TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Medientechnik

Quality of Experience-Driven Low-Delay Error-Resilient Video Communication

Yang Peng, M.Sc. (TUM)

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr.sc. Samarjit Chakraborty
Prüfer der Dissertation:	1.	Univ.-Prof. Dr.-Ing. Eckehard Steinbach
	2.	Univ.-Prof. Dr.-Ing. Klaus Diepold

Die Dissertation wurde am 15.11.2011 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 23.04.2012 angenommen.

To my parents and my brother

Abstract

With broadband access becoming increasingly common, even in mobile environments, the demand for video services, such as video streaming and video telephony, is growing rapidly. However, delivering satisfactory video services over today's communication networks is still a very challenging task.

The objective of this thesis is to achieve improved Quality of Experience (QoE) for video services with stringent delay requirements, delivered over time-varying error-prone communication channels. In the first part of the thesis, a low-delay error-resilient video transmission framework is developed for point-to-point communication with feedback information. The proposed framework consists of low-delay error-resilient video coding and delay-aware retransmission, where multi-dimensional video adaptation and joint source-channel resource allocation schemes are formulated to improve the channel adaptability. A heuristic approach is proposed to address the trade-off between spatial and temporal video quality. Experimental results show that the user-perceived video quality is significantly improved for a wide range of video content and channel characteristics. Based on the proposed framework, real-time software and software/hardware testbeds are implemented, which can be used to find the best system settings as well as the best hardware configurations for specific system requirements.

In the second part of the thesis, the respective impact of spatial and temporal impairments on the perceived video quality and their interaction are investigated. Specifically designed subjective tests are carried out. Based on graphical and statistical analysis of the subjective ratings, a full-reference video quality metric is proposed for QoE estimation in the presence of both spatial and temporal quality impairments. The proposed metric STVQM is based on PSNR, frame rate and selected spatial and temporal video content activity measures. Performance evaluation shows that STVQM is very accurate in estimating the subjective ratings, either significantly better than or as good as (with other advantages) related metrics in the literature.

With STVQM's ability to accurately estimate the QoE, a QoE-driven multi-dimensional video adaptation scheme is formulated and integrated into the low-delay error-resilient video transmission framework, which is presented in the third part of the thesis. Instead of using the

heuristic approach, the trade-off between spatial and temporal video quality is exploited in a QoE-optimized manner, where related quality/distortion estimation problems are addressed. Experimental results with various wireless channel models show that the resulting system delivers significantly improved QoE for a wide variety of video contents and channel conditions.

Kurzfassung

Mit zunehmendem Breitband-Zugang, auch in mobilen Umgebungen, wächst die Nachfrage nach Video-Diensten, wie zum Beispiel Video-Streaming und Video-Telefonie, rasant. Allerdings ist die Bereitstellung von zufriedenstellenden Video-Diensten über heutige Kommunikationsnetze weiterhin eine besondere Herausforderung.

Das Ziel dieser Arbeit ist es, die Nutzerzufriedenheit (QoE) für Video-Dienste mit strikten Verzögerungsanforderungen, die über zeitveränderliche und fehleranfällige Kommunikationskanäle bereitgestellt werden, zu verbessern. Im ersten Teil der Arbeit wird ein fehlerrobustes Videoübertragungssystem mit geringer Verzögerung für Punkt-zu-Punkt-Kommunikation mit Feedbackinformation vorgeschlagen. Das System verwendet fehlerrobuste Videocodierung mit geringer Verzögerung und verzögerungssensitive Rückübertragungen, wobei mehrdimensionale Videoanpassung und gemeinsame Quellen- und Kanalressourcenzuteilung für bessere Kanalanpassung formuliert werden. Ein heuristischer Ansatz wird vorgeschlagen, den besten Kompromiss zwischen räumlicher und zeitlicher Videoqualität zu finden. Experimentelle Ergebnisse zeigen, dass die wahrgenommene Videoqualität für eine breite Palette von Video-Inhalten und Kanaleigenschaften deutlich verbessert wird. Echtzeit-Software- und Software/Hardware-Testumgebungen basierend auf dem vorgeschlagenen System werden implementiert, die verwendet werden können, um die besten System-Einstellungen sowie die besten Hardware-Konfigurationen für spezifische Systemanforderungen zu finden.

Im zweiten Teil der Arbeit werden die jeweiligen Auswirkungen der räumlichen und zeitlichen Beeinträchtigungen auf die wahrgenommene Videoqualität und ihre Wechselwirkung untersucht. Speziell entworfene subjektive Tests werden durchgefhrt. Basierend auf grafischer und statistischer Analyse der subjektiven Bewertungen wird eine Full-Reference Videoqualitätsmetrik für die QoE-Schätzung in Gegenwart von räumlichen und zeitlichen Beeinträchtigungen entwickelt. Die vorgeschlagene Metrik STVQM basiert auf PSNR, Framerate und ausgewählten Maßen der räumlichen und zeitlichen Aktivität des Videoinhalts. Es wird durch formelle Evaluierung gezeigt, dass STVQM sehr genau die subjektiven Bewertungen vorhersagen kann, entweder deutlich besser als oder so gut wie (mit anderen Vorteilen) die

aus der Literatur bekannten Metriken.

