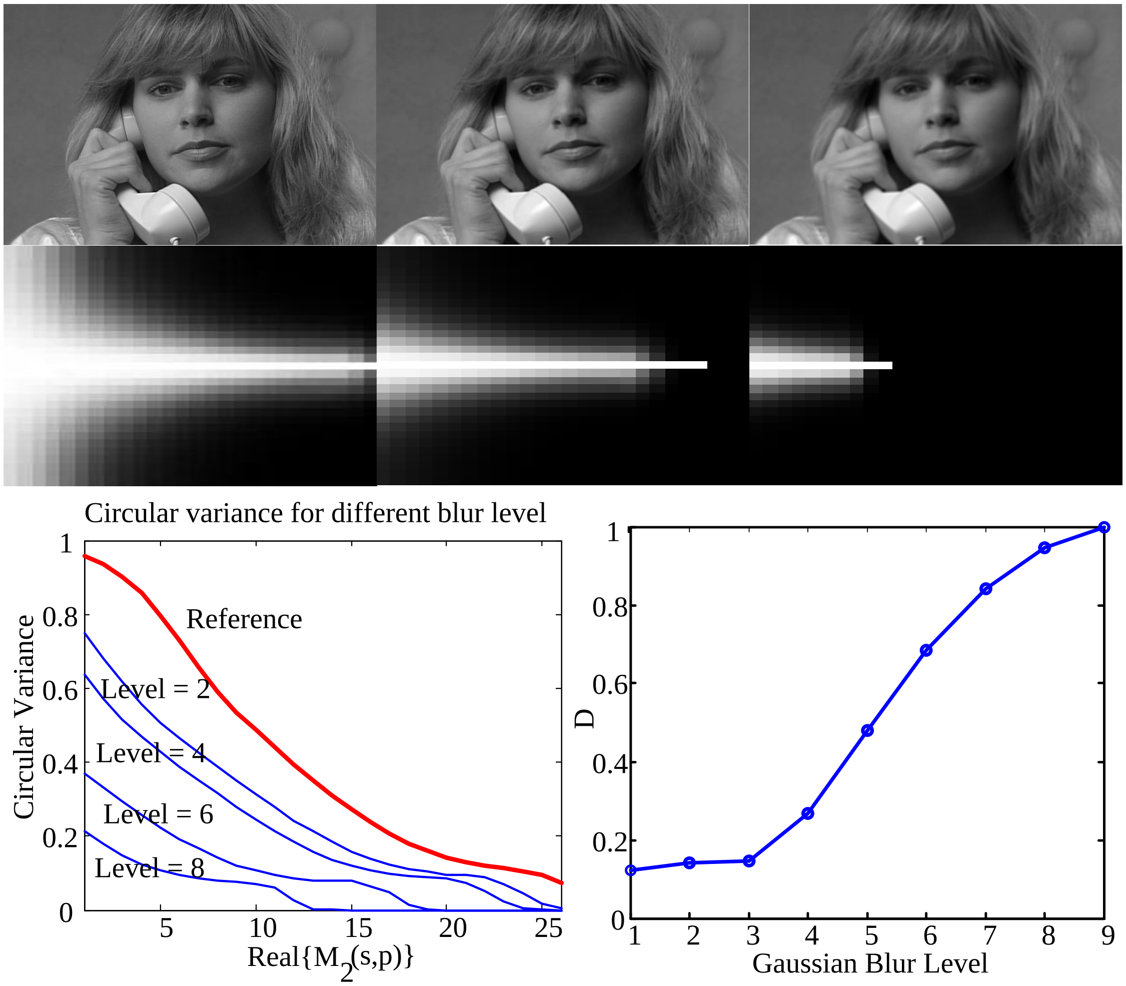


Figure 4.5: Experimental results for proposed RR-VQA approach on Gaussian blur. Top row: three images distorted by Gaussian blur at low, middle and high levels. Middle row: conditional histograms of $Imag\{L_2(s,p)\}$ versus $Real\{M_2(s,p)\}$ of Gaussian blur at low, middle and high distortion levels; Bottom left: circular variance as a function of $Real\{M_2(s,p)\}$ for the reference video sequence and distorted sequences at different distortion levels; Bottom right: proposed distortion measure as a function of Gaussian blur level.

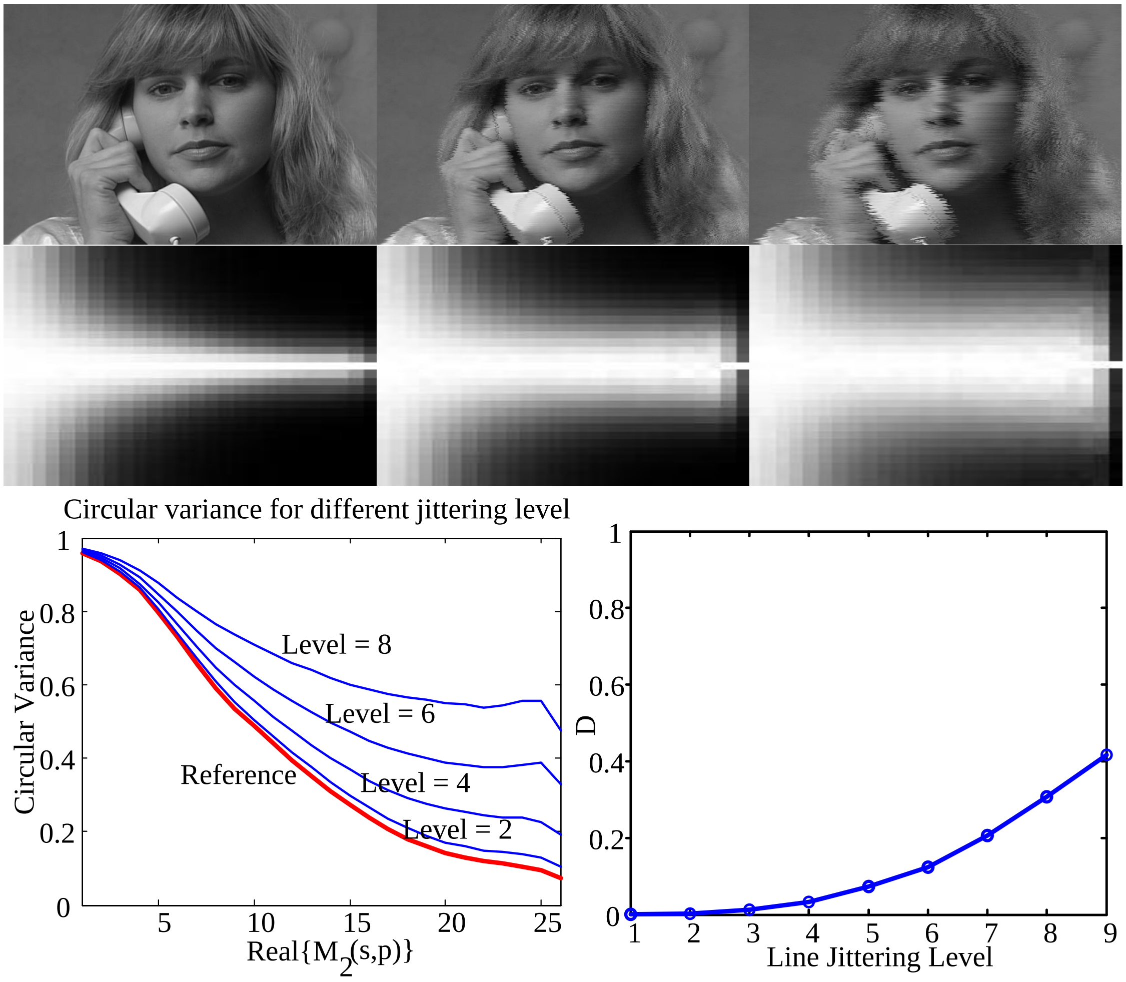


Figure 4.6: Experimental results for proposed RR-VQA approach on line jittering. Top row: three images distorted by line jittering at low, middle and high levels. Middle row: conditional histograms of $Imag\{L_2(s,p)\}$ versus $Real\{M_2(s,p)\}$ of line jittering at low, middle and high distortion levels; Bottom left: circular variance as a function of $Real\{M_2(s,p)\}$ for the reference video sequence and distorted sequences at different distortion levels; Bottom right: proposed distortion measure as a function of line jittering level.

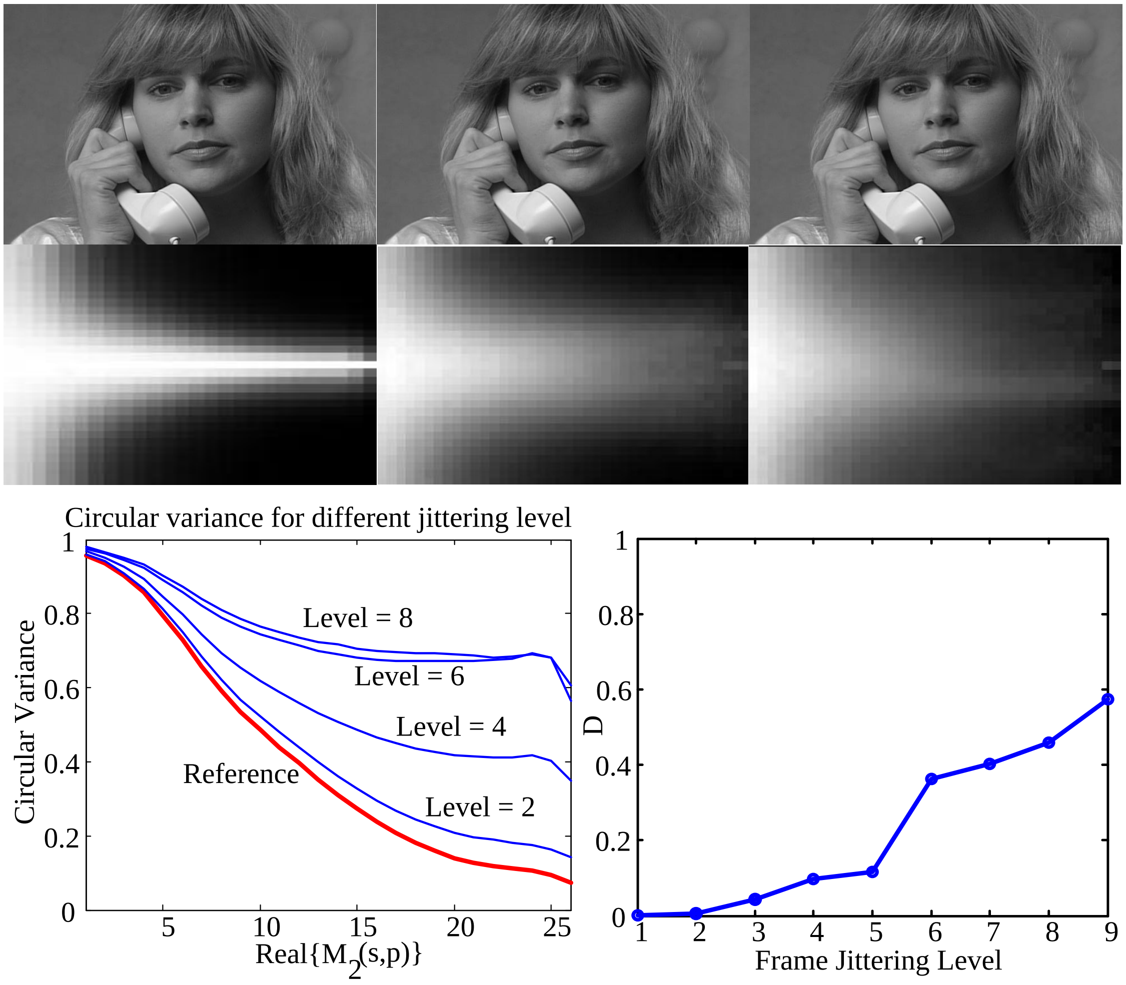


Figure 4.7: Experimental results for proposed RR-VQA approach on frame jittering. Top row: three images distorted by frame jittering at low, middle and high levels. Middle row: conditional histograms of $Imag\{L_2(s,p)\}$ versus $Real\{M_2(s,p)\}$ of frame jittering at low, middle and high distortion levels; Bottom left: circular variance as a function of $Real\{M_2(s,p)\}$ for the reference video sequence and distorted sequences at different distortion levels; Bottom right: proposed distortion measure as a function of frame jittering level.

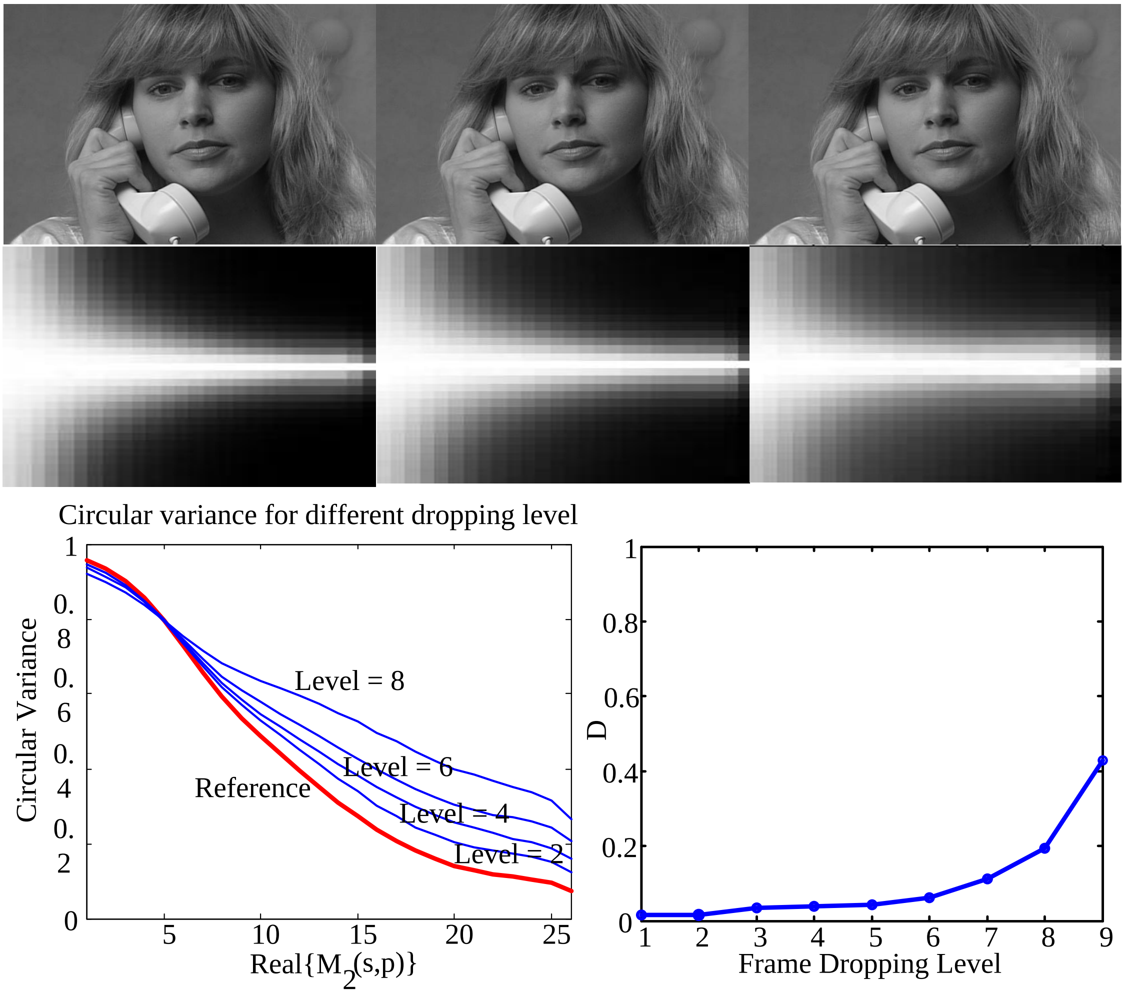


Figure 4.8: Experimental results for proposed RR-VQA approach on frame dropping. Top row: three images distorted by frame dropping at low, middle and high levels. Middle row: conditional histograms of $Imag\{L_2(s,p)\}$ versus $Real\{M_2(s,p)\}$ of frame dropping at low, middle and high distortion levels; Bottom left: circular variance as a function of $Real\{M_2(s,p)\}$ for the reference video sequence and distorted sequences at different distortion levels; Bottom right: proposed distortion measure as a function of frame dropping level.

Figure 4.4 to 4.8 shows the results of the experiment. First, it is interesting to observe that different distortions lead to different changes to the conditional histogram of $imag\{L_2(s,p)\}$ versus $real\{M_2(s,p)\}$. For example, noise contamination and jittering cause the histogram to spread, but Gaussian blur results in shrinkage of the histogram (as the energy reduces, especially at high frequencies). The observed changes are well captured by the departure of the CV curves of the distorted video sequence from the reference CV curves. Specifically, for each distortion type, the CV curve moves away from the reference CV curve with the increase of distortion level. This is further confirmed by computing the overall distortion measure D, which is monotonically increasing with the distortion level. From this experiment, we observe that the same objective distortion measure D works consistently for each individual type of distortion. This demonstrates the potential of the proposed method for general-purpose RR VQA, which is different from most approaches in the literature where ad-hoc features tuned to specific distortion types (such as blocking and ringing artifacts) are often used. Another interesting observation is regarding the frame jittering and frame dropping distortions. Notice that with these two types of distortions, the quality of each individual frame remains high quality, and thus frame-by-frame quality assessment approaches would give high quality scores to the image sequences undergoing these distortions, but the proposed method can capture them quite effectively without any specific change to the algorithm.

