PERCEPTUAL QUALITY ASSESSMENT OF VIDEOS AFFECTED BY PACKET LOSSES

by

Tao Liu

Advisor: Yao Wang

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy (Electrical Engineering)

January 2010

Due to rapid advance of various video applications and services, such as video telephony, mobile video broadcasting, and Internet Protocol television (IPTV), there is an increasing demand for accurate and effective quality assessment of underlying videos. Accurate video quality assessment is crucial to video codec development, network protocol planning, in-network quality monitoring, quality assurance of end users, etc.

This thesis develops several objective quality metrics for videos impaired by packet losses. Because transmission error is one of the main causes of quality degradations of networked video, a deep investigation on impacts of different attributes of packet losses on perceptual video quality is performed. Based on the observed relationships between perceptual video quality and various attributes of packet losses, e.g. error length, the loss severity, loss location, the number of losses, and loss patterns, we

incorporate a prior quality metric for coding artifacts and propose a novel video quality metric considering both coding and packet-loss artifacts. In the hope of improving the accuracy of quality metric for video sequences, we perform another study which focuses on quality assessment of single packet-loss-affected video frames. We evaluate the impacts of various properties of human visual system on quality of video frames, and develop quality metrics considering coding and packet-loss artifact, first separately and then jointly. In order to further improve the prediction performance of existing quality metrics, we exploit several methods of incorporating saliency into a video quality metric. The better performance of proposed saliency-aided quality metrics confirms the significant role of saliency in video quality assessment. Finally, we extend our study on saliency-aided video quality assessment to prediction of packet loss visibility. We update an existing loss visibility predictor with saliency-based features, and show that considering saliency can lead to improved prediction accuracy.

To explore relationships between various video attributes and perceptual video quality, we carefully design and perform extensive subjective video quality tests. The obtained subjective results not only confirm our assumptions about such relationships, but also inspire us to pursue our research in several novel directions. These subjective data also show fairly high correlations with the proposed objective quality measures.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Due to rapid advance of various video applications and services, such as video telephony, mobile video broadcasting, high definition television (HDTV), and Internet Protocol television (IPTV), there is an increasing demand for the accurate and effective quality assessment on underlying videos. An accurate video quality assessment is crucial to the video codec development, network protocol planning, in-network quality monitoring, quality assurance of end users, etc.

Since human are the very end users, the perceived video quality can be assessed in two different ways, i.e., *subjective quality evaluation* and *objective quality prediction*, either of which has its own merits and shortcomings. Subjective evaluation is to assess video quality from subjective ratings from a group of users. The advantage of this quality assessment method is that the obtained video quality ratings from user samples can be very close to actual assessment from a large amount of population, if the subjective tests are well designed and performed. On the other hand, one of the disadvantages of this quality assessment approach is that it is such a time- and effort-consuming process that

planning and performing one such test for a few of video clips requires several hours to several weeks and dozens of subjects, and hence quality assessment of a large number of test sequences in a batch fashion is not even possible. And it is even not feasible in some circumstances, e.g. real-time quality assessment.

Therefore, a more efficient quality assessment solution that requires no involvement of subjects once developed and is more applicable for more applications is badly needed. Objective quality prediction, as an alternative solution to subjective quality evaluation, has all these desirable features. Because it can emulate the humans' judgment on video quality based on mathematic models and can be easily applied to any test video sequence, it receives more and more attention in both industrial and academic communities. However, lack of thorough understanding about human visual perception system and large amount of possible video quality-affecting elements, both application-dependent and content-dependent, make the design of such effective objective video quality metrics a very challenging task.

During years of continuous efforts of researchers, a significant progress of study on video quality assessment has been made. A few objective video quality metrics have been standardized by ITU (International Telecommunication Union). However, because so far there is no single universal quality metric that can achieve satisfactory performance to replace the subjective assessment, there are two parallel active research directions being carried out in this field. One is to continue the study of effective and accurate subjective quality assessment schemes for different video applications so that they can not only provide improved quality judgment for various target videos, but also provide solid "ground truth" for the development and validation of objective quality metrics. The other

is to improve the performance and applicability of objective quality metrics so that they can robustly cope with various emerging video applications.

Due to the rapid development of network-oriented consumer multimedia electronics in recent years, the quality evaluation of videos transmitted over various networks becomes more and more pressing, and it hence becomes one active research field in the community of quality assessment. Among other artifacts, transmission loss is of the main causes of quality degradation of transmitted videos. Therefore, in this thesis, we attempt to investigate objective quality assessments of videos affected by packet losses from several different perspectives. Specifically, we first investigate the relationships between several attributes of packet losses and perceptual quality of the video sequence. Then we perform a study on quality evaluation of single frames of packet loss impaired videos. In addition, we also study how saliency, as one of the most important components of human visual system (HVS), affects subjective visibility of packet losses and perceptual video quality, and develop several saliency aided approaches to evaluate both the visibility of packet losses and the perceptual video quality.

In the remainder of this chapter, we will give reviews of both studies on subjective and objective quality assessments, respectively, where some standardized and some nonstandard methods are discussed.

1.1 Subjective Video Quality Assessment: Review of Standard and Nonstandard Protocols

When it comes to video quality evaluation, subjective quality assessment has been one essential research topic for a long time. Subjective quality evaluation usually requires

a subjective experiment where the quality of each tested video sequence is produced by the mean of opinion scores (MOS) of the video quality from subjects. As the most accurate and reliable way to evaluate the perceptual video quality, subjective quality assessment provides the "ground truth" for evaluation and validation of objective quality assessments.

To ease the design of subjective quality evaluation tests, and promotes the inter-usability of the results from different subjective test, based on the inputs of VQEG (Video Quality Expert Group), ITU suggested detailed protocols of conducting subjective experiments to measure perceptual video quality in its recommendations ITU-R Rec. BT.500 [1] for television system, and ITU-T Rec. P.910 [2] for multimedia application, in 2002 and 2008, respectively. These recommendations are similar and both suggested some most commonly used procedures for subjective quality assessment, such as viewing environment and procedures, criteria for the selection of viewers and test videos, and data analysis methods. Most of research on subjective quality assessment followed these standards.

According to the availability of reference video sequence presented in subjective tests, all the suggested test methods can be categorized as double-stimulus or single-stimulus scheme. The following are brief descriptions to some most common test schemes as well as their attributes, and please refer to [1] [2] for more details.

Double-stimulus Methods

Double-Stimulus Continuous Quality Scale (DSCQS) The viewers are presented with two videos, one of which is a source unimpaired sequence, and the other of is a test

(processed) version of that sequence. The sequence presentation orders are randomized. Viewers are asked to watch each video twice and evaluate the picture quality of both sequences using a grading scale (DSCQS, see Figure 1.1 and Figure 1.2) at the second presentation.

In previous VQEG studies that investigates contextual effects, it was shown that DSCQS was one of the most reliable methods. However, the shortcoming of DSCQS is its redundancy that limits the scale of test sequences.

Double-Stimulus Impairment Scale (DSIS) Viewers are presented with two sequences, first of which is source unimpaired sequence, and the second of which is a test (processed) version of that sequence. Viewers are asked to rate the level of impairment introduced in the test sequence with the first one as reference. The grading scale is from imperceptible (5), perceptible, but not annoying (4), slightly annoying (3), annoying (2), to very annoying (1).

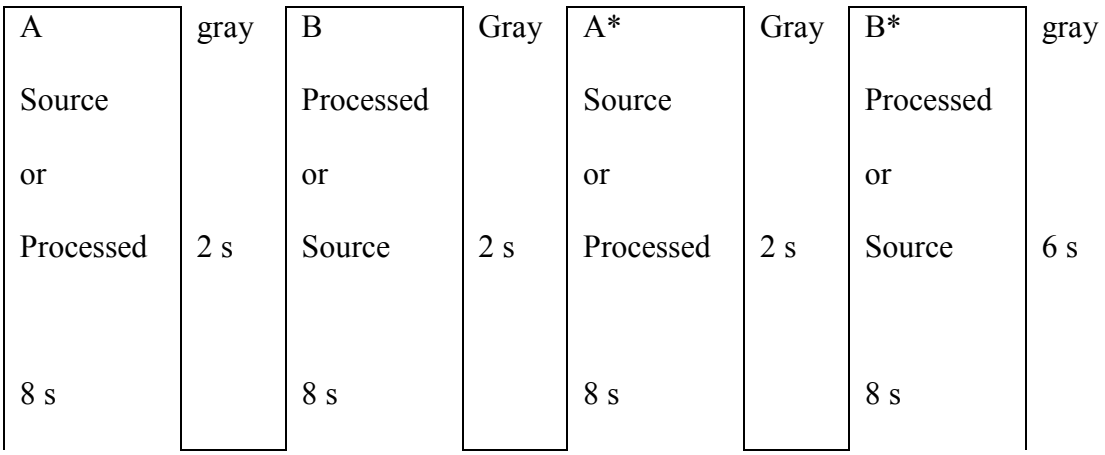


Figure 1.1 Presentation structure of test material for DSCQS (from ITU-T Rec. BT.500).

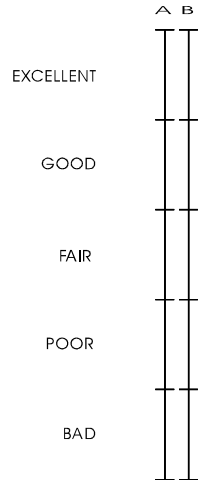


Figure 1.2 DSCQS grading scale (from ITU-T Rec. BT.500.)

Simultaneous double stimulus for continuous evaluation (SDSCE) In the previous methods, the viewing time duration of video sequences under evaluation is generally limited to 10 s, but it is not representative of much longer videos happening in real service. So SDSCE method was designed for this purpose.

In SDSCE, viewers are required to watch two sequences side-by-side in the same time (see Figure 1.3): one is reference distortion-free sequence, the other one is its test version. And they are requested to check the differences between the two sequences and to judge the fidelity of the test video with a slider. In order to make meaningful statistical analysis, duration of each test sequence should be at least 2 min. However, the drawback of this method is that viewers have to shift their attentions between two pictures from time to time.

Single-stimulus Methods

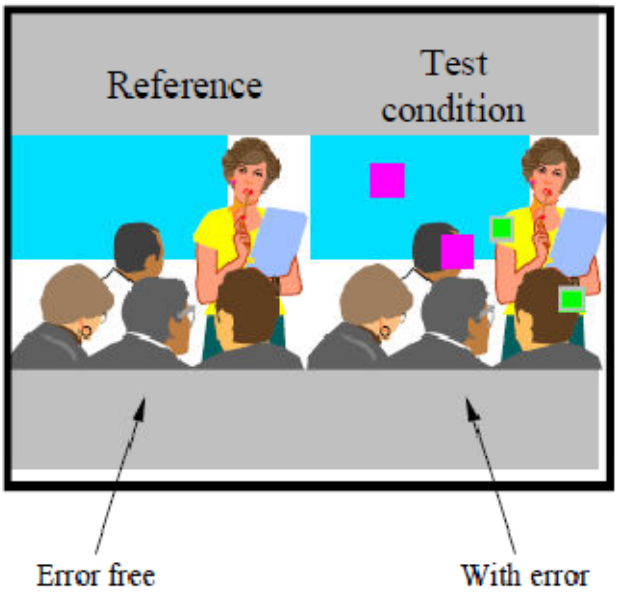


Figure 1.3 Example of display format for SDSCE (from ITU-T Rec. BT.500).

Absolute Category Rating (ACR) Since any of the double-stimulus methods cannot reproduce the real-world reference-free viewing conditions, ACR, as a single stimulus method, is designed to test the subjective quality scores given by viewers without explicit references.

ACR is a very efficient method and a large number of sequences can be tested in a relatively short time. Due to the lack of reference, it is assumed that all the references are perfect distortion-free video sequences. However, in most practical situations, some artifacts are inevitably introduced in the video capture phase, and hence these artifacts cannot be distinguished from the ones which are generated for testing purpose with this method.

To solve this issue, later on VQEG introduced *ACR-Hidden Reference (ACR-HR)* method [2], where the original unimpaired versions of test sequences are inserted randomly into the test dataset, and then also judged by viewers, but the viewers are

unaware of the existence of these references. Usually differential MOS (DMOS) between reference and test sequences is calculated to remove the reference effect, which is called *reference removal*.

Single-Stimulus Continuous Quality Evaluation (SSCQE) SSCQE was designed for measuring continuous subjective quality of longer video sequence without reference. Viewers are presented with one about 30 min sequence, just once and without any reference, and asked to give quality ratings instantaneously with a sliding bar as the video is playing.

However, this method is not used as frequently as double-stimulus method by other researchers. In order to produce a single quality score for test sequence, the continuous scores need to be calibrated first, rather than simply averaged over the time, because there has been shown that memory-based biases can prefer the sequence with noticeable impairments prior to last 10-15s of the sequence. Another issue associated with this method is the varying delay in different viewer response time which may influence the assessment results.

The main difference between the two protocols of double-stimulus and single-stimulus is that, with the former, the viewer can compare the reference and processed videos and hence be more precise in their judgments; whereas with the latter they see the video only once and have to judge its quality. Double-stimulus methods give a more precise quality rating in each vote but require a longer time to perform the tests than single stimulus methods. Single-stimulus methods enable the collection of more subjective judgments in a limited time and therefore permit more votes to be obtained to

increase the accuracy of the test. In the end, these two protocols have quite similar performances in terms of precision [1].

All of the methods described above have strengths and limitations and they are intended for different applications, but they share the same precautions that must be taken when designing the test schemes:

- Complex, which increases the difficulties for viewers to understand and perform the tests,
- Time consuming, which may cause viewers' fatigue and decrease of vote accuracy,
- Expensive, which is due to the need for dedicated test resources.

After the standardization of these subjective quality test methods, a significant amount of research efforts were made on various aspects of study of subjective quality assessment.

Some researchers have performed investigations on the relationships between subjective test results and subjective test protocols and proposed several approaches to improve the reliability and efficiency of the existing subjective quality assessment methods. A study [3] performed by NTIA/ITS (National Telecommunications and Information Administration/The Institute for Telecommunication Sciences) compared several aforementioned methods and concluded that SSCQE under proper design can produce quality estimates comparable to DSCQS. In a recent project report [4] of VQEG, subjective results obtained with ACR-HR method in different labs achieved very high consistency, which shows the effectiveness of this test method. Because of different ranges of rating scales of these recommended methods, a study on the impact of rating

scale on the subjective quality scores is performed [5]. The impact of clipping effect caused by the two extreme ends of rating scale is observed, but the assumption that discrete rating scales can increase the standard deviation of MOS is not supported by practical subjective test results. With respect to the issue of reusability of subjective results from different experiments, the work in [6] [7] propose to use the subjective scores from the common set of test sequences from different tests to map all the subjective scores onto a single scale so that available subjective data is greatly increased and hence the inter-test comparisons are enabled. In [8], the impact of different types of monitors, i.e. cathode ray tube (CRT) and liquid crystal display (LCD), and their resolutions used in the subjective tests on the subjective results are investigated, and the conclusion is reached that professional CRT and consumer LCD monitors can achieve statistically equivalent subjective results. The authors of [9] proposed a novel subjective quality testing plan to improve the testing efficiency by reducing the required number of test sequences. Specifically, the proposed algorithm can quickly find the optimal video quality operating point, in the space of coding and transmission parameters, by identifying the gradient ascent direction first, and use golden section line search to locate the point in the direction that approximate the best quality.

Another active research field of subjective quality assessment is to design specific subjective test plans for particular video applications or services. Using proper designs, some interesting relationships between various video attributes and perceptual quality are observed. It is found that humans tend to forget video contents displayed far enough from the current time instance due to the limitation of human memory capacity [10] [11]. Thus the quality ratings for that video segment are not determinative and the overall quality of

the entire clip is also affected. Therefore, the memory effect of human viewers is not only one of the concerns when designing and performing the subjective tests, but also a practical consideration when devising objective quality metrics, especially for quality assessment of longer video sequences. A study on the impact of memory on SSCQE results was performed [3], and the analyses indicated the last 9s to 15s of video content is critical for viewers to form their quality judgment on the entire clip. Another interesting subjective test example is the examination of impact of transmission conditions on the subjective video quality. The authors of [12] examined the subjective visibility of packet losses of encoded videos, where packet loss were deliberately inserted into video bitstreams and viewers were asked to indicate the occurrence of noticeable visual distortion by pressing the button on keyboard whenever they happen during the display. The design of this subjective test relives the burden of viewers to comprehend complex procedures to judge and record the quality of video sequences. In the work [13], the authors devised a program interface to compare the impacts of coding distortion and packet loss artifacts on subjective video quality. The viewers are presented with two video sequences side-by-side; one is the target sequence, usually distorted by coding (or packet loss) artifact, and the other one is the anchor sequence whose quality can be tuned by viewers with the slider controlling the amount of inserted packet loss (or coding) distortions. Viewers are asked to adjust the quality level of anchor sequence to match with that of target sequence. In [14], the author surveyed the opinions from a large amount of digital cable TV subscribers to investigate the impact of frequency of artifacts on subjective video quality. The consumers' vocabulary of artifact types and descriptions were first collected, and then in a web based questionnaire, based on the verbal

descriptions of artifacts, users were asked to indicate their preferences by choosing from different service prices.

1.2 Objective Video Quality Assessment

Objective quality metrics are algorithms designed to characterize the perceived video quality and predict viewers' opinion. In this section, classifications of existing objective video quality metrics will be addressed, and then their evolutions and state-of-the-art will be discussed.

1.2.1 Metric Classification

Based on the amount of access to the reference videos, metrics can be classified into full-reference (FR), reduced-reference (RR), and no-reference (NR) metrics [15].

- *Full-reference (FR) metrics* measure the quality of test video with respect to its original reference video. They require full access to every pixel of the reference video, and usually the two videos should be well aligned and calibrated before any further processing to ensure the exact match between corresponding pixels.
- *No-reference (NR) metrics,* also known as *reference-free metrics* or *blind metrics,* have access only to the test video. The complete lack of reference information makes the design of such metrics a very challenging task. Usually some assumptions about the attributes of both reference video content and possible distortion are necessary

- *Reduced-reference (RR) metrics* are compromise between FR and NR metrics in terms of access to reference information. To gain partial reference information, they first extract a number of features from the reference video, and then transmit them from sender to the receiver via reliable channel. The predicted quality of the test video is based on comparing the corresponding features from the both reference and test videos.

Until recently there are some derivations of the above metrics classifications. For different application usages, two sub-classifications of NR metrics are introduced: NR pixel domain metrics (NR-P) and NR bitstream domain metrics (NR-B) [12]. NR-P metrics ONLY have access to all the pixel information of test video, whereas NR-B metrics can ONLY access the bitstream of encoded test videos. Since in RR metrics, a back-channel for transmitting extracted features of reference videos increases the difficulty in implementing such system in practical situation, Quasi-NR metrics are proposed to relieve this problem [16], because it just requires very little amount of side information up to single value for each frame or even group of frames.

These three types of metrics have different operational uses. Due to the constraint of full access to the reference, FR metrics are suitable for offline video quality measurement for codec tuning or lab testing, NR and RR metrics are better suited for monitoring of in-service video at different points in the system, and NR can also be used to measure video quality of network end users.

There is another classification [17], with illustration shown in Figure 1.4, which focuses on different video representations in typical transmission systems:

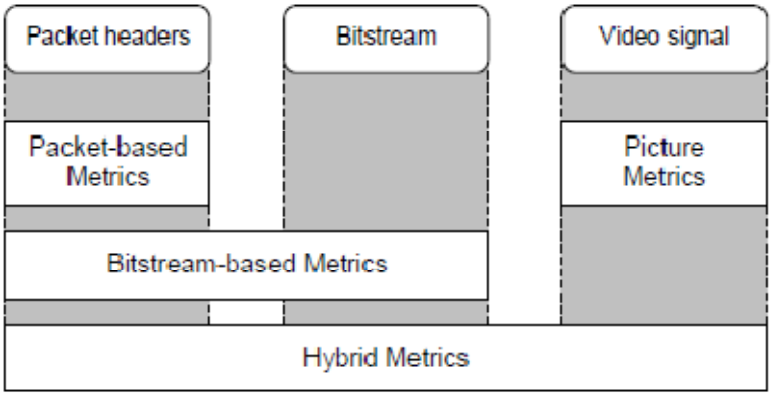


Figure 1.4 Classification of packet-based, bitstream-based, picture and hybrid metrics (from ITU-T).

- *Data metrics*, which measure the fidelity between original and processed videos by treating them as pieces of data, without considering their visual contents. As benchmarks, Mean Square Error (MSE) or Peak Signal-to-Noise Ratio (PSNR), and bit error rate (BER) or packet loss rate (PLR) fall into this category. The main advantage of this kind of metrics is that they are simple to calculate, whereas the drawback is that the meaning and visual importance of the pixels or packets associated with video contents are overlooked.
- *Picture metrics*, which treat the video data as the visual information that it contains. They take into account of the impacts distortions and content on perceived quality, either based on the properties of the human vision system, or based on the extraction and analysis of video features and artifacts.
- *Packet-based* or *bitstream-based metrics* for compressed video delivery over packet networks, which look at the packet header information and the encoded bitstream directly without fully decoding the video. Comparing to the metrics operating on the fully decoded videos, these metrics has the advantage of much

lower amount of data to be analyzed and hence lower processing requirements, so that they measure the quality of many video streams or channels simultaneously.

- *Hybrid metrics*, which use a combination of packet information, bitstream and/or decoded video pixels as inputs.

1.2.2 State-of-the-art and Trends

MSE and PSNR are two of the most popular video quality metrics. They measure the pixel-wise difference between two images. And the popularity of these two metrics is mostly thanks to their simplicity as well as many available mathematic tools to optimize them. Although people always consider them as the benchmark metrics of assessing video quality, they are not designed for image quality assessment; rather they are long-standing general measurements of the difference between any two signals. Because they just treat every pixel in image as an independent piece of data, the relations among the image pixels are always overlooked. Besides, they do not take into account any property of the HVS, so they are not expected to correlate with perceptual quality judgment very well. However, due to the lack of thorough understanding about HVS, the progress of research in this area is rather slow. In recent years, perhaps driving by dramatic developments of various video technologies, video quality modeling receives an increasing amount of attention from both industrial and academic communities.

A. VQEG Activities

Among others, some international telecommunication standardization bodies, such as ITU, address a growing number of issues related to both subjective and objective

video quality measurement. The validation and comparison of objective quality metrics is one of the study topics. For this reason, in 1997, a group of members from ITU-R and ITU-T study groups formed an expert team, called Video Quality Expert Group (VQEG), whose task is to perform validation testing of objective perceptual quality model, and provide inputs to standardization bodies for quality assessment standardization. Since its birth, VQEG has become the primary forum for algorithm developers and industry users to share ideas and collaborate. Currently, VQEG has performed several projects of validating objective quality algorithms for various video applications and services, and its work has resulted in ITU standardization of objective quality models designed for standard definition television and for multimedia applications. From these projects, we can gain a clear idea about the current status of research on objective quality assessment and some of its future trends.

To date VQEG has finished four projects of validating objective quality models for TV and multimedia applications, and is now performing or planning another three. Here we give a brief description about these projects, and discuss some interesting observations.

In 2000 and 2003, VQEG completed two tests on validating objective video quality metrics for television services. The first test, FR-TV Phase I, focused on out-of-service quality testing, which requires the reference videos; thus full-reference metrics were evaluated. The tested videos could be distorted by various levels of several artifacts encountered during production and distribution phases of standard-definition TV (SDTV), including MPEG-2 coding distortion, analog-to-digital conversion distortion, and transmission errors. Together with PSNR, another 8 quality metrics from different

institutes worldwide were evaluated. The results indicated that the perceptual quality prediction performances of all the 9 evaluated metrics were statistically equivalent [18]. This inconclusive result may result from too diverse artifact types tested in this project for the quality metrics at that time.

Although this test could not recommend any quality metric to be standardized, it brought up some potential issues when designing FR video quality metrics. For example, in some case, the video calibrations or normalizations before feeding videos into quality models are necessary. The calculation of PSNR was modified by allowing a pixel searching within a certain spatial and temporal ranges. In addition, standard evaluation metrics were first introduced.

In the second round of testing, FT-TV Phase II, VQEG emphasizes on secondary distribution of digitally encoded TV services by constraining the distortion types to MPEG-2 coding distortion. The best metrics in the test achieved correlations as high as 0.94 with MOS, thus significantly outperforming PSNR [19]. The top four (i.e. NASA, USA; Yonsei University, Korea; CPqD, Brazil; and NTIA, USA) out of six tested objective quality models were recommended in ITU-T Rec. J.144 [20] and ITU-R Rec. BT.1683 [21].

From the test results of both phases, we can find significant progress in development of FR metrics. However, only quality metrics focusing on MPEG-2 coding distortion achieved satisfactory performances.

In 2008, VQEG completed an evaluation of metrics for multimedia applications (MM Phase I), which is targeted for broadband Internet and mobile video streaming, at bitrates below 4 Mbps, with smaller frame sizes (QCIF, CIF, VGA). A wider range of

codecs and transmission conditions were considered in the test [4]. The MM Phase I set of tests was used to validate full-reference, reduced-reference, and no-reference objective models. Based on this test, two new standards for multimedia quality assessment were published, namely ITU-T Rec. J.247 [22], which defines four FR models (OPTICOM, Psytechnics, Yonsei University, and NTT), and ITU-T Rec. J.246 [23], which defines three RR models, with correlation up to 0.93 (all from Yonsei University, with different side channel bitrates from 1kbps to 128 kbps). However, NR models did not achieve satisfactory performance in this test.

In 2009, VQEG finished the reduced-reference and no-reference test for standard-definition television (RRNR-TV), which is an extension to the tests on FRTV, Phase I and II. MPEG- 2 and H.264 codecs were used, together with IP transmission errors. The final report [24] describes the performance of seven RR models; some top contenders (NTIA, and Yonsei University) may become part of a new ITU recommendation.

This test is originally designed to include both NR and RR quality metrics, however, all the five NR models were withdrawn from the record of the final report. Together with the MM Phase I test, no NR metrics were reported, which suggests that designing an accurate NR metric is still a very challenging problem and still open.

From 2004, VQEG started a project for the evaluation of models for high-definition television (HDTV) [25]. H.264 and MPEG-2 video codec are used, and distortion types include transmission error, pre- and post- video processing, and the bitrate is ranged from 1 Mbps to 30 Mbps. The test comprises full-reference, reduced-reference, and no-reference objective models. Currently, the validation test is finished, and the final reports is being compiled and yet to be published.

Currently, VQEG are planning another two projects. One is to evaluate perceptual quality models suitable for digital video quality measurement in video and multimedia services delivered over an IP network by making use of hybrid information from perceptual (pixel) domain and coded bitstream domain [26]. The scope of the test covers a range of applications including IPTV, internet streaming and mobile video. It considers much broader video quality ranges than all previous tests, from 16kbps to 30Mbps, and H.264, MPEG-2, and MPEG-4 video codec will be used, the considered video resolutions include SD, HD, QCIF, and QVGA, and distortion types include transmission error, frame rates, post-processing effects, live network conditions, and interlacing. Besides the constraints of access to reference information associated with FR, NR, and RR objective quality metrics, the types of inputs to the objective quality are to be restricted to ensure their easy implementations for measuring quality of networked videos. The concept of nonintrusive parametric model for the assessment of multimedia streaming (P.NAMS) is introduced; it uses only packet and codec information as inputs, but not any payload information. A follow-up project called P.NBAMS (B for bitstream) has similar goals, and P.NBAMS metrics will be allowed to use payload information. The other project that VQEG is planning is a second phase of Multimedia test. To extend the tests in MM Phase I, in Phase II audio-visual quality of multimedia contents will be assessed.

Furthermore, VQEG has also started to develop objective quality assessment metrics that combine multiple existing models. Hopefully this effort may lead to a reference objective metric and its implementation.

From results of all these VQEG projects, we can observe several attributes or trends of the development of objective video quality assessment in recent years as follows:

- Currently objective quality measures cannot replace subjective quality measurement;
- Quality assessment of transmission error and network conditions is an active research topic;
- Significant progress on development of FR metrics has been made, whereas NR and RR metrics are still at their infancy;
- Computational complexity of objective quality metrics starts to receive researchers' attention;
- Visual-audio quality assessment will be available in near future.

The descriptions of these VQEG projects are summarized in Table 1.1.