Im dritten Teil der Arbeit wird eine QoE-orientierte mehrdimensionale Videoanpassung formuliert und in das Videobertragungssystem integriert, bei der die Nutzerzufriedenheit durch STVQM geschätzt wird. Der Kompromiss zwischen räumlicher und zeitlicher Videoqualität wird in einer QoE-optimierten Weise gefunden, wobei Qualitäts-/Verzerrungsschätzungsprobleme addressiert werden. Experimentelle Ergebnisse mit verschiedenen Kanalmodellen zeigen, dass das QoE-basierte System deutlich verbesserte Nutzerzufriedenheit für eine Vielzahl von Video-Inhalten und Kanalbedingungen liefert.

Acknowledgments

First of all, I would like to express my particular gratitude to my advisor, Prof. Dr.-Ing. Eckehard Steinbach, for giving me the opportunity to join his research group and to pursue doctoral degree in an inspiring environment. He introduced me to the wonderful world of research and gave me the freedom to explore my own ideas. I greatly appreciate his continuing support, encouragement and guidance that have been instrumental in the completion of this thesis. It has been a great honor and a pleasure for me to work with him over the past years.

I am also grateful to Prof. Dr.-Ing. Klaus Diepold for being the second examiner of this thesis and to Prof. Dr.sc. Samarjit Chakraborty for chairing the examination committee.

I would like to thank all my colleagues in the Institute for Media Technology and in the Institute for Communication Networks for their cooperation and support. Especially, I enjoyed the fruitful joint work and discussions with Dr.-Ing. Ingo Bauermann, Shoaib Khan, Dr.-Ing. Wei Tu, Hu Chen and Fan Zhang on various research topics.

Finally, I would like to express my deepest gratitude to my family, especially my parents and my brother, whose unconditional love and support have carried me through this long journey and made the completion of this thesis possible.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description	Definition
ACK	positive ACKnowledgment	page 29
ACR	Absolute Category Rating	page 14
ACR-HR	Absolute Category Rating with Hidden Reference removal	page 16
AGOP	Adaptation Group Of Pictures	page 91
CBR	Constant BitRate	page 9
CI	Confidence Interval	page 79
DACAR	Delay-Aware Channel-Adaptive Retransmission	page 40
DMOS	Differential Mean Opinion Score	page 66
DSCQS	Double Stimulus Continuous Quality Scale	page 14
EPER	Estimated Packet Error Rate	page 40
ERME	Error Resilient Motion Estimation	page 34
FB	FootBall sequence	page 61
FEC	Forward Error Correction	page 25
FM	ForeMan sequence	page 61
FR	Full-Reference	page 18
GOB	Group Of Blocks	page 10
HVS	Human Visual System	page 2
LB	Lower Bound	page 79
MB	MacroBlock	page 8
MCP	Motion Compensated Prediction	page 25
MD	Mother&Daughter sequence	page 61
MDA	Multi-Dimensional Adaptation	page 4
MSE	Mean Square Error	page 19
NACK	Negative ACKnowledgment	page 29
NLOS	Non-Line-Of-Sight	page 102
NR	No-Reference	page 18
NR-B	No-Reference in Bistream domain	page 19
NR-P	No-Reference in Pixel domain	page 19
OR	Outlier Ratio	page 76

Abbreviation	Description	Definition
PCC	Pearson Correlation Coefficient	page 76
PER	Packet Error Rate	page 3
PSNR	Peak-Signal-to-Noise Ratio	page 2
PVS	Processed Video Sequence	page 61
QMDA	QoE-driven Multi-Dimensional Adaptation	page 87
QoE	Quality of Experience	page 2
QoS	Quality of Service	page 85
QP	Quantization Parameter	page 8
RD	Rate-Distortion	page 26
RMSE	Root Mean Square Error	page 76
RPER	Residual Packet Error Rate	page 40
RQ	Rate-Quantization	page 26
RR	Reduced-Reference	page 19
SAD	Sum of Absolute Differences	page 37
SAMVIQ	Subjective Assessment of Multimedia VIdeo Quality	page 16
SNR	Signal-to-Noise Ratio	page 101
SRC	SouRCe video sequence	page 61
SSIM	Structural SIMilarity	page 21
STVQM	Spatio-Temporal Video Quality Model	page 74
SVC	Scalable Video Coding	page 56
SVQM	Spatial Video Quality Model	page 69
SYNEC	SYNchronized Error Concealment	page 35
TVQM	Temporal Video Quality Model	page 72
UB	Upper Bound	page 79
VBR	Variable BitRate	page 1

Chapter 1

Introduction

With the increasing capacity in wireless communication systems and consumers' growing appetite for video contents, a soaring number of wireless video services are finding their way into our everyday lives and the growth will continue to accelerate. However, delivering video contents over wireless channels faces many technical challenges that have prevented wireless video services from reaching their full potential.