The effectiveness of the proposed temporal motion smoothness (TMS) measures for capturing temporal artifacts has been demonstrated above. They are useful novel RR features but do not take into account all distortions in VQA, which needs to include features that measures spatial distortions. The existing VQA databases usually include compression artifacts and transmission errors only and are not sufficient to fully test the usefulness of the current approach. In the next section, TMS will be combined with a spatial quality measure to construct a quality-aware video system..

4.3.2 Quality Aware Video

In our implementation, every 30 consecutive frames form a group of picture (GOP), where each frame is decomposed using a complex version [129] of a two-orientation steerable pyramid transform [130]. The subband statistics are carried out on the two orientation subbands at the finest scale by accumulating the coefficients of all frames in the GOP. These include the marginal statistics of real coefficients for intra-frame features and the statistics of the temporal correlation function conditioned on the energy function for inter-frame features. The intra- and inter-frame RR features are then extracted using the methods described in Section 4.2.1. This results in 8 features for each subband (3 intra- and 5 inter-frame features) and a total of 16 scalar features for both subbands. They are converted to 116 bits after 7-bit quantizations, and 256 bits after CRC and LDPC coding. The resulting encoded RR bitstream is then embedded into a 3D-DCT transform of the GOP using the method described in Section 4.2.2.

We simulated six types of distortions to test the proposed QAV system, which include 1) Gaussian noise contamination, where the distortion level is defined as the standard deviation of noise; 2) Gaussian blur, where the standard deviation of the blur filter defines the distortion level; 3) line jittering, simulated by shifting each line horizontally by a random number uniformly distributed between $[-S, S]$, and S defines the jittering level; 4) frame jittering, which is similar to line jittering except that the whole frame shifts together; 5) frame dropping, simulated by discarding every 1 out of N frames (empty frames are filled by repeating their previous frame) and 12-N defines the distortion level; and 6) MPEG2 compression, where the quantization parameter (QP) defines the distortion level. All distortion types are observed in real-world scenarios. For example, frame dropping occurs when the bandwidth of a real-time communication channel drops; and frame jittering is often caused by irregular camera movement such as hand shaking.

Figure 4.9 shows the test results for the robustness of information embedding, where the bit-error rates are calculated without LDPC correction, which can further improve the robustness. Compared with the traditional "3DDCT+QIM" method, "3DDCT+AQIM" leads to consistent improvement for all distortion types. As expected, the improvement is the most significant for blur distortions. Since information embedding alters the original video signal and thus its statistics, it is important to verify that such alteration does not have significant impact on the performance of the VQA algorithm. A comparison between the RR-VQA evaluation results with and without QAV information embedding is shown in Figure 4.10 for six types of distortions. It appears that the differences are generally small relative to the distortion measures. This may be explained by the fact that the VQA algorithm mostly relies on the variations of the statistics of the fine scale coefficients, while information embedding mainly affects relatively lower frequencies of the video content.

4.4 Summary

In this chapter, we first introduce a novel representation for motion information in natural video, termed temporal motion smoothness (TMS). The proposed measure is computed in complex wavelet transform domain and is demonstrated to be a potential solution for general-purpose RR-VQA. The TMS is supposed to capture the temporal artifacts only and should be combined with a spatial distortion measure to form a overall VQA method. This is also one of the reasons that TMS doesn't

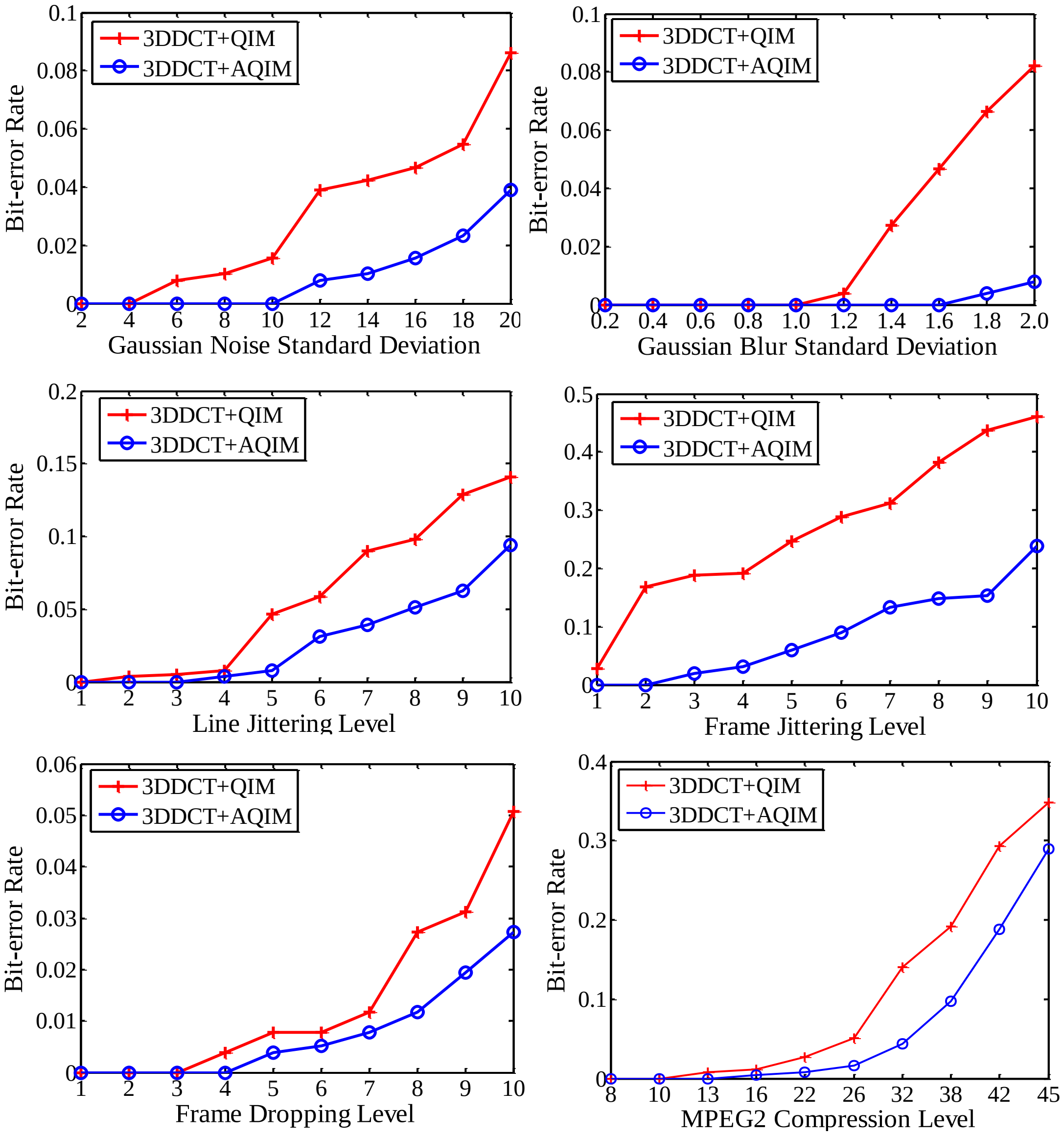


Figure 4.9: Robustness test of information embedding schemes.

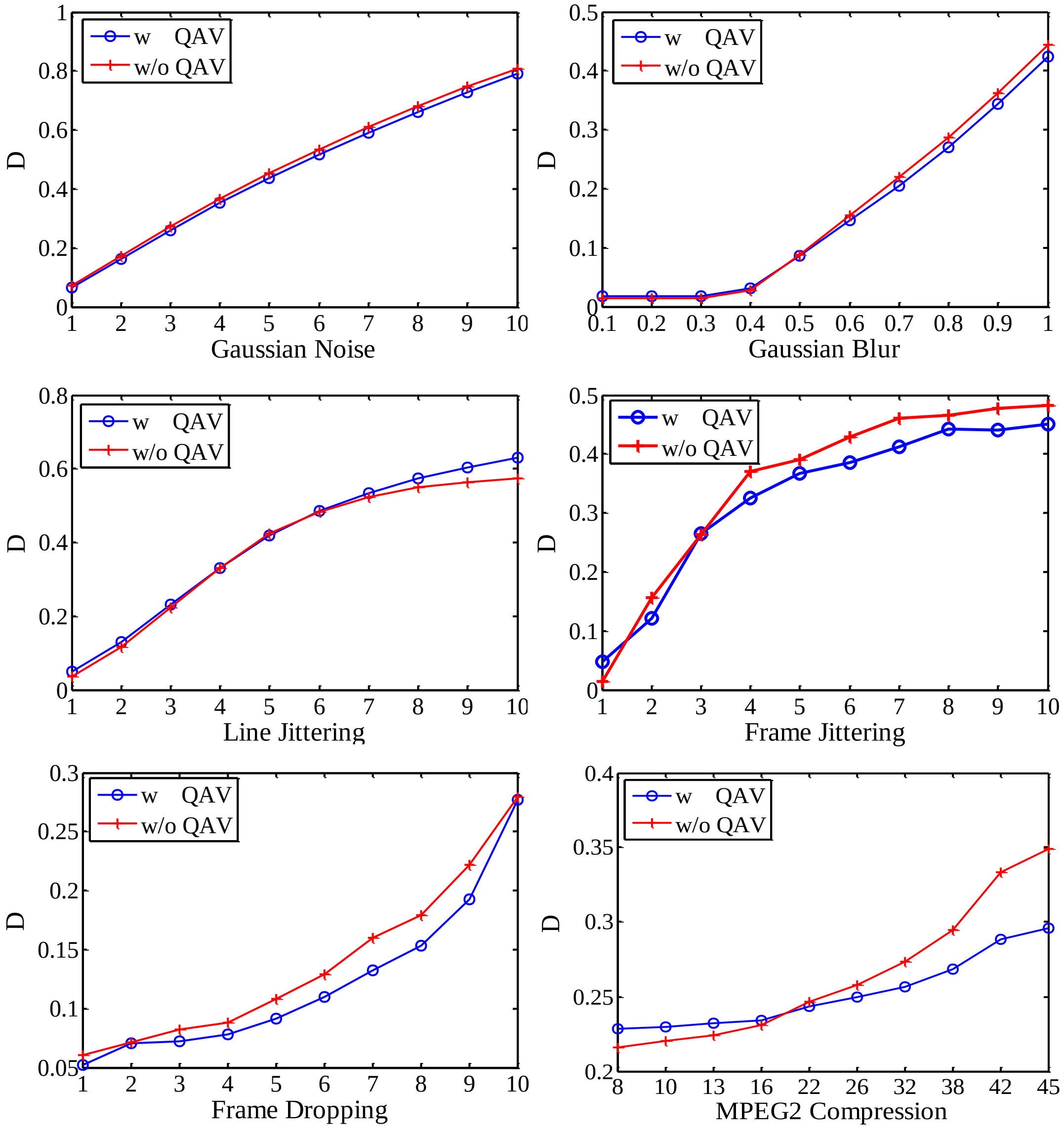


Figure 4.10: RR VQA consistency with and without QAV information embedding.

applied to any subjective VQA database. The tests on five simulated distortions demonstrated the usefulness of TMS. Because there is no assumption about the distortion type, TMS is applicable to measure the quality of video with any artifacts. Generally, motion estimation requires a time consuming search process [134] or solving simultaneous equations at every spatial location of the image [135], but TMS is able to capture the motion characteristics without explicit motion estimation. In addition, it has a very low RR data rate, which makes it easily adapted to practical visual communication system. Further improvement of the TMS can be gained by taking the frame rate into account, because higher frame rate generally leads to smoother motion between neighboring frames. Based on TMS and another quality feature, we propose a QAV system which also incorporates a novel robust information data hiding technique. This system does not need a error-free channel to transmit the RR features and does not require any changes of existing video compression and transmission systems. However, the performance of this system heavily relies on the watermarking technique because the RR features need to be recovered perfectly at the receiver side.

Chapter 5

Polyview Fusion for Video Denoising Enhancement

Instead of designing a video denoising algorithm, this section focuses on enhancing existing video denoising algorithm using a novel polyview fusion scheme. A video signal can be expressed as a 3D function $f(u, v, t)$, where u and v are the horizontal and vertical spatial indices and t is the time index, respectively. A video is typically played along the time axis. At any time instance $t = t_0$, the video is displayed as a 2D front-view image $g_{FV}^{(t_0)}(u, v) = f(u, v, t_0)$ and the image changes over time t. If we think of a video signal as 3D volume data, then it can also be viewed from the side or the top. This gives two other ways to play the same video $-$ a sequence of 2D top-view images $g_{TV}^{(u_0)}(v, t) = f(u_0, v, t)$ for different values of u_0 and a sequence of 2D side-view images $g_{SV}^{(v_0)}(u, t) = f(u, v_0, t)$ for different values of v_0. An example is given in Fig. 5.1, where the rarely observed side- and top-view images demonstrate some interesting regularized spatiotemporal structures.

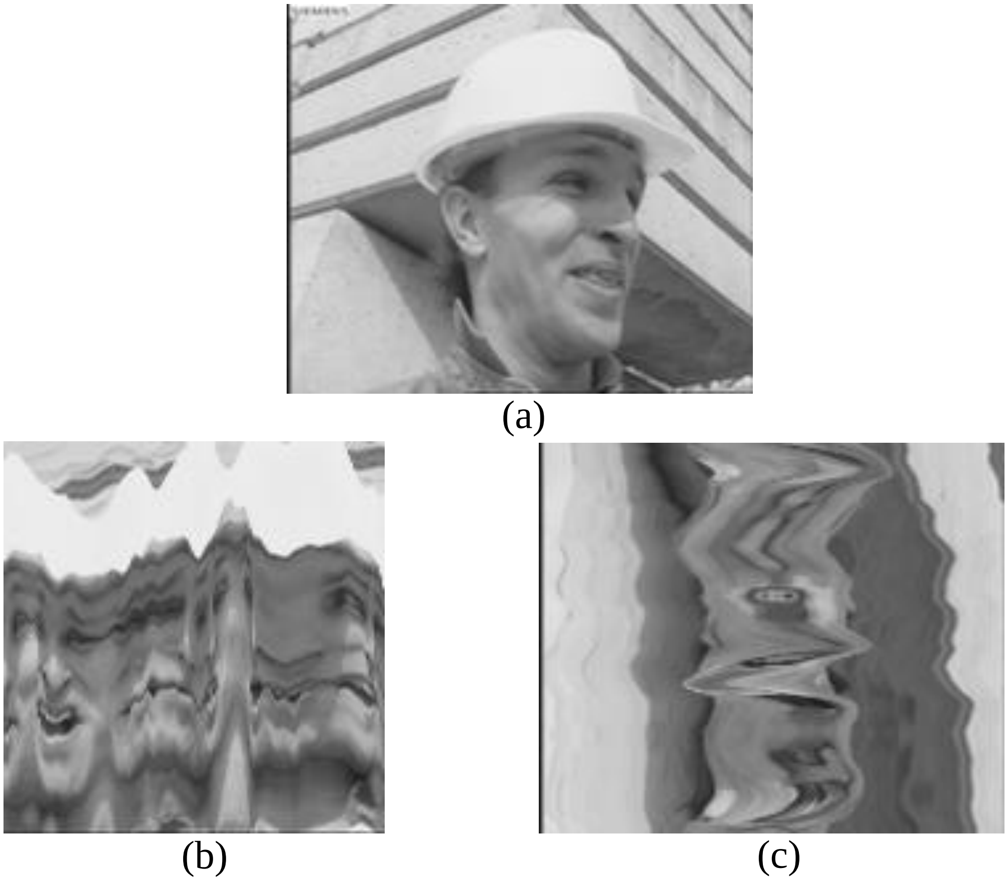

(a)
(b)
(c)

Figure 5.1: A video signal observed from (a) front view; (b) side view; and (c) top view.

5.1 Video Denoising Algorithm Enhancement by Polyview Fusion

Let x be an original noise-free video signal, which is contaminated by additive noise n, resulting in a noisy signal

$$y = x + n \,. \tag{5.1}$$

A video denoising operator D takes the noisy observation y and maps it to an estimator of x:

$$\hat{x} = D(y) \,, \tag{5.2}$$

such that the difference between x and $\hat{x}$ is as small as possible. How to quantify the difference between x and $\hat{x}$ is another subject of study. The most typically used ones are the mean squared error (MSE) and equivalently the peak-signal-to-noise ratio (PSNR). However, recent studies showed that the structural similarity index (SSIM) [1] may be a better measure in predicting perceived image distortion.