Table 1.1 Summary of VQEG projects

Project	FRTV_I	FRTV_II	NRRRTV
Focus	FR TV videos	Secondary distribution of digital encoded TV	Standard definition TV
Time	1997-2000	2000-2003	2000-2009
Subjective Test	DSCQS, 5 scale DMOS	DSCQS, 5 scale DMOS	5-scale ACR-HR DMOS
HRCs Considered	16 HRCs, Majority of MPEG2 and a few of H.263 and a few of analog videos, 625/50 and 525/60, with transmission errors, bitrate 768kbps to 50 mbps	10 HCR for 625/50 14 HCR for 525/60, All MEPG2 (except 1 H.263) videos, bitrate 768 kbps to 5 mbps	MPEG2 and H.264, with transmission error, bitrate 1 to 5.5 mbps
Model Types	FR	FR	NR (withdrawn) and RR
Proponents	**9 + PSNR** CPqDl Tektronix/Sarnoff NHK/Mitsubishi EPFL TAPESTRIES NASA KPN/Swisscom NTIA	**6 + PSNR** NASA British Telecom Yonsei University CPqD-IES Chiba University NTIA	**7 + PSNR** NEC Yonsei University NTIA
Evaluation Metrics	4 metrics (after polynomial or logistic mapping)	7 metrics (after logistic mapping)	Pearson Correlation, RMSE, Outlier Ratio (after polynomial mapping)
Winner(s)	8 of 9 perform equivalent to PSNR	2 models for 525 4 models for 625	**RR**: NTIA and Yonsei
Conclusion	VQEG not recommend any model in ITU Rec.	Some models good enough to be included in Recommendation; PSNR is worse than best models.	Some models good enough to be included in Recommendation; PSNR is worse than best models.
Comments	No model able to replace subjective testing; No model outperforms the others in all cases.	ITU-T J.144, ITU-R BT. 1683 standardized	

Project	MM_I	HDTV	HP/B
Focus	Mobile and broadband internet communication	HDTV application	Video and multimedia over IP network
Time	2004 -2008	2004 -ongoing	2007 -ongoing
Subjective Test	5-scale ACR-HR DMOS	5-scale ACR-HR DMOS	11-scale ACR-HR DMOS (MOS for NR)
HRCs considered	VGA/CIF/QCIF H.264/H.263/MPEG2/ MPEG4 , etc. compression artifacts, transmission error, pre-post-processing effects, live network conditions, interlacing problems, bitrates 16kpbs to 4 mbps, variable frame rates	1080i/p, MPEG2 and H.264, bitrate 1-30Mbps, compression artifacts, transmission error, pre- and post-processing, frame rate 25/30fps	SD/HD/QVGA/QCIF H.264/MPEG2/MPEG4, compression artifacts, transmission error, post-processing effects, live network conditions, interlacing problems, bitrates 16kpbs to 30 mbps, variable frame rates
Model Types	FR, RR, and NR	FR, RR, and NR	FR, RR, and NR P.NBAMS and P.NAMS
Proponents	**25 + PSNR** NTT Opticom Psytechnics SwissQual Yonsei Uni.	**? + PSNR**	**? + PSNR**
Evaluation Metrics	Pearson Correlation, RMSE, Outlier Ratio (after polynomial mapping)	Pearson Correlation, RMSE (after polynomial mapping)	Pearson Correlation, RMSE, Outlier Ratio (after polynomial mapping)
Winner(s)	**FR**: Psy., Opt., Yon., and NTT better than PSNR **RR**: Yon. better than PSNR **NR**: No model better than PSNR		
Conclusion	Some FR and RR models good enough to be included in Rec., NR model still to be improved.		
Comments	Winner models recommended in ITU-T J.246, ITU-T J. 247		P.NBAMS and P.NAMS are one topic of ITU Study Group 12 from 2009-2012

B. Quality Assessment of Video with Network Impairment

Besides VQEG and ITU, there are many groups looking into the problem of objective video quality assessment in different angles.

As transmission errors greatly impair video quality, the impact of packet losses on perceptual quality of videos transmitted in IP based network is an active research topic. Because of the nature of predictive video coding, distortions caused by packet losses usually can propagate to the neighboring video pixels; thus distort the video both spatially and temporally. Therefore, the development of objective quality metrics addressing packet losses is achieved in two directions. Works in [27] [28] [29] [30], investigated the temporal effect of packet loss, whereas [31] [32] [33] focus on evaluating their perceptual spatial distortions. In [27] and [28], the statistics of packet losses and delay jitters in videos were investigated and used to predict quality degradation levels. Whereas the authors of [29], believed that "fluidity" of video content directly affects its quality and proposed a metric to evaluate quality based on frame dropping. Furthermore, the authors of [30] proposed a more comprehensive metric to assess video quality, where several features were taken into account, including the amount of frame loss, object motion, and local temporal quality contrast. In the other direction, works in [31] [32] [33] evaluated the spatial effects of packet loss. In [31] and [32], two no-reference (NR) metrics are proposed for measuring block edge impairment artifacts caused by packet losses in decoded video. Their metric used strong spatial discontinuities as hints of packet losses, and evaluated perceptual distortions based on these strong discontinuities. In [33], the authors found strong correlation between network conditions and perceptual video quality and proposed a NR video quality assessment method.

In addition, a few works investigated the joint effect of both coding artifacts and packet losses at the same time. In 2002, NTIA/ITS developed a video quality metric (VQM) [34] to provide an objective quality measurement for video for a variety of encoding and transmission systems. It measures the perceptual effects of broad range of video impairments including blurring, jerky/unnatural motion, global noise, block distortion, color distortion, and packet loss. Independent tests by the VQEG have shown that the General Model of VQM on MPEG-2 and H.263 video has a high correlation with subjective video quality ratings [19]. This model has been recommended by ANSI as well as ITU-T as an objective video quality metric for secondary distribution of digitally encoded TV quality video [35] [20].

C. Quality Metrics Considering Properties of Human Visual Systems

Since humans are the very judge of video quality, the study of this research is inter- disciplinary topics for neuroscientist and physiologist, and understanding about the functions of human perceptual system is believed critically helpful in assess perceptual video quality. There has been physiological and psychological evidence that human beings do not pay equal attention to all exposed visual information, instead, they have excellent selectivity on what one sees in a scene [36] [37] [38] [39]. This high resolution vision due to fixation by the observer onto a region is called *foveal* vision, which is also known as focus of attention (FOA) or saliency region. In recent years, there has been a newly sparked interest in exploiting the visual attention model (VAM) or saliency detection mechanism for image and video quality assessment.

In [39] [40] [41] [42] the authors determine the regions of interests by a computational VAM, and use the resulting importance map to weight the visible distortion determined by a multi-channel early vision model. This method is used to assess the compression artifacts of coded images. The author in [43] extends the same concept to video quality assessment by constructing a perceptual quality significance map through extracting luminance, motion, and skin color features from video. But the consistency of these proposed models with visual attention is not sufficiently validated. The authors in [44] use eye-tracking experiments to determine the saliency map and propose a saliency-based quality metric. Surprisingly, results from their study show that considering the visual saliency does not lead to consistent improvement, at least for the JPEG and JPEG2000 compressed images they consider. In [45], the authors investigate two weighting strategies. One is to weight the SSIM quality index based on the visual importance, computed using a visual fixation predictor; the other assumes that pixels with large errors tend to attract visual attention and assigns more weight to pixels with lower SSIM scores. These strategies are tested on the images in LIVE database, which do not contain packet loss distortion. The results show that these strategies can significantly improve the correlations with subjective data. The wok in [46] proposes an embedded reference-free video quality metric based on the salient region, which is extracted based on color contrast, object size, orientation and eccentricity. Test results on JPEG2000 images show some improvement over PSNR in terms of correlation with subjective ratings, but the test results for video sequences are not clearly described and no comparison with other video quality

1.3 Dissertation Outline

This thesis is organized as follows. In Chapter 2 and Chapter 3, we investigate the impact of packet losses on perceptual video quality, and present objective quality metrics of packet loss impaired videos. In Chapter 4 and Chapter 5, we address the objective video quality assessment with aid of video saliency. Finally, in Chapter 6, we draw our conclusions and indicate possible future work. Due to the nature of studies on perceptual video assessment, subjective and objective quality assessments are parallel studies which interact with each other. So there are two parts in each chapter addressing the corresponding work.

In Chapter 2, we investigate the perceptual quality of video affected by packet losses. We first examine how several factors affect the video quality, and then by incorporating the existing quality metrics for coding distortion, we finally propose a full-reference video quality metric measuring the degradation due to both packet losses and lossy compression. The proposed metric correlates very well with subjective ratings, for a large set of video clips.

In Chapter 3, we perform an investigation on the impact of packet loss on the quality of individual video frames. In the study on quality assessment of packet loss impaired video described in Chapter 2, the quality of each individual video frame is measured simply by PSNR. In this chapter, in order to improve the accuracy of quality assessment of individual video frames, we incorporate human visual system components, i.e. masking effects, into the quality measurement and develop a more advanced quality metric of single video frames which considers both coding and packet loss artifacts.

In Chapter 4, we exploit the video quality assessment with aid of saliency information. We target the videos impaired by transmission loss, e.g. packet losses, in this work. By closely examining the relationships between attributes of visual saliency and packet loss and perceptual video quality, we propose three different schemes to incorporate saliency into quality assessment. To further improve prediction accuracy, we then use stepwise multiple linear regression analysis to combine multiple candidate metrics of all three types. The significant improvements of the final saliency-based quality metrics over their corresponding non-saliency metrics suggests the saliency information can be greatly helpful for assessing video quality.

In Chapter 5, we investigate how to improve visibility prediction of packet loss by incorporating the saliency information. Based on earlier findings about how saliency affects the perceptual quality of video with packet losses, we propose several saliency-based factors and incorporate them into a Generalized Linear Model (GLM) to predict loss visibility.

Chapter 2

Quality Assessment of Packet Loss Impaired Videos

This chapter discusses the perceptual quality of video affected by packet losses. We focus on low-resolution and low bit-rate video coded by the H.264/AVC encoder and the packet loss patterns likely in low bit-rate wireless networks. We examine the impact of several factors on the perceptual quality, including the error length (the error propagation duration after a loss), the loss severity (measured by the pixel difference between reference and distorted video in the area affected by a loss), loss location, the number of losses, and loss patterns. Based on our findings, we propose an objective metric for the quality degradation due to packet losses that considers all these factors. We also validate a prior metric relating the quality degradation due to compression artifacts and the peak-signal-to-noise ratio (PSNR). We finally propose a full-reference metric that measures the overall quality degradation due to both packet losses and lossy compression. The proposed metric correlates very well with subjective ratings for a large set of video clips with different loss patterns, coding artifacts, and scene contents.

2.1 Subjective Video Quality Tests

A. Test Design

Since perceptual video quality is affected by both packet loss and coding artifacts, spatially and temporally, we designed a series of tests to examine the impact of each factor, first separately and then jointly. By performing preliminary tests, we hypothesize that several factors may have significant impact on the perceptual quality: two (2) video codec related factors: coding artifacts (caused by lossy encoding, controlled by QP), and error concealment method; five (5) transmission loss related factors: the number of packet loss events, packet loss pattern (clustered or spread), duration of a loss-affected segment, severity of a loss, and loss position within a test video sequence.

Our target application is video delivery over 3G wireless networks, where frame rate, resolution, and bitrate are low, and transmission losses appear in bursts; hence the test video sequences, video encoding parameters, and packet loss patterns are chosen based on this type of application. We carefully designed several subjective tests examining the impacts of each of the aforementioned factors. The detailed descriptions of our test set-up are presented in the following subsections.

B. Test method

Two subjective rating methods are recommended by ITU-R BT.500 [1]: Double Stimulus Continuous Quality Scale (DSCQS) and Single Stimulus Continuous Quality Evaluation (SSCQE). Because we are interested in knowing the quality rating by a user when viewing a video sequence without seeing an error-free version, we choose to display a test video sequence without known reference. Although we are primarily

interested in the overall rating by a viewer for a sequence, we would like to investigate the immediate reaction of a viewer to a loss, and the impact of this reaction to the overall rating. Therefore, we choose to record both continuous rating and overall rating.

Specifically, a viewer is asked to give both continuous-time quality rating while the sequence is being displayed and an overall quality rating at the end of the sequence. The viewer uses a mouse to drag a scaling bar to give scores from 0 to 100 ("100" means best quality) both for continuous-time rating and overall rating. In each test, each viewer rated all sequences in a random order, different from others', and then repeated again in the same order, so that, for each video sequence, a viewer gave scores twice, which can be used to test viewers' self-consistency, discussed in detail in a later subsection. (Each run contains a large number of clips, so it is unlikely that ratings by a viewer in the second run will be affected by what he/she watches in the first run.) The entire procedure is controlled by an interactive rating software that we developed, whose interface is shown in Figure 2.1, and each session lasts less than 30 minutes without a break.

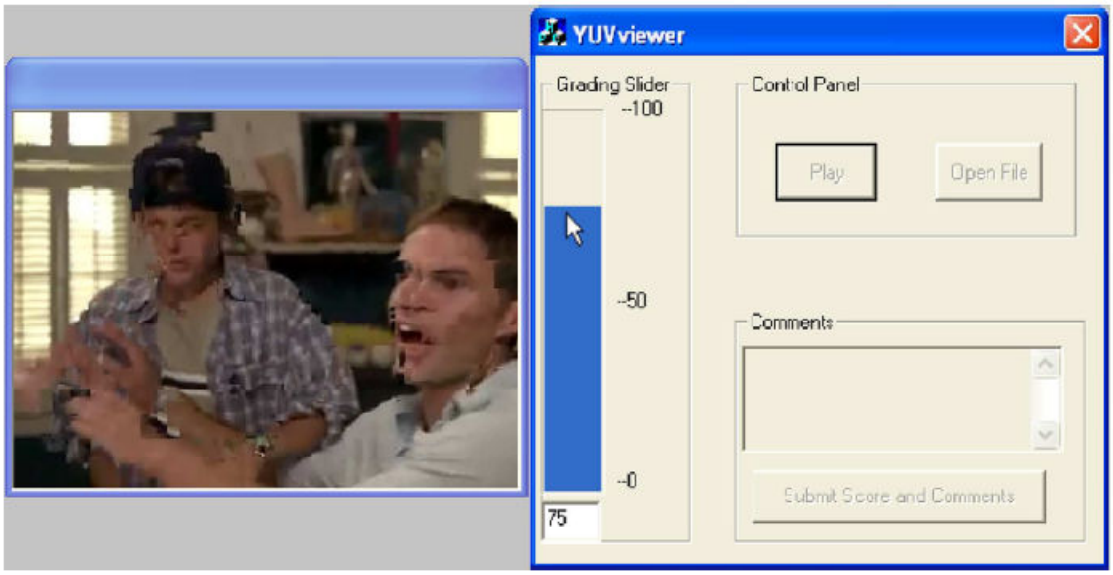


Figure 2.1 Subjective scoring software interface

C. Test materials

Five videos with different scene contents are used to generate a large set of test sequences. The scene contents cover indoor people interactions, outdoor sports games, with low to high motion, and plain to rich textures. In addition, these videos include a variety of camera motions. All the sequences are in QVGA (320x240) resolution, with encoding frame rate 12 fps or 15 fps, and clip durations from 20 s to 40 s. Table 2.1 provides a brief description of the sequence contents. All the testing sequences are encoded and decoded following the H.264 standard, using JM10.0 encoder/decoder (baseline profile, level 3, and IDRPP...P GOP structure) operating with a fixed QP. The QP ranges from 26 to 38, with corresponding bit rate from 40 kbps to 256kbps, and GOP lengths are 2 s.

Table 2.1 Content description of test videos

Title	Content Description
American Pie	Movie trailer, people talking and walking
F1 Car Race	Running cars, busy station crew
Interview	People sitting still. talking, close look
Paris	People sitting and talking, with body movements
Basketball Game	Players running, shooting and celebrating

Each frame is coded into a slice, which is then wrapped in a RTP packet. To simulate the loss characteristics in a 3G-like wireless network, where the loss of a program data unit at the link layer often leads to the loss of two consecutive IP packets, we purposely drop two consecutive frames in each loss position. Note that the distortion caused by each such frame loss usually propagates to the end of the GOP in which loss happens. We choose the location of the loss carefully to create sequences with different

loss position within the entire sequence, loss severity (measured by the drop between PSNR values of video frames with and without loss), propagation duration (or "error length"), as well as loss number and loss pattern.

In preliminary studies, we evaluated decoded videos using three error concealment methods: frame-copy (use the last correctly received frame to replace the lost frame), motion-copy (copy the last correctly received motion vectors for the lost frame, so that each block in the lost frame is copied from a displaced block in the last correctly received frame based on the copied motion vector), and frame-freeze (the last correctly received frame lasts to the end of GOP). With frame-copy and motion-copy error concealments, there are usually visible artifacts throughout the loss-affected segments of sequences, whereas, with frame-freeze method, packet losses can cause decoded videos to "pause" for a certain period of time [47]. It was concluded that frame-copy concealment gives overall more consistent results across viewers and loss patterns, and it gives highest Mean Opinion Score (MOS). Therefore, in the formal tests reported here, we only used the frame-copy method.

To explore the impacts of aforementioned quality-affecting factors, we produced totally 51 test sequences with different combinations of coding and packet loss configurations. We categorized them into four groups and we performed subjective tests on each of them separately with different purposes. Test 1 and Test 2 are designed to find out how is the perceptual quality affected by packet loss, whereas Test 3 concerns the impact of coding artifacts. The sequences used in the first three tests are all generated from the original video of "American Pie", which are used for exploring and training our proposed objective metric. The sequences used in Test 4 are generated from all five video

Table 2.2 A brief description of sequence sets for the four tests

Test 1(12 sequences)		Test 2 (13 sequences)	
Sequence Description	Factors Examined	Sequence Description	Factors Examined
3 sequences with same error at different positions	Loss position	3 sequences with same error at different positions	Loss position
3 sequences with different error lengths (GOP=2s)	Error length	2 sequences with short error lengths	Error length loss visibility
3 sequences with different loss patterns	Loss pattern	5 sequences with different PSNR drops	Loss severity Loss visibility
3 sequences with different loss numbers	Loss number	3 sequences with different error patterns	Loss pattern
Test 3 (7 sequences)		Test 4 (25 sequences)	
Sequence Description	Factors Examined	Sequence Description	Factors Examined
7 sequences encoded with different QPs	PSNR	25 sequences encoded with different QPs, and with random packet losses	Model Verifications

sources and contain both coding and loss artifacts, and they are used for verifying the proposed metric. Table 2.2 provides a brief description of sequence sets for the four tests.

Figure 2.2 illustrates the quality range of our test sequences. Here "quality" is defined in terms of differential MOS, or DMOS, which are the differences between the MOSs given to original (uncoded) sequences and those given to the corresponding processed sequences. We can see that our test sequences cover a wide spectrum of quality. Notice that the ratings presented here are obtained after performing necessary normalization processing, which will be discussed in Chapter 2.1.F.

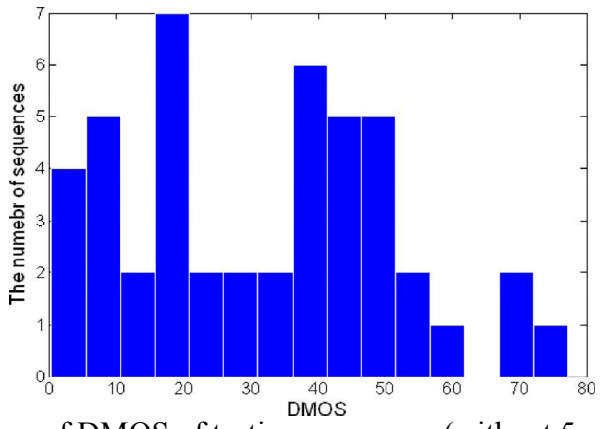


Figure 2.2 Histogram of DMOS of testing sequences (without 5 original sequences).

D. Viewing Condition

The tests were administered in the Image Processing Lab of Polytechnic University, following the ITU recommendation [1]. The computer used for display is Dell Dimension 910 with Dell 17 ” TFT LCD Flat Panel monitor (Model No.:E173FPf). The video window size is 3 inch diagonally (with monitor resolution at 1280x800), which approximates the display size of a typical PDA. The viewers sat in front of the monitor with comfortable distance and were allowed to move their heads as they wished.

E. Viewers

The viewers participating in our tests are mostly engineering students in Polytechnic Institute of New York University. There were a total of 60 viewers taking the tests. Some of them participated in one of the 4 tests and some performed multiple tests. (At least 15 valid viewers were involved in each test). Before the actual subjective tests, they all passed the visual acuity test (Snellen) and color blindness test (Ishihara). Totally, we obtained more than 2000 valid subjective quality rating samples.

F. Data Screening and Normalization

A viewer may be inconsistent with the majority of the viewers in its rating for the same testing sequence, or he/she may be inconsistent at different times when rating similar sequences. A viewer screening is conducted to eliminate the ratings by these viewers from further data analysis. We conducted the inter-viewer consistency test following [1]. In addition, to test the self-consistency of each viewer, we calculated the Pearson Correlation Coefficient between the two sets of overall scores given by this

viewer for the same set of sequences during a viewing session, and the data from viewers with low self-consistency (Pearson Correlation Coefficient lower than 0.6, here) were ruled out. After these two screening procedures, the data from 3 out 60 viewers were discarded.

Before performing data analysis on the overall scores by all viewers, the viewer scores are normalized. First, in each test, we set each viewer's lowest score to "0" and his/her highest score to "100", and linearly scale the rest of his/her scores. Then MOS is obtained by averaging scores over all the valid viewers. The reason for this normalization is that the viewers judged the qualities with their own criteria and the ranges of their ratings were different, not always from "0" to "100". Therefore, in order to perform meaningful averaging calculation this normalization processing is necessary.

Secondly, even though all the four tests were carried out in the same viewing conditions, with the same methodology, and in the same score range, the renormalization across the data in different tests is still necessary, because video quality ranges of different test are not equal [3]. Hence, by using some common sequences in all the tests, we linearly normalize the MOSs in different tests to a common scale.

2.2 Objective Video Quality Metrics

2.2.1 Quality Degradation due to Single Packet Loss

This section investigates the relations between several attributes of a single loss on the perceived quality degradation, including error length, error severity, and error location. The analysis in this section is based on subjective ratings obtained from Test 1

and Test 2, summarized in Table 2.2. The perceived degradation is denoted by differential MOS due to packet loss, or $DMOS_L$, which is the difference between the MOS given to a decoded sequence without any packet loss, and the MOS given to one with loss. Although all the test sequences in Test 1 and Test 2 have been subjected to some packet losses, the loss in one sequence has minimal impact on perceptual quality and was "invisible" by all viewers (as indicated from their continuous time ratings). So we used the MOS for that sequence as the reference when determining $DMOS_L$.

A. Error Length Effect

The error length is defined as the number of frames starting from the first lost frame to the end of the GOP or to the scene change within the GOP. To examine the impact of error length on the perceptual quality, we created several sequences with a single loss (losing two consecutive frames) in the middle of a sequence. The loss position is chosen to lead to different error propagation length. We constructed 3 sequences with similar PSNR drops (about 10 dB) but different error lengths in Test 1.

Figure 2.3 (a) shows the plots of PSNR vs. frame number for these three sequences, where a segment with reduced PSNR indicates the location and duration of the loss-affected segments. Figure 2.3 (b) shows the relation between the $DMOS_L$ and the error length. It is clear that the perceptual quality degrades as the error length increases.

B. Loss Severity Effect

In order to explore the impact of loss severity on perceptual quality, we tested 5 sequences with similar error length (12 frames) but different PSNR drops in Test 2. To measure the severity of a loss, we first determine the PSNR drop for each affected frame,

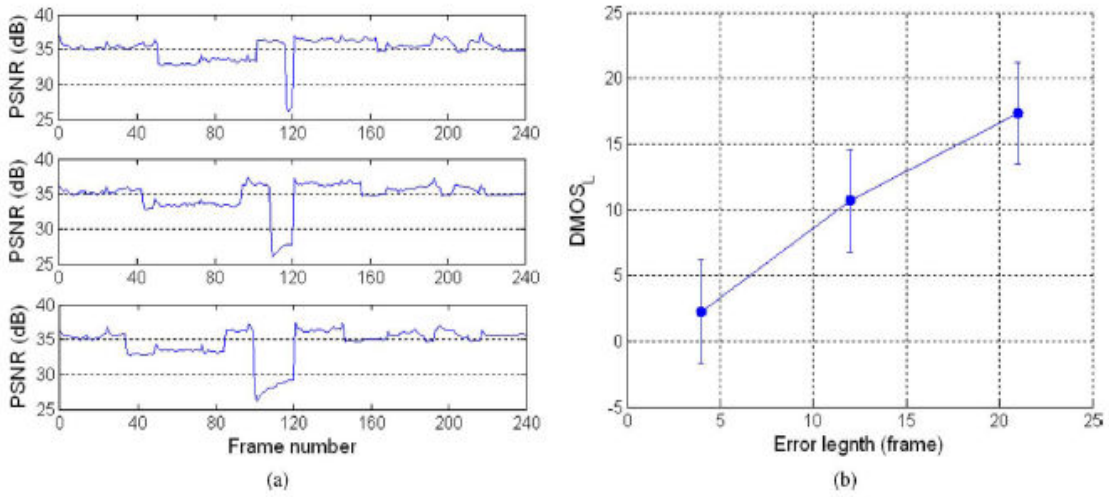


Figure 2.3 Error length effect on perceptual quality. (a) PSNR curves of three test sequences with different error lengths, and (b) relation between $DMOS_L$ and error length.

which is the difference between the PSNR of the frame decoded in the absence of the loss, and the PSNR of this frame decoded with packet loss. We then find the biggest PSNR drop among all affected frames, which is simply referred as PSNR drop.

Figure 2.4 (a) shows the plots of PSNR vs. frame number for these five sequences and Figure 2.4 (b) shows the relation between $DMOS_L$ and PSNR drop. We can see that perceptual quality decreases as the PSNR drop increases, but the relation is non-linear. The beginning flat portions in Figure 2.4 (b) suggests that viewers do not "see" the loss when PSNR drop is smaller than a certain threshold, and the end flat portion suggests that viewers think the qualities of sequences are equally "bad" once the PSNR drop exceeds a certain threshold.

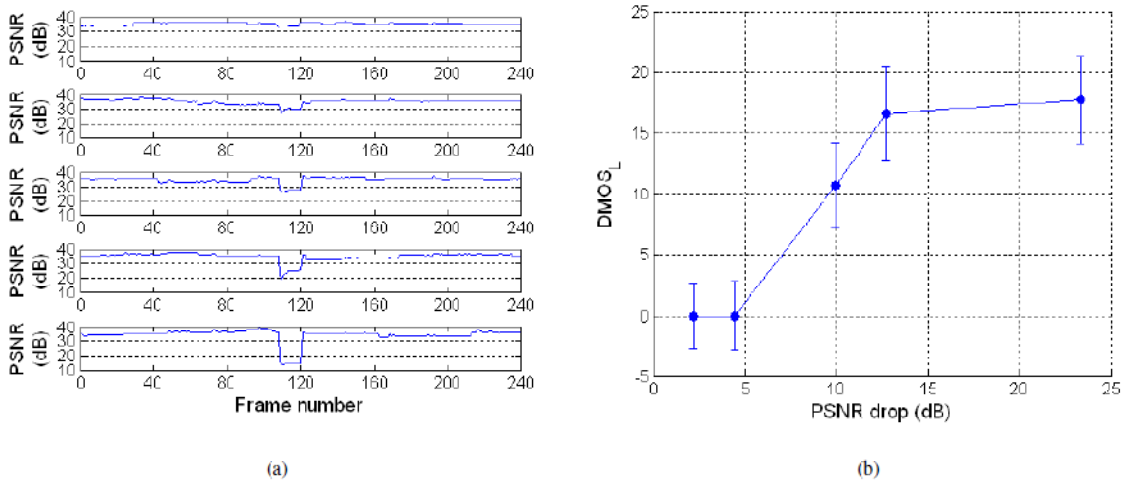


Figure 2.4 Loss severity effect on perceptual quality (a) PSNR curves of 5 test sequences with different loss severity, and (b) relation between DMOSL and PSNR drop.

C. Error Visibility

Figure 2.4 shows that when the PSNR drop associated with a loss is smaller than a certain threshold, a viewer tends not to notice the loss. This packet loss "invisibility" phenomenon in encoded video was discussed in the works [12] [48], however, the quality evaluation problem is not explicitly addressed. In order to further investigate the relation between error visibility and perceived quality in our case, we looked into the continuous score curves from viewers. We observed that, for some particular losses, most viewers did not lower their ratings after the loss (in their continuous-time rating), and those losses have either short error length or low PSNR drop. We measure the visibility of an error by the percentage of total number of times viewers who "saw" the error, as indicated by dips (with the magnitude over 15 units) in the continuous score curves.

Figure 2. (a), (b) and (c) show the relations between the visibility and the error length, PSNR drop and PSNR drop sum, respectively. Here, the PSNR drop sum is defined as the summation of the PSNR drops of all erroneous frames. From Figure 2.5,

we observe that, for our test set-up, the minimal PSNR drop is around 5 dB. Note that when two consecutive frames are lost due to a transmission loss, with frame-copy error concealment, the first two frames in each error duration are both copied from the last frame before the loss, which do not incur many noticeable artifacts. Visible artifacts usually start to be seen right after the two concealed frames.