Video service is the most demanding among all multimedia services. It generates a huge amount of data that need to be transmitted and processed in a timely manner, which would be impossible/infeasible without highly efficient compression schemes, especially for wireless communication where the bandwidth is scarce and expensive. The continued progress in digital video compression technologies has led to a burgeoning popularity of video services, but the compressed video data become highly sensitive to transmission errors that are very common in wireless communication channels. In addition, since most video services require real-time and continuous playback, the variable bitrate (VBR) nature of the compressed video stream and the time-varying nature of the wireless channel pose additional challenges for wireless video system design. Numerous error control tools (e.g., error resilient video coding, error concealment, channel coding, retransmission) have been proposed to address the error-sensitivity issue and various schemes (e.g., buffering) have been designed for adapting to the time-varying video content characteristics and wireless channel conditions. However, most of the previous designs have focused on video services such as video streaming that allow a relatively large end-to-end delay (i.e., several seconds) and therefore may not be suitable for interactive video services that have stringent delay requirements. For example, for conversational video services, such as video telephony and video walkie-talkie, the end-to-end delay is required to be below 150ms. Video-based teleoperation applications, where a teleoperator is to be maneuvered based on the video transmitted from a camera attached to it (e.g., car backup video system where video captured from a rear view camera is transmitted

to a front monitor to help backup the car), often requires end-to-end delay to be as low as 30-100ms. Therefore, designing video communication systems that are bandwidth-efficient, resilient to transmission errors, highly adaptive to time-varying video content characteristics and channel conditions, and at the same time meet the stringent delay requirements, is of great importance for the success of interactive video services. Motivated by this, one of the main focuses of this dissertation is to construct such a wireless video communication system design for scenarios where feedback information is available.

Typically, since video services are consumed by human users, maximizing the Quality of Experience (QoE) that human users receive should be the ultimate goal of the video communication system design. In order to autonomously provide the best possible QoE to the consumers, the system needs to be able to accurately estimate or predict the resulting QoE using an objective metric for any particular circumstances. In a sophisticated wireless video system design (e.g., with error control and/or adaptation schemes), user perceived QoE can be impaired by different visual quality degradations caused by various processing steps, including quantization, spatial and temporal resolution change, error concealment, and others. For example, quantization may lead to artifacts such as blocking or blurring, frame rate reduction may cause motion jerkiness, and error concealment may introduce artifacts that are different for different concealment methods. All those quality degradations are processed/perceived differently by the Human Visual System (HVS) and therefore impact the QoE in different manners. Those unique aspects of different quality degradations need to be considered by the objective metric used for QoE estimation. However, most of the previous work uses averaged Peak-Signal-to-Noise-Ratio (PSNR) between original and reconstructed video frames as the metric to estimate user perceived QoE when designing and/or optimizing video systems. Although PSNR has been used as the de facto image/video quality metric and has its unique merits, it does not consider the properties of HVS and has been shown not to correlate well with the subjective quality ratings, which are regarded as the most accurate and reliable QoE estimates. The performance of PSNR as a QoE metric is particularly poor when various types of quality degradations are involved, which as discussed above is common in wireless video systems. In recent years, a growing amount of attention has been devoted to research and standardization of better metrics for image/video quality assessment, but most of the work has focused on compression artifacts introduced by video codecs where quantization is the only major cause of quality degradation. Some of the work studied the impact of other processing steps on the QoE, but very few tried to design an objective metric that can accurately estimate the QoE in the presence of both spatial and temporal quality degradations and can be applied for dynamic system optimization. Therefore, another main focus of this dissertation is to design such a QoE metric so that the proposed wireless video communication system design, where various types (both spatial and temporal) of quality degradations may

be present, can always autonomously provide the best possible QoE.

Based on the designed QoE metric, a QoE-driven multi-dimensional video adaptation scheme is formulated and integrated into the proposed low-delay error-resilient system design. The trade-off between spatial and temporal video quality is adjusted in such a way that the resulting QoE, estimated by the designed QoE metric, is maximized. One of the main challenges here is that decisions need to be made before the video frames are encoded and transmitted, and as a result, the QoE metric can not be computed directly but needs to be estimated from the available data (e.g., statistics of the recently encoded/decoded frames, recently observed channel conditions, etc.). This estimation problem needs to be addressed for each type of quality degradation, as different related data are available for different types of quality degradation. In addition, because of the low-cost and low-power requirements of wireless devices, the solutions need to be light-weight. This low-complexity requirement is also closely considered for the aforementioned system design and QoE metric design.

1.1 Contributions of the Dissertation

The focus of this dissertation is to provide improved QoE for wireless video communication under stringent delay constraint. The main contributions are summarized as follows.

Low-Delay and Error-Resilient Design for Wireless Video Transmission

Error-resilient video transmission under stringent delay constraint is studied for point-to-point wireless communication with instantaneous feedback. A low-delay error-resilient video transmission framework is developed, where the available instantaneous feedback is utilized both for video coding and transmission. Many error control techniques are adopted, adapted or improved in the new framework, including error resilient coding, rate control, retransmission, joint source-channel resource allocation and error concealment. A multi-dimensional video adaptation scheme is formulated and integrated to improve channel adaptability and user perceived QoE, where a heuristic approach based on the packet error rate (PER) is proposed to address the spatial and temporal video quality trade-offs. Extensive experimental results with different video codecs, different error concealment schemes and for a wide range of video contents, transmission data rates and channel PERs are provided, which verify the effectiveness of the proposed framework.

Implementation of Real-Time Software and Software/Hardware Testbeds

A real-time software testbed is implemented based on the aforementioned video transmission framework. With a graphical user interface and abundant adjustable system parameters, the testbed is used to evaluate and demonstrate the performance of the proposed framework

under a wide variety of settings, also in comparison to conventional systems. This software testbed can also be used to find the best settings for specific system designs. A real-time FPGA-based software/hardware testbed is also implemented, which is used to find the best hardware configuration for specific system requirements.