The proposed ployview fusion (PVF) method relies on a base video denoising algorithm, which could be as simple as frame-by-frame spatially adaptive Wiener filtering (Matlab Wiener2 function) or as complicated as VBM3D [112]. The base denoiser is applied to the same noisy signal y multiple times but from different views, which yields multiple versions of denoised signal

$$\begin{aligned} z_1 &= D_1(y) \,, \\ z_2 &= D_2(y) \,, \\ &\ldots\ldots \,, \\ z_N &= D_N(y) \,. \end{aligned} \tag{5.3}$$

In this study $N = 3$, as we have three different views, but in principle the

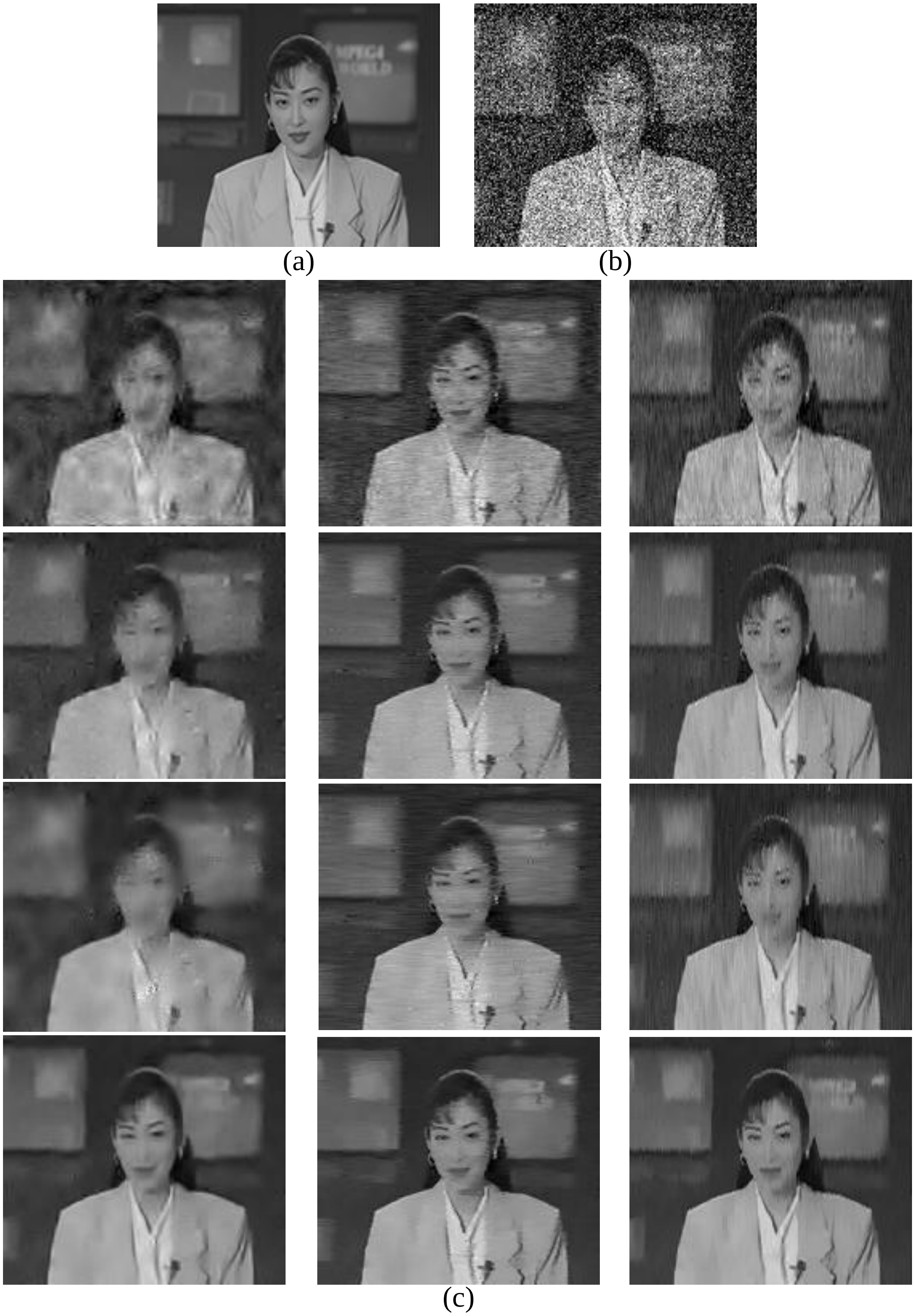


Figure 5.2: Denoised frames from three different views using different denoising algorithms. (a) Original frame; (b) Noisy frame with $\sigma_n = 50$; (c) Top to bottom: denoised frames by SURE-LET, BLS-GSM, K-SVD, and VBM3D; Left to right: denoised frames from front-, top-, and side-views, respectively.

general approach also applies to the cases of less or more views, or multiple denoising algorithms. Figure 5.2 shows sample denoised frames created by applying different denoising algorithms from three different views. It can be observed that the denoised frames have quite different appearances even when the same denoising method is applied (from different views). Some image structures preserved in one of the views may be missing in the other views, and some artifacts appear in one view may be absent from another. This suggests that the denoised frames from different views could complement each other, and fusing them (in appropriate ways) could potentially improve the denoising result. Let $\mathbf{z} = [z_1, z_2, ..., z_N]^T$ be a vector that contains all denoised results, then the final denoised signal $\hat{x}$ is given by applying a fusion operator F to $\mathbf{z}$:

$$\hat{x} = D(y) = F(\mathbf{z}) = F(D_1(y), D_2(y), ..., D_N(y))\,. \tag{5.4}$$

In the case that the base denoisers are predetermined, all the remaining task is to define the fusion rule F, which would be desired to achieve certain optimality. Here we employ a weighted average fusion method given by

$$\hat{x} = \mathbf{w}^T(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}) + \mu_x\,, \tag{5.5}$$

where $\mu_x = \mathbb{E}(x)$ (we use $\mathbb{E}$ to denote the expectation operator), $\boldsymbol{\mu}_{\mathbf{z}}$ is a column vector of expected values $[\mathbb{E}(z_1), \mathbb{E}(z_2), ..., \mathbb{E}(z_N)]^T$, and $\mathbf{w}$ is a column vector $[w_1, w_2, ...w_N]^T$ that defines the weight assigned to each denoised signal. To find the optimal weights $\mathbf{w}$ in the least-square sense, we define the following error energy function

$$E = \mathbb{E}[(x - \hat{x})^2] + \lambda\|\mathbf{w} - \frac{1}{N}\mathbf{1}\|^2\,, \tag{5.6}$$

where $\mathbf{1}$ is a length-N column vector with all entries equaling 1. The second term is to regularize the weighting vector towards all equal weights, and the parameter λ is used to control the strength of regularization. Taking the derivative of E with respect to $\mathbf{w}$ and setting it to zero, we obtain

$$(\mathbf{C_z} + \lambda\mathbf{I})\mathbf{w} = \mathbf{b} + \frac{\lambda}{N}\mathbf{1}\,, \tag{5.7}$$

where $\mathbf{I}$ denotes the $N \times N$ identity matrix, $\mathbf{C_z}$ is the covariance matrix

$$\mathbf{C_z} = \mathbb{E}[(\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})(\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})^T]\,, \tag{5.8}$$

and $\mathbf{b}$ is a column vector given by

$$\mathbf{b} = \mathbb{E}[(x - \mu_x)(\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})]\,. \tag{5.9}$$

We can then solve for optimal $\mathbf{w}$, which gives

$$\mathbf{w}_{opt} = (\mathbf{C_z} + \lambda\mathbf{I})^{-1}\left(\mathbf{b} + \frac{\lambda}{N}\mathbf{1}\right)\,. \tag{5.10}$$

Here the $\lambda\mathbf{I}$ term plays an important role in stabilizing the solution, especially when $\mathbf{C_z}$ is close to singular. It is a forgetting factor which may be optimized under information theoretic framework. The improvement of fusion performance is expected if this factor is adaptive to video content and changes over time. The computation of $\mathbf{b}$ requires the original signal x, which is not available. But by assuming n to be zero-mean and independent of $\mathbf{z}$, we have

$$\mathbf{b} = \mathbb{E}[(y - n - \mu_x)(\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})] = \mathbb{E}[(y - \mu_y)(\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})]\,. \tag{5.11}$$

When applying the above approach to real signals, the expectation operators would need to be replaced by sample means. In our implementation, we apply the weight calculation to individual non-overlapping 16 × 16 × 16 blocks, resulting in block-wise space-time adaptive weights in the 3D volume. Eq. (5.5) is then applied to each block to obtain the final denoised signal.

5.2 Variance Weighted Polyview Fusion

The previous section presents a PVF scheme which is optimal in the least-square sense. However, the estimation error of necessary statistics limits the final performance. In this section, we proposed an improved PVF, namely variance-weighted PVF (VPVF). Before the fusion step, we first apply a normalization process to each z_i. This is inspired by the SSIM index [1], which has been shown to be a much better predictor of perceived image quality than the MSE. Given two image patches, the SSIM index separate the similarity measure into the luminance, contrast and structure components. Since the luminance and contrast (measured by mean intensity and standard deviation, respectively) of an image patch can be adjusted freely without changing its structure, we can improve the SSIM measure by adapting the luminance and contrast of each z_i to match those of x while maintaining its structure. Specifically, we compute

$$\hat{z}_i = \frac{\sigma_x}{\sigma_{z_i}}(z_i - \mu_{z_i}) + \mu_x \, , \tag{5.12}$$

where μ_x and μ_{z_i}, and σ_x and σ_{z_i}, denote the means and standard deviations of x and z_i, respectively. The computation in (5.12) requires the mean and standard deviation of x, which is not available. Fortunately, we can estimate them from the noisy signal y using (5.1) and the known noise properties (independence, zero-mean, and known standard deviation) by

$$\mu_x = \mu_y \quad \text{and} \quad \sigma_x = \sqrt{\sigma_y^2 - \sigma_n^2}\,, \tag{5.13}$$

where μ_y and σ_y^2 are the mean and variance of y, respectively.

Our fusion rule is based on variance weighted averaging, which can be expressed as

$$\hat{x} = \frac{\sum_{i=1}^{N} \sigma_{z_i}^2 \hat{z}_i}{\sum_{i=1}^{N} \sigma_{z_i}^2}\,. \tag{5.14}$$

This is determined by our empirical studies on the relationship between the variance and quality of denoised video patches using state-of-the-art video denoising algorithms. Specifically, for three given 3D patches denoised by the same video denoising algorithm but from three different views, we compute their corresponding variances and PSNR values between the denoised and original patches. We then calculate the Spearman rank-order correlation coefficient (SRCC) between the three variance and three PSNR values. Table 5.1 shows the average SRCC values (over all patches) for nine video sequences denoised with four denoising algorithms. It can be seen that although a fairly large variations are observed (depending on both denoising algorithm and video sequence), the correlations are all positive. This suggests that the patches of larger variances tend to have better image quality, thus justifying variance-based weighting.

5.3 Experimental Results

We use publicly available video sequences to test the proposed algorithm, which include “Akiyo”, “Carphone”, “Miss America”, and “News”. The size of all sequences is $144 \times 176 \times 144$, and are contaminated by independent white Gaussian noise with standard deviation, σ, covering a wide range between 10 and 100. After the noisy

Table 5.1: SRCC between local variance and PSNR for $\sigma_n = 50$

	SURE-LET	BLS-GSM	K-SVD	VBM3D
Akiyo	0.436	0.658	0.718	0.747
Carphone	0.316	0.498	0.596	0.559
Mobile	0.645	0.882	0.891	0.748
Foreman	0.321	0.579	0.537	0.590
Miss America	0.288	0.418	0.470	0.581
Mother Daughter	0.439	0.721	0.746	0.820
News	0.566	0.767	0.779	0.772
Salesman	0.734	0.769	0.788	0.820
Suzie	0.291	0.458	0.531	0.420

sequences are denoised using a base denoiser along three different views, the noisy and denoised sequences are divided into 16×16×16 non-overlap 3D patches, within which sample means and variances are computed and employed in the normalization and fusion processes described in Section 5.1 and Section 5.2, respectively. The choices of non-overlapping patches and size 16 are based on compromises between the denoising performance and complexity.

All sequences are in YCrCb 4:2:0 format, but only the denoising results of the luma channel was reported here to validate the algorithm. In order to evaluate the quality of denoised video quantitatively, three objective criteria were employed: PSNR, SSIM [1], as well as 3D-SSIM developed in Chapter 3 of this thesis. PSNR is the most widely used method in the literature, but SSIM has been recognized as a much better measure to predict subjective quality measurement. 3D-SSIM has been proved to be a better video quality measure in Chapter 3. Assume that x and $\hat{x}$ are the noise-free and denoised images, respectively, and L is the dynamic range of intensity values, then

$$\mathrm{PSNR}(x, \hat{x}) = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}(x, \hat{x})} \right) . \tag{5.15}$$

The SSIM value between two image patches is computed as

$$\mathrm{SSIM}(x, \hat{x}) = \frac{(2\mu_x \mu_{\hat{x}} + C_1)(2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)} \tag{5.16}$$

where C_1 and C_2 are small positive constants to avoid instability when the means and variances are close to zero. This computation is applied at each location in the image using a sliding window that moves pixel-by-pixel across the image, resulting in an SSIM quality map, as demonstrated in Fig. 5.4. The SSIM value between two images is then computed as the mean of the SSIM map. Both PSNR and SSIM were computed on a frame-by-frame basis along the temporal direction and then averaged over all frames to yield the PSNR and SSIM values of the whole sequence. Meanwhile, 3D-SSIM will take a video as 3D-volume data and give one quality score.

Many state-of-the-art denoising algorithms are publicly available that facilitate direct comparisons. For simplicity, here we report our comparison results for 5 noise levels (σ equals 10, 15, 20, 50, and 100, respectively) using three base denoising methods with and without using our PVF and VPVF approach. The base algorithms are Matlab Wiener-2D, BLS-GSM [103] and VBM3D[112]. The denoising computations are conducted using the default parameter settings of the code available to the public at [136], [137], and [138], respectively. We have also applied our PVF approach to a list of other highly competitive algorithms, including NLM [111], K-SVD [105], and SURE-LET [110], and other popular test sequences, such as “Foreman”, “Salesman”, “Mobile”, and “Football”. Similar results were observed.

Table 5.2 and Table 5.3 show the comparison results using PSNR, SSIM and 3D-SSIM measures at 5 noise levels using 3 base denoising algorithms with and without PVF and VPVF. The average improvement over 4 test sequences is given

Table 5.2: PSNR, SSIM, and 3D-SSIM comparisons for three video denoising algorithms with and without PVF and VPVF for “Akiyo” and “Carphone”

Video Sequence	*Akiyo*					*Carphone*				
Noise std (σ)	10	15	20	50	100	10	15	20	50	100
PSNR Results (dB)										
Wiener-2D	33.22	30.38	28.33	21.58	15.94	32.66	29.84	27.86	21.35	15.86
with PVF	34.69	31.91	29.89	23.15	17.52	33.90	31.20	29.29	22.87	17.42
with VPVF	35.02	32.51	30.80	25.82	22.58	34.20	31.70	29.99	24.87	21.38
BLG-GSM	36.12	33.73	32.09	27.32	24.36	35.34	33.00	31.40	26.47	23.15
with PVF	39.95	37.58	35.88	30.78	27.43	37.01	34.92	33.50	29.02	25.81
with VPVF	40.13	37.81	36.20	31.22	27.76	37.11	35.05	33.63	29.33	25.26
VBM3D	42.01	39.76	37.91	30.79	24.39	38.50	36.64	35.35	29.82	23.30
with PVF	42.33	40.08	38.36	32.64	26.93	38.50	36.71	35.46	30.97	25.76
with VPVF	42.32	40.06	38.35	32.66	27.13	38.52	36.66	35.38	30.99	26.00
SSIM Results										
Wiener-2D	0.876	0.788	0.700	0.364	0.164	0.885	0.803	0.722	0.408	0.205
with PVF	0.906	0.833	0.757	0.432	0.213	0.909	0.840	0.771	0.472	0.255
with VPVF	0.917	0.864	0.814	0.615	0.470	0.923	0.876	0.830	0.634	0.477
BLG-GSM	0.952	0.924	0.898	0.765	0.636	0.951	0.927	0.902	0.773	0.627
with PVF	0.977	0.964	0.949	0.866	0.749	0.964	0.947	0.930	0.839	0.718
with VPVF	0.978	0.965	0.952	0.872	0.753	0.965	0.948	0.932	0.844	0.732
VBM3D	0.983	0.976	0.965	0.874	0.616	0.972	0.961	0.951	0.874	0.628
with PVF	0.986	0.978	0.967	0.903	0.684	0.972	0.961	0.952	0.892	0.691
with VPVF	0.986	0.978	0.968	0.904	0.697	0.972	0.962	0.952	0.893	0.703
3D-SSIM Results										
Wiener-2D	0.848	0.788	0.727	0.480	0.287	0.878	0.836	0.794	0.606	0.396
with PVF	0.916	0.874	0.826	0.563	0.298	0.937	0.906	0.874	0.678	0.416
with VPVF	0.935	0.904	0.870	0.658	0.390	0.947	0.924	0.900	0.746	0.503
BLG-GSM	0.922	0.892	0.859	0.649	0.403	0.926	0.907	0.887	0.744	0.523
with PVF	0.933	0.912	0.889	0.749	0.537	0.936	0.914	0.890	0.774	0.624
with VPVF	0.952	0.937	0.922	0.815	0.664	0.948	0.937	0.900	0.836	0.712
VBM3D	0.946	0.930	0.912	0.777	0.446	0.933	0.920	0.905	0.816	0.565
with PVF	0.954	0.945	0.921	0.806	0.501	0.946	0.938	0.931	0.850	0.613
with VPVF	0.958	0.947	0.933	0.858	0.660	0.947	0.939	0.932	0.873	0.728

Table 5.3: PSNR, SSIM, and 3D-SSIM comparisons for three video denoising algorithms with and without PVF and VPVF for “Foreman” and “Miss America”