Since both error length and PSNR drop affect the subjective ratings, it is natural to ask if there is some composite objective measure that can reflect the effect of both. We hypothesize PSNR drop sum may be used as one of the possible objective measures, and the relation between error visibility and PSNR drop sum is shown in Figure 2. (c). We can see from the figure, that generally speaking as the PSNR drop sum increases, error visibility increases too. And there is a jump at PSNRdrop_sum = 50 in this relationship.

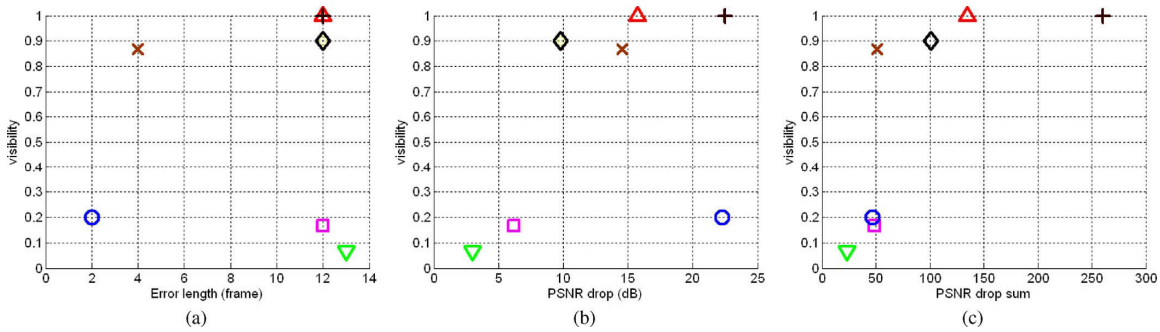


Figure 2.5 Relations between error visibility and (a) error length, (b) PSNR drop, and (c) PSNR drop sum.

D. Forgiveness Effect

Existing researches revealed that degradation in video materials can be “forgiven” or “forgotten” to some extent if the degradations are followed by good quality video [49] [10] [11] [49] [50] [51]. In other words, for similar perceptual distortion caused by frame

loss, the farther that loss locates away from the end of a sequence, the better will be the subjective rating. To verify the existence of such "forgiveness" effect and to investigate the specific relationship between the perceptual quality and the distance (in time) from the occurrence of the loss to the end of the sequence, we generated three 40 s long sequences, cut from original 60 s clip, each of which contains the same GOP with the same single loss (hence same PSNR drop and error length) in the middle. But the three cuts are shifted so that the GOP with loss is positioned in the beginning (15th s), middle (25th s), and end (35th s) of the three sequences, respectively.

The PSNR curves of these three sequences are shown in Figure 2. (a). Figure 2. (b) shows the relation between $DMOS_L$ and loss position (distance to the end of sequence). From the figure, we find that the overall score for the sequence with loss happening at the end is significantly lower than the overall scores of the other two cases, which received similar ratings.

Via paired-T significance test [52], we confirmed that the difference between the ratings for the sequence with end loss and the two sequences with beginning and middle losses are statistically significant (over 95%), whereas the difference between the ratings for the two sequences with beginning and middle losses are insignificant (less than 60%). This result substantiates the existence of the "forgiveness" effect, and shows that the increase of the overall score (or the "forgiveness factor") is non-linearly related with the distance. In our test, the "forgiveness factor" stays the same after 15 seconds, which means that viewers' memories do not differentiate well for losses happened 15 seconds before. This result correlates well with those of previous work on the memory effect of human visual systems [3].

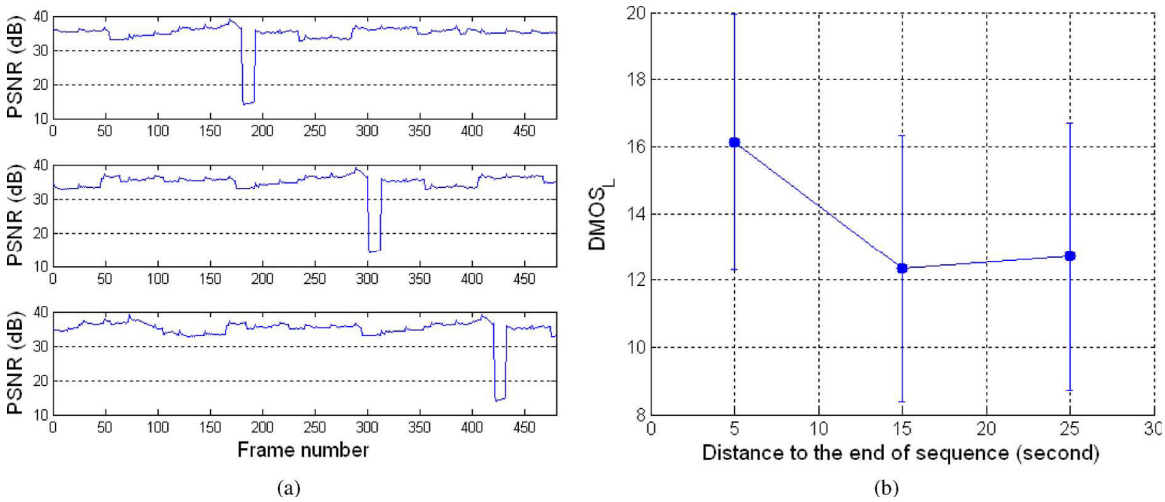


Figure 2.6 Forgiveness effect. (a) PSNR curves of sequences with different loss positions and (b) relation between $DMOS_L$ and loss position.

E. Proposed Video Quality Metric for Sequences with Single Loss

Taking into consideration all the aforementioned factors, we propose to use "PSNR Drop Sum" or PDS, which is the sum over the PSNR drops of all erroneous frames, formally defined as

$$PDS = \sum_{n=1}^{EL} PD_n \tag{2.1}$$

where *PDn* is the PSNR drop for frame *n*, with *n = 1* denoting the first lost frame. *EL* is the length (in terms of frame) of loss-affected video segment (or error length). To take into account the clipping phenomenon shown in Figure 2.4, we further modify the PSNR drop of each frame by defining:

$$\alpha(PD) = \begin{cases} 0, & PD < PD_{min} \\ PD - PD_{min}, & PD_{min} \le PD \le PD_{max} \\ PD_{max} - PD_{min}, & PD > PD_{max} \end{cases} \tag{2.2}$$

Considering the error length threshold of visibility, we only sum the PSNR drops for frames after a minimal error length, EL_{min}. The *Modified PSNR Drop Sum (MPDS)* is expressed as

$$MPDS = \sum_{n=EL_{min}}^{EL} \alpha(PD_{n)} \tag{2.3}$$

Then, we weight the contribution from a single loss based on its distance to the end of the sequence, to take into account of the "forgiveness effect". Based on the non-linear trend observed in Figure 2. (b), we use an exponential decay weighting factor $w(D) = e^{-r\,D}$, where D is the distance (in time by second) from last erroneous frame to the end of sequence, and r is a constant that we determine through least squares fitting of the subjective ratings to the model. This leads to the *Weighted MPDS (WMPDS)* measure:

$$WMPDS = e^{-r\,D} \sum_{n=EL_{min}}^{EL} \alpha(PD_{n)} \tag{2.4}$$

Lastly, the perceptual distortion also should be normalized by the sequence length L (in frames), because the same amount of distortion appearing in video with different lengths may cause different visual quality degradations, yielding the *Average WMPDS (AWMPDS)* metric:

$$AWMPDS = \frac{WMPDS}{L} \tag{2.5}$$

Figure 2.7 (a), (b), (c), and (d) show the relations between $DMOS_L$ and *PDS*, *MPDS*, *WMPDS*, and *AWMPDS*, respectively. Here we set *PDmin = 6, PDmax = 13, ELmin = 2,* and *r = 0.01*, by maximizing Pearson correlation coefficient. We can see that *AWMPDS* is more linearly related to the perceptual quality. To quantify how well a

model fits the measured perceptual ratings, we computed the linear correlation coefficient between the measured ratings and their corresponding *PDS, MPDS, WMPDS,* and *AWMPDS* values for all the testing sequences. The four measures have correlation coefficients of 0.8094, 0.9248, 0.9307, and 0.9555, respectively. When applying the paired-T test [52] to exam these four linear relationships in Figure 2.7, we find that the significance levels are over 99% in all cases. This indicates that the linear relations we observed between the $DMOS_L$ and *PDS, MPDS, WMPDS*, and *AWMPDS* respectively, are statistically significant.

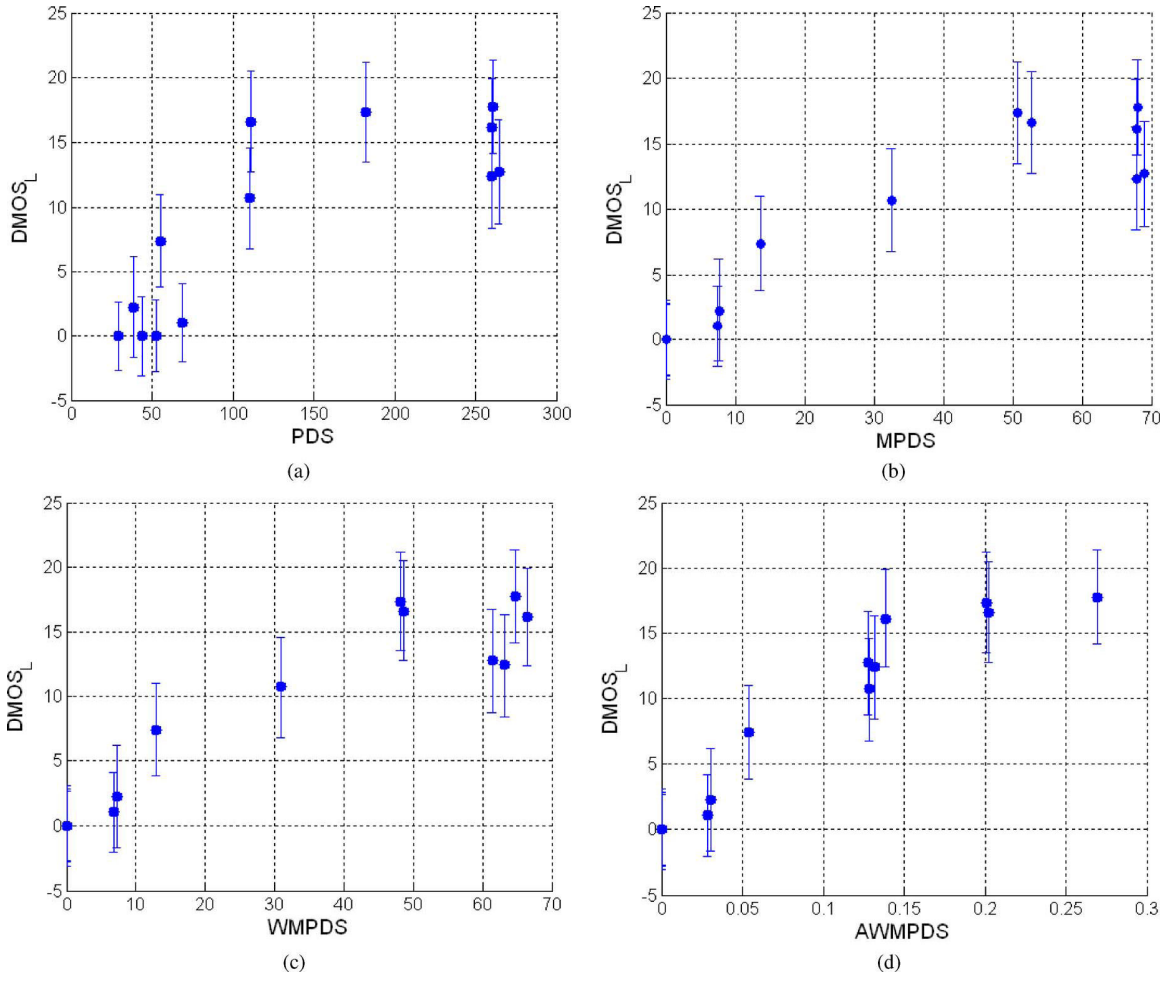


Figure 2.7 Relations between DMOSL and (a) PDS, (b) MPDS, (c) WMPDS, and (d) AWMPDS for single-loss data in Test 1 and Test 2 (Pearson correlations are 0.8094, 0.9248, 0.9307, and 0.9555, respectively).

2.2.2 Metric Proposed for Sequences with Multiple Losses

In the previous section, we investigated the effects of loss position, loss severity and error length, by examining sequences where only one loss (two consecutive frames) happens throughout the entire clip. The proposed metric correlates very well with subjective ratings. In this section, we focus on the situation where multiple losses exist in a sequence and explore the interactions among different individual losses and their group effect on the perceptual quality. Furthermore, we extend the proposed objective quality metric *AWMPDS* for single loss to multiple losses scenario, which happens more likely in real-world applications.

A. Effect of the Number of Loss Events

We first examine the effect of the number of loss events when each loss has similar error length and severity. Toward this goal, we constructed four test sequences containing 1, 2, 3, and 4 losses respectively and each loss has similar error length (12 frames), and PSNR drop (10 dB) approximately. The sequences are all 40 second long, and the losses are evenly spread out through the approximate middle 27 seconds. The PSNR curves of these four sequences are shown in Figure 2. (a). Figure 2. (b) shows the relation between $DMOS_L$ and the number of losses in a sequence. From the figure, we find that the overall score has a roughly linear relationship with the loss number. Because all losses have similar error length and PSNR drop, this also indicates that the overall score is linearly related to the total error length or PSNR drop sum, consistent with our observation from the single-loss experiment.

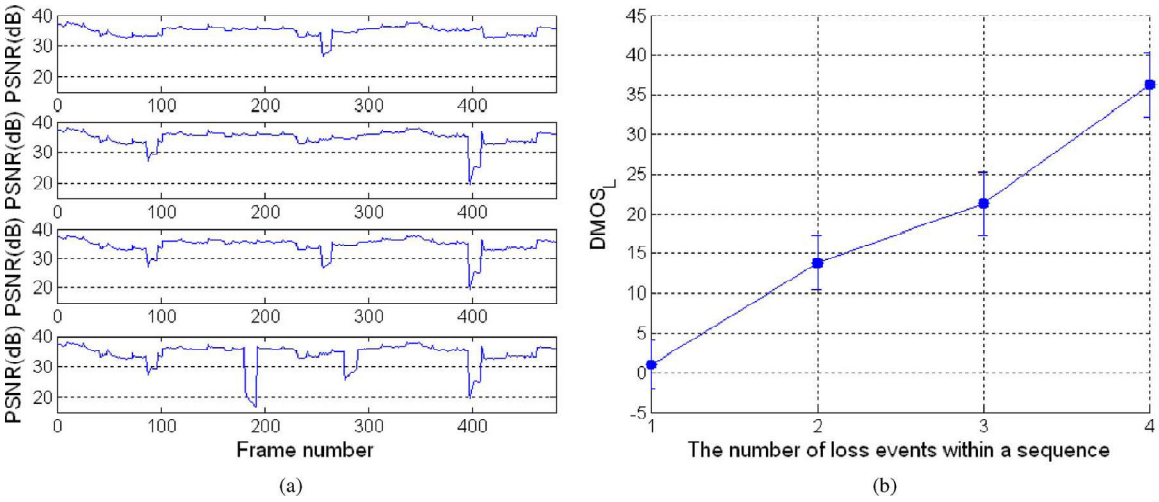


Figure 2.8 Effect of loss number on perceptual quality. (a) PSNR curves of the four test sequences with different numbers of losses. (b) Relation between DMOSL and the number of losses

B. Effect of Loss Patterns

Result in Figure 2. was obtained with sequences where the losses are distributed throughout the sequences in a uniform pattern. It is natural to ask, if the losses are distributed non-uniformly, how does the loss pattern affect perceptual evaluation of video sequences. To answer this question, we first characterize the loss pattern by defining the cluster degree (*CD*) as

$$CD = e^{-c\,L_{loss}}(1 - e^{-k\,N}) \tag{2.6}$$

where N is the number of losses, and L_{loss} is called "loss span", which is defined as the distance (in time by second) between the first lost frame to the beginning of the last loss. For the single loss case, L_{loss} is set to 0. c, and k are constants to be determined. The definition of CD is motivated by the fact that, for the same "loss span", as the number of losses increases, the losses appear more clustered; and for the same loss number, when the "loss span" increases, the loss become more spread.

We constructed three sequences with different loss patterns (or cluster degrees). In order to eliminate the effects by other factors, each sequence has similar total error length and with similar total PSNR drop sum (summation of PSNR drop sums for each loss). Figure 2. (a) shows the PSNR curves of these three sequences (all 40s long), ordered according to the cluster degrees of their losses: Sequence 1 contains two 12-frame loss-affected segments (caused by two separate losses and their propagated errors) spread far apart (least clustered), Sequence 2 includes two 12-frame loss-affected segments (caused by two separate losses and their propagated errors) close to each other in the middle (moderately clustered), and Sequence 3 contains four 6-frame loss-affected segments (caused by four separate losses and their propagated errors) evenly spread (most clustered).

The relation between $PDMOS_L$ and loss cluster degree is plotted in Figure 2. (b), where we have used $c = 0.002$, and $k = 0.8$, by maximizing Pearson correlation. From the

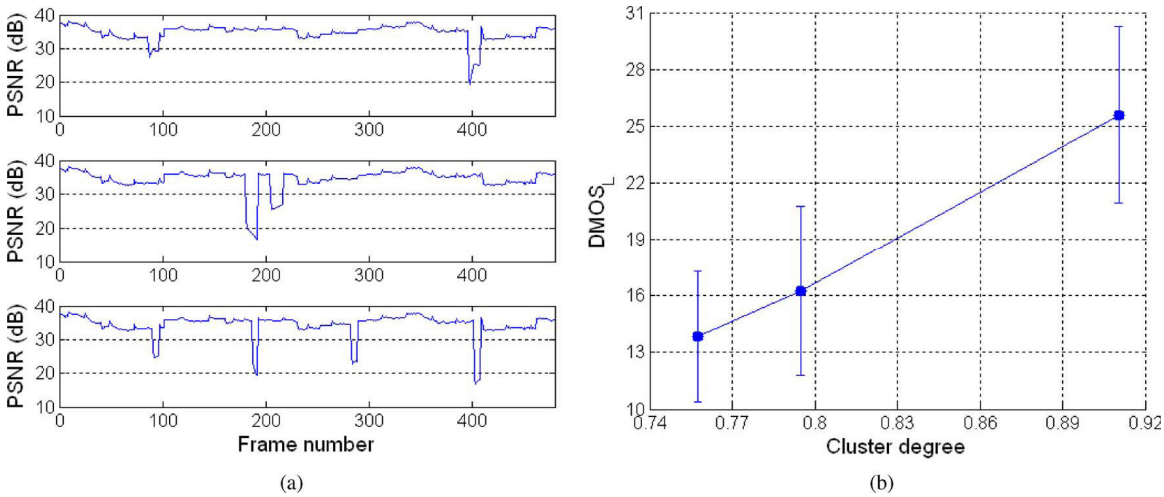


Figure 2.9 Effect of loss pattern on perceptual quality (a) PSNR curves of the three sequences with different loss patterns and (b) Relation between the $DMOS_L$ and loss cluster degree.

figure, we can see that the perceptual degradation increases linearly with the cluster degree. This indicates that the human eye is more disturbed by closely grouped losses, than by far apart losses. This may be in part due to the forgiveness effect discussed earlier.

C. Forgiveness Effect of Multiple Losses

In the previous section, we defined the "forgiveness effect" of packet loss solely based on examining on single loss sequences. For multiple losses case, it is not clear whether each loss still affects the overall viewer rating with the same forgiveness factor,

depending on the distance of the loss to the end of the sequence. Because there are multiple losses within the sequences, with different loss numbers and patterns, it is difficult to analyze the forgiveness effect of each loss individually. And the unknown "group effect" among different losses makes the task even harder.

Figure 2.9 indicates that the subjective rating for Sequence 1 (least clustered) is better than Sequence 2 (moderately clustered) in Figure 2. (b). One plausible explanation is that the inter-distance between the two losses in Sequence 1 is too large, so that the viewers almost forgot about the first loss when seeing the second loss, and rate the quality of this sequence mostly based on the second loss. This prompts us to use the inter-loss distance between losses to define the forgiveness factor. So in the definition of forgiveness effect for single loss sequences, the distance (in time by second) between a loss and the end of sequence is replaced by the distance between two consecutive losses or between the last loss and the end of sequence. The appropriateness of this modification in the definition of "forgiveness factor" will be proved in the later subsection.

D. Proposed Model for Quality Degradation Due to Multiple Losses

Based on the effect of the loss number, loss pattern, and forgiveness on perceptual rating individually reported in prior subsections, we propose to extend the AWMPDS metric developed for single loss sequence, to $PDMOS_L$, which is used to predict the overall quality degradation due to packet losses, which can contain single or multiple losses

$$PDMOS_L = CD \frac{1}{L} \sum_{i=1}^{N} W(D_i) MPDS_i \tag{2.7}$$

where N is the number of losses, and L is the length (in terms of frame) of video clip. CD is defined in Eq.(2.6), and $W(Di)$ is the exponential decay function that simulates the forgiveness effect of multiple losses, and is previously defined in Eq.(2.4). Noticing that the inter-distance D_i is closely related to the loss pattern, our definition for the forgiveness factor thus some what accounts for the impact of the loss pattern. $MPDS_i$ is defined in Eq.(2.3). Via maximizing correlation, we find the following parameters yield the best fit between the model and the subjective data: $PDmin = 4$, $PDmax = 13$, $ELmin = 3$, $r = 0.015$, $k = 0.8$, and $c = 0.002$. Figure 2.10 shows the relation between $DMOS_L$ and $PDMOS_L$ in Test 1 and 2. The Pearson correlation coefficient is 0.9626, which is supported by t-test with a significance level above 99%.

E. Verification of Proposed Quality Metric for Random Packet Losses

In order to verify the accuracy of our proposed model in Eq.(2.7) with more realistic packet losses, we designed and performed another test, Test 4. This test consists of all five chosen sequences, with each sequence having five different degraded versions: original (uncoded) sequence, encoded without any packet loss (with QP set to 26, 31, and

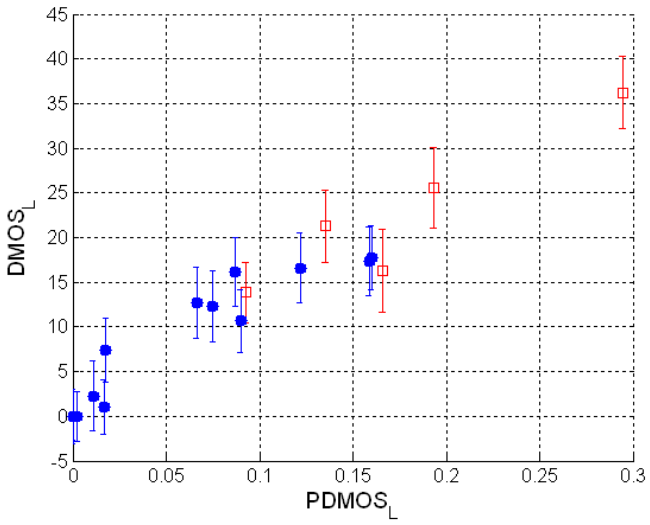


Figure 2.10 Relations between $DMOS_L$ and $PDMOS_L$ for data in Test 1 and 2. (Pearson correlation=0.9626). ("solid" points represents single loss data set, and "hollow" points correspond to multiple losses data set)

35), and three sequences with packet losses. We apply the Elliot-Gilbert (two-state Markov chain) [53] [54] to simulate the packet (or frame) loss characteristics of actual wireless network. Here we set transition probability from "Good" to "Bad" state to 0.02 and transition probability from "Bad" to "Good" state to 0.8. (We assume that the video frames transmitted in "Good" state are correctly received, whereas frames in "Bad" state are lost.)

The average packet loss rate for the sequence in Test 4 is up to 3% and average burst length is 2 frames. (We found that the severity caused by these packet loss parameters can yield enough artifacts to degrade the perceptual quality substantially but the resulting sequences are still acceptable for the viewers to give meaningful ratings.) The remaining testing conditions are kept same as previous tests. In this test, $DMOS_L$ for each sequence is the MOS difference from its no-loss encoded version.

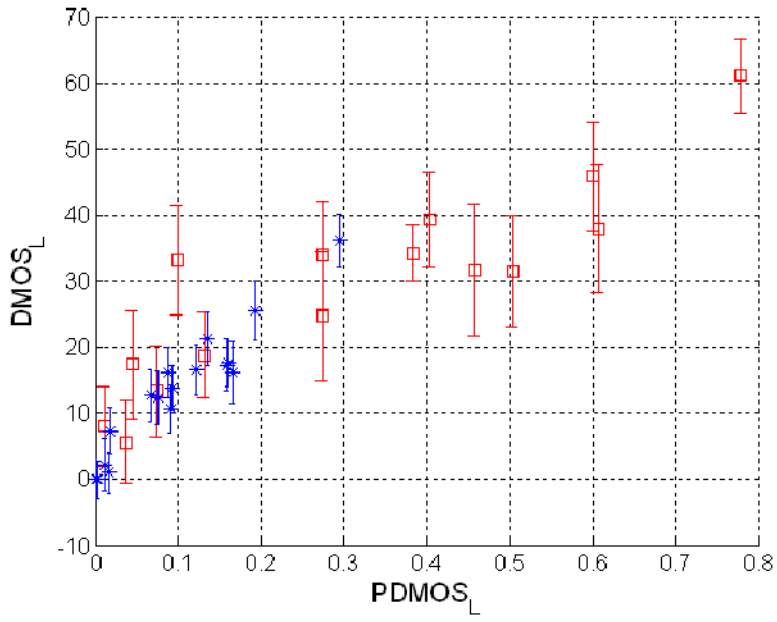


Figure 2.11 Relation between $DMOS_L$ and $PDMOS_L$ for Test 1, 2 and 4 data. (Pearson correlation = 0.9082) ("star" represents data in Test 1 and 2, and "square" represents data in Test 4.)

Figure 2. presents the relation between $DMOS_L$ and $PDMOS_L$ for the sequences with packet loss in Test 1, 2 and 4. The model computation is based on the same model parameters derived previously using data from Test 1 and Test 2 data, i.e. $PDmin = 4$, $PDmax = 13$, $ELmin = 3$, $r = 0.015$, $k = 0.8$, and $c = 0.002$. We can see that the calculated quality degradation has fairly high correlation with actual subjective data, with Pearson correlation 0.9082.

2.2.3 Verification of Quality Metric for Coding Artifacts

There are many proposed metrics evaluating the perceptual quality degradation due to coding artifacts. Here we adopt the PSNR-based model of VQM (VQM_P) proposed in [1], which is defined as

$$PDMOS_C = \frac{DMOS_{C,max}}{1 + e^{s(PSNR - PSNR_T)}} \quad (2.8)$$

where $PDMOS_C$ is the predicted perceptual quality degradation caused by coding artifacts, s is the roll-off factor of sigmoid function, and $PSNR_T$ is the transition value of that PSNR curve. $DMOS_{C,\,max}$ is maximum possible perceptual quality degradation due to coding artifacts.

In order to verify the performance of this model, we conducted Test 3, which consists of 7 loss-free test sequences, with the same 20s cut from "American Pie". These sequences are obtained by encoding with 7 different QP values: 26, 28, 30, 32, 34, 36, and 38. The GOP length is 2 s (24 frames), and the bit rate ranges from about 40 kbps to 256kbps. The other encoding/decoding configurations follow the same settings used in previous tests.

Although in this test only encoded sequences are tested, we can use its original sequence in Test 4 as a reference to obtain $DMOS_C$ for each sequence. Figure 2. (a) shows the relation between $DMOS_C$ and the PSNR values of encoded videos in Test 3, and we can see the relation closely follows the sigmoidal function. By using least square error fitting, we found the following model parameters work well, $s = 0.67$ and $PSNR_T = 33.4$, and $DMOS_{C;max} = 30$. Figure 2. (b) shows the scatter plot of $DMOS_C$ vs. $PDMOS_C$, which has a high correlation of 0.9991 with the significance level above 99%.