Investigation on QoE Impact of Spatial and Temporal Video Quality Impairments

The respective impact of spatial and temporal quality impairments on the QoE and their interaction are investigated, for which specifically designed subjective tests and formal statistical analysis (i.e., ANOVA) are carried out. The investigation shows how PSNR and video content affect the spatial quality perception, how frame rate and video content affect the temporal quality perception, and that an interaction exists between spatial and temporal quality perception.

A Full-Reference Objective Video Quality Metric for QoE Estimation

Based on the results from the aforementioned investigation, a full-reference objective video quality metric STVQM is developed for QoE estimation in the presence of both spatial and temporal quality impairments. The metric is based on PSNR, frame rate as well as selected spatial and temporal video content activity measures that can be easily computed from the source video. Due to its content-independency, high accuracy and low computational complexity, the proposed metric is highly applicable for dynamic QoE optimization in practical video transmission systems.

QoE-Driven Multi-Dimensional Adaptation

A QoE-driven solution is formulated to improve the PER-based heuristic solution for the multi-dimensional adaptation (MDA) scheme integrated into the aforementioned video transmission framework, where the QoE is estimated by the proposed STVQM metric when making the decisions. Related quality/distortion estimation problems are addressed, including the estimation of the source coding distortion as well as the channel-induced distortion. Extensive experimental results with various wireless channel models are provided, showing that the resulting system can autonomously deliver significantly improved QoE for a wide variety of video contents and a wide range of channel conditions.

1.2 Outline of the Dissertation

The rest of the dissertation is organized as follows. In Chapter 2, important background and related work on wireless video transmission and on video quality assessment are reviewed,

which intends to highlight the two main research areas that are related to the work presented in this dissertation and to provide readers with an overview to position the presented work.

In Chapter 3, a low-delay error-resilient video transmission framework is presented, in which the available instantaneous feedback is utilized in both video coding and transmission to provide high error resiliency with no or controlled impact on the end-to-end delay. A multi-dimensional video adaptation scheme is integrated into the framework to improve channel adaptability and perceptual video quality. Parts of this chapter have been published in [PZS10].

Chapter 4 is devoted to the investigation of the respective impact of spatial and temporal quality impairments on the overall perceptual video quality as well as their interaction. Based on the analysis of extensive subjective video quality evaluation results, a full-reference objective video quality metric is developed, which has high accuracy in estimating the perceptual video quality in the presence of both spatial and temporal quality impairments. Parts of this chapter have been published in [PS11].

In Chapter 5, a QoE-driven MDA scheme is formulated based on the objective video quality metric developed in Chapter 4. This adaptation scheme is integrated into the video transmission framework presented in Chapter 3, leading to significantly improved QoE with high adaptability to a wide range of video content characteristics and channel conditions. Parts of this chapter have been published in [PS11].

Chapter 6 concludes this dissertation with a summary of the results and recommendations of future work in the related areas.

Chapter 2

Background and Related Work

In this chapter, background and related work on wireless video transmission and on video quality assessment are reviewed. The review is intended to highlight the two main research areas that are related to the subsequent chapters and provide readers with an overview and basic understanding of the related work to position the work presented throughout this dissertation. Section 2.1 discusses the challenges and approaches for wireless video transmission as well as the resulting impact on the QoE, which are related to Chapter 3 and Chapter 5, while Section 2.2 gives an overview for video quality assessment, providing related background for Chapter 4.

2.1 Wireless Video Transmission - Challenges and Approaches

Wireless communication has a number of important advantages over its wired counterpart, including user mobility as well as high flexibility and low cost in deployment. However, when wireless communication meets video applications, many challenges arise. On one hand, video applications have distinct properties than conventional data applications, which impose additional requirements on the underlying communication systems, such as high data rate and low delay. On the other hand, the inherent properties of wireless channels, such as limited transmission data rate, error-prone transmission and time-varying characteristics, make it particularly difficult for wireless systems to meet these requirements. In this section, the challenges with wireless video transmission and the corresponding common approaches are reviewed. The review is kept very concise, as more thorough and detailed reviews can be found in various references, such as [WZ98, WWWK00, WHZ00]. The impact of those challenges and approaches on the user-perceived QoE is also discussed.

2.1.1 Challenges and Approaches

From the perspective of video applications, the following three properties of video data, coupled with characteristics of wireless communication systems, make wireless video communication a very challenging task.

High Data Rate

Video applications generate huge amount of data for transmission. Depending on the resolution, transmitting an uncompressed video requires a data rate of a few Mbps (e.g., QCIF@15fps) up to a few Gbps (e.g., 1080p@60fps). Considering that wireless communication systems have very limited capacity and typical user transmission data rates in today's 3G/4G mobile networks are in the range of a few hundred kbps (e.g., UMTS) to several Mbps (e.g., LTE), substantial rate reduction of the video data is crucial.