Video Sequence	*Foreman*					*Miss America*				
Noise std (σ)	10	15	20	50	100	10	15	20	50	100
PSNR Results (dB)										
Wiener-2D	32.22	29.49	27.55	21.17	15.77	34.36	31.35	29.17	21.91	16.07
with PVF	33.11	30.53	28.70	22.59	17.30	35.74	32.80	30.67	23.47	17.65
with VPVF	33.16	30.65	28.93	23.79	20.41	37.49	35.23	33.63	28.59	24.98
BLG-GSM	34.22	31.92	30.32	25.44	22.21	38.69	36.54	35.09	30.61	27.52
with PVF	35.83	33.65	32.12	27.36	24.05	41.03	38.99	37.59	33.16	30.02
with VPVF	35.89	33.73	32.24	27.66	24.35	41.14	39.16	37.76	33.30	29.82
VBM3D	37.37	35.50	34.12	28.47	22.46	41.93	40.19	38.81	33.55	26.57
with PVF	37.68	35.80	34.44	29.28	24.14	42.34	40.57	39.24	34.69	28.93
with VPVF	37.70	35.84	34.49	29.41	24.38	42.37	40.60	39.28	34.62	29.08
SSIM Results										
Wiener-2D	0.887	0.812	0.738	0.432	0.220	0.848	0.737	0.633	0.275	0.107
with PVF	0.906	0.843	0.778	0.488	0.267	0.879	0.785	0.692	0.331	0.138
with VPVF	0.911	0.856	0.802	0.578	0.414	0.935	0.899	0.865	0.709	0.567
BLG-GSM	0.938	0.910	0.884	0.746	0.591	0.958	0.939	0.922	0.841	0.751
with PVF	0.952	0.930	0.908	0.792	0.646	0.972	0.960	0.948	0.884	0.791
with VPVF	0.953	0.931	0.910	0.796	0.649	0.973	0.961	0.949	0.885	0.793
VBM3D	0.961	0.947	0.933	0.844	0.601	0.976	0.968	0.959	0.901	0.669
with PVF	0.962	0.948	0.934	0.857	0.643	0.978	0.970	0.962	0.915	0.685
with VPVF	0.962	0.948	0.935	0.858	0.648	0.978	0.970	0.962	0.915	0.703
3D-SSIM Results										
Wiener-2D	0.849	0.814	0.779	0.573	0.374	0.865	0.802	0.739	0.470	0.257
with PVF	0.921	0.897	0.869	0.662	0.399	0.938	0.903	0.866	0.638	0.352
with VPVF	0.928	0.909	0.886	0.722	0.479	0.950	0.924	0.897	0.711	0.432
BLG-GSM	0.889	0.876	0.860	0.741	0.537	0.935	0.903	0.866	0.666	0.429
with PVF	0.912	0.900	0.885	0.755	0.566	0.939	0.909	0.880	0.726	0.515
with VPVF	0.926	0.917	0.905	0.812	0.648	0.961	0.944	0.926	0.830	0.701
VBM3D	0.889	0.887	0.879	0.807	0.538	0.945	0.917	0.883	0.686	0.420
with PVF	0.908	0.909	0.907	0.836	0.566	0.964	0.952	0.936	0.835	0.557
with VPVF	0.915	0.911	0.908	0.848	0.658	0.966	0.954	0.942	0.871	0.721

in Table 5.4. It can be seen that the proposed PVF approach consistently leads to performance gain over all base denoising algorithms, for all test video sequences, and at all noise levels. And VPVF could further improve the denoising effect in most cases. The gain is especially significant at high noise levels, where the improvement can be as high as 2-3 dB in terms of PSNR over state-of-the-art algorithms such as VBM3D, which is among the best algorithms ever reported in the literature. We also observe that the gain is reduced for video sequences with significant amount of large motion. This is mainly due to the high complexity texture pattern in the top- and side-views, which leads to reduced performance of the base denoisers.

Another important observation is that VPVF consistently performs better than PVF on the tested sequences with different noise levels. It is not surprising that the SSIM results are better for VPVF because it introduces the SSIM-inspired normalization process. However, the same situation is also applied for PSNR results, which is counter-intuitive because PVF is designed to optimize the PSNR value. The reasons behind these phenomena might be two-fold: (1) one of the assumptions for PVF is that the noise is additive white noise and independent of the noisy and all denoised videos. This assumption might be too strong. In practice, the noise and noisy/denoised videos always correlates to a certain extent and may be more pronounced at high noise levels. (2) in order to limit the pixel values of all sequences to be within the range of [0 255], any value outside this range will be clipped to the nearest valid number. This is another reason that puts the assumption of PVF in question. On the other hand, the motivation of VPVF comes from our observation of the correlation between local variance and local quality, which is more realistic and may explain why VPVF achieves better PSNR performance than PVF.

To demonstrate the performance improvement for individual video frames, Figure 5.3 depicts PSNR and SSIM comparisons as functions of frame number for “Foreman” sequence. Because 3D-SSIM is based on 3D data and only offers a

Table 5.4: Average PSNR, SSIM, and 3D-SSIM improvement over all test sequences

Noise std (σ_n)		10	15	20	50	100
PSNR Improvement (dB)						
Wiener-2D	PVF	1.2450	1.3450	1.4100	1.5175	1.5625
	VPVF	1.8525	2.2575	2.6100	4.2650	6.4275
BLS-GSM	PVF	2.3625	2.4875	2.5475	2.6200	2.5175
	VPVF	2.4750	2.6400	2.7325	2.9175	2.4875
VBM3D	PVF	0.2600	0.2675	0.3275	1.2375	2.2600
	VPVF	0.2750	0.2675	0.3275	1.2625	2.4675
SSIM Improvement						
Wiener-2D	PVF	0.0260	0.0402	0.0513	0.0610	0.0443
	VPVF	0.0475	0.0887	0.1295	0.2642	0.3080
BLS-GSM	PVF	0.0165	0.0252	0.0322	0.0640	0.0748
	VPVF	0.0175	0.0262	0.0342	0.0680	0.0805
VBM3D	PVF	0.0015	0.0013	0.0018	0.0185	0.0473
	VPVF	0.0015	0.0015	0.0023	0.0193	0.0592
3D-SSIM Improvement						
Wiener-2D	PVF	0.0680	0.0850	0.0990	0.1030	0.0378
	VPVF	0.0800	0.1053	0.1285	0.1770	0.1225
BLS-GSM	PVF	0.0120	0.0143	0.0180	0.0510	0.0875
	VPVF	0.0287	0.0393	0.0453	0.1233	0.2083
VBM3D	PVF	0.0148	0.0225	0.0290	0.0603	0.0670
	VPVF	0.0183	0.0242	0.0340	0.0910	0.1995

single quality score for whole sequence, it is not applied for frame-based analysis. Again, consistent improvement is observed for almost all frames, indicating the robustness of the proposed PVF and VPVF approach.

Figure 5.4 provides visual comparisons of the denoising results of one frame extracted from “Akiyo” sequence, for which the original and noisy frames are given in Figure 5.2 (a) and (b), respectively. From left to right columns are 1) original denoised frame, 2) with PVF, 3) with VPVF, separately. From top to bottom odd rows are figures for 1) Wiener-2D, 2) BLS-GSM, 3) VBM3D, and even rows are corresponding SSIM quality maps, in which brighter pixels indicate higher SSIM values and thus better quality. Visual quality improvement by the proposed PVF and VPVF approach can be easily discerned at various locations in the denoised frames. The observation is also verified by the SSIM quality map, which provides a useful indicator of local image quality variations.

Furthermore, another experiment has been conducted to measure the computational complexity of the PVF and VPVF operation and how they compare with the complexity of the base denoisers. The results are reported in the Table 5.5, where the speed is measured in seconds based on Matlab implementations of the algorithms on a computer with Intel Core Duo CPU E8600 processor at 3.33GHz. Although the implementations are not speed-optimal, they give us a general idea about the amount of added complexities due to the PVF or VPVF process. As can be observed, generally the PVF/VPVF procedure is of low complexity relative to the base denoising algorithms. The percentage of time spent on PVF ranges from 0.0386% to 3.8471% of the overall denoising process (where a base denoiser needs to be run 3 times and thus the overall process increases the computational cost by a factor of 3 or more), and from 0.0431% to 4.2756% for VPVF. In conclusion, the complexity of the overall denoising algorithm mainly depends on the complexity of the base denoiser, and the PVF/VPVF portion is mostly negligible.

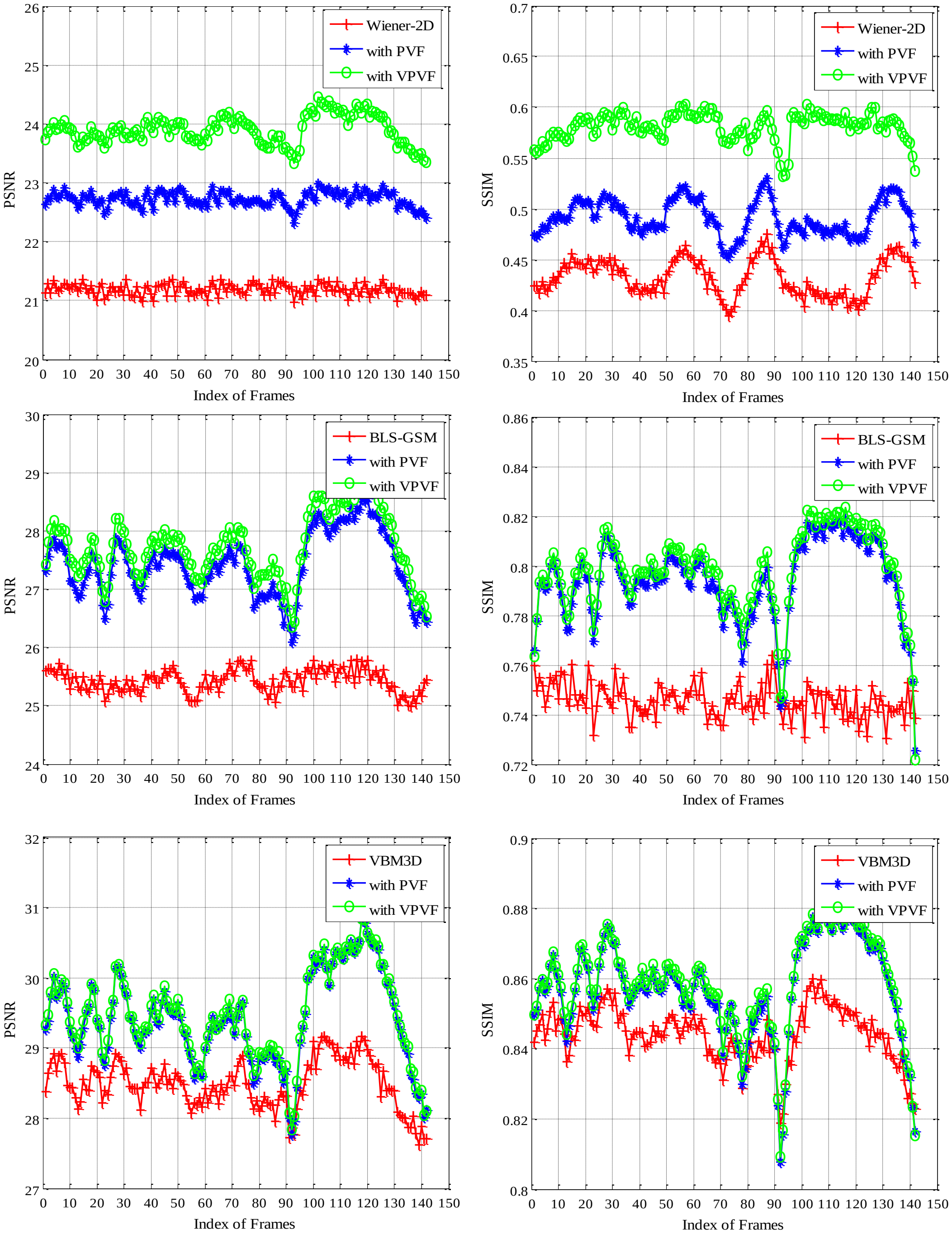


Figure 5.3: PSNR and SSIM comparisons as functions of frame number for "Foreman" sequence. Noise level $\sigma = 50$.

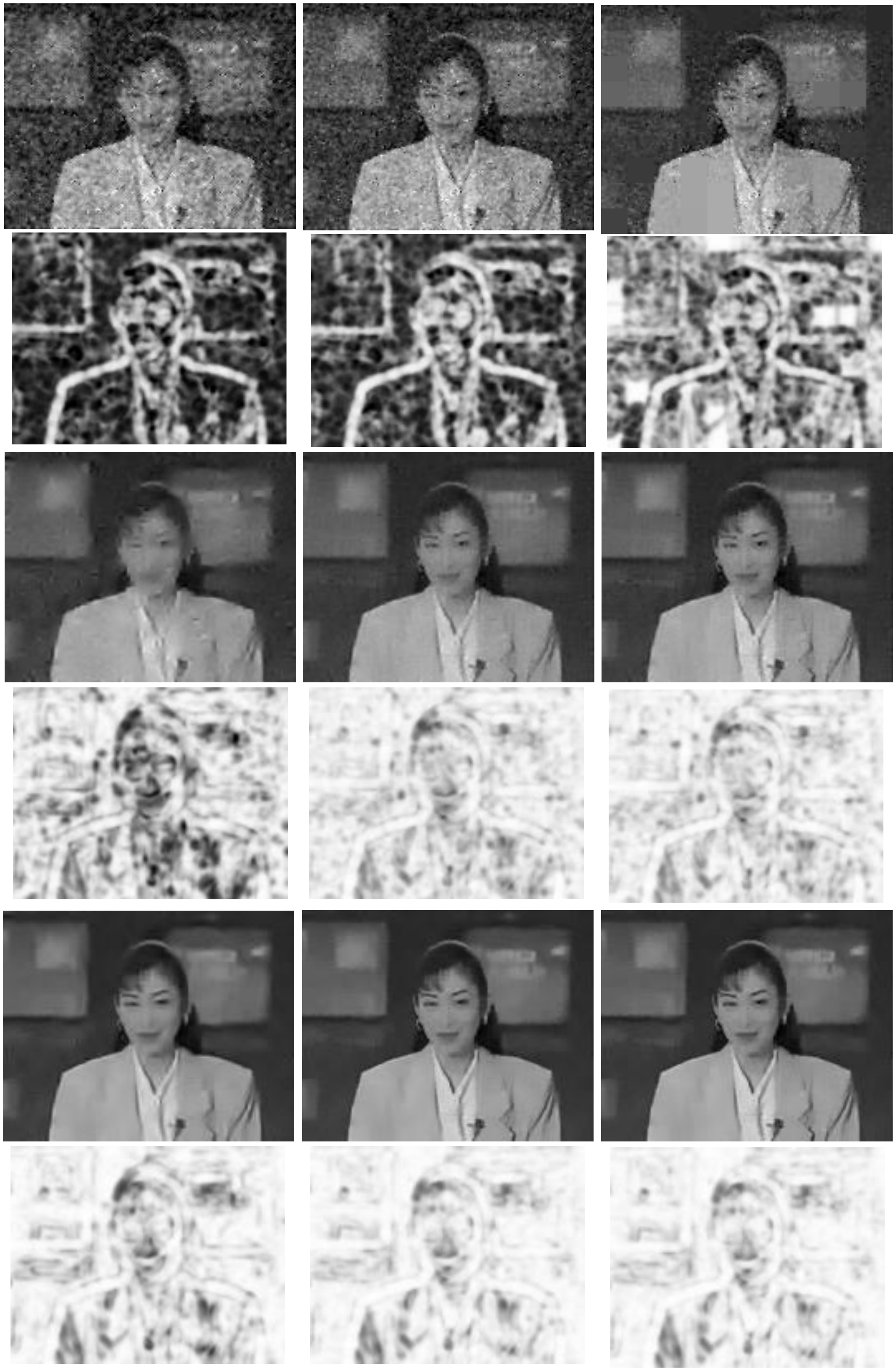


Figure 5.4: Comparison of one denoised frame from "Akiyo" sequence with and without PVF and VPVF using three base denoising algorithms.

Table 5.5: Computational complexity analysis

Base denoiser	One view denoising time (second)	PVF time (second)	PVF (%)	VPVF time (second)	VPVF (%)
Wiener-2D	1.353	0.1624	3.8471	0.1813	4.2756
BLS-GSM	140.3		0.0386		0.0431
VBM3D	8.791		0.612		0.6828

5.4 Summary

We propose two approaches that can improve video denoising performance of existing algorithms by fusing the denoising results from multiple views of video. The first one, PVF, was derived under a least-square framework to seeking an optimal solution for fusion. The performance was limited by the assumption of independence between noise and signal and estimation error of statistics. The second one, VPVF, was inspired by SSIM and successful fusion techniques. Variance based weighting scheme has also been justified by the correlation between local variance and quality. Our experimental results demonstrate consistent improvement over some of the best video denoising algorithms in the literature. The proposed method is conceptually simple, easy-to-use, and computationally efficient.

Chapter 6

Conclusion and Future Research

6.1 Conclusion

This thesis focused on two problems related to the perceptual quality of video: (1) video quality assessment and (2) video denoising for quality enhancement.

For full-reference VQA, a novel algorithm, namely 3D-SSIM, has been proposed in which a video signal is considered as a 3D volume image and a local SSIM-based quality measure is combined with information content and distortion weighted pooling methods. Based on the experimental results across four public VQA databases, the current implementation is computationally efficient and achieves superior performance compared with state-of-the-art VQA approaches. The low complexity mainly comes from the auto-scale process, that accounts for the influence of normal viewing distance for VQA, and the non-overlapping block-based scheme that significantly reduces the computational burden for weighted pooling. Compared with pixel- and frame-based VQA approaches, a potential disadvantage of the proposed strategy is the large amount of memory required to buffer 3D-volume data. However, this problem may be alleviated by simply dividing the whole video sequence into several segments or clips based on the adopted size of the 3D block involved in the computation. Then, a parallel computation scheme can also be employed to facilitate the VQA process.

For reduced-reference VQA, a complex wavelet transform domain temporal motion smoothness measure has been proposed and its potential for general-purpose RR-VQA demonstrated. The proposed algorithm has several useful properties:

- it is applicable to a wide range of practical distortion types;
- it captures relevant motion characteristics without explicit motion estimation, which often involves a complicated search procedure [134] or requires solving simultaneous equations at every spatial location of the image [135];
- it has a very low RR data rate (current implementation only uses 15 scalar features per video sequence).