2.2.4 Final Quality Metric Considering Both Packet Loss and Coding Artifact

Based on the results in previous sections, and the assumption that the degradations due to coding artifacts and loss-induced artifacts are additive, we propose

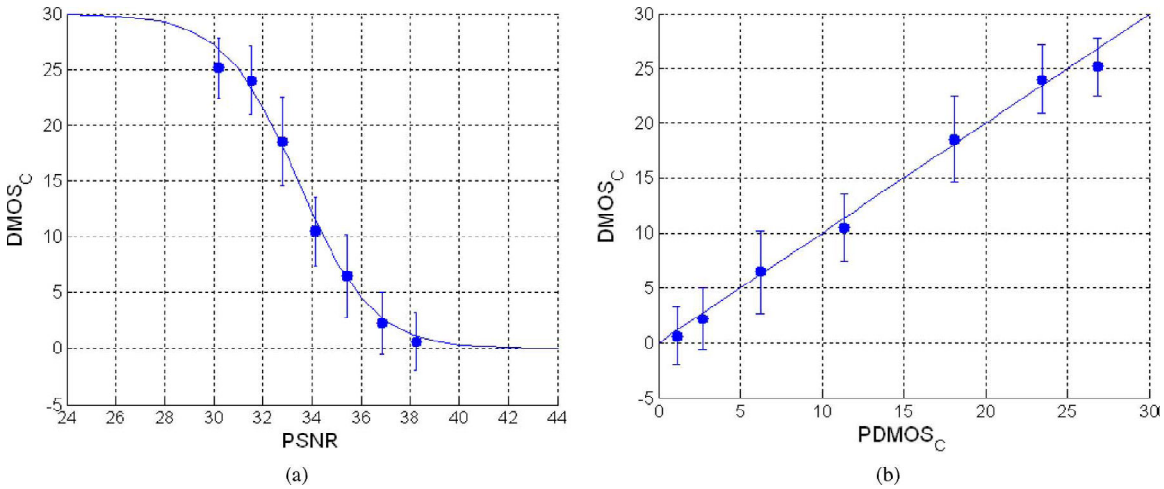


Figure 2.12 Coding artifacts impact on perceptual quality (a) relation between DMOSC and PSNR, and (b) relation DMOSC and PDMOSC, for the data in Test 3.

the following metric for predicting the overall quality degradation:

$$PDMOSCL = PDMOS_C + f\ PDMOS_L \tag{2.9}$$

$$PDMOS_L = \frac{1}{L} e^{-c\ L_{loss}} (1 - e^{-k\ N}) * \sum_{i=1}^{N} \left[e^{-r\ D_i} \sum_{j=EL_{min}}^{EL} \alpha(PD_j) \right] \tag{2.10}$$

$$PDMOS_C = \frac{DMOS_{C,max}}{1 + e^{s(PSNR - PSNR_T)}} \tag{2.11}$$

$$\alpha(PD) = \begin{cases} 0, & PD < PD_{min} \\ PD - PD_{min}, & PD_{min} \le PD \le PD_{max} \\ PD_{max} - PD_{min}, & PD > PD_{max} \end{cases} \tag{2.12}$$

where *PDMOSCL* is the predicated total quality degradation, and $PDMOS_C$ is predicted quality degradation contributed from coding artifacts of loss-free portion of encoded sequences. For readers' convenience, previously defined $PDMOS_L$ is given in Eq.(2.10), combining all previously defined relations in one expression. Note that Eq.(2.11) and Eq.(2.12) are the same as previous Eq.(2.8) and Eq.(2.2), respectively. The parameter *f*

serves to provide appropriate weighting between quality degradation caused by coding artifacts and that caused by loss artifacts.

A. Verification of the Final Proposed Metric

In order to verify the performance of our final proposed metric, the data set of 46 test sequences (5 original sequences are not included) from test 1 through test 4, including various packet loss and coding artifacts, are tested. DMOS here is determined as the differential MOS due to both packet loss and coding artifact, obtained by subtracting the MOS of a given sequence from the MOS of its uncoded original version. For all parameters except the new parameter f, we use the same parameter values derived in previous sections, i.e. $PDmin = 4;\ PDmax = 13;\ ELmin = 3;\ r = 0.015;\ k = 0.8;\ c = 0.002;\ s = 0.67;$ and $PSNRT = 33.4$. In order to determine f, we trained the model by using least square error fitting on 6 clips of two very different sequences, i.e. "American Pie" and "F1 Car Race" (with packet losses) in Test 4, which yields $f = 74$. The remaining 14 sequences (not including original sequences) in Test 4 are used as validating sequences. Figure 2. (d) shows the scatter plot of the DMOS scores vs. the predicted quality values for all the 46 test sequences, with Pearson correlation 0.9135.

B. Performance Comparison

To evaluate the performance of our proposed metric, we compare it with the classic measure PSNR and two well-known state-of-the-art metrics, including VQM (General Model) [34] and SSIM [55]. Specifically, PSNR and SSIM are calculated for each video frame and averaged over the entire sequence. To obtain the SSIM and VQM

(General model) scores of our test sequences, we used the public domain software, with default settings, in [56] and [57], respectively.

Figure 2. shows the relations between DMOS and the above four metrics, for all 46 test sequences in Tests 1-4. The performance of each metric is characterized by four evaluation metrics, i.e. Pearson Correlation Coefficient (PCC), Spearman Rank Correlation Coefficient (SRCC) [18] [19], Root Mean Square Error (RMSE) [4], and Outlier Ratio (OR) [4]. In order to compute RMSE and OR between DMOS and the three metrics, i.e. PSNR, SSIM, and VQM (General model), we first applied the nonlinear mapping function, suggested in [4], on the outputs of the three metrics, and it is defined as:

$$DMOS_P = \alpha_1(\beta_1 x^3 + \beta_2 x^2 + \beta_3 x) + \alpha_2 \quad (2.13)$$

where $DMOS_p$ is the mapped perceptual quality, x is the computed score by an objective quality metric, and α_1, α_2, β_1, β_2, and β_3 are constants to be determined. For each quality metric, we fit the function assuming $\alpha_1 = 1$, and $\alpha_2 = 0$ by maximizing the Pearson correlation between *DMOSp* and *DMOS*, yielding the values of β_1, β_2, and β_3. Then the values of α_1 and α_2 are determined by least root mean square error fitting. More detailed information about mapping processes can been found in [4]. The performances of the four compared metrics are summarized in the Table 2.3, and we can see that our proposed metric significantly outperforms the other three.

By looking at the data points corresponding to the clips with coding artifacts only in Figure 2., e.g. the data in Test 3, we see that the predicted quality by PSNR, SSIM and VQM has high correlation with subjective quality. While for clips with packet losses the predicted quality values do not follow a consistent trend with subjective ratings. This is

not surprising, as these measures were developed mainly to address coding artifacts. On the other hand, our proposed metric measures the coding and loss artifacts separately, and explicitly considers the effect of error length, error severity as well as "forgiveness" and group effects for loss artifacts, which contributes to the high correlation between our proposed metric and actual perceptual video quality.

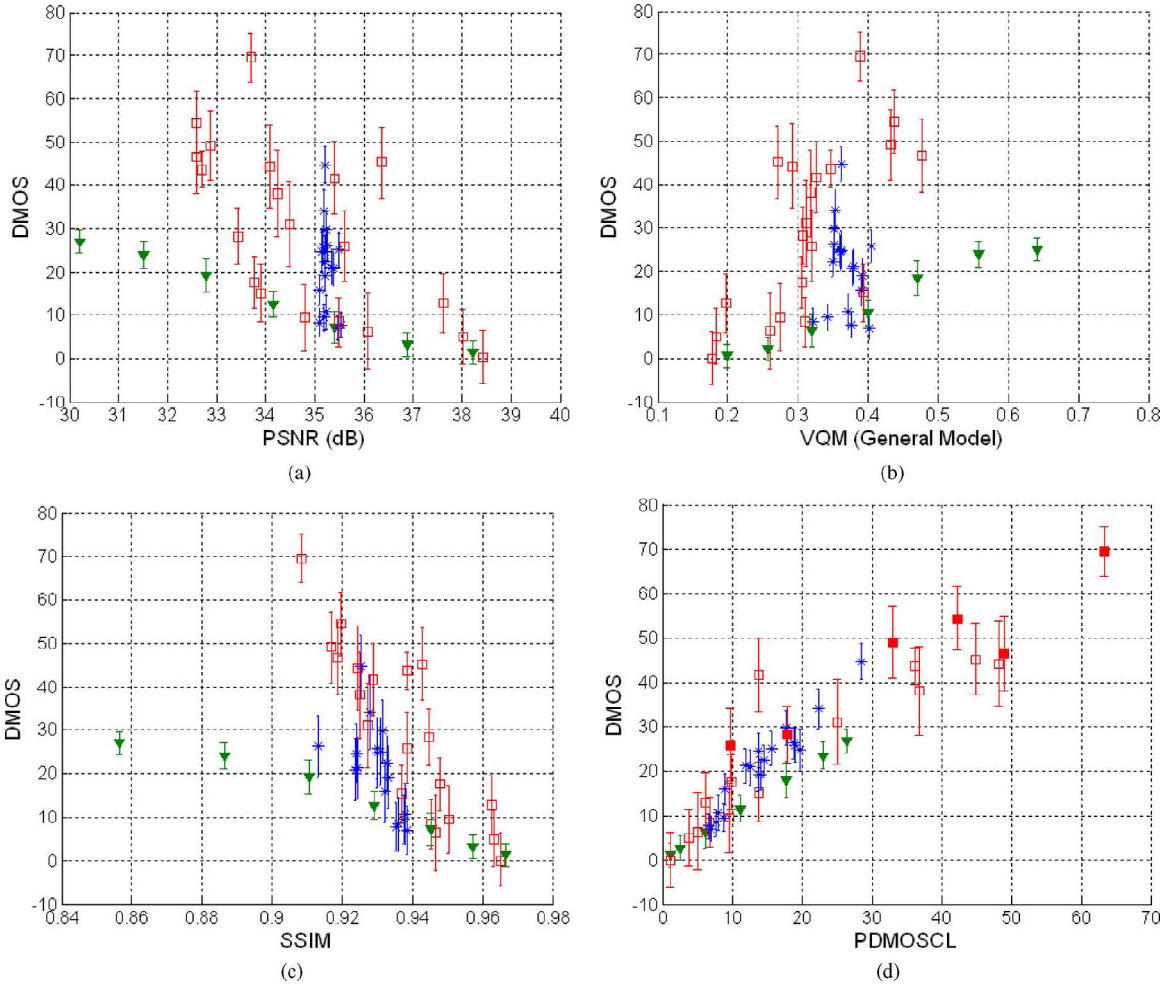


Figure 2.13 Relations between DMOS and different metrics for all test data: (a) PSNR, (b) VQM (General Model), (c) SSIM, and (d) our proposed metric

Table 2.3 Performance comparison of different metrics (for all tests)

Metric	PCC	SRCC	RMSE	OR
PSNR	-0.5221	-0.5338	13.1084	0.6739
VQM(General)	0.3404	0.3320	14.1470	0.6087
SSIM	-0.5022	-0.6786	12.2218	0.6957
Proposed Metric	0.9135	0.9292	8.1670	0.3261

2.3 Summary

In this chapter, we first examine the impact of error length, PSNR drop, and loss location on the perceptual quality of a decoded video subjected to a single transmission loss (which causes the loss of two consecutive frames). We find that an error is visible only if the error length or PSNR drop exceeds a certain threshold, and that the perceptual quality is approximately linearly related to the error length and PSNR drop, subject to some clipping in the beginning and end portion. Based on this finding, an objective measure, MPDS is proposed. Then to take into account the forgiveness effect as well as the clip length factor, we further propose AWMPDS by weighting the MPDS of each loss with a distance-dependent weighting factor, followed by a normalization process by sequence length.

When a sequence contains multiple losses, the perceptual rating depends on the PSNR drop sums of individual losses, the number of losses, the loss pattern, and the inter-distances between losses (which is one attribute of loss pattern). By analyzing the effect of these factors on the subjective rating, we extend our AWMPDS metric to $PDMOS_L$. Finally, by incorporating a prior model for quality degradation due to coding

artifacts given in [1], we proposed a combined metric $PDMOS_{CL}$ for predicting the overall quality degradation due to both compression and packet losses. This metric provides a high correlation (Pearson correlation = 0.9135) with subjective ratings for a large set of sequences with different video content, coding artifacts and loss patterns, significantly higher than some other widely accepted metrics.

Chapter 3

Quality Assessment of Packet Loss Impaired Video Frames

As the basis of quality assessment of video sequences, the accurate measurement of perceptual quality of individual video frames is of significant importance. However, there is a lack of effective quality metrics designed for single video frames affected by packet losses. And our proposed video quality metric in Chapter 2 is built on the fact that PSNR is used to measure the quality of individual frames. As a first attempt to improve the proposed metric, in this chapter, we address the problem of perceptual quality measurement on individual coded video frames with distortions caused by both lossy source coding and lossy transmission.

We also focus on low bitrate H.264 coded video with error concealment and propagation due to packet losses. Identifying Human Visual System (HVS) masking effect as one of the most important factors affecting the perceptual quality, we investigate its impacts on the two aforementioned artifacts in individual video frames respectively, and then evaluate their corresponding Just-Noticeable-Difference (JND) based perceptual

distortions. Finally, we combine the two distortions together and propose a block-wise full-reference quality metric for a single distorted video frame. Our subjective test results show that the proposed metric correlates fairly well with actual subjective quality ratings.

3.1 Subjective Quality Test of Individual Video Frames

A. Test Design

In order to focus on the assessment of spatial quality of videos impaired by packet loss, we choose to subjectively test the quality of single distorted video frame. The benefit of this design is multifold. First, this subjective spatial quality focused test constrains the number of considered quality-affecting features, and thus simplifies the complex problem. Second, we believe the assessment of the quality of video sequence is decided by the quality of individual frames pooled by some temporal attributes. With this subjective test, we can deeply diagnose the intrinsic cause of video quality degradation due to packet losses artifact.

B. Testing Materials

We select 29 video sequences from the standard video database, and they consist of various types of scenes, whose motion and activity levels range from low to high, so that the bias on any particular video scene or category is avoided. In order to eliminate the effect of scene change within individual video, every clip only contains one scene, which lasts about 8~ 10 second.

All the testing sequences are first encoded and decoded following H.264 standard, using JM10.0 encoder/decoder (baseline profile, level 3, and IPP...P GOP structure). The sequences are encoded at QVGA(320x240) resolution and 12 fps frame rate, with the GOP length of 2 sec long. The encoded sequences have various QPs ranging from 30 to 38, so that the spectrum of tested sequences is wide enough. We are mainly interested in wireless video applications over the

3G wireless networks, where payload packets are transported in program data units (PDU) of fixed length, without alignment with the payload packets, so that the loss of a PDU often leads to the loss of two payload packets. With H.264, each frame is coded into one packet. Therefore, to simulate typical losses in a 3G wireless channel, we randomly drop two consecutive P-frames within one selected GOP. We use frame-copy as error concealment method.

Instead of simply evaluating the individual video frames as isolated still images, we would like them to be assessed in the context of video sequences, so that viewers can have certain reference information to judge the test frames before they appear, which better resembles the real-world situation. Therefore, we present each test video frame in following way. Firstly, we extract the last frame in the loss-affected GOP as the test frame, because it usually has severe quality degradation due to error propagation after packet loss occurs. Then, the test sequence is generated by truncating the original uncoded sequence up to the frame immediately before the test frame, and appending the test frame at the end. Totally we designed and tested 46 such processed clips. Half of them are used for training the model and the other half is used for testing the performance of this trained model. The illustration of this process is shown in Figure 3.1.

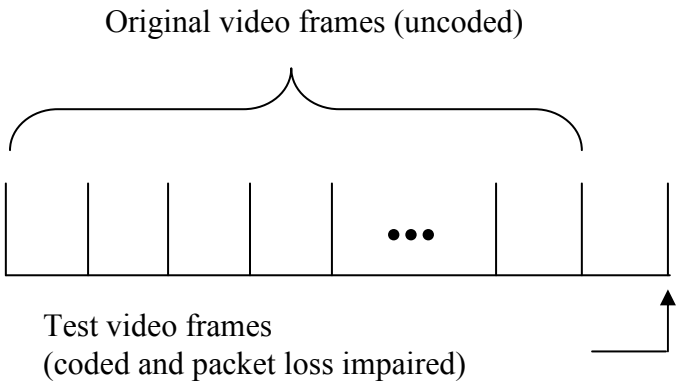


Figure 3.1 Illustration of generating a test sequence

C. Testing Procedures

The subjective test was carried out by following the Single Stimulus Continuous Quality Evaluation (SSCQE) standard in [1]. According to the result in [58], because of memorization effect, displaying duration does not affect viewers' quality ratings on still images very much. Therefore, we performed our test as follows. The test sequence is played normally and the picture is frozen at the last (i.e. test) frame, then viewers are asked to rate the quality on that frame only. Totally 15 viewers participated in the test. The quality was rated by using a mouse to control a scaling bar to give scores from 0 to 100 ("100" means best quality).

3.2 Objective Quality Metrics of Single Video Frames

3.2.1 Distortion Analysis and Classification

Quality degradation from video coding is mainly due to quantization of motion compensated prediction errors between video frames. Because of the powerful

deblocking filter in H.264 decoder, the "blocking" artifact is not obvious, at least at not very low bitrate. So the dominant visual distortion is "blurring", which can be perceived on the edges of objects or textured areas of an image. They impair the perceptual quality of frames globally. On the other hand, the main artifacts from error propagation (due to packet loss) are local spatial chaos caused by some small mispositioned image fractions, usually around the edges of objects with motions. In this work, we treat these two kinds of distortion separately, and then combine them into one final quality metric.

In order to differentiate these two artifacts, the attributes of their incurred distortions are investigated. Base on our observation, generally speaking, packet losses cause the artifacts in the form of displacements of small image fragments, but within them the texture of most pixels are preserved, which can be seen from the sails of the boats in Figure 3.2 (b). One the other hand, the coding distortion just blurs the image details, and hence its texture is damaged, which is obvious in the regions of water waves in Figure 3.2 (b). Because of this difference between the two artifacts, the averaged absolute intensity differences between corresponding blocks in original and distorted frames caused by packet loss is usually larger than those due to coding artifacts. Therefore, all the blocks in a distorted video frame can be classified as either coding-impaired or loss-impaired blocks, according to

$$Distortion\ Type = \begin{cases} coding\ artifacts, & D_{blk}(j,k) < p \\ propagation\ error, & D_{blk}(j,k) \geq p \end{cases} \tag{3.1}$$

with

$$D_{blk}(j,k) = \frac{1}{n^2} \sum_{(x,y)\in blk_{j,k}} \left| D_{pix}(x,y) \right| \tag{3.2}$$

where $D_{pix}(x, y)$ is the pixel intensity difference at position (x, y) within block (j, k) between original and distorted frames, and n x n is block size. We set $n=8$, and $p=10$ in our experiment. Note that we assume that the coding distortion within the blocks where packet losses occur can be ignored, since usually packet loss artifacts are the dominant distortion.

(a) (b)

Figure 3.2 Sample of (a) original video frame and (b) decoded video

3.2.2 Perceptual Distortion of Error Propagation

For individual frames, in order to evaluate the perceptual distortion due to packet loss and its error propagation, we take into account the spatial masking effects of the HVS [59]. There are two kinds of significant masking effects with respect to the visual perception of the distortions. The first is the luminance masking, or luminance adaptation. Usually, human eyes are more sensitive to errors in the mid-gray luminance areas, but less sensitive to errors in very bright or dark areas. A piece-wise linear approximation to the JND threshold is shown in Figure 3.3 [60]. The second is activity masking, or texture masking, which occurs in the areas with complex textures. Generally

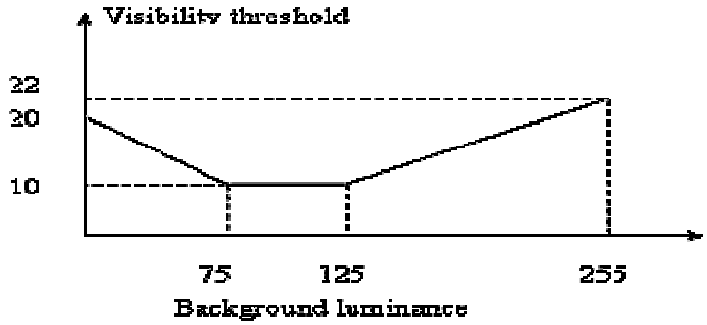


Figure 3.3 Illustration of relation between the error visibility threshold and the background luminance

speaking, distortions occurring in textual regions are less noticeable than those occurring in the smooth regions.

Because the texture activity is presented in the form of edges, we propose to use the density of strong edges in an area as an indication of the texture richness or activity level. Extraction of edge information can be achieved by using any edge detection method, and herein, the popular Sobel filter is used and only strong edges (with magnitude above 15) are kept. Since isolated edge points can be considered as noise, they are deleted by morphological operations. The edge density in a region is calculated as the ratio of the number of edge pixels to the number of all pixels in the region. Because of the confined size of error propagation, the incurred quality degradation can be affected by both luminance and texture masking effects. In other words, a packet loss with its artifact in small regions can be masked out if it is below the visibility threshold of combined masking effects. Although this "Just Noticeable" attribute of some packet losses is contrary to our intuition, it can be applied in some particular circumstances, e.g. the regions with very dark intensity or bushes with rich texture.

In order to calculate the combined visibility threshold, we extend the concept of pixel-wise JND to block-wise JND, where the visibility of the distortion in a center block

is dependent on the masking effect of its surrounding blocks. Following the concept in [59], for each 8x8 block, its visibility threshold is defined as:

$$JND_P(j,k) = \max\left(b * den\_edge(j,k), luma\_th(j,k)\right) \tag{3.3}$$

where *den_edge(j,k)* is the maximum edge density across 8 neighboring blocks of the center block at *(j,k)* in the reference frame, and *luma_th(j,k)* is the visibility threshold due to luminance masking effect, which is illustrated in Figure 3.3, where the background luminance is the average of pixel values of the 8 neighboring blocks in the reference frame. *b* is a scaling parameter such that *b*den_edge* properly represents the visibility threshold due to texture making effect, and we set *b = 200* in our experiment.

From this JND profile, the distortion of a block can be normalized and converted into JND unit. The total perceptual distortion over a frame caused by error propagation is calculated in Minkowski fashion as follows:

$$D_P = \left(\sum_{(j,k)\in I} D_{P,blk}(j,k)^{\alpha} \right)^{1/\alpha} \tag{3.4}$$

with

$$D_{P,blk}(j,k) = \max\left[\frac{D_{blk}(j,k)}{JND_P(j,k)} - 1, 0\right] \tag{3.5}$$

where $D_{blk}(j,k)$ is the block-wise intensity deviation previously defined in Eq.(3.2), $D_{p,blk}(j,k)$ is the JND-scaled distortion due to packet losses, and α is the parameter to be determined. In addition, to avoid overweighting the extreme large values of $D_{p,blk}(j,k)$ during the pooling process, it is clipped (to 2 here), so that the proposed metric is well correlated to our subjective data.

3.2.3 Perceptual Distortion of Coding Artifacts

Although there are many existing pixel- and subband-based JND metrics for evaluating image/video coding distortion [59], most of them are not compatible with our proposed block-wise packet-loss-distortion metric. Therefore, we try to propose a specialized coding distortion measurement for this purpose.

Based on our observations in our subjective experiments, the blurring effect is the dominating distortion caused by coding artifacts, and it usually spreads throughout the entire images. However, the quality degradations across the images are not even. The reason is twofold. First, smooth regions are usually relatively better coded so that the distortions in these regions are objectively small. Second, the regions with high contrast (or rich in strong edges) are less affected by coding artifacts than those with low contrast, although their objective distortions are similar. Taking Figure 3.4 as an example, it is one of the JPEP2000 images from LIVE image dataset [61]. We can clearly see that the highly contrasted regions around the eye of the bird are relatively less blurred, while the distortions, at similar level, on its feathers around the neck and on the wing badly blur those objects, where the local contrasts are relatively low.

Figure 3.4 Sample (a) encoded image, and (b) absolute difference between original and distorted images (brighter areas indicate higher distortion).

Based on our observations and inspired by the work [62], we propose a simple full-reference block-wise JND-based quality metric for coding artifacts, based on the local contrast information.

Firstly, we differentiate smooth regions from edgy regions by computing local contrast within 8x8 block. Within smooth regions, since distortions are usually unnoticeable, or the distortion visibility is very high, for simplicity we set JND threshold as infinity; within edgy regions, we set JND threshold proportional to the local contrast, which is defined as:

$$JND_C(j,k) = \begin{cases} \dfrac{contrast(j,k)}{F}, & contrast(j,k) \geq t \\ \infty, & contrast(j,k) < t \end{cases} \tag{3.6}$$

where *contrast(j,k)* is the contrast of block at *(j,k)*, and for simplicity, it is defined as the difference between maximum intensity and minimum intensity values in that block in original frame. *F* is the proportional factor linking contrast and JND value. Based on the experimental results in [63], noticeable luminance difference is roughly proportional to the luminance edge height, despite of slight variation of their slopes (around 8) in different background luminance. Here we set *F=8*. *t* is a threshold to separate smooth from edgy regions. Via close subjective examination on the test images, we find that *t=16* is a good threshold.

Secondly, we convert the pixel-wise distortion between original and coded images to JND scale as follows:

$$D_{C,pix}(x,y) = \max\left[\frac{D_{pix}(x,y)}{JND_C(j,k)} - 1, 0\right] \tag{3.7}$$

where $D_{pix}(x,y)$ is the pixel-wise distortion within the block *(j,k)*, and $D_{C,pix}(x,y)$ is the corresponding converted JND scaled distortion.

Thirdly, based on the results in [67] that the highest distortion (in JND scale) among all pixels within a block can be used to indicate the distortion level of that block, we define the distortion of each 8x8 block as:

$$D_{C,blk}(j,k) = \max_{(x,y)\in blk_{j,k}} [D_{C,pix}(x,y)] \tag{3.8}$$

Lastly, we pool the distortions of all blocks as follows:

$$D_C = \left(\sum_{(j,k)\in I} D_{C,blk}(j,k)^{\beta} \right)^{1/\beta} \tag{3.9}$$

where β is a parameter to be determined.

Before we integrate the proposed distortion measurement for coding artifacts into final quality model, it is validated with 169 JPEG2000 coded images from LIVE image dataset, where blurring is the most annoying artifact. We also compare the performances of MSE and SSIM with our proposed metric. For our proposed metric we set $\beta=2$, which produces its highest correlation with the tested LIVE subjective data; for SSIM, we use the default parameters in the Matlab program [56]. Figure 3.5 shows the scatter plot between Differentiate Mean Opinion Sores (DMOS) and the three comparing metrics.

Table 3.1 lists the Root Mean Squared Errors (RMSE) between subjective data and their logistics fittings, along with the Pearson correlation coefficients (after logistics mapping according to their fittings).

Table 3.1 Performance comparison of different metrics

	MSE	**1-SSIM**	**Proposed**
RMSE	7.498	5.815	5.733
Correlation	0.84462	0.9002	0.92964

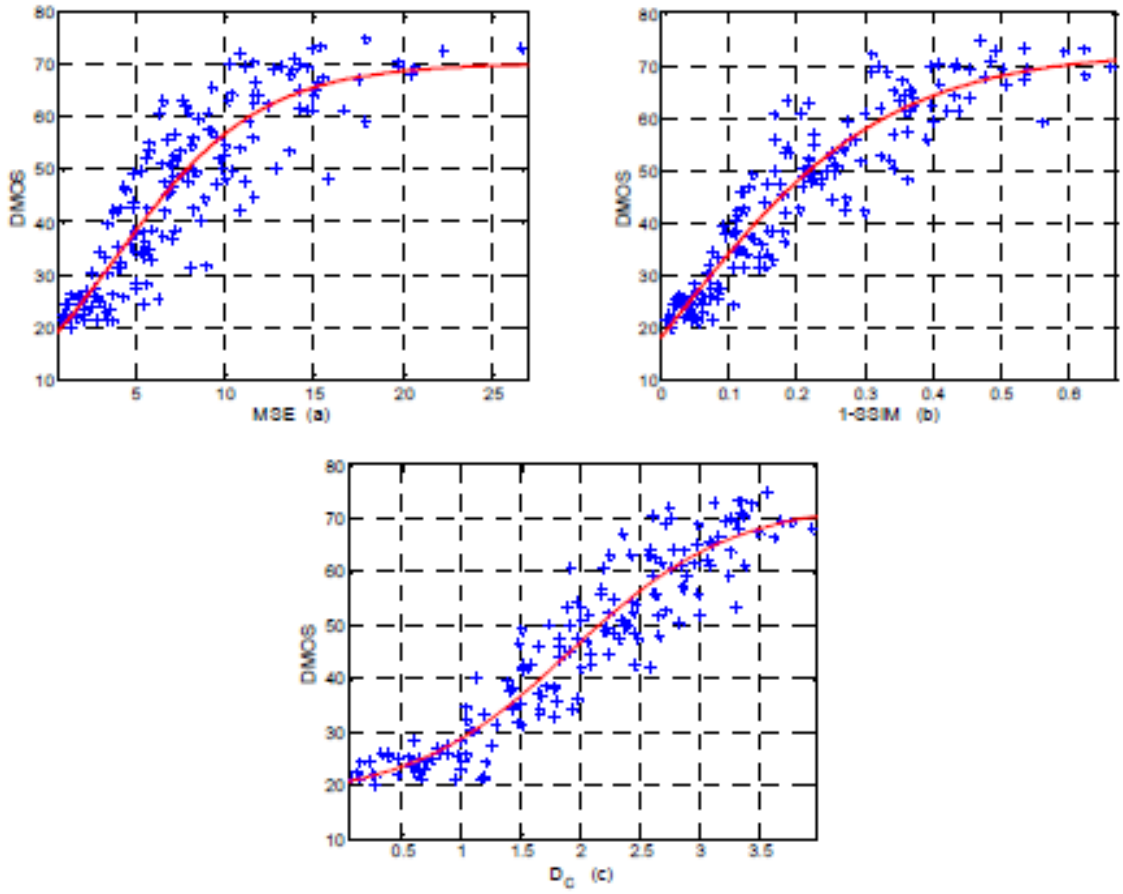


Figure 3.5 Scatter plots of the relations between DMOS caused by coding distortion and (a) MSE, (b) 1-SSIM, (c) the proposed metric. ("Stars" represent data points of LIVE database, and curves represent logistics fittings.)