Fortunately, advances in video compression technologies have made it possible for video data rate to be reduced significantly. All the common video compression standards (i.e., MPEG-x and H.26x standards [ITU05, ISO04, JVT03]) follow the same block-based hybrid video coding structure (see Figure 2.1) to reduce the size of the video data by 1) quantization, which introduces a certain amount of distortion depending on the quantization level that is specified by the quantization parameter (QP), and by 2) removing redundancies in the video data, which is realized by predictive coding in both spatial and temporal (i.e., motion compensated prediction) directions followed by entropy coding. Generally, a video frame is divided into macroblocks (MB) and each MB can be encoded in one of the following three modes: 1) I-mode, the MB is encoded based only on the blocks in the same frame; 2) P-mode, the MB is predicted from previous frames; 3) B-mode, the MB is predicted from both previous and following frames. An I-frame is encoded entirely in I-mode, resulting in a significantly larger size than a P-frame or a B-frame, where MBs may be encoded in P-mode or B-mode, respectively. A B-frame has smaller size than a P-frame, but introduces additional delay because of the frame reordering caused by the bi-directional temporal prediction. This dissertation deals entirely with video codecs adopting this common encoding structure.

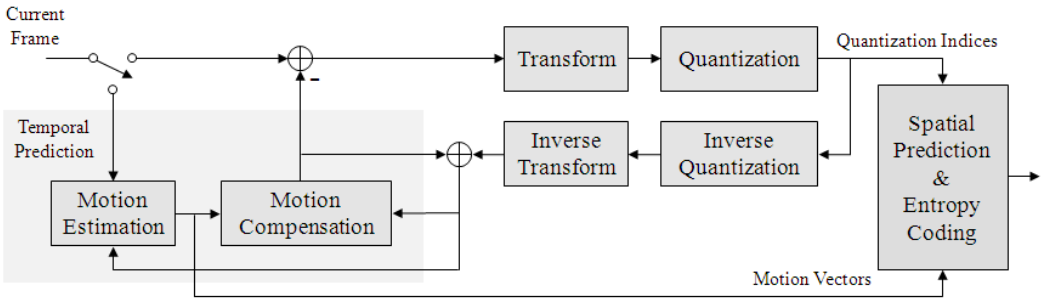


Figure 2.1: Block-diagram of the hybrid video coding structure. Adopted from [WOZ02].

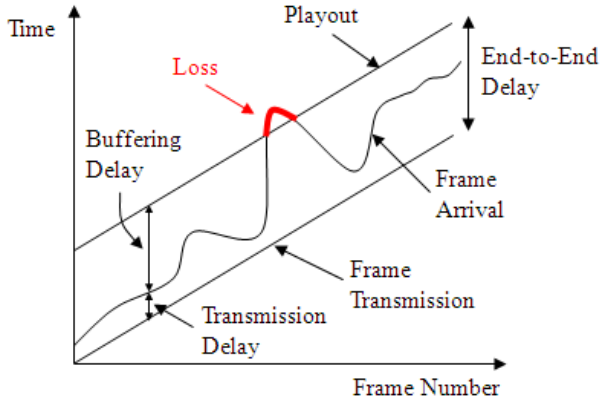


Figure 2.2: End-to-end delay of transmitting a VBR video over a CBR channel with receiver buffering. The variation of the frame arrival time is in this example due to the video bitrate variation. For illustration, delay and delay jitter caused by other factors are not included.

Delay Sensitivity

Video applications are delay sensitive. Since a video needs to be played out continuously, each frame has to be available (i.e., received and decoded) by a certain time deadline to be useful. Therefore, if a video packet does not arrive at the receiver on time, it can be considered lost. This end-to-end delay constraint is typically required to be constant for the entire duration of a video session, as display devices operate on a constant frame rate. Different applications have different requirements on the end-to-end delay [ITU01]. For example, video streaming applications may tolerate delay in the order of several seconds, while conversational and interactive applications require very low delay, generally less than 150ms.

The delay sensitivity of video applications is particularly challenging for communication systems with constant bitrate (CBR) due to the variable bitrate nature of the compressed video bitstream. In such a system, the end-to-end delay can be kept constant by buffering at the transmitter and/or the receiver [LOR98], but additional delay will be introduced by the buffering, resulting in a larger end-to-end delay as depicted in Figure 2.2. The more data are buffered, the larger variation of the video bitrate can be tolerated (i.e., less loss due to late arrival), but the larger the end-to-end delay will be. However, the buffering delay, which is usually in the order of several frame times (i.e., multiples of 33ms at 30fps), may not be tolerable for applications with low-delay requirements. In this case, the video coding rate may have to be adjusted so that the smallest frame size meets the available transmission bitrate, which without a very accurate rate control would decrease the video source coding rate significantly. Otherwise, many video packets would be lost due to late arrival, which may cause even more severe damage to the video quality.

This dissertation addresses the challenging scenario of low-delay video transmission over CBR channel. Since I-frames have significantly larger sizes and B-frames require frame re-

Figure 2.3: Example of error propagation in both temporal and spatial direction.

ordering, they may lead to higher end-to-end delay. Therefore, low-delay applications usually encode the video frames as IPP...P without the typical group of pictures (GOP) structure. This encoding structure is adopted throughout this dissertation.

Error Sensitivity

Videos compressed using the common hybrid coding structure are highly sensitive to transmission errors. In the spatial direction, because of the variable length entropy coding and the spatial predictions, one single bit error during the transmission may result in the loss of an entire video frame. This problem is often mitigated by adopting the slice structure [JVT03], which is also referred to as the group of blocks (GOB) [ITU05]. A slice, consisting of a number of MBs, can be decoded independently from other slices within the same frame, which is enabled by including resynchronization information in the slice header and by stopping spatial predictions between slices. Typically, one slice is encapsulated into one packet for transmission, during which a transmission error results in the loss of the entire packet/slice, affecting only a smaller region of the frame. The lost slices are usually concealed at the decoder based on the available spatially and/or temporally neighboring MBs before the frame is displayed. The downside of the slice structure is that the compression efficiency is reduced due to the resynchronization overhead and the lack of prediction between slices.