All these properties make it an attractive approach in real-world visual communication applications. For example, it can be directly adopted in a quality-aware video system [128]. The proposed approach may fail under scene changes or very large motion (where distances of moving objects between frames are beyond the coverage of the wavelet filter envelopes) due to the locality of the wavelet-based approach in the measurement of temporal motion smoothness. Therefore, to create a practical VQA system, such measurement needs to be combined with intra-frame quality measures, such as [9].

A quality-aware (QAV) system has been proposed that incorporates novel RR-VQA algorithms with a novel robust information data hiding approach. Such a QAV system has a number of attractive properties:

- It provides the useful functionality of "quality-awareness" without affecting the conventional use of the video content;

- It avoids the necessity of an ancillary channel in the deployment of RR-VQA schemes;
- It allows the video content to be converted and distributed using any existing or user-defined formats, provided the embedded messages are not corrupted during lossy format conversion;
- It also provides an opportunity at the receiver side to partially "repair" the distorted video signal using the embedded RR features.

Two new approaches, PVF and VPVF, have been proposed to improve the video denoising performance of existing algorithms by fusing the denoising results from multiple views. The experiments detailed in Section 5.3 demonstrate significant and consistent improvement over existing video denoising methods. The proposed methods are conceptually simple, easy-to-use, and computationally efficient. The complexity of the whole algorithm mainly depends on that of the base denoising method, but not the PVF or VPVF procedure. In principle, the PVF and VPVF strategies could be applied to any existing video denoising algorithm, but the major intention here is to apply it to 2D approaches (Categories 1 and 2 described in Section 2.4). Because the denoising results obtained by applying 2D approaches from different views tend to complement one another. By contrast, 3D approaches (Category 3 in Section 2.4) such as those using 3D patches have already considered the dependencies between neighboring pixels from all directions. Thus applying them from different views may lead to similar results that would not complement each other to any significant extent. In practice, to apply PVF or VPVF, one would need to store all video frames involved in the denoising and fusion processes in the memory. This may be a problem in practical systems, especially when the video sequence is long. It is therefore preferable to divide long sequences into segments along the temporal direction, and then denoise each segment independently. By adjusting the length of the segments, the memory requirement can be controlled.

6.2 Future Research

The approaches described in this thesis can be further improved in many aspects by employing advanced mathematical models or technologies. In the future, the proposed full-reference VQA method will achieve better performance by incorporating more accurate statistical models in the estimation of local information content. Currently, the derivation of the local information content model requires the assumption of Gaussian source and additive Gaussian noise model, which may not be consistent with real/practical signals, because video is usually a non-stationary non-Gaussian distributed signal, and the type of noise also depends on the specific application scenario. In addition, more advanced adaptive strategies for local distortion weighting can be further explored. Superior VQA performance may be obtained by using local distortion to control the weighting function with more freedom.

The proposed RR-VQA approaches maybe improved and extended in several ways. First, higher-order temporal correlation functions may be employed to characterize the smoothness of higher-order motion (such as acceleration). Second, appropriate adjustments are needed to accommodate the cases of scene changes and very large motion (which may be solved by adopting a multi-scale, coarse-to-fine strategy). Third, temporal motion smoothness is only one aspect that affects perceived video quality. Other RR features (such as intra-frame statistical features [9]) may be incorporated under a unified framework to provide a full solution to the problem of RR-VQA. For QAV, future work includes improving the performance of both the accuracy of RR-VQA and the robustness of information embedding, and providing meaningful video quality evaluations when RR features cannot be fully recovered (for example, by relating the decoding error rate to perceived video quality).

For video denoising, the performance of our current PVF and VPVF approaches may be further improved by incorporating more advanced base denoising algorithms or by improving the fusion method. Future work may also attempt to fuse the denoising results not only from multiple views but also by multiple algorithms. Although the current implementation only fuses the denoising results by the same base denoiser applied along three views, the general PVF and VPVF approaches facilitate fusing the results of any finite number of denoising algorithms. Two issues are critical to the success of this approach. First, the denoising algorithms need to be complementary to one another. Second, the fusion algorithm needs to select the best denoising result among many or optimally assign weights to multiple denoising results. In our current experiment, we observe that 2D approaches from different views tend to be more complementary to each other than 3D approaches, which have already considered the dependencies between neighboring pixels from all directions. Since the structural regularities exhibited in the top- and side-views are substantially different from those in the front-view (as can be observed in Fig.5.2), it is preferable to use different denoising methods best suited to the corresponding views before fusing the results. Currently, no denoising algorithm specifically tuned to denoise from top- and side-views has been developed. This gap suggests another interesting topic for future study.

Finally, the idea of the proposed PVF and VPVF strategies can be extended to solve the video frame rate up-conversion problem using image interpolation algorithms. The frame rate up-conversion technique, which increases the frame rate of the moving pictures by inserting newly generated frames into the original sequence, is highly desirable in the video industry, especially for high-definition TV. If a side- or top- view video frame is considered as an image, the problem of frame rate up-conversion can be transformed to estimate the missing column/row of that image, which is an image interpolation issue. In this case, existing image interpolation algorithms (such as the curvature interpolation method (CIM) [139] and edge-directed interpolation [140]) can be applied for frame rate up-conversion and be combined with the PVF and VPVF fusion methods for enhancement. In addition, because of the different characteristics between front-view and side-/top- view frames, new interpolation approaches tuned to the latter will be an interesting research topic to explore.

Appendices

Appendix A

Objective Quality Assessment in Video Compression

In response to the development of multimedia communication systems and video technology, the amount of data for video signals is exponentially increasing according to a forecast white paper by Cisco Systems Inc. [141]. Video compression or coding plays a critical role in this process and deserves a huge global market. Significant progress has been made recently towards the next generation video-coding standards by the joint collaborative team on video coding (JCT-VC). Recently reported preliminary subjective tests, conducted by JCT-VC members, show that the test model of high efficiency video coding (HEVC) draft codec HM5.0 achieves an average of more than 50% rate savings over H.264 JM18.3 codec without sacrificing subjective quality. Here we study the performance of well-known objective video quality assessment (VQA) models and find that state-of-the-art models, including the structural similarity (SSIM) [1], the multi-scale SSIM index (MS-SSIM) [2], the video quality metric (VQM) [87], and the motion-based video integrity evaluation index (MOVIE) [45], all provide significantly better predictions of subjective video quality than peak signal-to-noise ratio (PSNR) model. Surprisingly, compared with subjective evaluation scores, all objective VQA models systematically underestimate the coding gain of HEVC-HM5.0 upon H.264-JM18.3. We carried out further subjective tests to study this somewhat unexpected phenomenon by comparing JM18.3 and HM5.0 coded videos in terms of frame-level and sequence-level quality, as well as flickering and ghosting effects. The results provide new insights for the future development of subjective/objective VQA and perceptually-tuned video coding methods.

A.1 Introduction

Since the official joint call for proposals (CfP) [142] on the next generation video compression standard was announced in January 2010 by ISO/IEC moving picture experts group (MPEG) and ITU-T video coding experts group (VCEG), the JCT-VC has made significant progress in developing the test model, HEVC, which targets reducing the 50% bit-rate of the MPEG4/H.264 AVC standard while maintaining the same level of subjective quality. Recently, a preliminary subjective test was conducted by JCT-VC members to quantify the rate-distortion (RD) gain of the HEVC draft codec HM5.0 against a similarly-configured H.264/AVC JM18.3 codec [143]. The results show that an average RD-gain of 57.1% is achieved based on the subjective test data in the form of mean opinion scores (MOSs). A more detailed objective and subjective evaluation of HM5.0 was reported in [144], which again suggested that HM5.0 has achieved the target of 50% RD gain over H.264/AVC and that the actual savings can be even higher. Although these subjective tests and evaluations were on random access coding configuration only and more comprehensive tests are still to be conducted, it is speculated that similar improvement may be achieved under other test conditions, and thus HEVC is very likely to achieve its initial RD performance target.

While subjective quality assessment is essential in fully validating the performance of video codecs, it is also highly desirable to know how the existing objective image and video quality assessment (IQA/VQA) models predict the subjective test results and the coding performance. In recent decades, objective IQA/VQA models have been an active research topic, in which aimed to develop ways to automatically predict perceived image and video quality of human subjects. These models are useful in real world applications to control and maintain the quality of image/video processing and communication systems on the fly, where subjective quality assessment is often too slow and costly. They may also be embedded into the design and optimization of novel algorithms and systems to improve perceived image/video quality. Compared with IQA, VQA is a much more challenging problem because of the additional complications due to temporal distortions and our limited understanding of motion perception and temporal visual pooling. Traditionally, PSNR has been used as the "default" criterion in the video coding community in the design, validation and comparison of video codecs. Although PSNR is widely criticized for its poor correlation with perceived image quality and many perceptual objective IQA/VQA models have been proposed in the literature [145], currently PSNR is still the primary objective quality reference in codec development (such as HEVC) mostly by convention and its low complexity.

Given the subjective test data in the form of MOSs collected by JCT-VC members that compare H.264-JM18.3 and HEVC-HM5.0 [143], here we reexamine well-known objective VQA algorithms that emerged in the past decade by observing how well they predict the subjective scores of compressed video sequences and how well they predict the RD-gain between HEVC-HM5.0 and H.264-JM18.3. Moreover, we carry out further subjective tests to exploit the relationship between frame-level and sequence-level subjective quality, and to investigate special temporal coding artifacts created by standard video codecs. This study may help the video coding community select useful VQA models for their future validation and comparison of novel video codecs, may provide new insights about the perceptual aspects of H.264 and HEVC coding schemes and how they may be further improved, and may also help VQA researchers discover the problems in the current subjective testing methodologies and objective VQA models and find ways to improve these models.

A.2 Test of Objective Video Quality Assessment Models

Five existing objective VQA models are being examined here:PSNR, VQM [87], SSIM [1, 37] (As in [1], a preprocessing step of spatial downsampling by a factor of 2 is applied to each frame before the SSIM index is computed), MS-SSIM [2], and MOVIE [45]. All five models are well-known in the IQA/VQA and video coding communities. In particular, VQM has been recommended by the video quality experts group (VQEG) and adopted as a north America standard. SSIM (together with PSNR) is commonly included in popular video codecs such as x264 and VP8 as a quality index automatically computed after the video frames are encoded. MOVIE achieved superior performance in the widely noted LIVE video database [16].

In the subjective data given in [143], a total of 72 HM5.0 and JM18.3 compressed video sequences were tested, which were generated from 9 original source video sequences, including 5 Class B sequences of 1080p resolution (1920 × 1080) and 4 Class C sequences of WVGA resolution (854×480). The encoding configuration of HM5.0 was set as random-access high-efficiency (RA-HE), and for fair comparison, the JM18.3 configuration was adjusted accordingly to best match that of HM5.0. No rate control scheme has been applied to either JM18.3 or HM5.0 encoding. The specific details of coding configurations can be found in [143]. The subjective test results were recorded in the form of MOS for each test video sequence.

Table A.1: Quality prediction performance comparison of PSNR, VQM, MOVIE, SSIM and MS-SSIM

VQA Model	PLCC	MAE	RMS	SRCC	KRCC
PSNR	0.5408	1.1318	1.4768	0.5828	0.3987
VQM [87]	0.8302	**0.7771**	0.9768	0.8360	0.6243
MOVIE [45]	0.7164	0.9711	1.2249	0.6897	0.4720
SSIM [1]	0.8422	0.8102	0.9467	0.8344	0.6279
MS-SSIM [2]	**0.8526**	0.7802	**0.9174**	**0.8409**	**0.6350**

Table A.2: Complexity and coding gain prediction performance comparison of PSNR, VQM, MOVIE, SSIM and MS-SSIM

VQA Model	Computational Complexity (normalized)	RD-gain (Class B)	RD-gain (Class C)	RD-gain (Average)
PSNR	1	-45.0%	-34.1%	-39.6%
VQM [87]	1083	-43.1%	-31.9%	-38.6%
MOVIE [45]	7229	-36.4%	-25.1%	-33.8%
SSIM [1]	5.874	-45.5%	-32.8%	-39.2%
MS-SSIM [2]	11.36	-46.8%	-34.6%	-40.7%
MOS	-	-66.9%	-47.2%	-57.1%

The following criteria were used to evaluate the quality and coding gain prediction performance of each objective VQA model, as well as the relative complexities.

- Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficient (SRCC), which were introduced in Section 3.2.
- Mean absolute error (MAE) is calculated using the converted objective scores after the nonlinear mapping described above:

$$\text{MAE} = \frac{1}{N} \sum |q_i - o_i| \,. \tag{A.1}$$

- Root mean-squared (RMS) error is computed similarly as

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum (q_i - o_i)^2}\,. \tag{A.2}$$

- Kendall's rank correlation coefficient (KRCC) is a non-parametric rank-order based correlation evaluation measures, which is given by

$$\mathrm{KRCC} = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}\,, \tag{A.3}$$

  where N_c and N_d are the numbers of concordant and discordant pairs in the data set, respectively. It is independent of any fitting function that attempts to align the scores.

- Because speed is often a major concern in real-world applications of VQA models, the computational complexities of the VQA models, which are reported as their relative computation time normalized by the computation time of PSNR (this should be considered as only a crude estimate of the computational complexities of the VQA models because no algorithm and/or code optimization has been conducted to accelerate the speed).

- the RD-gain of HM5.0 over JM18.3 is estimated for each source video sequence by comparing the RD curves of HM5.0 and JM18.3, where R denotes bit-rate and D denotes the distortion measure based on the specific VQA model and each RD curve is created by piecewise linear interpolation of the rate and distortion values of four coded video sequences generated by the same coding scheme [143]. The average RD-gain of HM5.0 over JM18.3 is then computed as the average of the RD-gains of all source videos.

The premium performance of objective quality models is represented by higher PLCC, SRCC, and KRCC, and lower MAE and RMS values for quality prediction, less computation time, and better RD-gain prediction compared with that of subjective score.

The quality prediction performance of the objective models over all test video sequences are showed in Table A.1, where the best performances are highlighted with bold face. In Table A.2, the normalized complexity and coding gain prediction performance for each VQA model is summarized. The scatter plots of objective scores versus MOSs are shown in Fig. A.1. From Table A.1, A.2, and Fig. A.1, it can be observed that all four state-of-the-art VQA models clearly outperform PSNR in terms of PLCC, MAE, MSE, SRCC and KRCC, where on average MS-SSIM obtains slightly better results than the other three. On the other hand, VQM and MOVIE are extremely expensive in computational cost, while SSIM and MS-SSIM achieves a much better balance between quality prediction accuracy and computational complexity.

Table A.3 reports the paired statistical significance comparison (*t*-test), which assumes that the MOS residuals are Gaussian distributed, using the approach introduced in [126], where a symbol "1" denotes the objective model of the row is statistically better than that of the column, "0" denotes that the column model is better than the row model, and "-" denotes that the two objective models are statistically indistinguishable. The Wilcoxon-Mann-Whitney test (*h*-test) [146], which is a non-parametric test and does not require a Gaussian distribution of MOS residuals, is also conducted to measure the statistical significance among different VQA methods. Exactly the same results as those in the *t*-test are obtained.