From these results, we can see that our proposed simple metric effectively predicts the image quality degradation caused by coding artifacts, and for the tested dataset, it significantly outperforms MSE, and achieves slightly better performance than SSIM (with default parameters).

3.2.4 Proposed Quality Metric Considering both Coding and Packet Loss Artifacts

The impacts of error propagation and coding artifacts on perceptual video quality are assumed additive, and the overall perceptual distortion is defined as

$$D_{PC} = D_P + w\, D_C \tag{3.10}$$

where D_{PC} is total distortion due to the two artifacts, and w is their weighting factor.

The high-level flowchart of the entire process is shown in

Figure 3.6 The flowchart of entire process of the proposed metric.

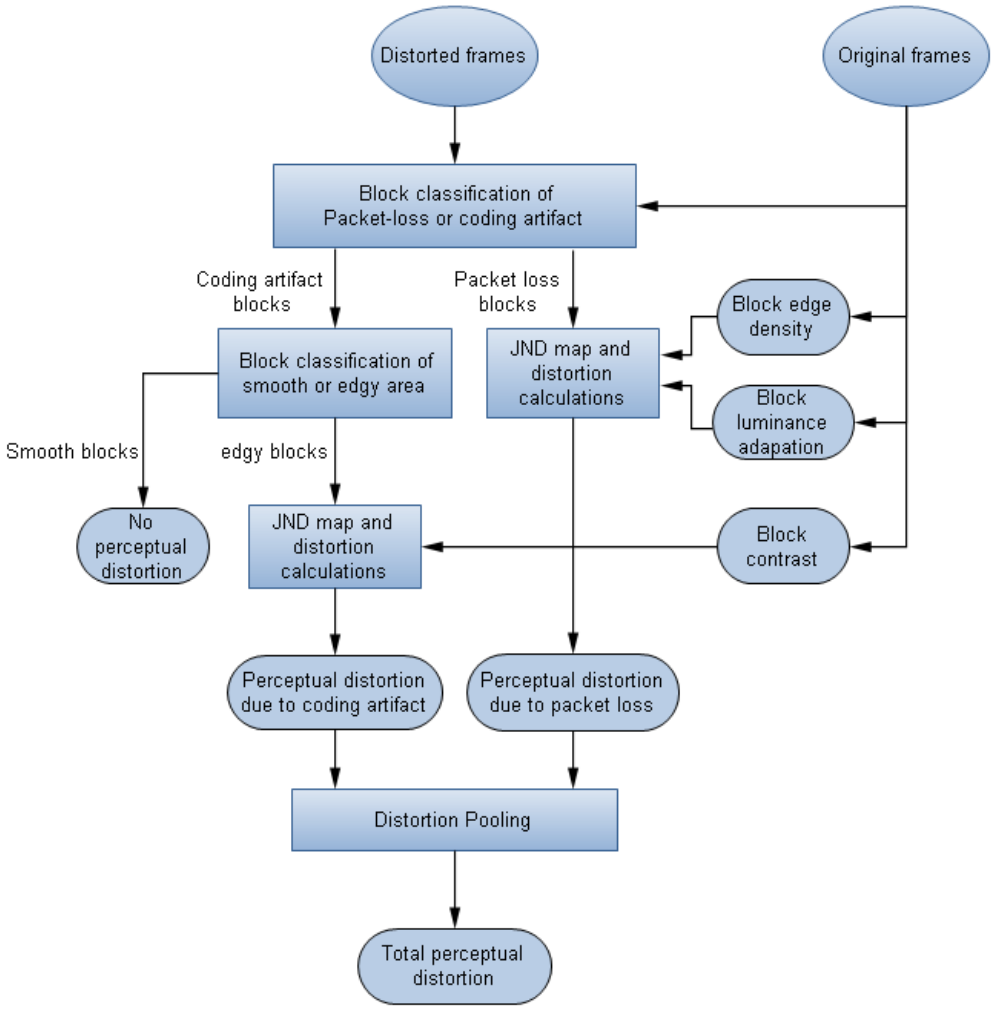


Figure 3.6 The flowchart of entire process of the proposed metric

After post-screening the raw data in the way suggested by [1], Mean Opinion Score (MOS) was calculated. By maximizing the correlation between MOS and proposed metric, after cubic polynomial mappings, over the training dataset, we set $\alpha=2$, $\beta=1$, $w=0.003$, with the other parameters kept same as before. Testing dataset is used to validate this trained model. Figure 3.7 shows the relationships between MOS and our proposed metric, for both training and testing datasets. For comparison purpose, the scatter plots between MOS, and MSE and SSIM are also shown in Figure 3.7. The Pearson correlation coefficients between MOS and the three metrics (after cubic polynomial mappings), are listed in Table 3.2. Note that SSIM is calculated using the Matlab code in [56], with default parameters.

We can see that our proposed perceptual quality degradation metric can accurately predict the actual subjective quality of decoded video frames distorted by joints artifacts of coding and packet loss, while the generic metrics MSE and SSIM do not perform well in this case.

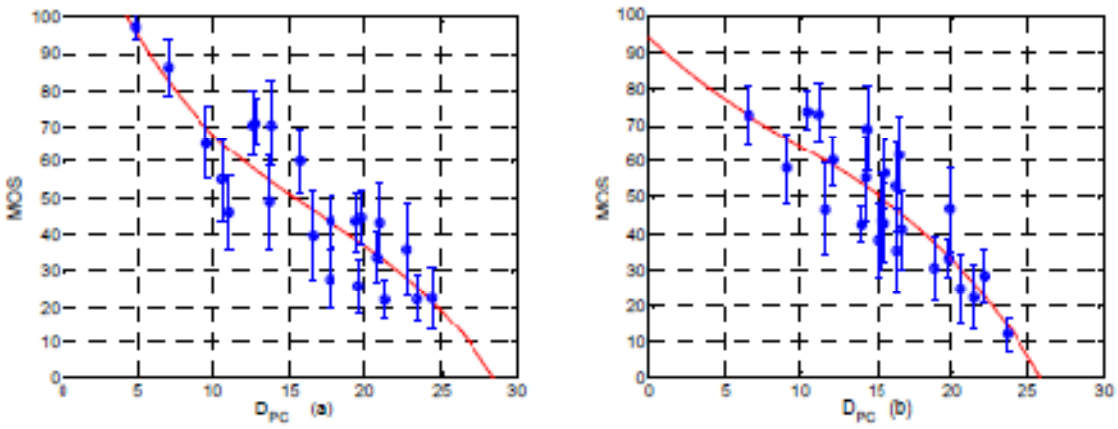


Figure 3.7 Relationship between MOS and proposed metric (a) Training dataset, and (b) Testing dataset, with RMSE of fittings 10.57 and 9.866, respectively. (Vertical bars around data points represent 95% confidence interval.)

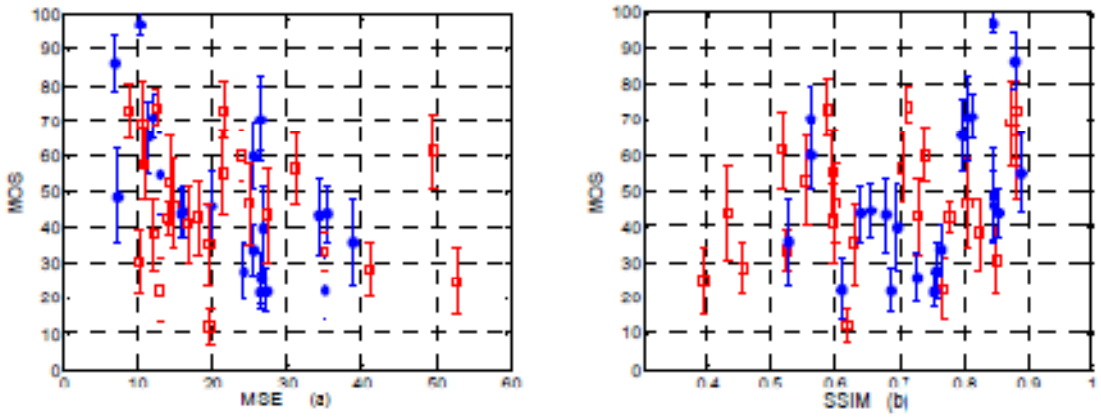


Figure 3.8 Relations between MOS and (a) MSE, and (b) SSIM ("solid dots" represent the training data, and "empty squares" represent the testing data).

Table 3.2 Pearson correlations of the compared metrics

	MSE	SSIM	Proposed
Training	0.6315	0.5695	0.8860
Testing	0.2665	0.4286	0.8626

3.3 Summary

In the work described in Chapter 2, we mainly focus on investigating the impact of loss patterns (length, position, pattern) on video perceptual quality, whereas we use PSNR drop in individual frames for simplicity to evaluate the severity of a loss. In the work presented in this chapter, we attempt to improve quality prediction accuracy for individual frames by incorporating the properties of the human visual system into measures of distortions in loss-affected frames.

Towards this goal, we evaluate the perceptual quality of individual decoded video frames, which are distorted by both coding artifacts and error propagation due to packet loss. By taking into account of different masking effects of the HVS, we propose two

JND profiles for coding and packet loss artifacts. Furthermore, we propose a combined metric that evaluates the perceived quality degradation due to both artifacts. The predicted perceptual distortions by our proposed metric have fairly high correlation with subjective quality ratings.

Ideally, the video quality metric described in Chapter 2 can be improved by replacing PSNR drop with the proposed metric in this chapter for measuring the quality of individual video frames. It is one of our future work.

Chapter 4

Video Quality Assessment with Aid of Saliency

As one of the most important components of human visual system, saliency is studied in many different image and video processing applications. In this chapter, we try to investigate whether and how saliency can be used to improve objective quality assessment.

Traditional objective approaches to video quality assessment such PSNR and MSE, consider the error at all pixels equally and overlook the uneven distribution of visual importance, hence their predicted and perceived video qualities are usually not correlated very well with subjective ratings [4] [55] [64] [65]. This discrimination is even more obvious in the presence of packet losses, because packet losses can usually cause significantly different visual impacts in different video segments. Along with the rapid development of network technologies, packet loss is becoming one of the most annoying distortions in videos delivered over the network. However, the research work on quality evaluation for video impaired by packet loss is still at beginning phase. Therefore, in the

hope that saliency may significantly improve the performance of existing quality metrics, we perform our research work discussed in this chapter.

In this chapter, we exploit the application of saliency information for perceptual quality assessment of packet-loss-impaired videos. We propose and validate three approaches of saliency-based objective metrics for video quality assessment, i.e., saliency-weighted error, saliency-fidelity, and saliency temporal variation metric. The saliency-weighted error metric uses a weighted average of pixel errors between original and distorted video, where the weight at a pixel depends on its visual saliency. The saliency-fidelity metric measures the change of saliency distributions between the original and distorted video as quality indicator. The saliency temporal variation metric considers the temporal variation of the saliency map of the distorted video and uses the product of this temporal variation with each metric in the first two categories. Validated by our subjective test data, each of the three saliency-based metrics can significantly improve quality prediction accuracy over conventional non-saliency based metrics.

To further improve prediction accuracy, we combine multiple factors from the previous three categories using stepwise multiple linear regression analysis. The final metric for video that includes four factors provides additional significant gain over using the best single factor.

4.1 Saliency Measurement

Psychology evidence suggests that the most important function of human selective visual attention is to orientate rapidly towards salient objects in a cluttered

visual scene. Authors of [36] [37] [38] [39] who work on computational VAM, have agreed that a unique "saliency/importance map" that topographically encodes for stimulus conspicuity over the visual scene is an efficient and plausible bottom-up control strategy. In the models they developed, several early vision features are first extracted in the bottom-up manner, then either "center-surround" structure based on multi-scale feature maps (as in [36] [37] [38]) or contrast computation between important region and the background (as in [39]) is applied to obtain the local conspicuous regions that simulate the visual receptive fields of human. The resulting topographic feature maps are then combined across scales and channels to form a "saliency/importance map". Human eye movements are recorded by the eye-tracking device to examine the correlation of the developed models with subjects. Among these VAMs, Itti's bottom-up saliency based visual attention model (SVAM) [37] [38] has demonstrated high correlation with human eye movements over static images and been used in various applications successfully. This encourages us to integrate his SVAM into video quality evaluation.

4.1.1 Saliency and FOA Detection Methods for Images

Itti's SVAM attempts to simulate which location in the image will automatically and unconsciously attracts our attention. With SVAM, an input image is decomposed into a set of multiscale "feature maps" of color, intensity and orientation, in a massively parallel manner, using linear filtering at eight spatial scales. Color perception is built on red-green (R-G) and blue-yellow (B-Y) opponent color space. Local orientation information is obtained using oriented Gabor pyramids in 4 directions (0°, 45°, 90°, and 135°). Center-surround differences between a "center" fine scale and a "surround"

coarser scale yield the feature maps (color, intensity, orientation), which is the simulation of retina high and low contrast sensitivity. Each feature map is then endowed with non-linear spatially competitive process and linearly combined with some pre-defined weights into a unique scalar "saliency map". The saliency map is a gray scale image, with a high gray level at a pixel indicating that the pixel attracts more visual attention. The size of the saliency map is determined by a chosen pyramid level, with default level being 5, which is 1/16 size of the original image. Subsequently the interplay between the winner–take–all and inhibition–of–return is applied to the saliency map to draw the FOA towards the locations in the order of decreasing saliency, which generates the model's output in the form of spatio–temporal attention scanpaths. Generally, this step can produce multiple binary FOA maps which are calculated based on the saliency map, indicating detected FOA areas, in the time order of humans' gazes across the image. In this work, we used the SaliencyToolbox 1.0 [66] in MATLAB developed by Dirk B. Walther, which implements Itti's SVAM.

In the original SaliencyToolbox 1.0, the three basic feature maps are given equal weight when generating the overall saliency score at each pixel. In the video frames with packet loss artifacts, loss-affected regions often lead to discontinuity across block boundaries. In order to guide the saliency detection algorithm toward these regions, we experimented with different weights between color, intensity and orientation by closely looking into each feature map. We also explored the impact of the number of iterations used in the non-linear spatially competitive dynamics computation, which determines the size of detected saliency regions. We found that the following setting leads to the detected saliency regions that best match with our subjective examination. We set

weights to 0.5, 0.5, 1 for color, intensity and orientation respectively, and 2 for iteration number. Figure 4.1 shows the boundaries of the first 5 detected FOA regions of two sample packet-loss impaired frames as well as their reference versions.

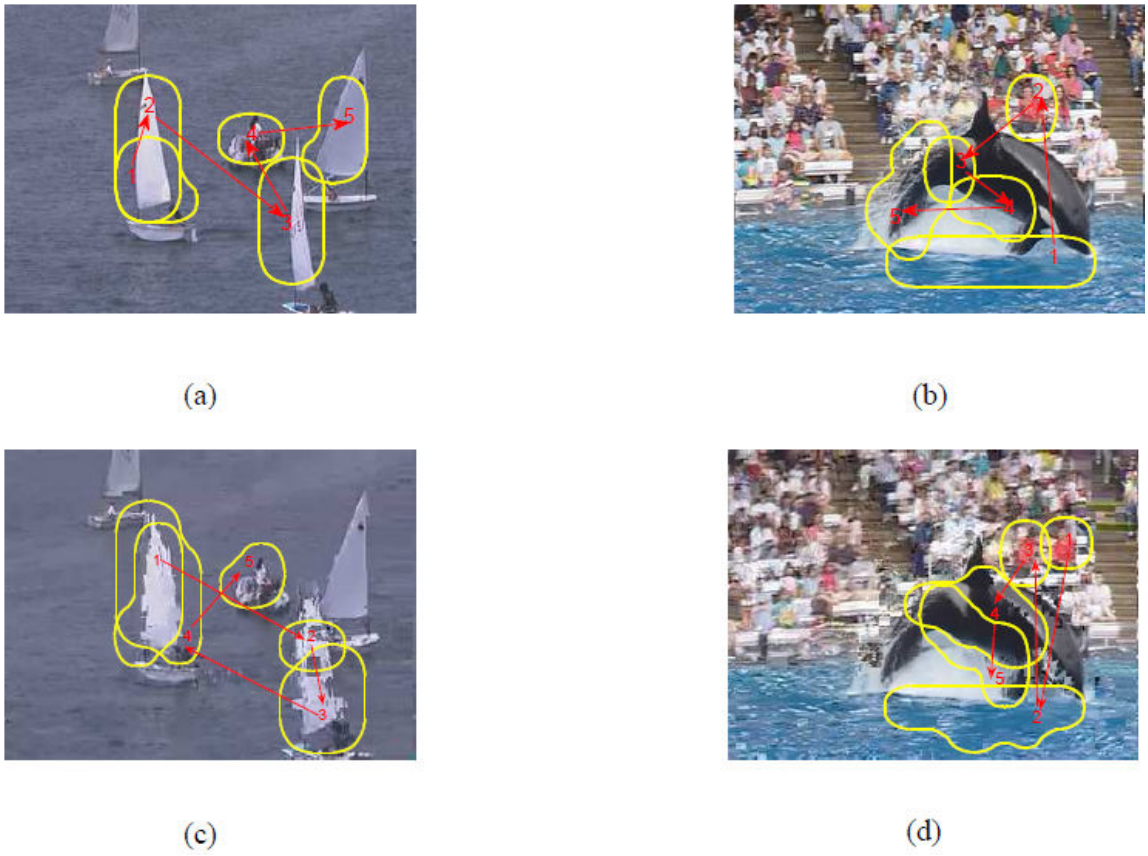


Figure 4.1 Reference frames from the sequence (a) “optis” and (b) “whale”; distorted frames from the sequence (c) “optis” (with MOS=2.53), and (d) “whale” (with MOS=2.73). The circles in each video frame shows the detected first 5 FOAs using the saliency model, the arrow gives the attention scanpaths.

4.1.2 Incorporating Motion As Saliency Feature For Video

Motion is one of the most important features that distinguish a natural video sequence from still images, and it is an important factor when evaluating perceptual video quality. However, the original SaliencyToolbox is developed for detecting saliency on a single image, and hence motion is not considered. Following the general idea of saliency

detection, the motion feature that draws visual attention should be distinct from their surroundings. It should signal a perceived change of the motion characteristics of a region in contrast with its surrounding. Hence, unlike traditional motion detectors for video, the detector in saliency computation needs to find where there is a significant change in motion characteristics, but not exactly how much the target moves. Following [67], we implement a biologically inspired elementary visual motion detection method, referred as "Hassenstein–Reichardt (HR) correlation-based motion detector" [68]. Since Hassenstein and Reichardt first described the visual system of the beetle *Chlorophanus* using correlation of signals between adjacent ommatidia with a time delay in 1956 (Figure 4.2), visual motion detectors based on this principle have been widely developed [69] [70] [71].

The aim of the model is to study the ways that change in motion attributes alter the way in which natural moving stimuli are likely to be perceived. Here we incorporate this model by making use of the existing multi-scale representation of video frames. For every pyramid level, we compute four opponency maps for motion in four directions: left, right, up and down, by one pixel each between two consecutive frames. Then we model the receptive fields for motion perception with the existing center surround mechanism in SVAM by inputting the four motion feature maps into the saliency map computation, as with the other features (color, intensity and orientation). Recall that a movement of 1 pixel at level σ in the pyramid corresponds to a movement of 2^{σ} at the original level 0 image. Thus the overall motion feature examines changes in motion in four directions over a magnitude range of 1 to 2^8.

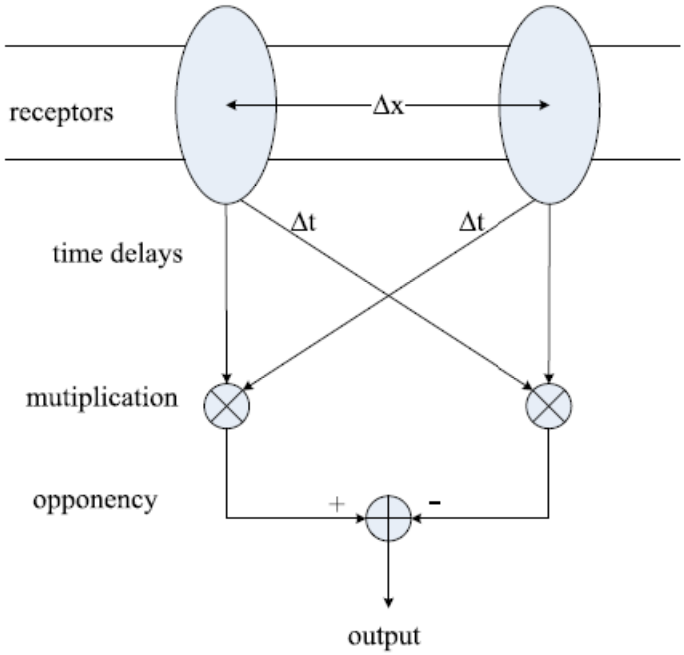
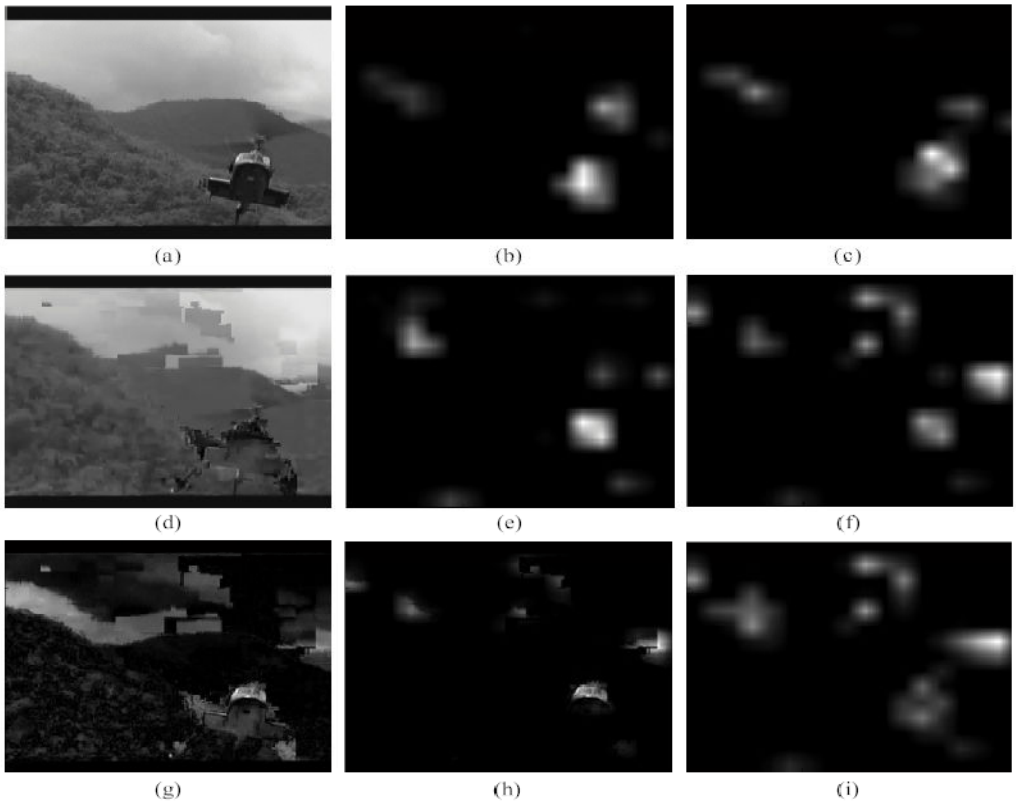


Figure 4.2 Schematic of Hassenstein - Reichardt Correlation-Based Motion Detector

In order to better understand the role played by the motion feature, we compare the saliency maps with and without motion feature in Figure 4.3 (b) and (c), (e) and (f) in each subimage of three sample frames. We can see motion is a very useful cue for visual attention detection, not only for the case where there are multi-objects moving simultaneously, such as the dangling leaves in “leaf”, but also can increase saliency intensity when the salient objects have movement as in “aircraft” and “optis”.

Furthermore, motion is very helpful in catching the artifacts-affected areas, such as in “aircraft”. These observations motivate us to include the motion feature into the SVAM and assign a higher weight to it when combining all the feature maps to derive the saliency map. Through trials-and-errors, we found that the weights of 0.3, 0.3, 0.7 and 1.0 for color, intensity, orientation and motion respectively, and using one iteration, yielded saliency maps which best match with our subjective examination.

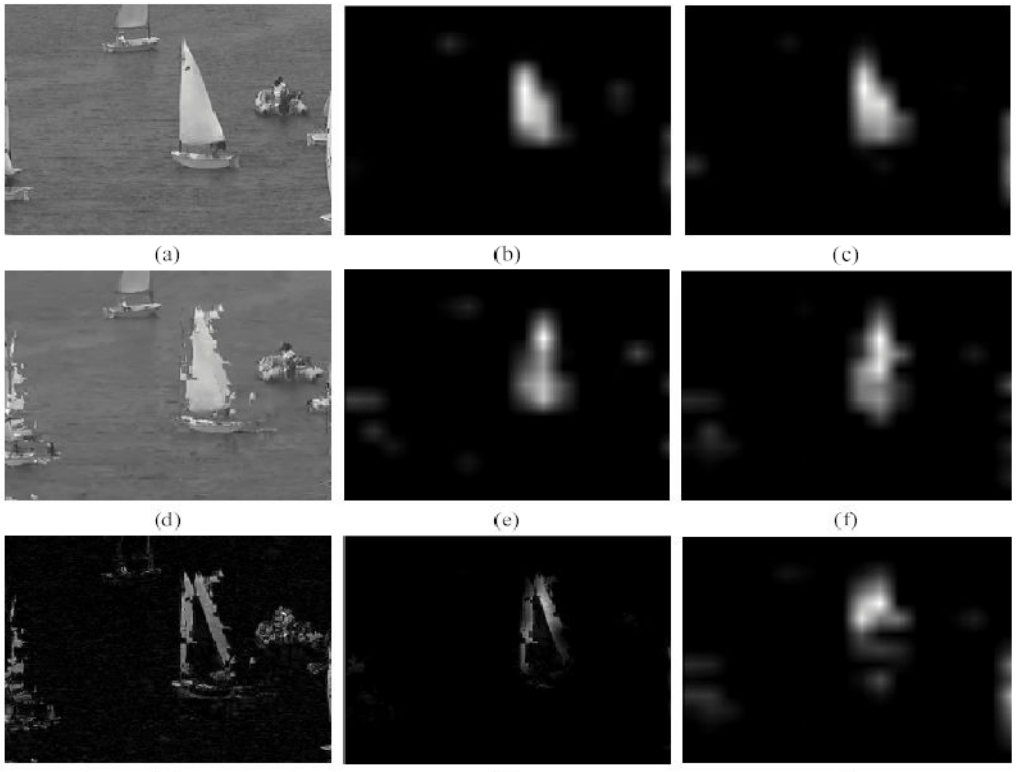


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

"aircraft"



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

"optis"

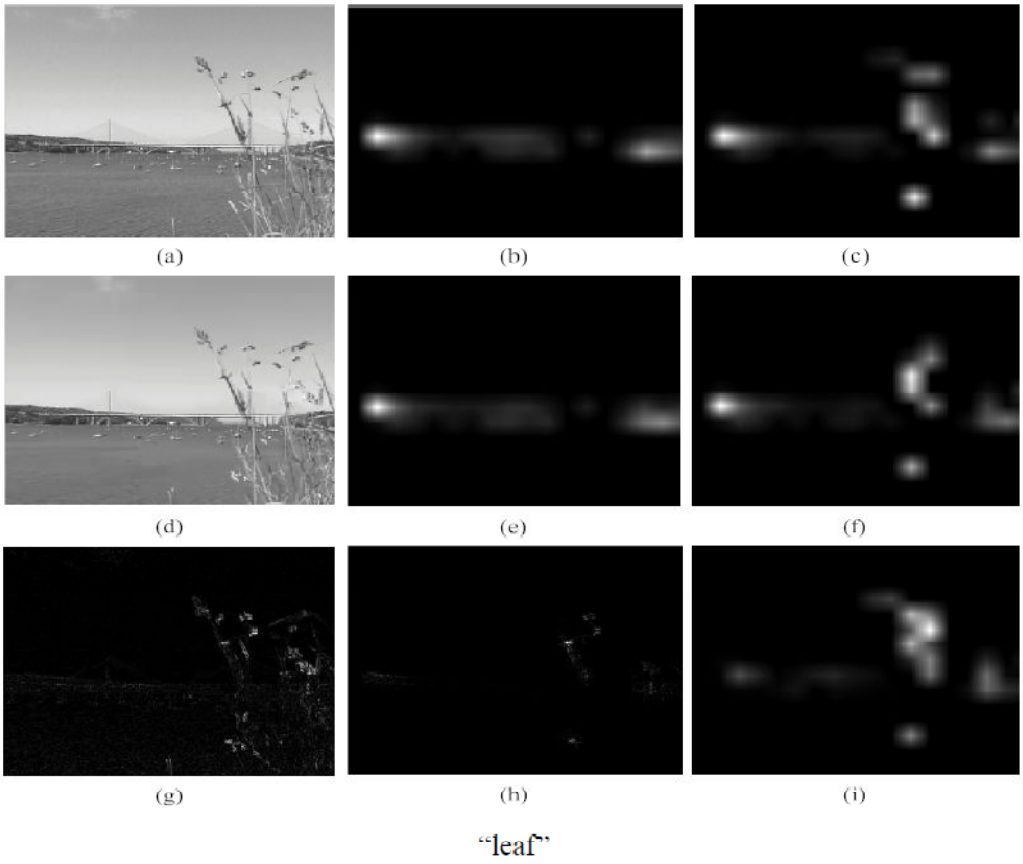


Figure 4.3 Demonstration of proposed methods taking the example of "aircraft" (MOS = 2.45, the 11th frame is shown), "optis" (MOS = 2.36, the 12th frame is shown) and "leaf" (MOS = 5, the 12th frame is shown). Included in each subimage are: (a) the reference video frame, (b) the reference saliency map without using motion feature, (c) the reference saliency map using motion feature, (d) the distorted video frame, (e) the distorted saliency map without motion, (f) the distorted saliency map with motion, (g) the absolute difference between reference and distorted video frame, (h) the saliency-weighted absolute difference image using motion features, (i) the absolute difference between the saliency maps obtained using motion features.