The motion compensated temporal prediction is another major cause for the error sensitivity of a compressed video. Although a transmission error only affects a single slice in the current frame, the error may propagate to successive frames and remain visible for a long period of time. An example of error propagation from one single slice loss is illustrated in Figure 2.3, where it can be seen that the error not only propagates in the temporal direction, but also spreads in the spatial direction due to motion compensation, which makes the resulting artifacts particularly annoying. With error concealment, the error may become less visible in the current frame, but the spatio-temporal propagation of the error still degrades the video quality significantly. Inserting an I-frame can stop the error propagation, but increases the resulting bitrate significantly.

The error sensitivity of compressed videos, coupled with the error-prone and time-varying characteristics of wireless channels (due to shadowing and multi-path fading effects as well as user mobility), poses probably the most challenge for wireless video transmission, especially under delay constraints. A large number of studies have been carried out to improve the performance of video transmission over error-prone channels. General overviews of error-resiliency and error concealment for video transmission can be found in [WZ98, WWWK00, WHZ00, WHZ+01, EY05]. Overviews of error resilient tools in the latest H.264/AVC standard are available in [SHW03, KXMP06]. An overview of feedback-based error control approaches is given in [GF99]. Error-resilient video transmission under stringent delay constraint is one of the main focuses of this dissertation.

2.1.2 QoE Impairments

Due to the challenges in wireless video transmission and the applied approaches to tackle these challenges, QoE of video applications may be impaired in many different ways.

Limited transmission data rate in a wireless channel may require a strong lossy compression of the video data, which would result in various visually annoying artifacts [YW98] in the video presented to the user. Such lossy compression schemes may involve high quantization level as well as reduction of spatial [BEK03] and temporal [LK05] resolution, each introducing different types of artifacts. The high quantization level could lead to artifacts such as blocking, blurring, and many others. Postprocessing techniques have been developed to mitigate compression artifacts [SK98], which may introduce new artifacts (e.g., de-blocking filtering to remove blocking may lead to blurring). The process of down- and up-sampling of the spatial resolution may lead to ringing or blurring artifacts. Temporal resolution reduction (i.e., frame rate reduction) may cause motion jerkiness for video scenes with high motion.

The end-to-end delay in a video transmission system can have a significant impact on the QoE, which manifests itself quite differently from that of the compression. Instead of introducing visible artifacts in the displayed video, delay deteriorates the QoE of video applications in terms of impaired human interactivity. For non-interactive applications such as one-way video streaming, the QoE impairment is perceived by the user as an initial delay from the request to the start of video playout, in which case a delay less than 10 seconds is usually deemed as tolerable. In the context of interactive applications (e.g., video telephony), the impact of delay is much more critical. For example, delay in a video telephony session would cause callers to talk over each other, rendering the conversation unbearable. Even a delay in the order of a few hundred milliseconds could easily lead to user's frustration or even failure of a conversational communication or an interactive task. Delay may also vary from packet to packet, which in general can cause additional fixed delay due to buffering or result in packet losses caused by late arrival.

Transmission errors, including bit/packet errors in the wireless channel as well as packet losses due to late arrival, results in losses of video slices, which are concealed at the receiver before display. Depending on the video content, the loss pattern and the concealment method applied, error concealment may introduce visible artifacts in the displayed video (see Figure 3.12 for some examples). Since the artifacts caused by error concealment are typically more structured and localized, they could be perceptually more annoying than compression artifacts. Even more detrimental to the QoE, the error between the actual slice and its concealment may propagate both temporally and spatially, causing the artifacts stay for a long time and the affected image area become larger.

2.2 Video Quality Assessment - An Overview

As discussed in Section 2.1.2, various quality impairments may be present in a video transmitted over a wireless communication channel. In order to deliver the best possible QoE to end-users, it is essential to understand how these impairments affect the overall video quality perceived by human beings, which goes to the fundamental question of how to assess the perceptual quality of a video with quality impairments. Generally, there are two primary ways to assess video quality: subjective quality assessment and objective quality assessment. Subjective quality assessment uses human subjects to evaluate the perceived quality of the videos under test. The advantage of subjective assessment is that the quality ratings given by the test subjects (samples) can yield a reliable estimate of the actual quality assessment on a large-scale population level. However, subjective assessment is complex and time-consuming, has to be carefully designed and performed to achieve meaningful results, and can not be adopted in real-time applications.

On the other hand, objective quality assessment evaluates video quality based on physical parameters of the video and/or mathematical models of the HVS. It requires no human involvement, can be easily applied to any video, and therefore is applicable in applications where autonomous video quality measurement or prediction is desired, such as in-service video quality monitoring and dynamic optimization of video systems. The downside of objective quality assessment is that objective quality metrics do not always provide accurate and reliable estimates of the perceptual quality. With so many factors that may affect the perceptual video quality, including fidelity of the video, characteristics of the video content, properties of the HVS, as well as application-specific factors such as display properties, user expectations, etc., objective quality metrics often have good performances in some situations but fail in others. Therefore, it is very important to be aware of the limitations of an objective quality metric before applying it to estimate perceptual video quality.