Perhaps the most surprising results here is in the RD-gain columns in Table A.2 – the five objective VQA models predict the average RD-gain of HM5.0 against JM18.3 to be between 33.8% to 40.7% , which largely underestimates the 57.1% gain obtained from subjective scores. Similar behaviors are also observed for individual test classes. This suggests that all objective VQA models are systematically in favor

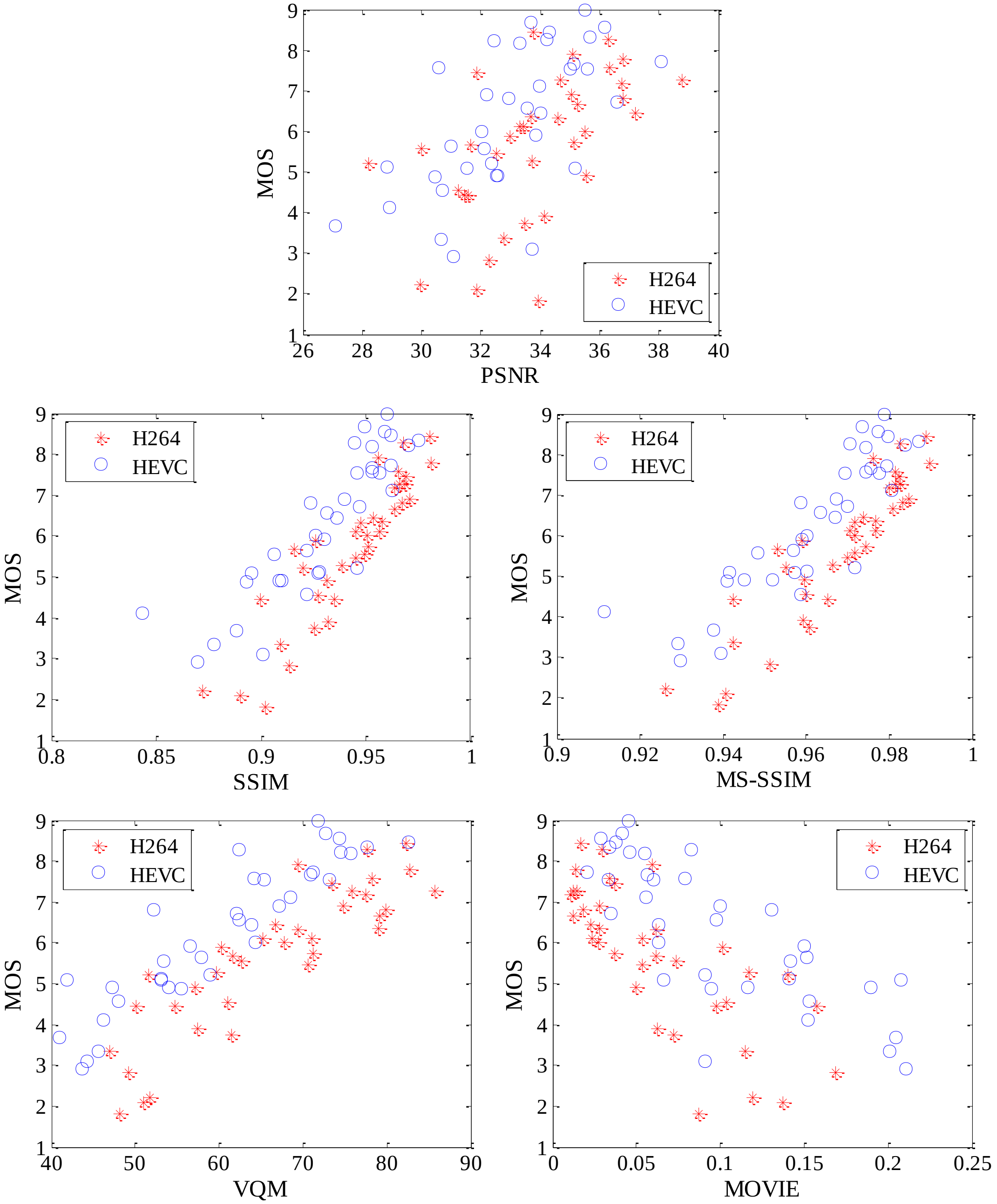


Figure A.1: Scatter plots of VQA measure vs. MOS

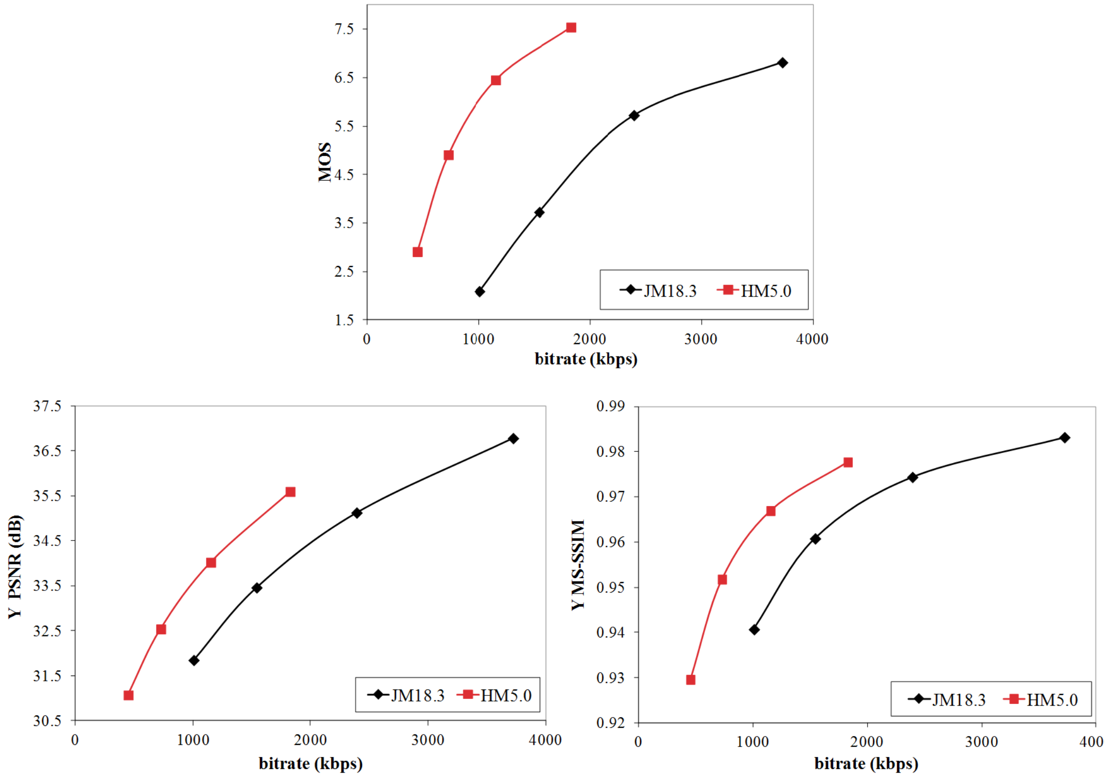


Figure A.2: Rate-quality comparison of JM18.3 and HM5.0 compressed 1080p "ParkScene" sequence, where the quality measures are MOS (left), PSNR (middle) and MS-SSIM (right), respectively. The RD-gain of HM5.0 upon JM18.3 computed using MOS, PSNR, and MS-SSIM are -63.6%, -36.8%, and -39.4%, respectively.

Table A.3: Statistical significance test for PSNR, MOVIE, VQM, SSIM and MS-SSIM

	PSNR	MOVIE	VQM	SSIM	MS-SSIM
PSNR	–	–	0	0	0
MOVIE	–	–	–	0	0
VQM	1	–	–	–	–
SSIM	1	1	–	–	–
MS-SSIM	1	1	–	–	–

of H.264 JM18.3 while human subjects tend to prefer HEVC HM5.0. This can also be seen in Fig. A.1, where in all scatter plots, the clusters of HM5.0 and JM18.3 coded video sequences are visually separated (though with overlaps), and HM5.0 sequences tends to have higher MOS values. Fig. A.2 provides an example using 1080p "Parkscene" sequence, where we can observe how subjective and objective video quality measures change as a function of bit rate. Again, it can be seen that the gap between the HM5.0 and JM18.3 MOS-rate curves is significantly larger than those of the PSNR-rate and (MS-SSIM)-rate curves. Similar phenomena had been observed partially in previous studies. In [144], it was reported that PSNR accounts for 39% rate savings of HM5.0 over JM18.3, as compared to more than 50% by human subjective scores. Similar results are also found in [147]. In [148], the coding performance of HM5.0 and JM16.2 was compared under the RA-HE test conditions over 15 test sequences in terms of perceptual quality index (PQI) [50], PSNR and SSIM [1], and the results showed that the predicted RD-gain by all VQA models are almost the same.

A.3 Subjective Study of Spatial and Temporal Video Quality

To better understand the significant bias of objective VQA models towards H.264-JM18.3 as opposed to HEVC-HM5.0, we carried out a series of subjective experiments to inspect the quality of coded video sequences at both frame and sequence levels. Ten compressed sequences (5 by JM18.3 and 5 by HM5.0) were selected and 5 frames were chosen randomly from each sequence, resulting in totally 50 still image frames. 17 naïve observers participated in the subjective assessment session. The test method conforms with ITU-T BT.500 [149]. Absolute categorical rating (ACR) was adopted to collect the MOS which is the average of subjective opinion from all observers. Four tests have been carried out. The first test is to assess frame-level image quality, where the subjects give scores regarding the quality of the 50 individual still image frames. The second test is on sequence level, where the subjects report a single score for each test video sequence. In the third and the fourth tests, the subjects are asked to evaluate the flickering and ghosting effects of the test video sequences, where flickering refers to the discontinuities of local average luminance over time, and ghosting refers to the traces of video content in previous frames that are remained in the current frame (often created by the Skip mode in the video codec).

From our subjective test, we have the following observations. First, there are significant conflicts between frame-level and sequence-level quality assessment. This can be seen from the top plot in Fig. A.3, where frame-level MOSs (computed by averaging all still frame MOS values of a sequence) and sequence-level MOSs obtained in our subjective experiment do not correlate well with each other. In addition, there is a clear tendency that HM5.0 coded videos obtain higher sequence-level MOSs and lower frame-level MOSs in comparison with JM18.3. A visual example

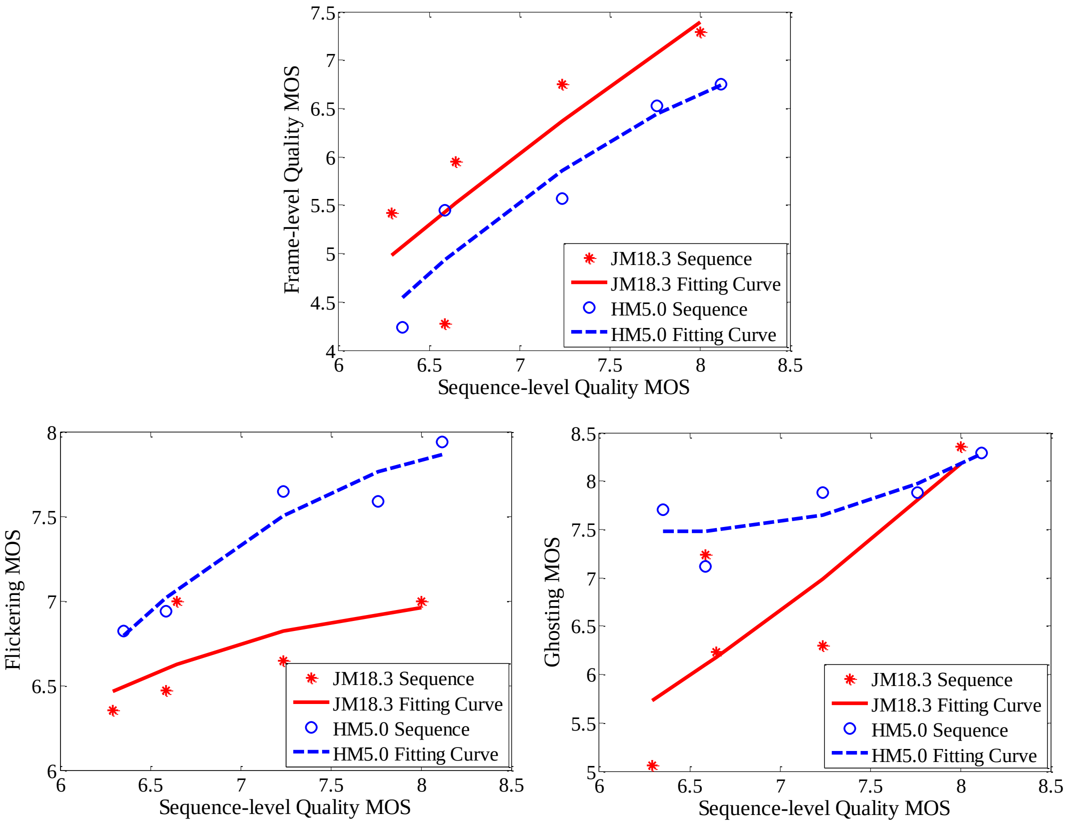
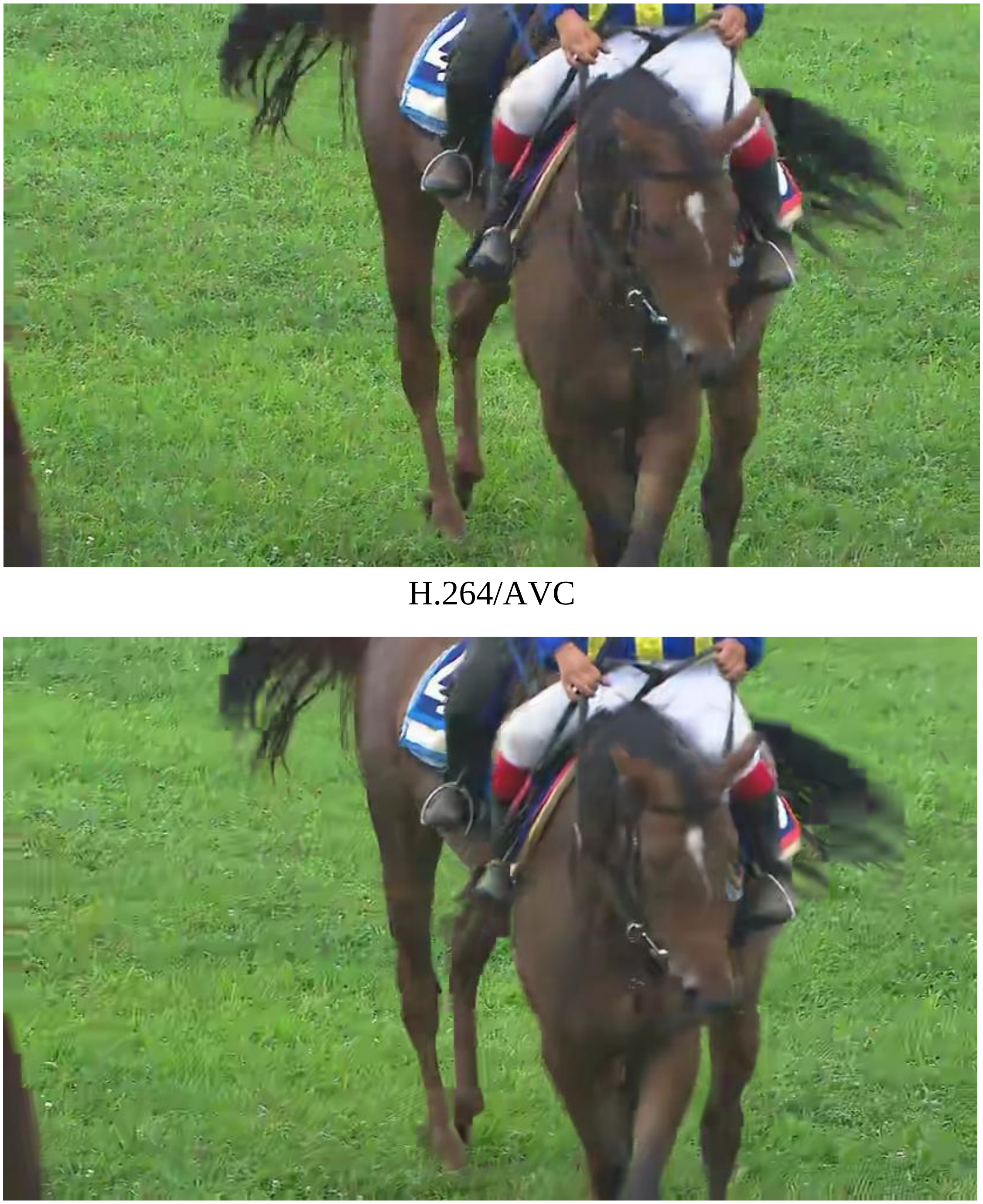


Figure A.3: Relationship between subjective test results for JM18.3 and HM5.0 coded sequences. Left: sequence-level MOS vs. average frame-level MOS; middle: sequence-level MOS vs. flickering MOS; right: sequence-level MOS vs. ghosting MOS.

is given in Fig. A.4, which shows a still frame extracted from a JM18.3 and an HM5.0 coded “Horse” sequences. They are both B-frames with double bit-rate for H.264 than HEVC encoded sequences. On a high quality monitor, the JM18.3 frame appears to better preserve the image details and thus has better quality. The same phenomenon has been observed in all frames throughout the whole video sequences. By contrast, the sequence-level MOS of the HM5.0 video is significantly higher than that of the JM18.3 video. Fig. A.5 depicts an example of the contradiction between frame-level objective quality score and sequence-level MOSs. This observation, combined with the fact that frame-based objective VQA measures often well predicts frame-level MOS (in our experiment, the SRCC between still frame MOS and MS-SSIM is 0.8627), provides an explanation for why objective VQA tends to underestimate sequence-level subjective quality.

Second, significant annoying temporal artifacts may appear in coded video sequences that may dominate subjective evaluation of video quality. We have included flickering and ghosting assessment in our subjective tests. The scatter plots of sequence-level MOS versus flickering and ghosting are shown in the middle and right plots of Fig. A.3, respectively, where higher flickering or ghosting MOS indicates less flickering or ghosting effect. From these plots, we observe that JM18.3 coded sequences have clearly stronger flickering and ghosting effects than HM5.0 sequences. This is in clear contrast to the left plot in Fig. A.3 and provides strong support of the conjecture that compared with frame-level quality, temporal artifacts contribute strongly to the overall sequence-level quality. Third, there is significant spatial and temporal quality non-uniformity of coded video sequences. Such non-uniformity is partially predicted by the objective VQA models (for example, using the SSIM maps) and is more evident in JM18.3 coded video sequences.

The observations above give us useful insights to address several issues in subjective tests. First, the past experience of the subjects and the context of the



H.264/AVC

HEVC

Figure A.4: An example of visual comparison between H.264-JM18.3 and HEVC-HM5.0 coded videos. Top: H.264 frame, PSNR = 28.36dB, SSIM = 0.8012, MS-SSIM = 0.8601. Bottom: HEVC frame, PSNR = 27.64dB, SSIM = 0.7437, MS-SSIM = 0.8259. When comparing individual frames, H.264 frame appears to have clearly better visual quality, but when the video is played at normal speed, the H.264 video receives a significantly lower quality score likely due to strong temporal artifacts.