4.2 Subjective Video Quality Tests

In order to examine the role of saliency in the perceived overall quality of video impaired by packet losses, we perform the following subjective test.

In this test, the overall qualities of video sequences impaired by packet losses are subjectively evaluated. There are totally seventeen (17) sequences in this test, listed in

Table 4.1, (all the test sequences are available online [72]). We code each sequence with JM H.264 encoder (baseline profile) with IPPP GOP structure with GOP length of 2 second. The packet losses are deliberately inserted so that 3rd and 4th frames of a selected GOP are dropped. The remaining part of the GOP suffers from impairments due to error propagation. We cut out only the GOP with packet loss impairment for subjective viewing, thus each test clip is 2 second long.

One may argue that 2 second is not long enough for the viewers to give confident quality ratings. However, when designing our preliminary test plan, we observe that the quality ratings from majority of the viewers are fairly consistent, and the feedbacks from viewers also suggest that viewers are comfortable with rating the 2 s clips. We believe this may be due to the fact that they are well instructed and have enough time to familiarize themselves with the test procedures. On the other hand, the error propagation during this period can cause a reasonable level of quality degradation, but not totally destroy the sequences.

In this test we use ACR protocol recommended in [2]. The viewers are told that rating is on packet loss distorted sequences and are asked to rate the quality of each video in the range from 1 (bad) to 5 (excellent). Each viewer, who can freely adjust their viewing distance, is asked to give overall rating after one video is played completely. In each test session, a viewer is shown all 17 test sequences, with the first 5 used for training, and remaining 12 for testing (randomly ordered), as described in

Table 4.1.

Table 4.1 Description of video clips used for subjective test

1	f1	a racing car running on the racetrack
2	car	a car passing by fast
3	bottles	several bottles moving on the product line
4	wave	water waving
5	plane	a plane flying across the sky
6	bus	a bus passing by the street
7	leaf	leaves dangling in breeze, a bridge at a distance
8	optis	several sailboats moving on the sea
9	ship	a moving ship on the sea
10	stockholm	bird-eye view of city, buildings, streets and cars
11	whale	a whale performing in the center, audiences on the background
12	livingroom	camera moving, furniture and decoration
13	liberty	liberty statue, and a ship passing by
14	mobile	a red ball pushed by a toy train
15	bedroom	camera moving, a bed and table in the front
16	boat	a boat crossing the sea with a man standing on it
17	aircraft	a helicopter flying quickly on the sky towards the camera

During each viewing session, the test lasts about 2 minutes without break. All the tests are conducted in a well lit room using the same monitor and settings with test described in Chapter 2. Each set of video sequences is evaluated by 32 viewers (12 video experts, 20 non-experts). Other aspects of the test set-up closely follow the description of ITU-R BT500-11 recommendation [1] and the subjective test discussed in Chapter 2.1. The range of rating for video is shown in Figure 4.4 (training sequences are not included).

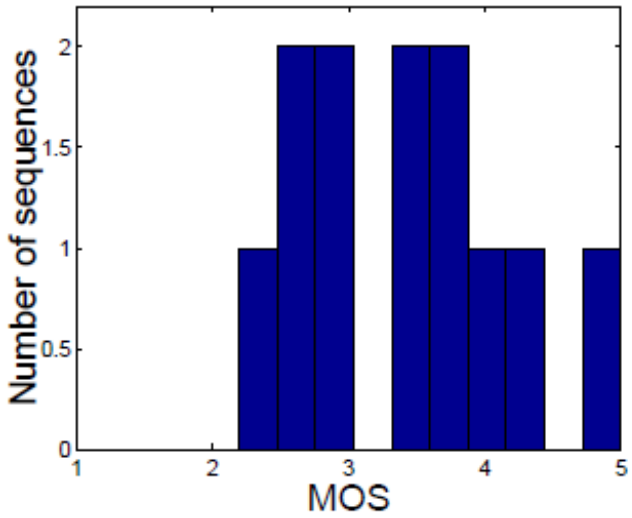


Figure 4.4 Histograms of video quality ratings

4.3 Saliency-based Video Quality Modeling

In this section, we first present two categories of saliency-based error factors for measuring the perceptual video distortion.

4.3.1 Quality Assessment Using Saliency Weighted Pixel Errors

Intuitively, errors at pixels that belong to saliency/FOA regions are more visually important than that in non-saliency/FOA area. Traditional error measurements treat the error of every pixel equally, and some of them from non-salient area are considered in the quality evaluation. We propose saliency weighted quality measurement which gives the errors at pixels that belong to saliency regions higher weighting. The proposed framework for saliency weighted objective quality assessment is illustrated in Figure 4.5.

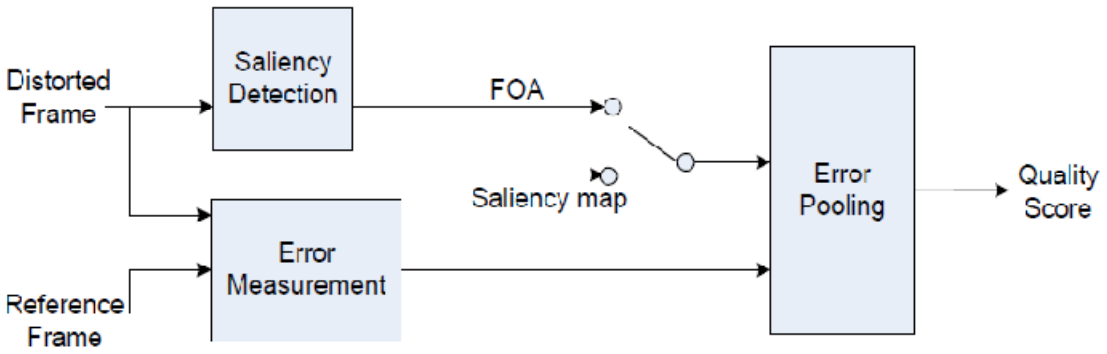


Figure 4.5 Using saliency weight pixel-wise error for objective quality measurement

First, Itti's SVAM is applied to the distorted frame extracted from video sequence affected by packet loss. The gray-scale saliency maps, and the multiple binary FOA maps are alternatively considered as output. Simultaneously the Error Measurement block compares the distorted frames with the original frames, and computes a pixel-wise error map for each frame (measured by square, absolute or structure similarity difference). The Error Pooling block assigns weights to the errors in different pixels based on the saliency or FOA map, and averages the weighted error map spatially and temporally to obtain a final distortion score. Because generating FOA maps needs extra time for each frame, we only consider using saliency map for the weighting.

Specifically, conventional error measurements MSE, MAD, and Structure Similarity Index (SSIM [56]) are calculated pixel-wise. Here, because SSIM measures similarity instead of difference, and its range is from 0 to 1, to be consistent with the MSE and MAD based measures, we use (1 − SSIM) (denoted by $\overline{SSIM}$) in the proposed saliency based SSIM measure.

Let $I_1(x, y, t)$, and $I_2(x, y, t)$ denote the original frame pixel, and $S_1(x, y, t)$, and $S_2(x, y, t)$ denote the saliency values at position (x, y) and time t for the reference and

distorted video, respectively. The saliency weighted pixel error metrics for video sequences can be defined as:

$$SWMSE = \mathbf{E}_{(t)}\left\{\mathbf{E}_{(x,y)}\{S_2(x,y,t)[I_1(x,y,t) - I_2(x,y,t)]^2\}\right\} \quad (4.1)$$

$$SWMAD = \mathbf{E}_{(t)}\left\{\mathbf{E}_{(x,y)}\{S_2(x,y,t)|I_1(x,y,t) - I_2(x,y,t)|\}\right\} \quad (4.2)$$

$$SWSSIM = \mathbf{E}_{(t)}\left\{\mathbf{E}_{(x,y)}\{S_2(x,y,t)\overline{SSIM}[I_1(x,y,t) - I_2(x,y,t)]\}\right\} \quad (4.3)$$

where $\mathbf{E}_{(x,y)}\{\cdot\}$ is the 2-D mean operator averaging over all pixels in frame *t*, and $\mathbf{E}_t\{\cdot\}$ is the mean operator averaging over time in the whole sequence.

Figure 4.6 (a) and (b) show the saliency maps of the two distorted images shown in Figure 4.6 (c) and (d), we can see the detected saliency areas correspond to most of the regions we would pay attention to. The absolute error images between original and distorted frames for the two frames are shown in Figure 4.6 (c) and (d), and the error maps weighted by the saliency maps are shown in Figure 4.6 (e) and (f). We can see that for "optis", saliency weighted pixel error gives more weights to salient errors on the white sail, but for "whale", it conceals most of the non-salient errors in the background.

We have compared the saliency maps with and without motion in Chapter 4.1, taking example of the three sample frames ("aircraft", "optis", "leaf") in Figure 4.1 (b), (c), (e) and (f) in each subimage, and found that motion is an important feature for saliency detection on packet-loss impaired videos. Hence, motion feature is included and assigned a higher weight (as described in Chapter 4.1) in saliency computation for the saliency-based error factors for video sequences. Original absolute difference images, with and without saliency weighting of the three frames, are shown in Figure 4.1 ((g) and (h)) for each example clip. We can see from the three examples, most of the visually

annoying errors are located in the detected saliency regions, especially, for sequence "aircraft", motion feature helps to capture the noticeable artifacts in the background.

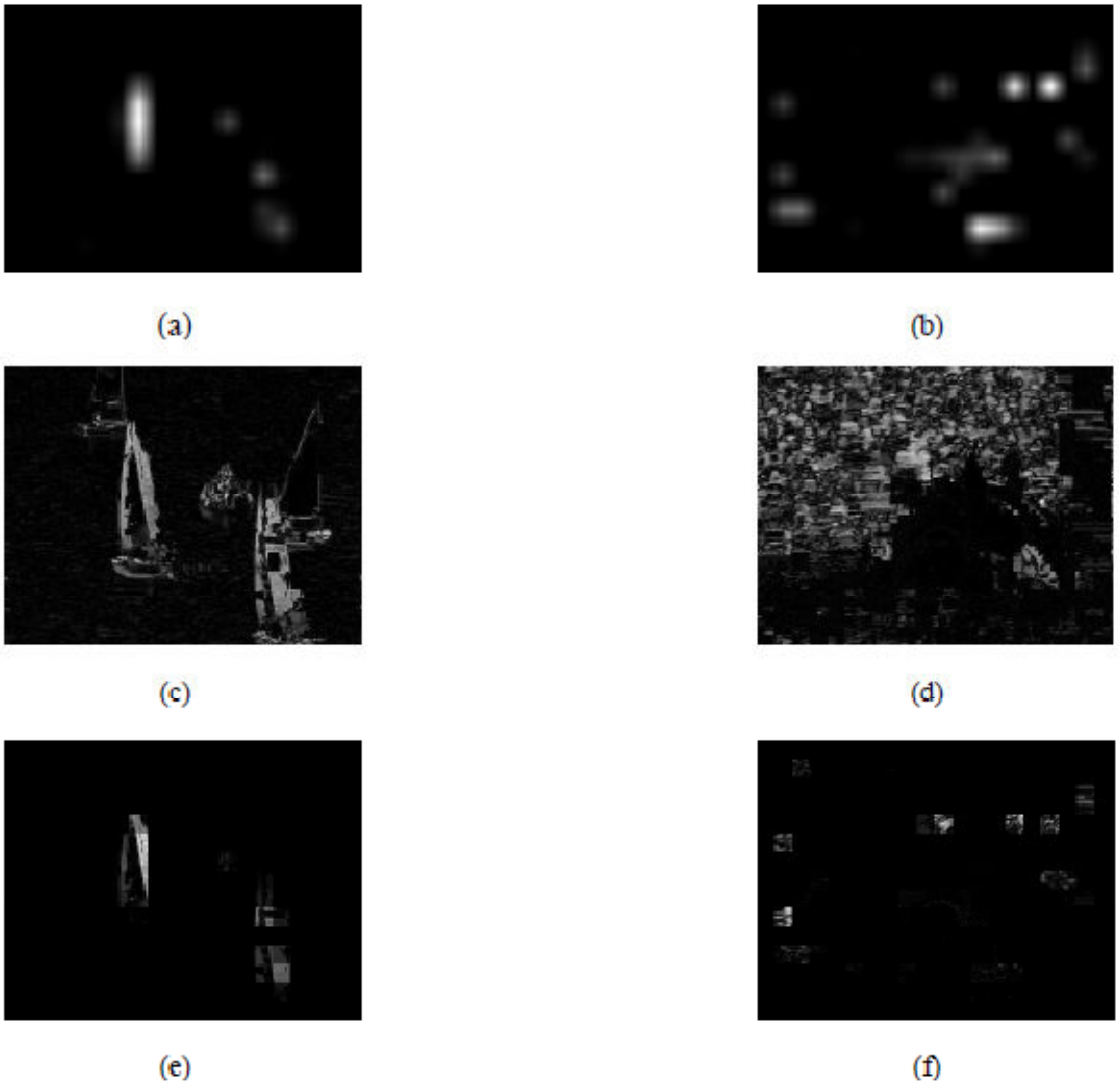


Figure 4.6 Demonstration of proposed methods taking the example of "optis" and "whale": (a)-(b) Saliency maps of the distorted frames. (c)-(d) Original absolute difference images. (e)-(f) Weighted absolute difference images using saliency maps.

4.3.2 Quality Assessment Based on Saliency Variations

So far, the presented saliency based quality assessment method follows the general idea of weighting the pixel error based on their perceptual importance. However,

besides the deviation in pixel-domain, we have observed that packet losses can bring significant changes to the distributions of saliency maps of distorted video frames, both spatially and temporally. We believe that this may be caused by the fact that human eyes tend to be attracted to packet loss artifacts that are not present in original videos. The more changes in these maps often signal the presence of more visible artifacts.

A. Saliency Deviation from Reference Video

In order to investigate the impact of saliency deviation on video quality, we select the videos with either very low or very high subjective quality ratings, and closely examine their saliency maps and FOA scanpaths from both reference and packet-loss-impaired versions. We observe that the packet loss causes noticeable changes in their saliency maps for “bus”, “whale” and “optis” in Figure 4.7, whose quality ratings are relatively low. On the other hand, the saliency of “liberty” in Figure 4.7, is less affected by packet loss, and it receives a high rating. Similarly, the relationship between video quality and changes in video FOA follow the same trend (see Figure 4.8).

Based on the above observations, we hypothesize that a method that measures the difference between the saliency or FOA maps of the original video and distorted video may be able to predict the perceived quality well. Hence instead of quantifying pixel errors in the distorted video frames, we can evaluate the changes in the saliency or FOA maps. In order to reduce computation, we only focus on exploiting the deviation of saliency map in this work.

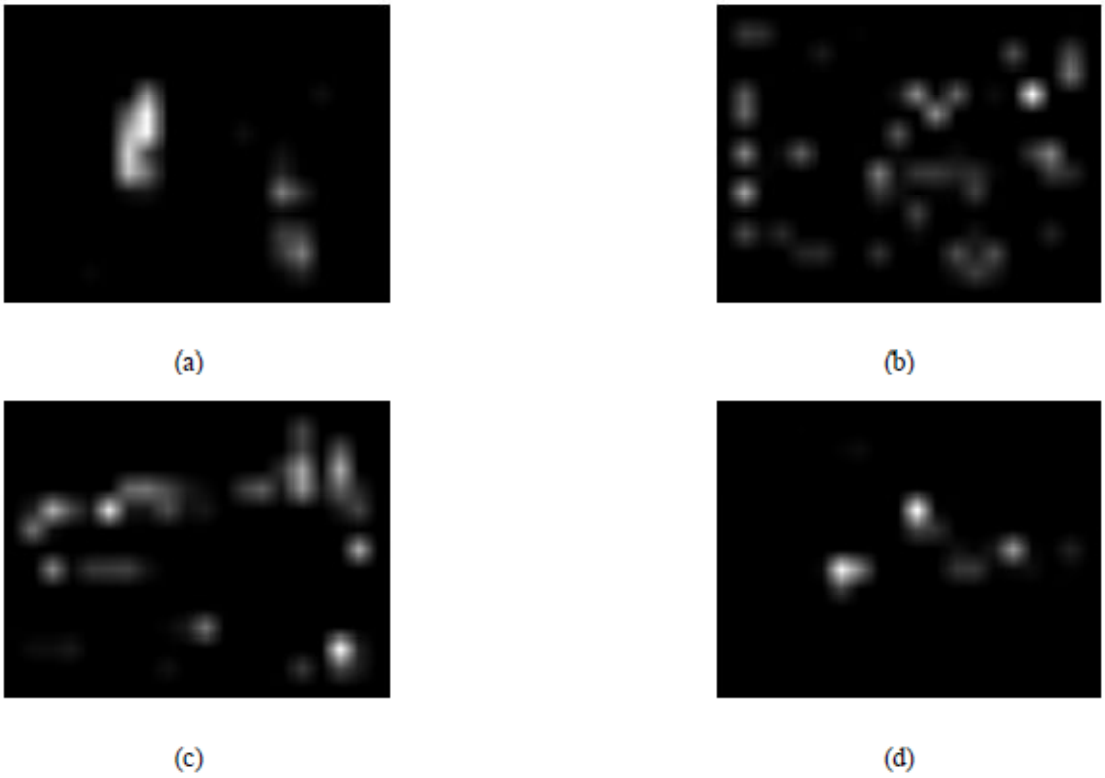


Figure 4.7 The absolute difference map between the saliency maps of original frame and distorted frame for sample images: (a) “optis” (MOS = 2.53). (b) “whale” (MOS = 2.73). (c) “bus” (MOS = 2.8). (d) “liberty” (MOS = 3.93).

Figure 4.9 shows the diagram of the proposed saliency-deviation-based quality error factor. First, saliency maps of reference and distorted frames are extracted separately through Saliency Detection processing. Then the differences between these two maps are calculated through the Error Measurement block, which measures the effect of packet loss on the visual attention change for individual frames.

The saliency deviation based factor for video is defined by measuring the difference between saliency map of each original video frame and that of the corresponding distorted frame, then averaging over time. As before, three error measurements are used to measure the saliency difference — the absolute, squared and dissimilarity based on SSIM measure, yielding:

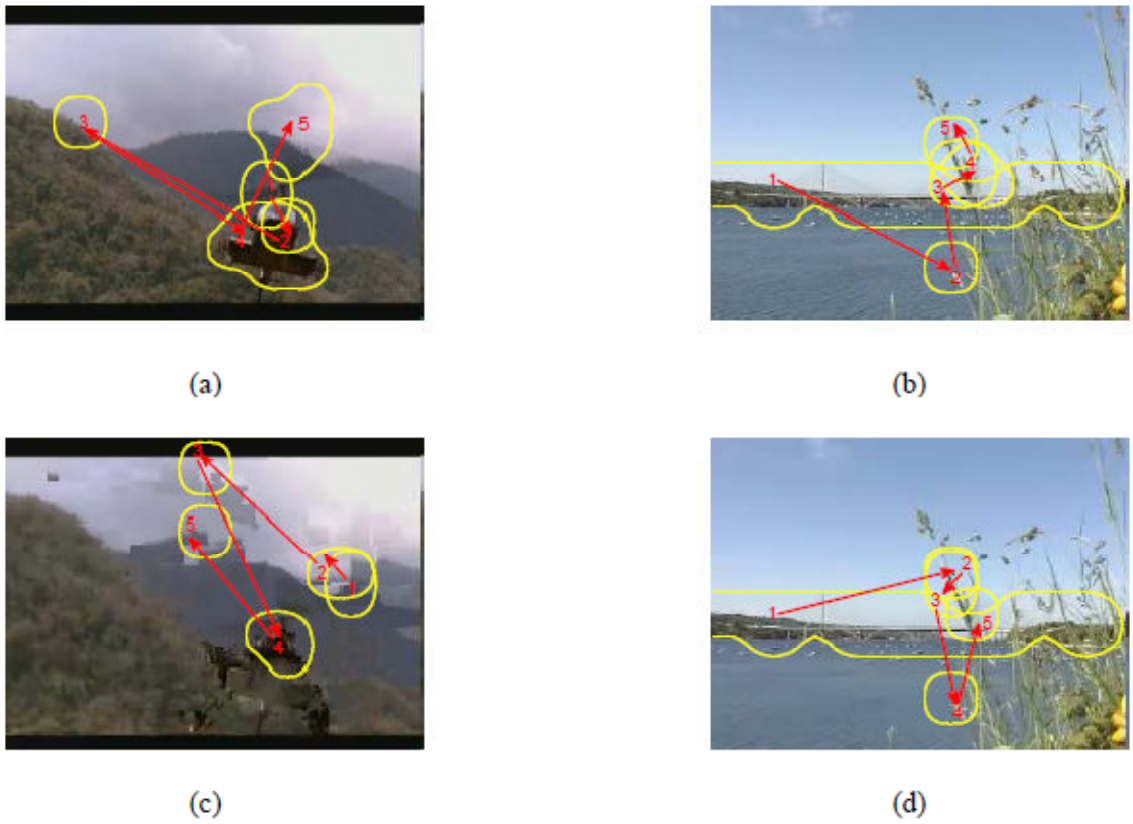


Figure 4.8 Reference frames from the sequence “aircraft” (a, the 11th frame is shown) and “leaf” (b, the 12th frame is shown); Distorted frames from the sequence “aircraft” (c, MOS=2.45), and “leaf” (d, MOS=5), the boundary in each shows the detected first 5 FOAs using the saliency model including motion feature, the arrow gives the attention scanpaths.

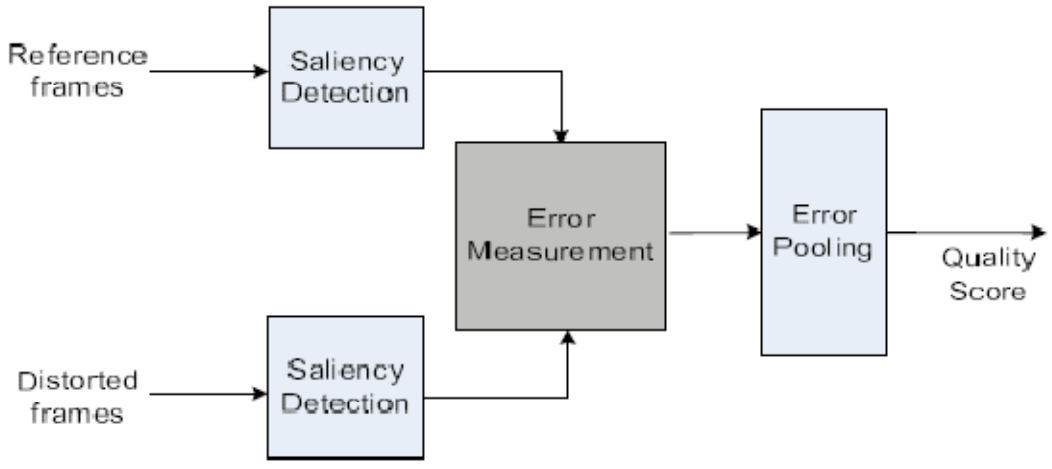


Figure 4.9 Using saliency spatial variation for objective quality assessment

$$SMSE = \mathbf{E}_{(t)}\left\{\mathbf{E}_{(x,y)}\{[\mathrm{S}_1(\mathrm{x},\mathrm{y},\mathrm{t}) - \mathrm{S}_2(\mathrm{x},\mathrm{y},\mathrm{t})]^2\}\right\} \quad (4.4)$$

$$SMAD = \mathbf{E}_{(t)}\left\{\mathbf{E}_{(x,y)}\{|\mathrm{S}_1(\mathrm{x},\mathrm{y},\mathrm{t}) - \mathrm{S}_2(\mathrm{x},\mathrm{y},\mathrm{t})|\}\right\} \quad (4.5)$$

$$SSSIM = \mathbf{E}_{(t)}\left\{\mathbf{E}_{(x,y)}\{\overline{\mathrm{SSIM}}[\mathrm{S}_1(\mathrm{x},\mathrm{y},\mathrm{t}), \mathrm{S}_2(\mathrm{x},\mathrm{y},\mathrm{t})]\}\right\} \quad (4.6)$$

B. Saliency Temporal Variation of the Distorted Video

In addition to deviation in individual saliency maps in each frame of the distorted video compared to the reference video, we also observe that distorted videos that are rated low quality tend to have rapid temporal changes in the saliency maps, this may be because packet losses and subsequent error propagation introduces randomly moving artifacts. As indicated in [73], temporal changes are stronger predictors of human saccade than static feature in video visual processing. Motivated by this finding, we explore the impact of the temporal variation of the saliency map on the perceived video quality. We first define *Saliency Mean* in frame *t* as:

$$SM_i(t) = \mathbf{E}_{(x,y)}[\mathrm{S}_\mathrm{i}(\mathrm{x},\mathrm{y},\mathrm{t})] \quad (4.7)$$

where *i*= 1 or 2 referring to the reference and distorted videos respectively. Figure 4.10 (a,b,c) give $SM_i(t)$ for three sequences:"aircraft", "leaf", "optis". We see that for "aircraft", and "optis'", the two videos that are given low MOS, both $SM_1(t)$ and $SM_2(t)$ changed rapidly in time, but for "leaf", which is given higher rating, its $SM_1(t)$ and $SM_2(t)$ are very smooth. This comparison encourages us to further explore whether temporal attention changes has certain relation with quality evaluation. To quantify this relation, we compute standard deviation of *SMi* for each test sequence to measure the *saliency temporal variation (STV)*

$$STV_i = \mathbf{STD}_t\{\mathrm{SM_i(t)}\} \tag{4.8}$$

where $\mathbf{STD}_t\{\cdot\}$ is the standard deviation operator over time for the whole sequence, and i = 1 or 2 denotes the reference and distorted video respectively.

We plot the relation between STV_2 and MOS in Figure 4.10 (d). From the figure, we found that there is a rough linear relation between *STV* of distorted sequences and their MOS. However, although there is a high correlation between STV_2 and MOS, we can not use the temporal variation of the saliency mean of a video to predict its quality. This is because the temporal variation of the saliency map of the reference video (i.e. STV_1) is often as high as that of the distorted video (i.e. STV_2), which can be seen from Figure 4.11. Specifically, for the sequences that have higher quality ratings including "leaf", "liberty" (MOS = 4.36), "ship"(MOS = 4), "bedroom" (MOS = 3.90), STV_1 and STV_2 are equally small; On the other hand, for sequences with very low quality ratings, including "aircraft", "mobile" (MOS = 2.63),

STV_1 and STV_2 are equally large. Therefore, if we just use *STV* of a sequence to rate its quality, we would rate the original version of the "aircraft" as bad as the distorted one. To summarize, *STV* can be interpreted as one of the characteristics of video, which measures its temporal activity, but it is not sensitive to spatial artifacts. Therefore, STV_2 alone is not sufficient for quality evaluation.