Both subjective and objective quality assessment are essential components in video quality assessment. Subjective quality assessment, being the most reliable way of evaluating percep-

tual video quality, provides "ground-truth" quality ratings for understanding how humans perceive quality impairments, as well as for the design, evaluation and validation of objective quality metrics. Once the reliability of an objective quality metric is verified by subjective evaluations, it can be applied to applications where subjective assessment is either too costly or infeasible. This section provides reviews of important aspects and related studies for both subjective and objective quality assessment, respectively.

2.2.1 Subjective Video Quality Assessment

Subjective video quality assessment is to evaluate video quality using human subjects. For applications where videos are to be viewed by human beings, it is obviously the most reliable way to determine the actual video quality perceived by real end-users. Subjective assessment usually requires a specifically designed subjective test, in which subjective quality ratings of the test videos are collected from a number of test subjects. The average subjective rating for a particular video under test, often referred to as the Mean Opinion Score (MOS), is used to measure the perceptual quality of this video.

There are many aspects that must be considered in order to design and conduct a subject test that can provide reliable and reproducible subjective data. Several international standards by ITU, including ITU-R Rec. BT.500 [ITU02] for television systems as well as ITU-T Rec. P.910 [ITU99] and ITU-R Rec. BT.1788 [ITU07] for multimedia applications, provide detailed guidelines for designing and conducting subjective tests. Some of the important aspects are summarized and discussed in the following.

2.2.1.1 Test Material

Content is one of the most important factors to be considered when it comes to video quality assessment. The same level of impairment can have a very different impact on the perceptual quality for different video contents. For example, reducing the temporal resolution is more visually annoying for a video with high motion than for one with little or no motion. Therefore, various types of source video material should be used in a subjective test so that the results can be generalized. It is recommended by ITU to select the source material based on the spatial and temporal perceptual information, which are parameters that measure the spatial and temporal complexity of a video scene, respectively. The two parameters are defined as

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\} \tag{2.1}$$

$$TI = max_{time}\{std_{space}[F_n - F_{n-1}]\} \tag{2.2}$$

where F_n denotes the video frame at time n and $Sobel(F_n)$ represents the frame filtered by the Sobel filter. More details on how to calculate SI and TI can be found in [ITU99]. Test video

sequences with different scene characteristics can be obtained from a number of organizations such as ITU, VQEG, 3GPP and others.

2.2.1.2 Test Subject

In general, there are two types of test subjects: expert and non-expert. Experts are people who are directly involved with image/video quality evaluation as part of their work or have extensive experience assessing image/video quality. Experts know what they are looking for and can finish a test fast, but they usually have a pre-determined way of looking at a test video, which an average end-user may not have. Therefore, only non-experts should be selected for tests whose results are to be generalized to the average end-user. In order to produce statistically valid results, a large number of test subjects should be used. It is recommended in [ITU07] to have at least 15 test subjects. As an example, VQEG requires subjective ratings from 24 valid test subjects for its multimedia project [VQE08a]. All test subjects should have normal or corrected-to-normal visual acuity and normal color vision.

2.2.1.3 Test Method

A test method describes how the test videos are presented to the test subjects and what scale to use for the subjective ratings. Particular test methods should be used to address particular assessment problems. For video quality assessment, where the subjects are asked to assess the overall perceived quality of any given presentation, there are three most commonly used test methods that are standardized by ITU:

- *Double Stimulus Continuous Quality Scale (DSCQS)* [ITU02]: In DSCQS, two videos are presented to the viewer in one test case. One is the unprocessed source video, and the other is a processed version of that source. The order of the two videos is randomized so that the viewer would never know which one is the source video (hidden reference). As illustrated in Figure 2.4(a), the videos are presented twice, and during the second repetition, the viewer rates both videos on a continuous quality scale shown in Figure 2.4(b). The difference between the two ratings is used to indicate the quality of the processed video.

- *Absolute Category Rating (ACR)* [ITU99]: ACR is a single stimulus method, where the test videos are presented one at a time (see Figure 2.5(a)) and rated independently on a five-point quality scale (see Figure 2.5(b)). The presentation order should be randomized for each subject to reduce the contextual effect*. Oftentimes, an unprocessed version

*The contextual effect refers to the tendency for the test subjects to rate the quality of a presentation depending on the quality level of the most recent presentations. A detailed discussion on the contextual effect can be found in [CGHS99]. Single stimulus test methods are generally more susceptible to the contextual effect because of the lack of an immediate reference.