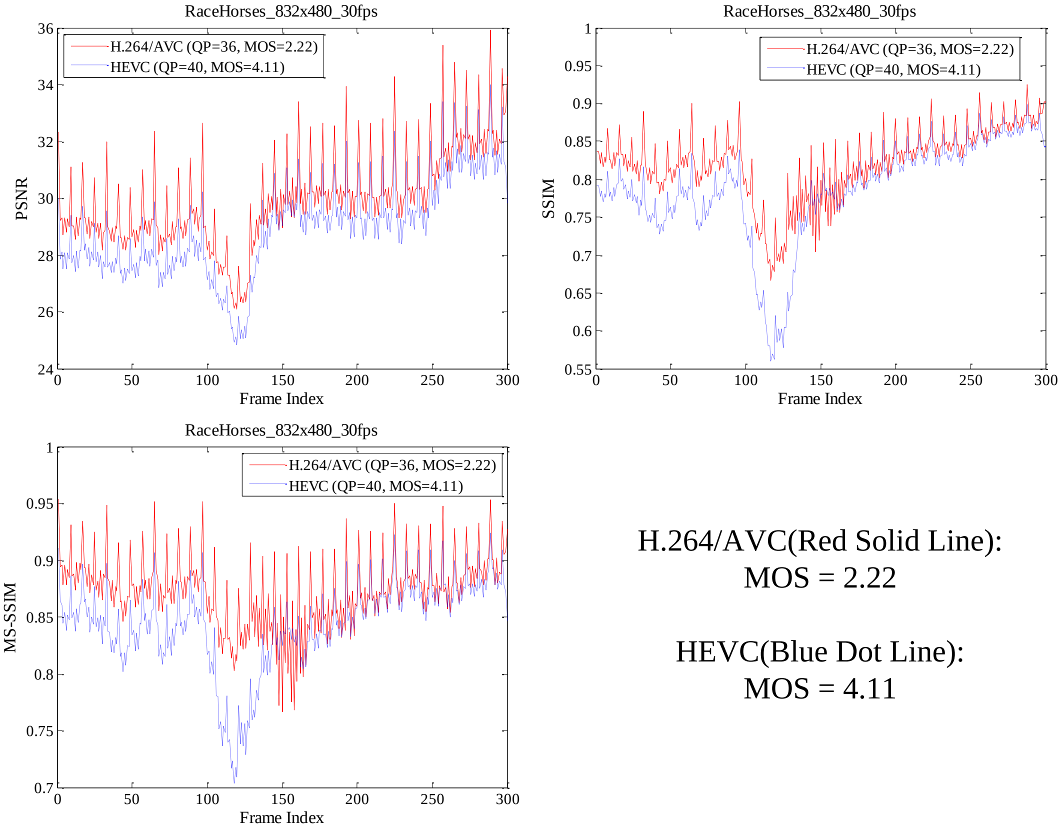


Figure A.5: An example of objective QA results for each frame, which is contradictory with MOS

subjective experiment need to be better taken into account. For example, is the subject a naïve observer or a video quality expert? Is there an pre-training session before the test and what videos are shown in the pre-training phase? What instructions/tasks are given to the subjects − to tell the story behind a movie or to pick artifacts (possibly specified during pre-training) from the video? The subject quality scores could be extremely sensitive to these contexts. Second, questions may be asked to the subjects about what strategies they use to make an overall decision on an entire video sequence that has significant quality non-uniformity over space and/or time. For example, one highly undesirable artifact may appear at a specific location for a short time period, and an subject may give a low quality score to the whole video sequence regardless of the good quality in the rest of the video, but if the subject is attracted into other content and does not see the artifact, then the video may end up with a high subjective quality score. Third, it is desired to record eye movement in the subjective experiments. The importance is not only to detect the regions of interest (ROIs) in the video content, but also to study whether compression artifacts change eye fixations and how the context (e.g., tasks given to the subjects) affects visual attention − are the subjects trying to understand the story of the video content or to detect the distortion artifacts? Previous studies suggest that compression artifacts generally have little impact on visual attention [150], but is this still true when extremely annoying artifacts occur?

A.4 Further Discussion

The observations in the current study raise new questions that need to be answered in the development of objective VQA models. First, there is a strong need to develop novel approaches to capture specific temporal artifacts (such as flickering and ghosting) in compressed video. PSNR, SSIM and MS-SSIM are completely IQA methods, where no inter-frame interactions are considered. It is not surprising that temporal artifacts are missing from these models. However, both VQM and MOVIE consider temporal features, but are still not fully successful in capturing and penalizing the temporal artifacts. Second, many VQA models such as SSIM and MS-SSIM generate useful quality maps that indicate local quality variations over space and time. In the case of significant spatial and temporal non-uniformity in these quality maps, how to pool the maps into a single quality score of the entire video is not a fully resolved problem. There have been attempts to use non-linear models and temporal hysteresis for temporal pooling [151, 152]. However, our current test shown in Table A.4 indicates that they only lead to small improvement over MS-SSIM, and the large gap between subjective and objective RD-gain predictions still exists. Third, it would be useful to incorporate visual attention models. These attention models may be saliency predictors based on both low-level and high-level vision features, and may also be based on detections of severe visual artifacts.

Meanwhile, what we learned from this study may help us improve the design and implementation of video coding technologies. It is useful to be aware of and to avoid certain temporal artifacts such as flickering and ghosting effects, which may vastly change subjects' opinions about the quality of the entire video sequence. Many of these artifacts occur when quantization parameters are not carefully chosen and when Skip mode is selected in low- to mid-energy regions with slow motion. Moreover, rate control and rate-distortion optimization (RDO) schemes may be adjusted not only to achieve the best average quality over the whole video sequence, but also to reduce significant quality fluctuations across both space and time.

Table A.4: The impact of temporal pooling strategies on MS-SSIM method

VQA Model	PLCC	MAE	RMS	SRCC	KRCC	RD-gain (Average)
MS-SSIM [2]	0.8526	0.7802	0.9174	0.8409	0.6350	-40.7%
MS-SSIM with min temporal pooling [151]	0.8670	0.6859	0.8749	0.8645	0.6663	-43.2%
MS-SSIM with temporal hysteresis pooling [152]	0.8544	0.7498	0.9123	0.8467	0.6400	-42.4%
MOS	-	-	-	-	-	-57.1%

A.5 Conclusion and Future Work

In conclusion, based on recently published comparative results regarding the subjective quality of HEVC and H.264/AVC coded sequences, our study about the performance of popular objective VQA models shows that advanced models clearly outperform conventionally used PSNR/MSE in terms of predicting quality scores given by human subjects. One consequence of this observation is that PSNR/MSE may result in incorrect direction for video compression. This also suggests that the video coding community and the standard development body may consider replacing PSNR with a perceptually more meaningful VQA model in not only the testing but also the development phases of novel video codecs. This could lead to substantial changes in the structural design and system optimization of the next generation video codec. In terms of RD-gain prediction, however, none of the objective VQA models aligns well with the subjective test results. We conjecture that this may be due to one (or the combination) of the following issues:

- The ambiguities in subjective testing methodology lead to unreliable or unstable subjective benchmark scores;
- The ability of current VQA models to capture specific types of temporal

artifacts (such as flickering and ghosting) is limited;

- A good spatio-temporal pooling strategy able to account for human perceptual importance weighting is still missing;
- An effective saliency- or artifacts-based visual-attention model may needs to be embedded.

The current discussions are non-conclusive but hopefully could inspire future improvement in both VQA and video coding methodologies.

In terms of VQA in the context of video coding, the analysis conducted in our work mainly contributes to the future development of both VQA models and video coding schemes. For example, it has been shown that an advanced VQA approach capable of capturing specific temporal artifacts (such as flickering and ghosting) in compressed video is urgently needed. An effective spatial and temporal pooling strategy, which is able to combine the highly non-uniform local quality score into one more meaningful final mark, would also be appreciated. In addition, for a successful VQA model, it would be useful to incorporate visual attention models. These attention models may be saliency predictors based on both low-level and high-level vision features, and may also be based on detection of very obvious visual artifacts.

Published Papers

Kai Zeng, Abdul Rehman, Jiheng Wang and Zhou Wang, "From H.264 to HEVC: Coding gain predicted by objective video quality assessment models", *International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM'13)*, Scottsdale, AZ, U.S.A., January 30 - February 1, 2013.

Kai Zeng and Zhou Wang, "3D-SSIM for video quality assessment", *IEEE International Conference on Image Processing (ICIP'12)*, Orlando, FL, U.S.A., September 30 - October 2, 2012.

Kai Zeng and Zhou Wang, "Polyview fusion: a strategy to enhance video denoising algorithms", *IEEE Transaction on Image Processing*, vol. 21, no. 4, pp. 2324-2328, Apr. 2012.

Kai Zeng and Zhou Wang, "Enhancing video denoising algorithms by fusion from multiple views", *International Conference on Image Analysis and Recognition*, Burnaby, BC, Canada, June 22-24, 2011.

Kai Zeng and Zhou Wang, "Quality-aware video based on robust embedding of intra- and inter-frame reduced-reference feature", *IEEE International Conference of Image Processing (ICIP'10)*, HongKong, China, September 26-29, 2010.

Kai Zeng and Zhou Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, U.S.A., March 14-19, 2010.

Wen Lu, **Kai Zeng**, Dacheng Tao, Yuan Yuan and Xinbo Gao, "No-reference Image Quality Assessment in Contourlet Domain," *Neurocomputing (Elsevier)*, vol.73, no.4-6, pp.784-794, 2010.

Wen Lu, Xinbo Gao, **Kai Zeng** and Lihuo He, "Image quality evaluation metrics based on HWD," *Journal of Infrared and Millimeter Waves*, vol. 28, no. 1, pp. 72-76, 2009.

References

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004. 2, 6, 9, 28, 33, 34, 36, 63, 67, 69, 86, 89, 90, 95

[2] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems and Computers*, November 2003, vol. 2, pp. 1398 – 1402. 2, 9, 86, 89, 90, 103

[3] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Processing*, vol. 20, no. 5, pp. 1185 – 1198, May 2011. 2, 9, 26, 28, 29

[4] M.P. Sampat, Z. Wang, S. Gupta, A.C. Bovik, and M.K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. Image Processing*, vol. 18, no. 11, pp. 2385 – 2401, November 2009. 2, 9

[5] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Trans. Image Processing*, vol. 20, no. 8, pp. 2378 – 2386, Augest 2011. 2, 9

[6] H.R. Sheikh and A.C. Bovik, "Image information and visual quality," *IEEE Trans. Image Processing*, vol. 15, no. 2, pp. 430 – 444, Feburary 2006. 2, 9

[7] A. C. Bovik, *Handbook of Image and Video Processing (Communications, Networking and Multimedia)*, Academic Press, Inc., Orlando, FL, USA, 2005. 2, 4

[8] Z. Wang and Q. Li, "Statistics of natural image sequences: Temporal motion smoothness by local phase correlations," in *Human Vision and Electronic Imaging XIV, Proc. SPIE*, January 2009, vol. 7240. 5, 39, 41, 42

[9] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Processing*, vol. 15, no. 6, pp. 1680–1689, June 2006. 5, 19, 38, 42, 45, 46, 80, 82

[10] K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment," in *IEEE Inter. Conf. Acoustics, Speech & Signal Proc.*, March 2010. 5, 45

[11] F. Ourique, V. Licks, R. Jordan, and F. Perez-Gonzalez, "Angle QIM: a novel watermark embedding scheme robust against amplitude scaling distortions," in *IEEE Inter. Conf. Acoustics, Speech and Signal Proc.*, March 2005, vol. 2, pp. 797–800. 5, 47

[12] Ulrich Reiter, Jari Korhonen, and Junyong You, "Comparing apples and oranges: assessment of the relative video quality in the presence of different types of distortions," *Eurasip Journal on Image and Video Processing*, vol. 2011, no. 1, pp. 1–10, 2011. 7

[13] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, Morgan & Claypool Publishers, March 2006. 7, 9, 18, 19, 38, 42

[14] ITU-T P.910 (09/99), "Recommendation: Subjective video quality assessment methods for multimedia applications," September 1999. 7

[15] Marcus Barkowsky, Margaret Pinson, Romuald Pépion, and Patrick Le Callet, "Analysis of freely available dataset for hdtv including coding and transmission distortions," in *Fifth International Workshop on Video Processing and Quality Metrics*, Scottsdale, January 2010. 8

[16] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L.K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010. 8, 12, 89

[17] A.K. Moorthy, Lark Kwon Choi, A.C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012. 8

[18] 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 *International Workshop on Quality of Multimedia Experience, 2009. QoMEx 2009.*, 2009, pp. 204–209. 8

[19] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet, "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*, Kyoto, Japan, Sept. 2008, p. 6. 8

[20] Stefan Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012. 8

[21] J. Lubin, "A visual discrimination model for imaging system design and evaluation," *in Visual Models for Target Detection and Recognition, E. Peli, rkEd. Singapore :: World Scientific*, pp. 207–220, 1993. 9

[22] M. Miyahara, K. Kotani, and V. Algazi, "Objective picture quality scale (PQS) for image coding," *IEEE Trans. Communications*, vol. 46, no. 9, pp. 1215–1226, 1998. 9

[23] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 636–650, 2000. 9

[24] Arthur A. Webster, Coleen T. Jones, Margaret H. Pinson, Stephen D. Voran, and Stephen Wolf, "An objective video quality assessment system based on human perception," *Proc. SPIE Human Vision, Visual Processing, and Digital Display IV, San Jose, CA, USA*, vol. 1913, pp. 15–26, 1993. 10

[25] Christian J. van den Branden Lambrecht and Olivier Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. SPIE, Digital Video Compression: Algorithms and Technologies*, 1996, vol. 2668, pp. 450–461. 10

[26] S. Olsson, M. Stroppiana, and J. Baina, "Objective methods for assessment of video quality: State of the art," *IEEE Trans. Broadcasting*, vol. 43, no. 4, pp. 487–495, December 1997. 10

[27] K. T. Tan, M. Ghanbari, and D. E. Pearson, "An objective measurement tool for MPEG video quality," *Signal Process.*, vol. 70, no. 3, pp. 279–294, Nov. 1998. 10

[28] Andrew B. Watson, "Toward a perceptual video-quality metric," pp. 139–147, 1998. 10

[29] Stefan Winkler, “A perceptual distortion metric for digital color video,” in *in Proc. SPIE*, 1999, pp. 175–184. 10

[30] Stephen Wolf and Margaret H. Pinson, “Spatial-temporal distortion metric for in-service quality monitoring of any digital video system,” pp. 266–277, 1999. 10

[31] Xin Tong, David J. Heeger, and Christian J. Van den Branden Lambrecht, “Video quality evaluation using st-cielab,” *Proc. SPIE 3644, Human Vision and Electronic Imaging IV*, pp. 185–196, 1999. 10

[32] K. T. Tan and M. Ghanbari, “A multi-metric objective picture-quality measurement model for MPEG video,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 7, pp. 1208–1213, October 2000. 11

[33] Zhenghua Yu and H.R. Wu, “Human visual system based objective digital video quality metrics,” in *Signal Processing Proceedings, 2000. WCCC-ICSP 2000. 5th International Conference on*, 2000, vol. 2, pp. 1088–1095 vol.2. 11

[34] Ann Marie Rohaly, Philip Corriveau, John Libert, Arthur Webster, Vittorio Baroncini, John Beerends, Jean-Louis Blin, Laura Contin, Takahiro Hamada, David Hathson, Andries Hekstra, Jeffrey Lubin, Yukihiro Nishida, Ricardo Nishihara, John Pearson, Antonio Franca Pessoa, Neil Pickford, Alexander Schertz, Massimo Visca, Andrew Watson, and Stephan Winkler, “Video Quality Experts Group: Current results and future directions,” in *in Proc. SPIE, Visual Comminications and Image Processing*, King N. Ngan, Thomas Sikora, and Ming-Ting Sun, Eds., 2000, vol. 4067, pp. 742–753. 11

[35] Ohjae Kwon and Chulhee Lee, “Objective method for assessment of video quality using wavelets,” in *IEEE International Symposium on Industrial Electronics*, 2001, vol. 1, pp. 292–295 vol.1. 11

[36] A.P. Hekstra, J.G. Beerends, D. Ledermann, F.E. de Caluwe, S. Kohler, R.H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss, "PVQM - a perceptual video quality measure," *Signal Processing: Image Communication*, vol. 17, no. 10, pp. 781–798, 2002. 11

[37] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication,* Special issue on objective video quality metrics, vol. 19, no. 2, pp. 121–132, February 2004. 11, 25, 89

[38] J. Guo, M. Van Dyke-Lewis, and H. R. Myler, "Gabor difference analysis of digital video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 302–311, September 2004. 11

[39] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America A*, vol. 24, no. 12, pp. B61–B69, December 2007. 11, 26

[40] K. Yang, C. C. Guest, K. El-Maleh, and P. K. Das, "Perceptual temporal quality metric for compressed video," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1528–1535, November 2007. 11

[41] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From psnr to hybrid metrics," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 660–668, September 2008. 12

[42] 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," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 280–293, April 2009. 12

[43] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, April 2009. 12, 25

[44] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, April 2009. 12

[45] K. Seshadrinathan and A.C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Processing*, vol. 19, no. 2, pp. 335–350, February 2010. 12, 26, 33, 34, 36, 86, 89, 90

[46] A.K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A.C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 4, pp. 587–599, April 2010. 12

[47] A.K. Moorthy and A.C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1653–1658, 2010. 12

[48] Q. Huynh-Thu and M. Ghanbari, "Modelling of spatio-emporal interaction for video quality assessment," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 535–546, 2010, Special Issue on Image and Video Quality Assessment. 13

[49] J. You, J. Korhonen, A. Perkis, and T. Ebrahimi, "Balancing attended and global stimuli in perceived video quality assessment," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1269–1285, December 2011. 13, 26, 33, 34, 36

[50] Y. Zhao, L. Yu, Z. Chen, and C. Zhu, "Video quality assessment based on measuring perceptual noise from spatial and temporal perspectives," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 21, no. 12, pp. 1890–1902, December 2011. 13, 25, 95

[51] C. Yim and Alan C. Bovik, "Evaluation of temporal variation of video quality in packet loss networks," *Signal Processing: Image Communnication*, vol. 26, no. 1, pp. 24–38, January 2011. 13