C. Combinations of Spatial Metrics and Temporal Saliency Variation

As discussed above, the saliency temporal variation of the distorted video alone cannot be used to predict the quality of a video. However, if both the *STV* of the distorted sequence and the spatial error between this sequence and the reference video are high,

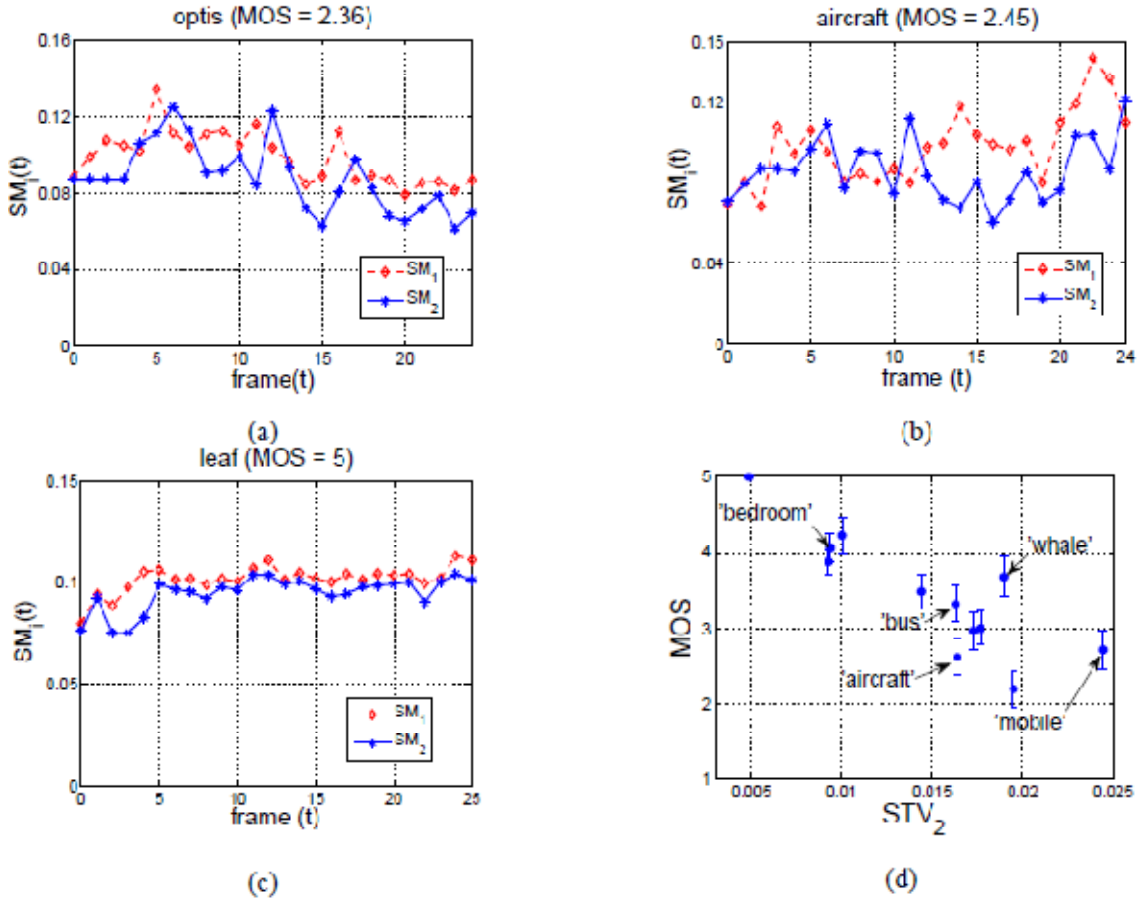


Figure 4.10 $SM_1(t)$ (dash line) and $SM_2(t)$ (solid line) of (a) "optis"; (b) "aircraft"; (c) "leaf"; (d) STV_2 vs. MOS. The vertical bar indicates the 95% confidence interval.

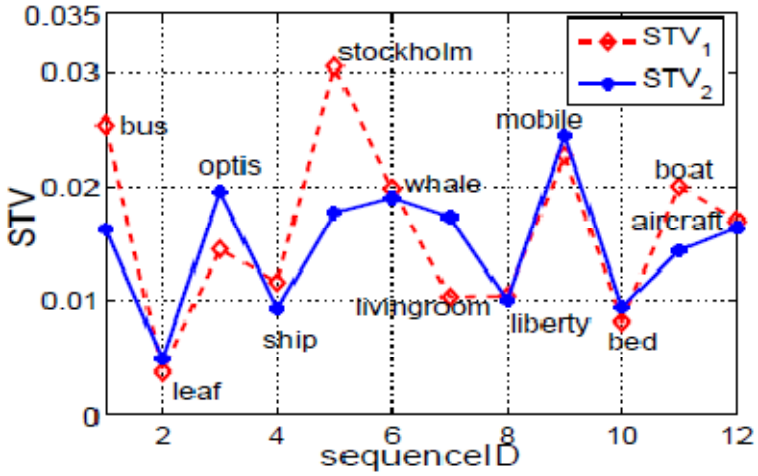


Figure 4.11 STV_1 (dash line) vs. STV_2 (solid line)

this sequence typically has low quality rating. Motivated by this observation, we examine the effectiveness of the product of the *STV* (used to represent STV_2) with each of the previously defined spatial errors as additional saliency-based factors. Specifically, the following factors are considered: *MSE*STV, SWMSE*STV, SMSE*STV, MAD*STV, SWMAD*STV, SMAD*STV, SSIM*STV, SWSSIM*STV,* and *SSSIM*STV*.

4.3.3 Performance Comparison of Different Error Factors

In order to gain a clear picture of the performance of the various saliency-based error factors, we draw the scatter plots of all the factors vs. MOS in Figure 4.12. The three types of error measures (MSE, MAD, SSIM) are shown in three columns separately. The factors are all normalized to the range of [0,1], based on the actual minimum and maximum values of each feature among all test videos.

Overall, these results indicate that both saliency weighted pixel errors and deviation in saliency maps are useful for quality evaluation of packet loss affected videos. Saliency weighted method may work better when packet loss does not cause visible errors in non-salient regions; whereas saliency deviation based factor is more effective when pronounced errors appearing in originally inconspicuous areas. Note that, in terms of complexity, saliency deviation based measurement needs to compute saliency maps for both the original and distorted images (1/16 size of the original image), hence saliency weighed method is more feasible in some special applications with computation or power limitation.

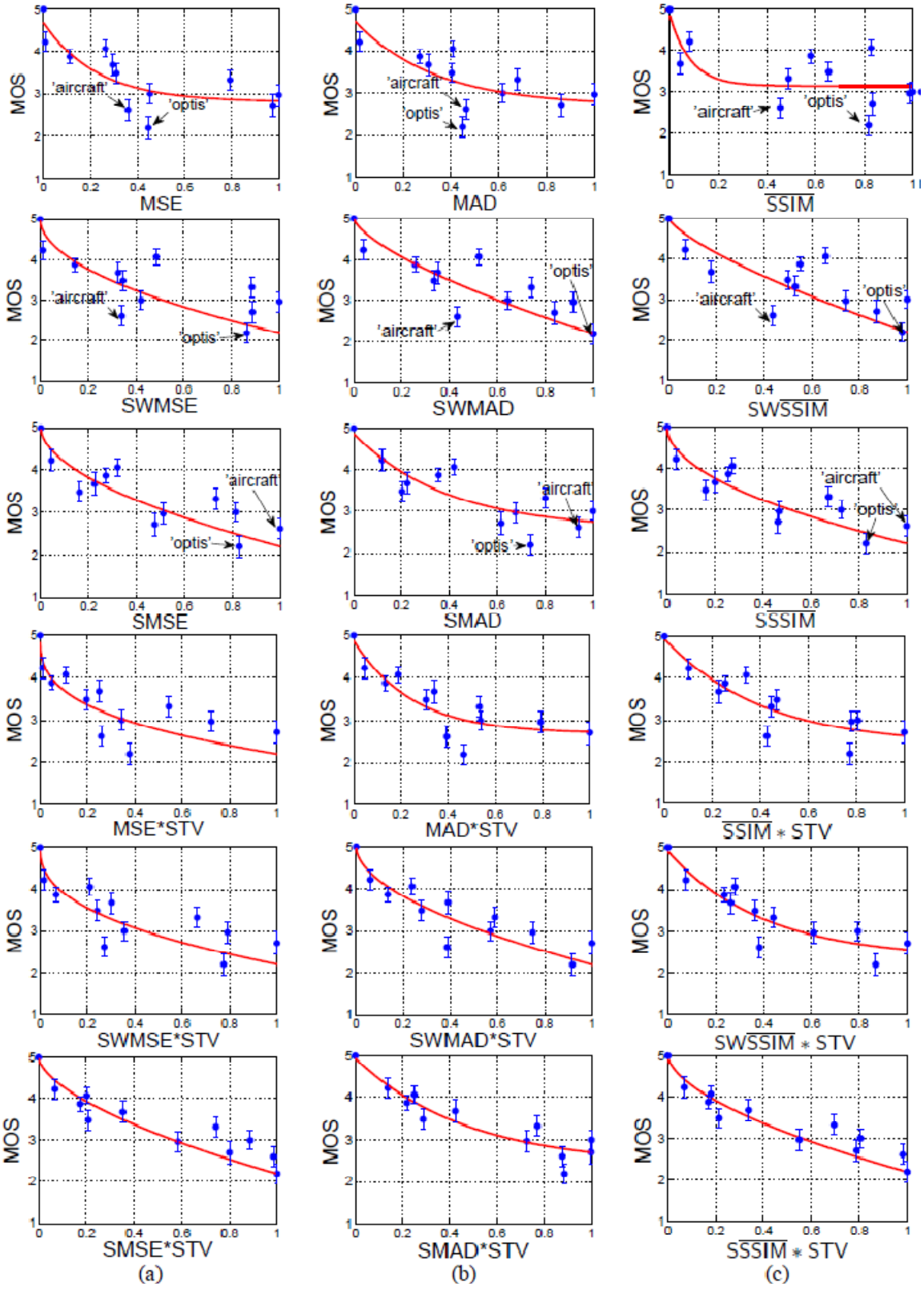


Figure 4.12 Scatter plots with the best mapping curve of proposed error factors for video sequences. Column (a): MSE based factors; Column (b): MAD related factors;

Column (c): SSIM related factors. In each plot, points are measured MOS, the curveis the predicted MOS using the best mapping function for that factor. The vertical bar indicates the 95% confidence interval.

The products of the saliency temporal variation and all spatial metrics, including conventional non-saliency factors, bring forth the benefits of both measures. This improvement can be seen obviously by comparing the last three rows with the previous three in Figure 4.12. Especially, the combination of Saliency-Deviation error with *STV* correlates with MOS pretty well. Therefore, we can conclude that with the same spatial saliency map difference, a high *STV* is typically associated with a poor perceptual quality; while with the similar *STV*, a larger change in the saliency map tends to indicate a lower quality.

To quantify the performance of each quality predictor, we apply nonlinear mapping on all of them so that the mapped factors can have as much linear relationship with the MOS. After examining, the exponential and power functions are used as candidate mapping functions for each factor, and the best mapping form of each single factor is then analyzed separately. To reduce the number of parameters, for the power form, we use $b2 = MOSMAX - MOSMIN = 2.806$, and $c2 = MOSMIN = 2.194$ for video dataset; while for exponential form, all the three coefficients are determined using the least square fitting. Finally, for each factor, the mapping form that gives the minimum Prediction Error (in terms of MSE) (PE) in the "leave-one-out" cross validation process is determined as the best mapping form.

The best mapping curve for each factor is shown in its scatter plot in Figure 4.12, and the actual mapping forms of all factors are indicated in Figure 4.13. The bar plot of PE of all the factors mapped with the best form, in the ascending order, is shown in

Figure 4.13 (the black bar using the left axis). Also shown on this figure is the PC of all mapped factors using the gray bar and right axis. We can see clearly that the mapped factors with lower PE typically also have higher PC. Importantly, both saliency-weighted and saliency-variation based factors perform better than conventional non-saliency factors. In particular, the combinations of the saliency spatial errors and the saliency temporal variations are quite useful, with most of them getting much smaller PE than other factors.

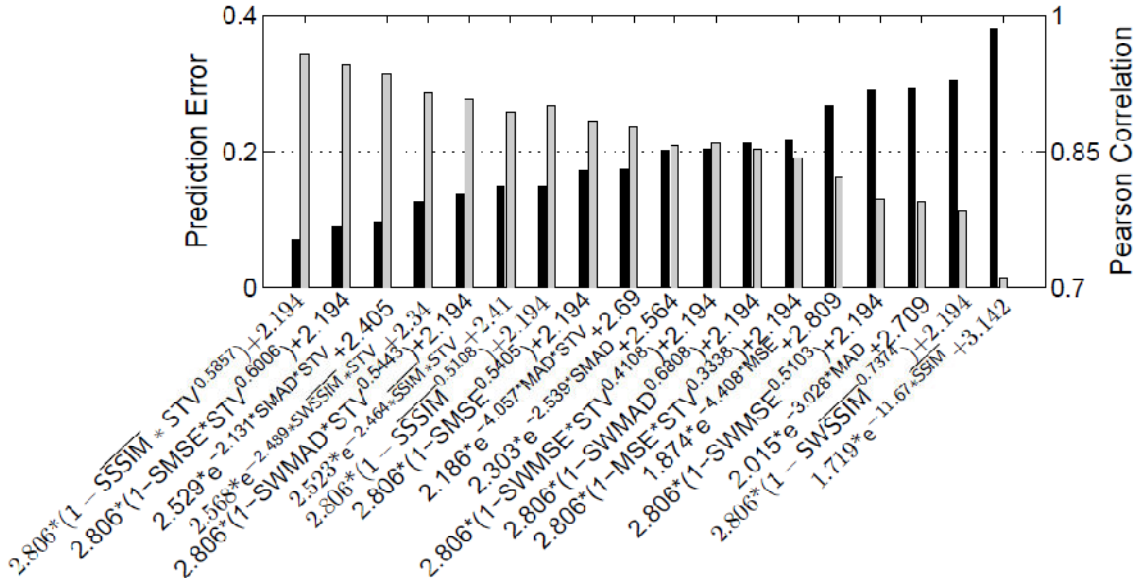


Figure 4.13 Prediction performance of different factors (mapped with the best form) for video sequences, in the ascending order of Prediction Error (MSE), in the black bar with the left axis. Note that 2.194 = MOSMIN, 2.806 = MOSMAX-MOSMIN. The gray bar using the right axis of Pearson correlation

4.3.4 Proposed Metric Combining Multiple Factors

In order to further improve the prediction accuracy, we try to linearly combine multiple mapped factors in the final proposed metric for video quality assessment. Here, the critical problem is how to determine which factors to include. Because some of the

factors are correlated, to select factors into the final metric, the stepwise multiple linear regression approach is applied. The linear model (LM) using M factors for predicting y is described by

$$\boldsymbol{Y} = \gamma + \sum_{m=1}^{M} X_m \beta_m \tag{4.9}$$

where $\boldsymbol{Y}$ denotes the N-dimensional vector containing the MOS values for N test videos, X_m represents the N-dimensional vector containing the values of factor m computed for the N test video, γ and β_m and are the model parameters. The factors to be included and the model parameters are determined using the following two steps:

Step 1: In this step, we try to add one factor at a time, starting from a null set (i.e. $M = 0$). Suppose we already have M factors, we try to add each of the remaining factors, evaluating the prediction error with leave-one-out cross-validation. The factor that leads to the maximum reduction in the PE is then chosen. This process is repeated until we can no longer reduce the prediction error by adding a single factor, then the model is preliminarily established.

Step 2: We check if there are any interactions (products) between any two factors chosen in Step 1 that can further reduce the prediction error. If there are multiple candidate pairs, the decisions of their inclusion are made in the similar fashion as step 1.

Following this procedure, we find that when choosing from all 18 factors, four factors are included in the final linear model, including the mapped forms of *SSSIM∗ STV, SMSE ∗ STV, SSIM∗ STV, MAD*.

We call this final metric the Saliency-Based Video Quality Metric (SVQM). Using the same procedure, we also derive a metric that only chooses among 3 non-

saliency-based factors, i.e. mapped forms of MSE, MAD, and SSIM. The final model is called Non-Saliency-Based Video Quality Metric (NSVQM), and includes only the factors MSE and SSIM.

To compare the prediction performance of NSVQM and SVQM, We draw the bar plots of the selected factors vs. model prediction error during the stepwise regression procedure, in the order of their inclusions, in Figure 4.14.

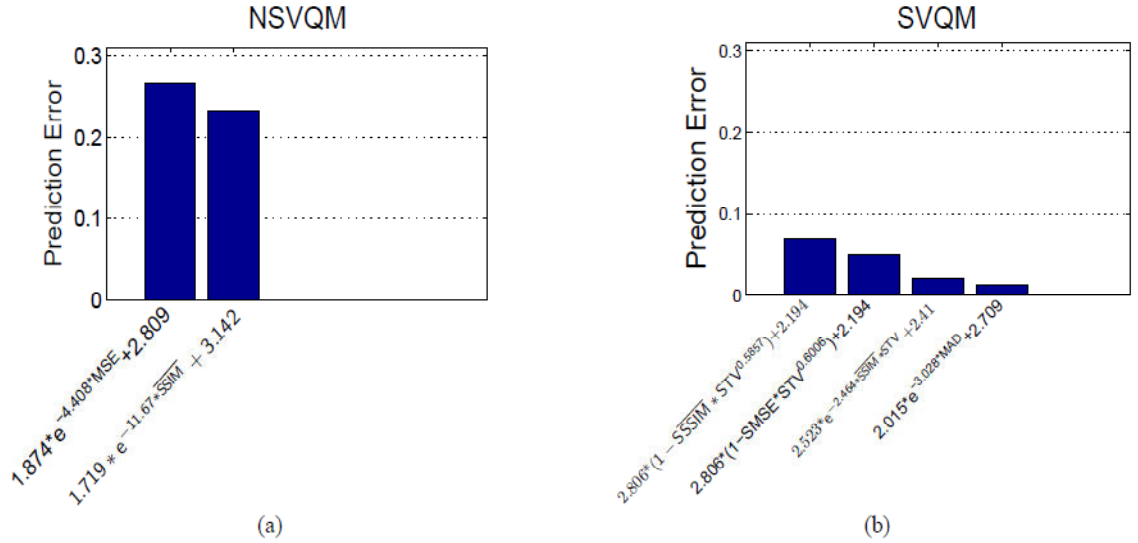


Figure 4.14 The factor inclusion order (from left to right) and the corresponding average PE. (a) NSVQM; (b) SVQM. Note that 2.194 = MOSMIN, 2.806 = MOSMAX-MOSMIN.

Table 4.2 summarizes the final best metrics found for NSVQM and SVQM and compares their prediction accuracy.

Table 4.2 Comparison of video quality metrics with and without using saliency

Metrics	Predictor form	PE	PC
NSQVM	$2.842 + 1.453 * e^{(-4.408*\mathrm{MSE})} + 0.6355 * e^{-11.67*\overline{\mathrm{SSIM}}}$	0.2331	0.8411
SQVM	$2.181 + 11.23 * (1 - \mathrm{S}\overline{\mathrm{SSIM}} * \mathrm{STV}^{0.5857}) - 9.364 * (1 - \mathrm{SMSE} * \mathrm{STV}^{0.6006})$ $+1.486 * e^{-2.464*\overline{\mathrm{SSIM}}*\mathrm{STV}} - 0.5606 * e^{-3.028*\mathrm{MAD}}$	0.0123	0.9947

We can see that NSVQM uses two non-saliency-based factors and the prediction error is 0.2331, while SVQM uses three saliency based factors and one non-saliency based factor and has a much reduced prediction error of 0.0123. Also in terms of correlation performance, SVQM obtains a very high PC (0.9947), which is significantly better than NSVQM (0.8411). Figure 4.15 shows the scatter plots of NSVQM and SVQM vs. MOS respectively. We can see clearly that SVQM correlates with MOS very well, while NSVQM has many outliers. The large improvement offered by SVQM proved that considering visual saliency can provide substantial gain in assessing the perceptual quality of video with packet-loss impairments.

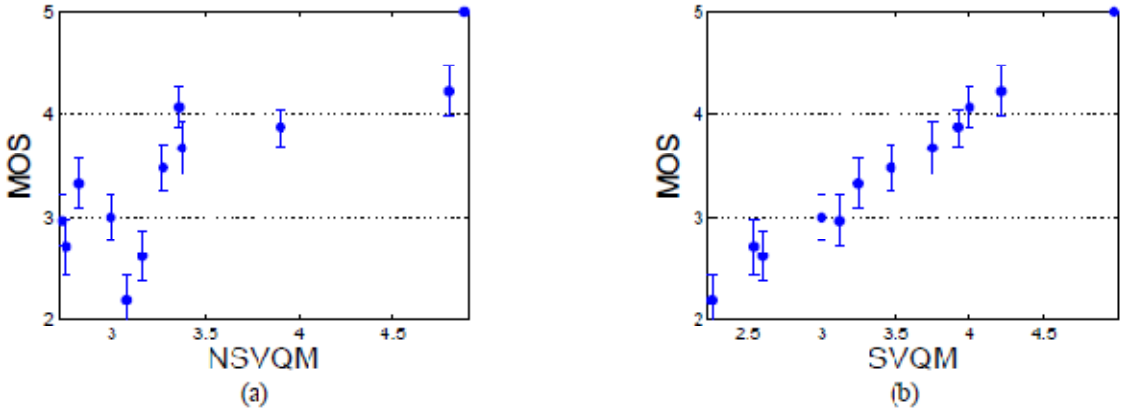


Figure 4.15 Scatter plots of (a) NSVQM vs. MOS; (b) SVQM vs. MOS; The vertical bar indicates the 95% confidence interval.

4.4 Summary

In this chapter, we investigate the role of saliency in video quality assessment. We first calculate visual saliency by extending a widely accepted image saliency detection algorithm with motion information. Then present three saliency-aided objective video quality assessment methods. Based on the assumption that human eyes are more sensitive

to the errors occurring in saliency/FOA regions, we propose the first scheme by giving the errors at pixels in saliency/FOA regions more weights when pooling a distortion score. We observe that humans' saliency can be significantly changed by packet losses, and this change is closely related to the perceptual quality of underlying video, so we develop our second and third saliency-based quality assessment schemes by calculating the fidelity between the saliency maps of original and packet loss affected videos and temporal variation of the saliency of packet loss affected video, respectively.

We show that all the saliency-based metrics can achieve greatly better quality prediction accuracy than conventional non-saliency metrics, and the improvement can be further increased by combining different saliency-based metrics

We should point out that although the performance improvement of saliency-aided over non-saliency metrics is very significant for our test data which only contains 12 test sequences, the performance gain may not be as significant for larger subjective dataset containing more variety of packet loss induced distortion patterns.

One way to improve the proposed quality metric described in Chapter 2 may be by replacing the PSNR drop sum of packet loss affected video segment with the proposed saliency-aided metrics in this chapter. This is one of our future research directions.

Chapter 5

Saliency Inspired Modeling of Visibility of Packet Loss

There is an interesting phenomenon that different packet losses may not cause equal video quality degradation, and in fact some may not be visible to human viewers. For example, a packet loss within a fast moving scene, e.g. a scene with large global motion caused by camera panning, is very obvious, whereas a packet loss occurring in a scene with little motion, e.g. a person is talking on the phone, is almost not visible to average viewers. In fact, the visibility of packet loss of decoded video is investigated by some prior work [12] [48], and it depends on various factors and their complicated interactions, such as loss severity and duration, characteristics of background signal, and its distance to scene change. However, the role of visual attention or saliency in predicting the visibility of packet loss is not studied in the existing literature. In this chapter, we investigate how to improve visibility prediction by incorporating the saliency information. Based on the findings in Chapter 4 about how saliency affects the perceptual quality of video with packet losses, we propose several saliency-based factors and

incorporate them into a Generalized Linear Model (GLM) to predict loss visibility. Test results with 1080 MPEG-2 packet losses indicate that saliency information can help improve the prediction accuracy about 12% over nonsaliency based model, and that saliency-weighted mean-square error and variation of saliency information are promising metrics.

5.1 Subjective Test on Visibility of Packet Loss

The subjective data used in this work was from the work presented in [12]; to be self-contained, we describe it briefly here. This subjective test was designed not to assess the quality of video at a given packet loss rate, but instead to learn about what affects the visibility of impairments caused by individual packet losses. The test videos shown were compressed with MPEG-2 at 720x480 resolution and 30 fps frame rate, with various scene contents and different camera motions, using 13-frame GOPs with 2 B-frames before every P-frame at a bitrate of around 4Mbps.

One isolated packet loss was randomly inserted into the video in every 4-second window. The 1080 packet losses affect either one slice, two slices, or an entire frame. The decoder applied Zero-motion error concealment (copying macroblocks from the closest reference frame) when losses occurred.

Each test video clip is 6-minutes long and watched by 12 viewers, whose task were to indicate each time they saw an artifact by hitting the space bar on the computer. The “ground truth” of “visibility” of each packet loss was defined as the percentage of

viewers who indicated they saw the loss. For the detailed information about this subjective test, please refer to [12].

5.2 Objective Assessment on Visibility of Packet Loss

In this section, we investigate how to improve visibility prediction by incorporating the saliency information. Based on earlier proposed several saliency-based factors, we incorporate them with other prior non-saliency factors into a Generalized Linear Model (GLM) to predict loss visibility.

5.2.1 Non-saliency factors affecting visibility

In [12] [48], there were totally 20 non-saliency factors (and their variations) proposed, which covers the characteristics of both videos contents and packet loss impairments. Furthermore, the interaction between them on a scene-level was considered as well. Here we only briefly discuss them in each category, and please refer to [12] [48] for more detailed descriptions.

A. Error characteristics

Mean Squared Error (MSE) and Structural Similarity Index Metric (SSIM) are two widely accepted quality metrics. In the loss-visibility scenario, for the sake of easy calculation, two simplified variations of each metric are used to predict the visibility of packet loss: those measurements in the initial frame in the loss-affected segment, **IMSE**

and **ISSIM**; and the extreme values of IMSE and ISSIM at macroblock level in the initial frame, **MaxIMSEmb** and **MinISSIMmb.**

When there is a single slice loss (as opposed to the loss of an entire frame), the impact of discontinuities caused by lost slices on video quality can be measured by the Slice-Boundary Mismatch (SBM), first proposed in [33] and modified in minor details in [48]. Only SBM on the initial frame of loss-affected segment, **ISBM**, is considered.

Additionally, some important content-independent measures, such as spatial extent, or **SXTNT** (the number of slices lost in one frame), **HGT** (the average height of the lost slices), **Duration** (duration of the loss-affected segment), are also considered.

B. Video characteristics

Motion is one of the most important characteristics of videos. Therefore, the mean and variance of the magnitudes of the motion vectors across all macroblocks initially affected by a loss, **MotMean** and **MotVar** can also be used to predict the visibility of a loss.

SigMean, and **SigVar**, the mean and variance of intensity values of the initial frame of loss-affected segment, and **ResidEng** (residual energy after motion compensation) of that frame are also effective.

C. Scene-level characteristics

In addition to the factors in two first categories, some high-level characteristics are also considered. The authors of [48] show that the relative position between scene change and packet loss impairment has great influence on its visibility. Therefore, **D2R**

(the distance between the current frame (with packet loss) and the reference frame used for concealment), **DistFromCut (**the distance in time between the first frame affected by the packet loss and the nearest scene cut, either before or after) and its threshold versions, **AtScene**, **beforeScene**, and **afterScene,** are considered in this work.

Since these non-saliency factors were proved capable to predict the loss visibility, we use them as candidate factors to design a GLM.

5.2.2 Saliency Inspired Modeling of Packet Loss Visibility

A. Saliency-based Factors

Based on the findings in the study discussed above, we propose to supplement the IMSE factor by saliency-weighted IMSE, denoted by IMSE_Sal. We also consider the saliency weighted MSE computed over all loss-affected frames, yielding MSE_Sal. We also use SMSE, which measures the changes (in terms of MSE) between saliency maps of original and loss-impaired frames (only in the position where loss happens); STV, which measures temporal variation of the saliency map of loss-impaired frames, respectively.

Table 5.1 summarizes all the proposed saliency-based factors, as well as non-saliency-based factors, which are used to build the final GLM quality model.

Note that, for saliency computation, we tested two methods, one using color, orientation, and intensity information only, as in the original Itti's model [38] (which results in the saliency-based factor in Table 5.1 denoted as "no motion"); another one further using motion information with the motion features computed following [67]. From our previous study, the second method produces saliency maps that are more

consistent with our visual inspections, although it requires extra computation over the former one in saliency detection. Therefore, we focus on the latter one in this work.