[52] D. Ćulibrk, M. Mirković, V. Zlokolica, M. Pokric, V. Crnojević, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Trans. Image Processing*, vol. 20, no. 4, pp. 948–958, 2011. 13

[53] M. Narwaria and Weisi Lin, "Machine learning based modeling of spatial and temporal factors for video quality assessment," in *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2513–2516. 13

[54] Songnan Li, Lin Ma, and King Ngi Ngan, "Full-reference video quality assessment by decoupling detail losses and additive impairments," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1100–1112, 2012. 13

[55] M. Narwaria, Weisi Lin, and Anmin Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 525–535, 2012. 13

[56] Mikolaj Leszczuk, Lucjan Janowski, Piotr Romaniak, and Zdzislaw Papir, "Assessing quality of experience for high definition video streaming under diverse packet loss patterns," *Signal Processing: Image Communication*, , no. 0, pp. –, 2012. 13

[57] Yue Wang, Tingting Jiang, Siwei Ma, and Wen Gao, "Novel spatio-temporal structural information based video quality metric," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 989–998, 2012. 13

[58] Jincheol Park, K. Seshadrinathan, Sanghoon Lee, and A.C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Processing*, vol. 22, no. 2, pp. 610–620, 2013. 14

[59] M.C.Q. Farias and S.K. Mitra, "No-reference video quality metric based on artifact measurements," in *IEEE Inter. Conf. Image Processing, 2005, ICIP 2005.*, September 2005, vol. 3, pp. III–141–4. 14

[60] F. Yang, S. Wan, Y. Chang, and H. Wu, "A novel objective no-reference metric for digital video quality assessment," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 685–688, October 2005. 14

[61] Y. Kawayoke and Y. Horita, "NR objective continuous video quality assessment model based on frame quality measure," in *15th IEEE Inter. Conf. Image Processing, 2008, ICIP 2008.*, October 2008, pp. 385–388. 15

[62] Shu Tao, J. Apostolopoulos, and R. Guerin, "Real-time monitoring of video quality in ip networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 5, pp. 1052–1065, 2008. 15

[63] M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for H.264/AVC coded video," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 932–946, August 2009. 15

[64] C. Keimel, T. Oelbaum, and K. Diepold, "No-reference video quality evaluation for high-definition video," in *IEEE Inter. Conf. Acoustics, Speech and Signal Proc., 2009. ICASSP 2009*, April 2009, pp. 1145–1148. 15

[65] M.A. Saad and A.C. Bovik, “Natural motion statistics for no-reference video quality assessment,” in *Inter. Workshop Quality of Multimedia Experience, 2009. QoMEx 2009.*, July 2009, pp. 163–167. 15

[66] Dubravko Ćulibrk, Dragan Kukolj, Petar Vasiljević, Maja Pokrić, and Vladimir Zlokolica, “Feature selection for neural-network based no-reference video quality assessment,” in *Proceedings of the 19th International Conference on Artificial Neural Networks: Part II*, Berlin, Heidelberg, 2009, ICANN ’09, pp. 633–642, Springer-Verlag. 15

[67] Quan Huynh-Thu and M. Ghanbari, “No-reference temporal quality metric for video impaired by frame freezing artefacts,” in *16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2221–2224. 15

[68] T. Brandão and M.P. Queluz, “No-reference quality assessment of H.264/AVC encoded video,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1437–1447, November 2010. 15

[69] S. S. Hemami and A. R. Reibman, “No-reference image and video quality estimation: Applications and human-motivated design,” *Signal Processing: Image Communication*, vol. 25, pp. 469–481, August 2010. 15

[70] F. Yang, S. Wan, Q. Xie, and H. Wu, “No-reference quality assessment for networked video via primary analysis of bit stream,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1544–1554, November 2010. 16

[71] Ning Liao and Zhibo Chen, “A packet-layer video quality assessment model based on spatiotemporal complexity estimation,” *Proc. SPIE, Visual Communications and Image Processing*, vol. 7744, pp. 77441K–1–77441K–10, 2010. 16

[72] T. Kawano, K. Yamagishi, K. Watanabe, and J. Okamoto, "No reference video-quality-assessment model for video streaming services," in *18th International Packet Video Workshop (PV)*, 2010, pp. 158–164. 16

[73] Tao Liu, G. Cash, Wen Chen, Chunhua Chen, and J. Bloom, "Real-time video quality monitoring for mobile devices," in *44th Annual Conference on Information Sciences and Systems (CISS)*, 2010, pp. 1–6. 16

[74] S. Argyropoulos, A. Raake, M. N Garcia, and P. List, "No-reference video quality assessment for SD and HD H.264/AVC sequences based on continuous estimates of packet loss visibility," in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, 2011, pp. 31–36. 16

[75] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet, "A metric for no-reference video quality assessment for HD TV delivery based on saliency maps," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–5. 16

[76] Xingang Liu, Min Chen, Tang Wan, and Chen Yu, "Hybrid no-reference video quality assessment focusing on codec effects.," *TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS*, vol. 5, no. 3, pp. 592–606, 2011. 16

[77] Hui Shi and Xiuhua Jiang, "No-reference quality assessment for video streams based on china mobile multimedia broadcast," in *International Conference on Multimedia Technology (ICMT)*, 2011, pp. 584–587. 16

[78] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "No-reference pixel video quality monitoring of channel-induced distortion," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 4, pp. 605–618, 2012. 16

[79] Caihong Wang, Xiuhua Jiang, Yuxia Wang, and Fang Meng, “Quality assessment for mpeg-2 video streams,” in *Recent Progress in Data Engineering and Internet Technology*, Ford Lumban Gaol, Ed., vol. 157 of *Lecture Notes in Electrical Engineering*, pp. 453–458. Springer Berlin Heidelberg, 2012. 17

[80] Xiangyu Lin, Xiang Tian, and Yaowu Chen, “No-reference video quality assessment based on region of interest,” in *2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2012, pp. 1924–1927. 17

[81] H. Boujut, J. Benois-Pineau, T. Ahmed, O. Hadar, and P. Bonnet, “No-reference video quality assessment of H.264 video streams based on semantic saliency maps,” in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, Jan. 2012, vol. 8293 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series.* 17

[82] Jie Yao, Yongqiang Xie, Jianming Tan, Zhongbo Li, Jin Qi, and Lanlan Gao, “No-reference video quality assessment using statistical features along temporal trajectory,” *Procedia Engineering*, vol. 29, no. 0, pp. 947 – 951, 2012, ¡ce:title¿2012 International Workshop on Information and Electronics Engineering¡/ce:title¿. 17

[83] C. Bailey, M. Seyedebrahimi, and Xiao-Hong Peng, “Pause intensity: A no-reference quality assessment metric for video streaming in TCP networks,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 818–823. 17

[84] Fan Zhang, Weisi Lin, Zhibo Chen, and King Ngi Ngan, “Additive log-logistic model for networked video quality assessment,” *IEEE Trans. Image Processing*, vol. 22, no. 4, pp. 1536–1547, 2013. 17

[85] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2013. 17

[86] Bumshik Lee and M. Kim, "No-reference PSNR estimation for hevc encoded video," *IEEE Trans. Broadcasting*, vol. 59, no. 1, pp. 20–27, 2013. 18

[87] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, September 2004. 19, 33, 34, 36, 86, 89, 90

[88] P. Le Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Networks*, vol. 17, no. 5, pp. 1316–1327, 2006. 19

[89] T. Oelbaum and K. Diepold, "A reduced reference video quality metric for AVC/H.264," in *15th European Signal Processing Conf., Poznan, Poland, 2007, EUSIPCO 2007*, 2007, pp. 1265–1269. 19

[90] I. P. Gunawan and M. Ghanbari, "Reduced-reference video quality assessment using discriminative local harmonic strength with motion consideration," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 18, no. 1, pp. 71–83, January 2008. 19

[91] I. P. Gunawan and M. Ghanbari, "Efficient reduced-reference video quality meter," *IEEE Trans. Broadcasting*, vol. 54, no. 3, pp. 669–679, September 2008. 20

[92] Wen Lu, Xuelong Li, Xinbo Gao, Wenjian Tang, Jing Li, and Dacheng Tao, "A video quality assessment metric based on human visual system," *Cognitive Computation*, vol. 2, no. 2, pp. 120–131, 2010. 20

[93] M. N Garcia and A. Raake, "Parametric packet-layer video quality model for IPTV," in *10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA)*, 2010, pp. 349–352. 20

[94] Caihong Wang, Xiuhua Jiang, and Yuxia Wang, "Content-related features for video quality assessment based on bit streams," in *Advances in Automation and Robotics, Vol. 2*, Gary Lee, Ed., vol. 123 of *Lecture Notes in Electrical Engineering*, pp. 223–230. Springer Berlin Heidelberg, 2012. 20

[95] Chun-Yu Yang, Hsin-Ho Yeh, and Chu-Song Chen, "Video aesthetic quality assessment by combining semantically independent and dependent features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1165–1168. 20

[96] Yuzhen Niu and Feng Liu, "What makes a professional video? a computational aesthetics approach," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1037–1049, 2012. 21

[97] Lin Ma, Songnan Li, and King Ngi Ngan, "Reduced-reference video quality assessment of compressed video sequences," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 10, pp. 1441–1456, 2012. 21

[98] Fuzheng Yang, Jiarun Song, Shuai Wan, and Hong Ren Wu, "Content-adaptive packet-layer model for quality assessment of networked video services," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 672–683, 2012. 21

[99] Luigi Atzori, Alessandro Floris, Giaime Ginesu, and Daniele Giusto, "Streaming video over wireless channels: Exploiting reduced-reference quality estimation at the user-side," *Signal Processing: Image Communication*, vol. 27, no. 10, pp. 1049–1065, November 2012. 21

[100] B. Karacali and A. S. Krishnakumar, "Measuring video quality degradation using face detection," in *35th IEEE Sarnoff Symposium (SARNOFF)*, 2012, pp. 1–5. 21

[101] Yen-Fu Ou, Yuanyi Xue, and Yao Wang, "Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *Computing Research Repository*, vol. abs/1206.2320, 2012. 21

[102] J. S. Lim, *Two-Dimensional Signal and Image Processing*, Englewood Cliffs, NJ, Prentice Hall, 1990. 22

[103] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Processing*, vol. 12, pp. 1338–1351, 2003. 22, 70

[104] A. Buades, B. Coll, and J. M. Morel, "Nonlocal image and movie denoising," *Inter. Journal of Computer Vision*, vol. 76, pp. 123–139, February 2008. 22

[105] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, November 2006. 22, 70

[106] T. Blu and F. Luisier, "The SURE-LET approach to image denoising," *IEEE Trans. Image Processing*, vol. 16, no. 11, pp. 2778–2786, November 2007. 22

[107] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Processing*, vol. 16, no. 8, pp. 2080–2095, August 2007. 22

[108] V. Zlokolica, A. Pizurica, and W. Philips, "Wavelet-domain video denoising based on reliability measures," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 16, no. 8, pp. 993–1007, August 2006. 22

[109] G. Varghese and Zhou Wang, "Video denoising based on a spatiotemporal Gaussian scale mixture model," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 7, pp. 1032–1040, July 2010. 22, 23

[110] F. Luisier, T. Blu, and M. Unser, "SURE-LET for orthonormal wavelet-domain video denoising," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 20, no. 6, pp. 913–919, June 2010. 22, 23, 70

[111] A. Buades, B. Coll, and J.M. Morel, "Denoising image sequences does not require motion estimation," in *IEEE Conf. Advanced Video and Signal Based Surveillance, 2005. AVSS 2005*, September 2005, pp. 70–74. 23, 70

[112] K. Dabov, A. Foi, and K. Egiazarian, "Video denoising by sparse 3D transform-domain collaborative filtering," in *Proc. of 15th Euro. Signal Processing Conf.*, Poznan, Poland, September 2007. 23, 63, 70

[113] M. K. Ozkan, M. I. Sezan, and A. M. Tekalp, "Adaptive motion-compensated filtering of noisy image sequences," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 3, no. 4, pp. 277–290, August 1993. 23

[114] G. R. Arce, "Multistage order statistic filters for image sequence processing," *IEEE Trans. Signal Processing*, vol. 39, no. 5, pp. 1146–1163, May 1991. 23

[115] J. Kim and J. W. Woods, "Spatiotemporal adaptive 3-D Kalman filter for video," *IEEE Trans. Image Processing*, vol. 6, pp. 414–424, 1997. 23

[116] J. C. Brailean and A. K. Katsaggelos, "Recursive displacement estimation and restoration of noisy-blurred image sequences," vol. 5, pp. 273–276, April 1993. 23

[117] I. W. Selesnick and K. Y. Li, "Video denoising using 2D and 3D dualtree complex wavelet transforms," in *Proc. SPIE, Wave.: App. in Signal and Image Process. X*, San Diego, November 2003, vol. 5207, pp. 607–618. 23

[118] M. Protter and M. Elad, "Image sequence denoising via sparse and redundant representations," *IEEE Trans. Image Processing*, vol. 18, no. 1, pp. 27–35, January 2009. 23

[119] X. Li and Y. Zheng, "Patch-based video processing: A variational bayesian approach," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 19, no. 1, pp. 27–40, January 2009. 23

[120] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993. 25

[121] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Spatio-temporal quality pooling accounting for transient severe impairments and egomotion," in *18th IEEE Inter. Conf. on Image Process. (ICIP)*, Sept. 2011, pp. 2509–2512. 26, 29

[122] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *2006 IEEE Inter. Conf. on Image Process.*, Oct. 2006, pp. 2945–2948. 29

[123] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005. 29, 31

[124] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "The SSIM index for image quality assessment," `http://www.cns.nyu.edu/~lcv/ssim/`. 31, 33, 34, 36

[125] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Apr. 2000, available at http://www.vqeg.org/. 32, 33

[126] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006. 32, 92

[127] N. Ponomarenko, F. Battisti, K. Egiazarian, J. Astola, and V. Lukin, "Metrics performance comparison for color image database," in *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, Arizona, USA, Jan. 2009. 32

[128] B. Hiremath, Q. Li, and Z. Wang, "Quality-aware video," in *IEEE Inter. Conf. Image Processing*, San Antonio, TX, September 2007. 38, 42, 80

[129] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Inter. Journal of Computer Vision*, vol. 40, no. 1, pp. 49–71, December 2000. 42, 56

[130] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Information Theory*, vol. 38, no. 2, pp. 587–607, 1992. 42, 56

[131] N. I. Fisher, *Statistical analysis of circular data*, Cambridge University Press, New York, 2000. 42

[132] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Human Vision and Electronic Imaging X, Proc. SPIE*, San Jose, CA, January 2005, vol. 5666. 45

[133] T. K. Moon, *Error Correction Coding: Mathematical Methods and Algorithms*, Wiley-Interscience, 2005. 47

[134] F. Dufaux and F. Moscheni, "Motion estimation techniques for digital TV: A review and a new contribution," *Proceedings of IEEE*, vol. 83, no. 6, pp. 858–876, June 1995. 60, 80

[135] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys*, vol. 27, no. 3, pp. 433–467, September 1995. 60, 80

[136] "http://www.mathworks.com/help/toolbox/images/ref/wiener2.html," . 70

[137] "http://www4.io.csic.es/pagspers/jportilla/portada/software," . 70

[138] "http://www.cs.tut.fi/˜foi/gcf–bm3d/bm3d.zip," . 70

[139] Hakran Kim, Youngjoon Cha, and Seongjai Kim, "Curvature interpolation method for image zooming," *IEEE Trans. Image Processing*, vol. 20, no. 7, pp. 1895 – 1903, July 2011. 84

[140] Xin Li and M.T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Processing*, vol. 10, no. 10, pp. 1521 – 1527, October 2001. 84

[141] Cisco System Inc., "Cisco visual networking index: Forecast and methodology, white paper, 2011-2016," 2012. 86

[142] ITU-T Q6/16 Visual Coding and ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Audio, "Joint call for proposals on video compression technology," in *MPEG Document N11113*, Kyoto, Japan, January 2010. 87

[143] V. Baroncini, J. R. Ohm, and G. J. Sullivan, "Report on preliminary subjective testing of hevc compression capability," in *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, San José, CA, USA*, 2012. 87, 88, 89, 90, 91

[144] T. K. Tan, A. Fujibayashi, and J. Takiue, "Ahg8: Objective and subjective evaluation of hm5.0," in *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, San José, CA, USA, 2012. 87, 95

[145] Z. Wang and Alan C. Bovik, "Mean squared error: love it or leave it? - a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, pp. 98 – 117, 2009. 88

[146] Michael P. Fay and Michael A. Proschan, "WilcoxonMannWhitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules," *Statistics Surveys*, vol. 4, pp. 1–39, 2010. 92

[147] B. Li, G. J. Sullivan, and J. Xu, "Comparison of compression performance of hevc working draft 5 with avc high profile," in *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, San José, CA, USA, Febuary 2012. 95

[148] Y. Zhao and L. Yu, "Coding efficiency comparison between hm5.0 and jm16.2 based on pqi, psnr and ssim," in *JCT-VC of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, San José, CA, USA, Febuary 2012. 95

[149] ITU-R BT.500-12, "Recommendation: Methodology for the subjective assessment of the quality of television pictures," November 1993. 96

[150] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?," *Signal Processing: Image Communication*, vol. 25, no. 8, pp. 597 – 609, 2010. 101

[151] C. Keimel and K. Diepold, "Improving the prediction accuracy of psnr by simple temporal pooling," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010, vol. 2009. 102, 103

[152] K. Seshadrinathan and A.C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, May 2011, pp. 1153 – 1156. 102, 103