Table 5.1 List of all the non-saliency and saliency-based factors

1	IMSE	14	D2R
2	MaxIMSEmb	15	SigVar
3	ISSIM	16	DistFromCut
4	MinSSIMmb	17	AtScene
5	ISBM	18	BeforeScene
6	SXTNT	19	AfterScene
7	HGT	20	FarConceal
8	Duration	21	IMSE_Sal (no motion)
9	ResidEng	22	IMSE_Sal
10	CameraMotion	23	MSE_Sal
11	SigMean	24	S_MSE
12	MotMean	25	STV
13	MotVar		

B. Generalized Linear Model Fittings

As [12] [48], we model the probability of visibility using a GLM [74], which is a development of linear models to accommodate both non-normal response distributions and transformations to linearity in a straightforward way. It is defined as follows

$$\log\left(\frac{P}{1-P}\right) = \gamma + \sum_{m=1}^{M} X_m \beta_m \tag{5.1}$$

where P denotes the N-dimensional vector containing the visibility for N test packet losses, X_m represents the N-dimensional vector containing the values of factor m computed for the N test packet losses, γ and β_m and are the model parameters. logit() as the link function.

With the help of statistical software [75], the model is fit with an iteratively re-weighted least-square method to generate a maximum-likelihood estimate. The GLM

fitting is performed in a similar fashion as what we discussed in previous section with subjective data described in Chap. 4. We perform 10-fold cross-validation in the process of building up the model. Specifically, we divide the entire data set of 1080 losses into 10 groups of equal size and choose the data from 9 out of the 10 sets as a training set. The remaining data set is used for testing. We repeat this process 10 times, each time choosing a different set for training. The average prediction error is used as the performance measure.

In order to test the impact of saliency information on modeling the loss visibility, we fit the same subjective data with two different models, one containing only non-saliency-based factors (**Model 1**); the other one containing all the aforementioned factors (**Model 2**). We note that our model differs from that in [48] because it uses just one subjective dataset.

The factors and the coefficients of both models are summarized in Table 5.2. To test the significance of each factor in the model, including the interaction terms added in the third step, we re-fit the models stepwise. The new order provides a ranking of the significance of each factor and interaction term. This allows the two models to be compared if we limit each to having the same number of predictive factors. We draw the bar plots of factors and model prediction errors of this stepwise procedure, in the order of their inclusions, in Figure 5.1 (a) and (b).

C. Model Comparison

To compare the prediction performances of Model 1 and Model 2, we show the relationship between the number of factors used in the model and prediction error

reduction ratio (Model 2 to Model 1) in Figure 5.1. We can see that the overall prediction error of Model 2 (0.027449) is about 12% less than that of Model 1(0.03196)! When both

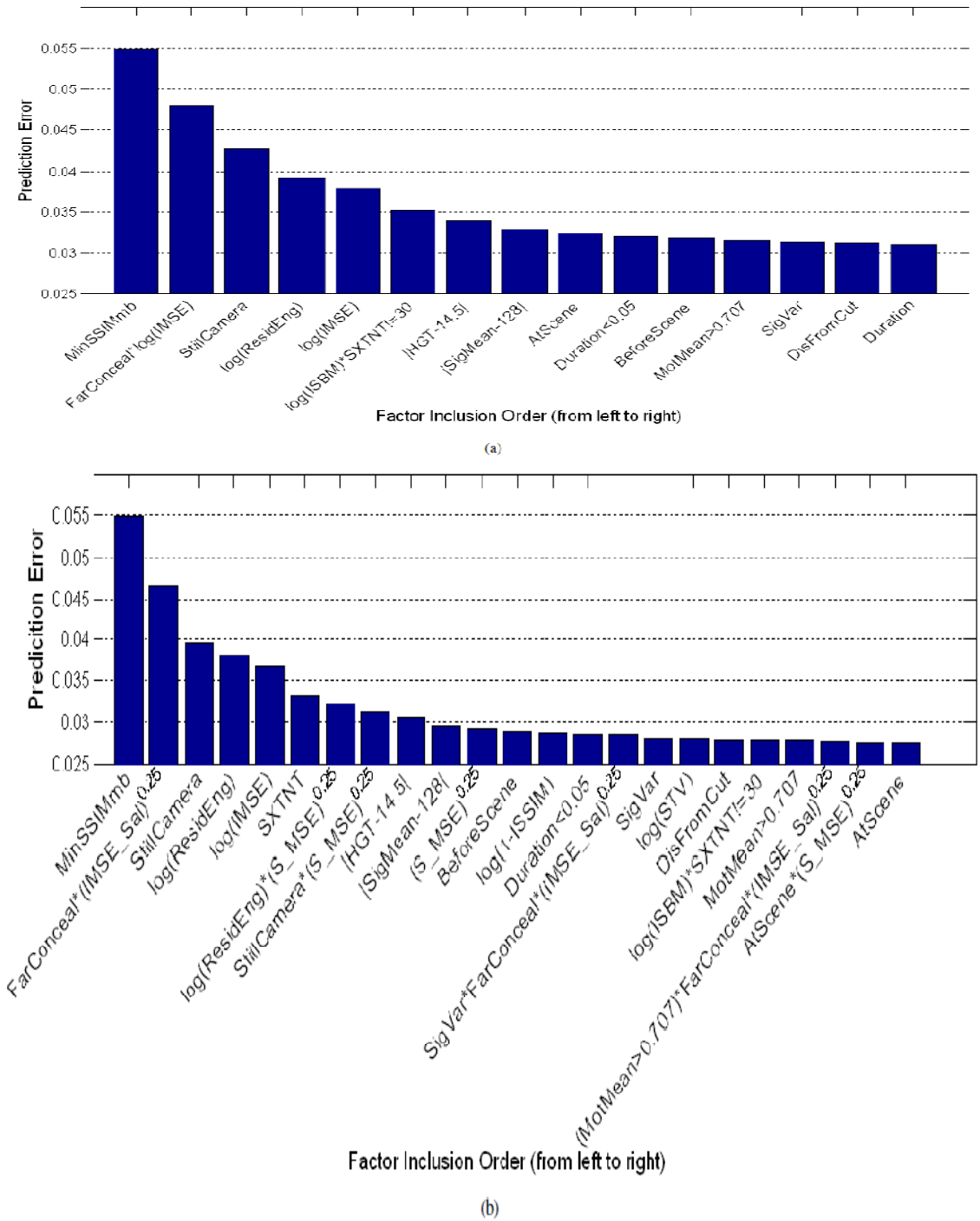


Figure 5.1 Factor inclusions of (a) Model 1; and (b) Model 2.

Table 5.2 Coefficients of Model 1 and Model 2

Model 1		Model 2	
factor	coef.	factor	coef.
MinSSIMmb	-1.141e+00	MinSSIMmb	-9.794e-01
FarConceal* log(IMSE)	1.929e-01	FarConceal* $(IMSE\_Sal)^{\frac{1}{4}}$	7.604e-01
StillCamera	-6.825e-01	StillCamera	-1.229e+00
log(ResEng)	-7.015e-01	log(ResEng)	-1.058e+00
log(IMSE)	7.209e-01	log(IMSE)	7.090e-01
log(ISBM)* (SXTNT!=30)	6.590e-02	SXTNT	-1.296e-01
|HGT-14.5|	-7.248e-02	log(ResEng)* $(S\_MSE)^{\frac{1}{4}}$	3.027e+00
|SigMean-128|	-1.012e-02	StillCamera* $(S\_MSE)^{\frac{1}{4}}$	-2.937e+00
AtScene Duration<0.05	1.404e+00 -3.689e-01	|HGT-14.5| |SigMean-128|	-6.205e-02 -8.870e-03
BeforeScene	-7.798e-01	$(S\_MSE)^{\frac{1}{4}}$	-1.036e+01
MotMean>0.707 SigVar	3.317e-01 -4.433e-04	BeforeScene log(1-ISSIM)	-5.802e-01 2.915e-01
DistFromCut	-1.033e-02	Duration<0.05	-5.966e-01
Duration	1.578e+01	SigVar* FarConceal*$(IMSE\_Sal)^{\frac{1}{4}}$	-8.087e-04
		SigVar	8.905e-04
		log(STV)	1.715e-01
		DistFromCut	-9.092e-03
		log(ISBM)* SXTNT!=30	1.284e-02
		MotMean>0.707	5.519e-01
		(MotMean>0.707)* FarConceal*$(IMSE\_Sal)^{\frac{1}{4}}$	-3.517e-01
		AtScene * $(S\_MSE)^{\frac{1}{4}}$	1.472e+01
		AtScene	-4.448e+00

models are limited to use 15 factors, Model 2 still outperforms Model 1 by about 9%. In addition, we can see that, no matter how many factors are used to fit the models, our model with saliency-based factors always outperform those without, except for the case of only one factor, since that factor (MinSSIMmb) is the same for both models. Therefore, we conclude that saliency information significantly boosts the visibility prediction performance.

To gain a clearer picture of the contribution of each individual saliency-based factor, we examine the inclusion orders (or significance rank) of saliency-based factors in

Model 2 in Figure 5.1 (b). There are 4 saliency factors in the first half of the 23 factors in the model: FarConceal*(IMSE_Sal)$^{0.25}$, log(ResidEng)*S_MSE$^{0.25}$, StillCamera*(S_MSE)$^{0.25}$, and S_MSE$^{0.25}$. If we expand our focus to the first 15 factors, FarConceal*(IMSE_Sal)$^{0.25}$*SigVar is also present. Therefore, saliency-weighted pixel-wise error and the difference of saliency maps of original and distorted video caused by packet loss are two significantly helpful factors in modeling of packet loss visibility.

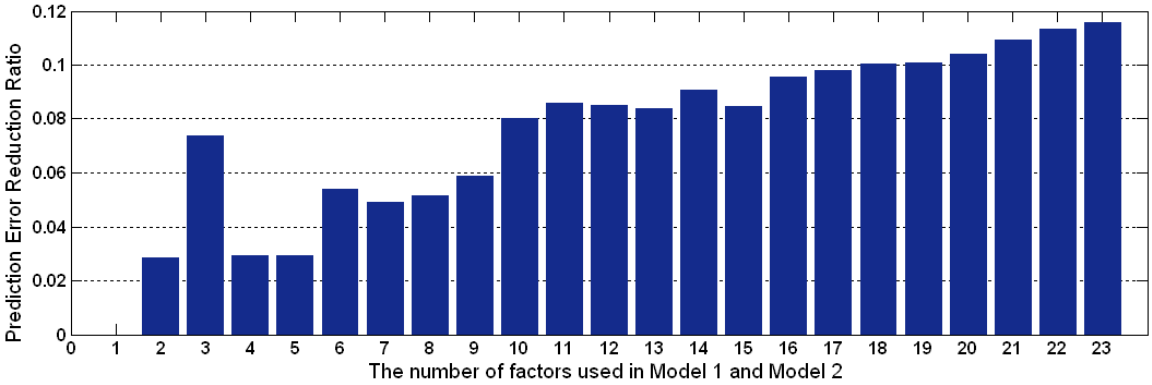


Figure 5.2 Performance comparison between Model 1 and Model 2.

5.3 Summary

In this chapter, based on the saliency-based approaches proposed in Chapter 4, we present a practical solution to improve the prediction of the visibility of packet losses. Together with a variety of non-saliency based factors, we fit the GLMs with and without our proposed factors using existing subjective data, and the results show that saliency-based factors significantly improve the performance in loss-visibility modeling.

Our work described in Chapter 4 about perceptual quality prediction and the current work on modeling visibility of packet loss have shown that saliency information is helpful for both scenarios.

Chapter 6

Conclusions and Possible Future Work

This thesis has conducted various research related to objective quality assessments of transmitted videos. In this chapter, we first summarize our contributions, and then suggest some possible future works.

6.1 Summary of Major Contributions

While perceptual video quality assessment has flourished in recent years, objective quality assessment on packet loss impaired videos is in its early phase. The research discussed and models proposed in this thesis are important supplements to the existing research work in literature. The major contributions of this thesis are listed below.

1. We examine the impact of several attributes (duration, severity, location, pattern, etc.) of packet losses on perceptual video quality, and then propose quality

metrics for videos affected by single loss and multiples loss respectively. Finally, by incorporating a prior model for quality degradation due to coding artifacts, we proposed a combined metric for predicting the overall quality degradation due to both compression and packet losses. This metric provides a high correlation with subjective ratings for a large set of sequences with different video content, coding artifacts and loss patterns, significantly higher than some other widely accepted metrics.

2. We investigate the perceptual quality of individual video frames affected by packet losses and coding artifacts. We first classify the two kinds of artifacts occurring at one video frame simultaneously, and then, by taking into account of different masking effects of the HVS, we propose two block-wise JND profiles for coding and packet loss artifacts, respectively, and then combine them into one metric that can evaluate the perceived quality degradation due to both artifacts. The predicted perceptual distortions by our proposed metric have fairly high correlation with subjective quality ratings.

3. We study the role of saliency in video quality assessment. Based on our observations that human eyes are more sensitive to the errors occurring in saliency/FOA regions, and packet losses may significantly change the distribution of saliency over both space and time, we propose three saliency-aided quality assessment schemes, i.e. saliency-weighted error, saliency-fidelity, and saliency temporal variation schemes. We show that all the saliency-based metrics can achieve greatly better quality prediction accuracy than conventional non-saliency metrics, and the improvement can be further increased by combining different saliency-based metrics

4. Based on the saliency-based approaches proposed for video quality assessment, we present a metric which can predict the visibility of packet losses by updating various non-saliency based factors and combining them with GLM. As a result, saliency-based

factors greatly improve the performance in loss-visibility modeling, which confirms that saliency is helpful to predict visibility of packet losses.

5. For the purpose of objective quality modeling, we design and perform several subjective quality tests for specific video applications. The available subjective data and well-designed test plans, which are already uploaded to the public domain, as well as the software quality grading interface we developed, can be valuable for future studies.

6.2 Possible Future Works

There are a number of possible extensions and applications for the work presented in this thesis. Some suggestions are as follows.

First, we propose a video quality metric, PDMOSCL, which considers both coding and packet loss artifacts in Chapter 2. It can be improved by incorporating the results of the studies described in other chapters. Specifically, the current quality measurement of individual video frames in the metric PDMOSCL is based on PSNR drop which can be replaced by the more advanced quality metric proposed in Chapter 3. And the saliency-aided video quality assessment schemes proposed in Chapter 4 can be applied to PDMOSCL to improve its performance by converting PSNR drop to saliency weighted PSNR drop, or saliency variation based metrics. In addition, PDMOSCL can benefit from the general idea of combining multiple quality factors in Chapter 4 and 5 so that it can be upgraded to a hybrid quality metric with superior performance.

Second, the study on objective quality evaluations on packet loss impaired videos is performed on the condition that both frame rate and resolution are fixed. In order to

better assess the video quality in multimedia applications in practical situation where those quality-affecting factors may also vary, we may need to find a solution that can balance the visual annoyances of the two types of distortions and combine them into one single quality metric.

Third, for the study of saliency aided video quality assessment, an accurate and efficient saliency detection algorithm is always desirable. However, the performances in both detection accuracy and computation complexity of current widely accepted saliency detection systems are still far away from acceptable in practical situations. Due to these reasons, we alternatively use the output of a face detection algorithm as the saliency information. And we have performed a preliminary study of incorporating detected human face into PDMOSCL by using the saliency-weighted PSNR drop. However, this specialized saliency cannot bring significant improvement to the modified quality metric for our test data. This may be due to the inaccuracy of face detection itself or the fact that human faces cannot be used to represent the entire saliency. One of our ongoing works is to exploiting a better saliency detection and estimation model.

Finally, in this thesis, we have mainly focused on investigating the impact of loss patterns (length, position, and pattern) and saliency on video perceptual quality, while we mostly used PSNR based metrics to assess the deviation of the distorted video from the original video. Since PSNR is a full reference metric, and it is not an accurate enough metric for quality evaluations, in the future, we may need to devise more efficient and accurate metrics that operate in a RR or NR fashion.

Bibliography

[1] BT.500-11, ITU-R Recommendation, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[2] P.910, ITU-T Recommendation, "Subjective video quality assesment methods for multimedia applications," 2008.

[3] M. Pinson, and S. Wolf, "Comparing Subjective Video Quality Testing Methodologies," *SPIE Video Communications and Image Processing Conference*, Lugano, Switzerland, Jul. 2003.

[4] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, Phase I," 2008.

[5] S. Winkler, "On the properties of subjective ratings in video quality experiments," *First International Workshop on Quality of Multimedia Experience,* San Deigo, CA, Jul. 2009.

[6] M. Pinson, and S. Wolf, "An Objective Method for Combining Multiple Subjective Data Sets," *SPIE Video Communications and Image Processing Conference*, Lugano, Switzerland, Jul. 2003.

[7] M. Pinson, and S. Wolf, "Techniques for Evaluating Objective Video Quality Models Using Overlapping Subjective Data Sets," National Telecommunications and Information Administration Technical Report TR-09-457, Nov. 2008.

[8] M. Pinson, and S. Wolf, "The Impact of Monitor Resolution and Type on Subjective Video Quality Testing," NTIA Technical Memorandum TM-04-412, Mar. 2004.

[9] S. Voran, and A. Catellier, "Gradient Ascent Paired-comparison Subjective Quality Testing," *First International Workshop on Quality of Multimedia Experience*, San Deigo, CA, Jul. 2009.

[10] R. Hamberg and H. Ridder, "Time-varying Image Quality: Modeling the Relation between Instantaneous and Overall Quality," *SMPTE Journal*, vol. 108, pp. 802-811, Nov. 1999.

[11] D. Hands, "Temporal characterization of forgiveness effect," *Electronic Letter*, vol. 37, pp. 752-754, 2002.

[12] S. Kanumuri, et al., "Modeling packet-loss visibility in MPEG-2 video," *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 341-355, Apr. 2006.

[13] U. Reiter, and J. Korhonen, "Comparing Apples and Oranges: Subjective Quality Assessment of Streamed Video with Different Types of DistortionU," *First International Workshop on Quality of Multimedia Experience*, San Deigo, CA, 2009.

[14] G. Cermak, "Subjective video quality as a function of frequency of artifacts," *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Sccotsdale, AZ, 2009

[15] S. Winkler, "Perceptual video quality metrics – a review," in *Digital Video Image Quality and Perceptual Coding* H. R. Wu and K. R. Rao, Ed. CRC Press, 2005.

[16] V. Baroncini, et al., "Quasi-blind on line video quality tracking based on polar edge coherence," *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, Jan. 2009.

[17] S. Winkler, *Digital Video Quality--Vision Models and Metrics*, John Wiley & Sons, 2005.

[18] VQEG, "Final report on the validation of objective models of video quality assessment, FRTV Phase I," 2000.

[19] VQEG. "Final report on the validation of models of video quality assessment, FRTV Phase II," 2003.

[20] J.144, ITU-T Recommendation, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," International Telecommunication Union, Geneva, Switzerland, 2004.

[21] BT.1683, ITU-R Recommendation, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," International Telecommunication Union, Geneva, Switzerland, 2004.

[22] J.247, ITU-T Recommendation, "Objective perceptual multimedia video quality measurement in the presence of a full reference," International Telecommunication Union. Geneva, Switzerland, 2008.

[23] J.246, ITU-T Recommendation, "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," International Telecommunication Union, Geneva, Switzerland, 2008.

[24] VQEG, "Final report from the Video Quality Experts Group on the validation of reduced-reference and no-reference objective models for standard definition television, Phase I," Jun. 2009.

[25] VQEG, "Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content," 2009.

[26] VQEG. "Hybrid Perceptual/Bitstream Testplan," 2009.

[27] M. Claypool and J. Tanner, "The effects of jitter on the perceptual quality of video," *ACM Multimedia Conference*, 1999.

[28] L. Lu and X. Lu, "Quality assessing of video over a packet network," *Second Workshop on Digital Media and its Application in Museum & Heritage*, pp. 365-369, 2007

[29] R. Pastrana-Vidal and J. Gicquel, "Automatic quality assessment of video fluidity impairments using a no-reference metric," *Workshop Video Process. Quality Metrics for Consumer Electron,* Jan. 2006.

[30] K. Yang, C. Guest, K. El-Maleh, and P. Das, "Perceptual temporal quality metric for compressed video," *IEEE Trans. Multimedia*, issue 7, vol. 9, pp. 1528–1535, Nov. 2007.

[31] S. Qiu, H. Rui, and L. Zhang, "No-reference perceptual quality assessment for streaming video based on simple end-to-end network measures," *International Conference on Networking and Services*, Jul. 2006.

[32] R. Babu, A. Bopardikar, A. Perkis, and O. Hillestad, "No-reference metrics for video streaming applications," *International Workshop on Packet Video*, Dec. 2004.

[33] H. Rui, C. Li, and S. Qiu, "Evaluation of packet loss impairment on streaming video," *Journal of Zhejing University SCIENCE*, vol. 7, pp. 131–136, Jan. 2006.

[34] M. Pinson, and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transcation on Broadcasting*, issue 3, vol. 50, pp. 312–322, Sep. 2004.

[35] "American National Standard for Telecommunications—Digital transport of one-way video signals—Parameters for objective performance assessment," American National Standard Institute, T1.801.03, 2003.

[36] L. Itti, and C. Koch, "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, issue 3, vol. 2, pp. 194-203, Mar. 2001.

[37] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, issue 1, vol. 42, pp. 107-123, Jan. 2002.

[38] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, issue 11, vol. 20, pp. 1254-1259, Nov. 1998.

[39] W. Osberger, and A. Maeder, "Automatic identification of perceptually important regions in an image using a model of the human visual system," *International Conference on Pattern Recognition*, vol. 1, pp. 701-704, Aug. 1998.

[40] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating high level perceptual factors," *IEEE International Conference on Image Processing*, pp. 414-418, Oct.1998.

[41] W. Osberger, A. Maeder, and D. Mclean, "A computational model of the human visual system for image quality assessment," *Digital Image Computing: Techniques and Applications*. pp. 337-342, Dec.1997.

[42] W. Osberger, A. Maeder, and N. Bergmann, "A technique for image quality assessment based on a human visual system model," *Ninth European Signal Processing Conference (EUSIPCO-98)*, Sep.1998.

[43] Z. Lu, et al., "Perceptual-quality significance map and its application on video quality distortion metrics," *International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3, pp. 617-620, Apr. 2003.

[44] A. Ninassi, et al., "Does where you gase on an image affect your perception of quality? Applying visual attention to image quality metric," *International Conference of Image Processing*. vol. 2, pp. 169-172, Oct. 2007.

[45] A. Moorthy, and A. Bovik, "Visual Importance Pooling for Image Quality Assessment," *IEEE Journal of Selected Topics in Signal Processing*, issue 2, vol. 3, pp. 193-201, Apr. 2009.

[46] C. Oprea, I. Prinog, C. Paleologu and M. Udrea, "Perceptual Video Quality Assessment Based on Salient Region Detection," *Fifth Advanced International Conference on Telecommunications*, pp. 232-236, May 2009

[47] S. Bandyopadhyay, Z. Wu, P. Pandit, and J. Boyce, "Frame loss concealment for H.264/AVC," ISO/IEC MPEG and ITU-T VCEG, Jul. 2005.

[48] A. Reibman, and D. Poole, "Predicting packet-loss visibility using scene characteristics," *International Workshop on Packet video*, 2007.

[49] V. Seferidis, et al., "Forgiveness effect in subjective assessment of packet video," *Electronic Letter*, vol. 28, no.21, pp. 2013-2015, 1992.

[50] D. Pearson, "Viewer response to time-varying video quality," *SPIE Human Vision and Electronic Imaging III*, vol. 3299, pp. 16-25, Jan. 1998.

[51] A. Rohaly, J. Lu, N. Franzen, and M. Ravel, "Comparison of temporal pooling methods for estimating the quality of complex video seqeuence," *SPIE Human Vision and Electronic Imaging IV*, vol. 3644, pp. 218-225, Jan. 1999.

[52] D. Freedman, *Statistics,* 4th. ed. New York : Noton, 2007.

[53] J. Ebert and A. Willig, "A Gilbert–Elliot bit error model and the efficient," Tech. Rep. TKN-99-002, Tech. Univ. of Berlin, Berlin, Germany, 1999.

[54] H. Wang and N. Moayeri, "Finite State Morkov Channel—A Useful Model for Radio Communication Channel," *IEEE Transaction on Vehicular Technology,* vol.44, no.1, pp.163-171, Feb. 1995.

[55] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image qualitu assessment From error visibility to structural similarity," *IEEE Transaction on image processing*, issue 4, vol. 13, pp. 600-612, Apr. 2004.

[56] Structural Similarity (SSIM) Index Software: *http://www.ece.uwaterloo.ca/~z70wang/research/ssim/.*

[57] Video Quality Metric (VQM) Software: *http://www.its.bldrdoc.gov/n3/video/VQM_software.php.*

[58] M. Barkowsky, et al., "Influence of the Presentation Time on Subjective Votings of Coded Still Images," *IEEE International Conference on Image Processing*, 2006

[59] W. Lin, "Communication Model for Just-Noticeable Difference," *Digital Image Quality and Perceptual Coding*, CRC, Nov. 2005.

[60] W. Lin, et al., "Visual Distortion Gauge Based on Discrimination of Noticeable Contrast Changes," *IEEE Transaction on Circuit and System for Video Technology*, issue 7, vol. 15. Jul. 2005.

[61] LIVE image dataset: www.live.ece.utexas.edu/.

[62] R. Ferzli, and L. Karam, "A No-reference Objective Image Sharpness Metric Based on Just-Noticeable Blur and Probability Summation," *International Conference of Image Processing*, 2008.

[63] A. Netravali, and B. Haskell, *Digital Picture: Representation, Compression and Standards.* 2nd ed., Springer, 1995.

[64] Z. Wang, H. Sheikh and A. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications,* CRC, 2003, pp. 1041-1078.

[65] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, issue 2, vol. 78, pp. 231-252, Oct. 1999.

[66] SaliencyToolbox1.0: http://www.saliencytoolbox.net.

[67] D. Walther, "Interactions of visual attention and object recognition: computational modeling, algorithms, and psychophysics," PhD dissertation, California Institute of Technology, Pasadena, CA, 2006.

[68] A. Borst, "Models of motion detection," *Nature neuroscience supplement*, vol. 3, pp. 1168, Nov. 2000.

[69] A. Pallus, and L. Fleishman, A two-dimensional visual motion detector based on biological principles, *http://bioengineering.union.edu/motion/maindocument.htm.*

[70] E. Chong, C. Lim, "Elementary motion detection with selective attention," *International Conference on Knowledge-Based Intelligent Information Engineering Systems*, pp. 365-368, Dec. 1999.

[71] Z. Alvidrez, and C. Higgins, "Contrast saturation in a neuronally-based model of elementary motion detection," *Elsvier on Neurocomputing*, vols. 65-66, pp. 173-179, 2005.

[72] PolyVideoLab dataset: *http://vision.poly.edu/index.html/index.php?n=HomePage.PerceptualVideoQualityInPresenceOfPacketLoss.*

[73] L. Itti and P. Baldi, "A principle approach to detecting surprising events in video," *IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, Jun. 2005.

[74] P. McCullagh and J. Nelder, *Generalized Linear Models.* 2nd. ed., Chapman & Hill, 1989.

[75] R software . *http://www.r-project.org/.* [Online]

List of Publications

[1] X. Feng, T. Liu, D. Yang, and Y. Wang, " Saliency Inspired Modeling of Packet-loss Visibility in Decoded Videos," *IEEE Trans. Image Processing*, 2009. under preparation

[2] Y. Ou, Z. Ma, T. Liu, and Y. Wang, "Perceptual Quality Assessment of Video Considering both Frame Rate and Quantization Artifacts," *IEEE Trans. Circuits and Systems for Video Technology*, 2009. under review

[3] T. Liu, Y. Wang, J. Boyce, H. Yang, and Z. Wu, "A Novel Video Quality Metric for Low Bitrate Video Considering Both Coding and Packet-loss Artifacts," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Visual Media Quality Assessment*,Vol.3 No. 2, pp.280~293, April 2009.

[4] Z. Liu, T. Liu, D. Gibbon, and B. Shahraray, "Effective and Scalable Video Copy Detection," *ACM SIGMM International Conference on Multimedia Information Retrieval*, Mar. 2010.

[5] T. Liu, H. Yang, A. Stein, and Y. Wang, "Perceptual Quality Measurement of Video Frames with Error Propagation Artifact Due to Packet Loss," *IEEE International Workshop on Quality of Multimedia Experience*, Jul. 2009.

[6] T. Liu, X. Feng, A. Reibman, and Y. Wang, "Saliency Inspired Modeling of Packet-loss Visibility in Decoded Videos," *Fourth International workshop on Video Proc. And Quality Metrics*, Jan. 2009.

[7] X. Feng, T. Liu, D. Yang and Y. Wang, "Saliency Based Objective Quality Assessment of Decoded Video Affected by Packet Losses," *IEEE, International Conf. Image Proc.*, Oct. 2008.

[8] Y. Ou, T. Liu, Z. Ma, Z. Zhao, and Y. Wang, "Modeling the Impact of Frame Rate on Perceptual Quality of Video," *IEEE, International Conf. Image Proc.*, Oct. 2008.

[9] T. Liu, Y. Wang, J. Boyce, Z. Wu, and H. Yang, "Subjective Quality Evaluation of Decoded Video in the Presence of Packet Losses," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2007.