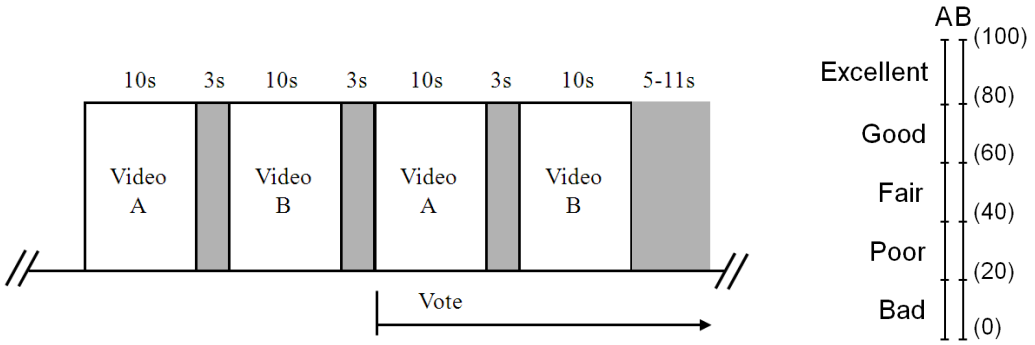


Figure 2.4: Double-Stimulus Continuous Quality Scale (DSCQS) [ITU02]: (a) presentation structure; (b) continuous quality scale.

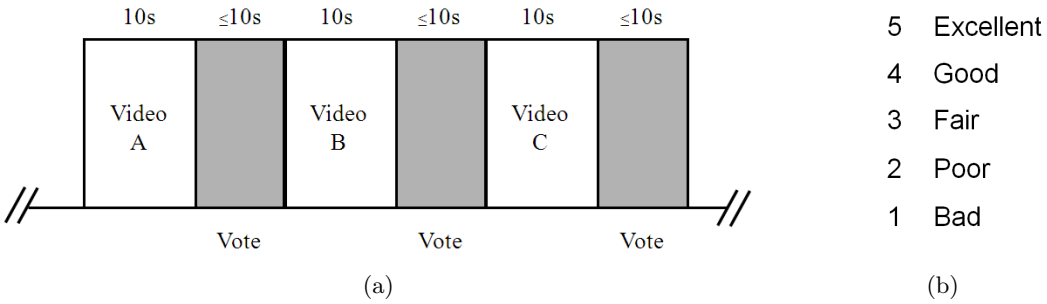


Figure 2.5: Absolute Category Rating (ACR) [ITU99]: (a) presentation structure; (b) five-point quality scale.

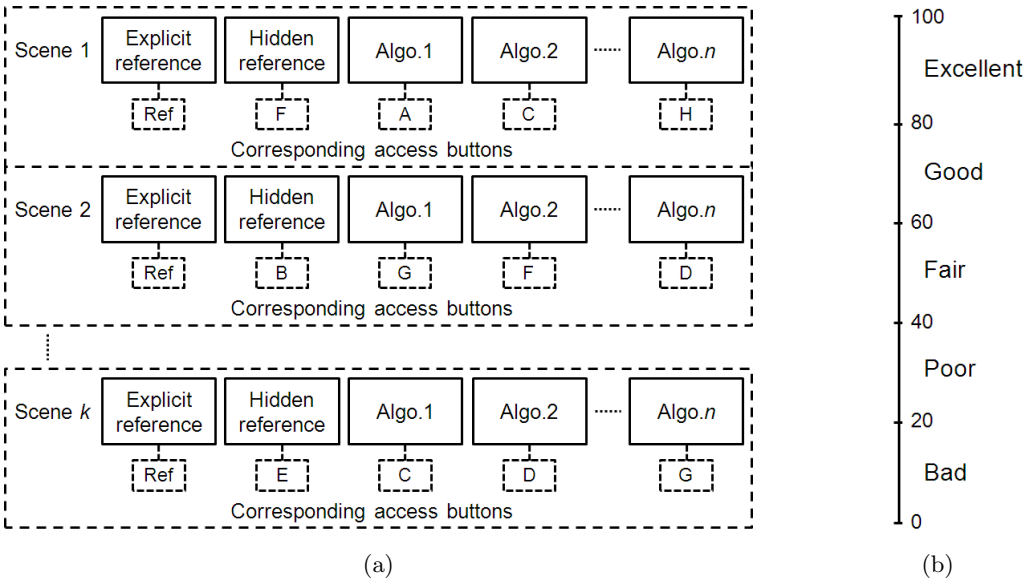


Figure 2.6: Subjective Assessment of Multimedia Video Quality (SAMVIQ) [ITU07]: (a) presentation structure; (b) continuous quality scale.

of each source video is included in the test as a hidden reference, and the difference between the rating of a processed video and its hidden reference is used to indicate the quality of this processed video. This variation of ACR is referred to as ACR with hidden reference removal (ACR-HR).

- *Subjective Assessment of Multimedia VIdeo Quality (SAMVIQ)* [ITU07]: SAMVIQ is specifically designed for multimedia contents and differs significantly from conventional test methods that are originally designed for television systems (e.g., DSCQS, ACR). Unlike DSCQS and ACR, in which the test videos are presented sequentially, SAMVIQ allows the viewer to randomly select the test videos through a computer graphic interface. As illustrated in Figure 2.6(a), the test is carried out scene by scene. For each scene, the test videos (including an explicit reference and a hidden reference) are rated one at a time on a continuous quality scale shown in Figure 2.6(b). The viewer may access each test video (through an access button) several times, compare between the test videos as well against the explicit reference, and adjust the ratings accordingly. The test videos and the buttons are randomly associated to reduce the contextual effect. The difference between the rating of a processed video and its hidden reference is used to indicate the quality of this processed video.

Each of the above three test methods has its merits and disadvantages. DSCQS is widely accepted as an accurate and reliable test method for television systems and VQEG has used DSCQS in its FR-TV tests [VQE00, VQE03]. Since each test video is paired with an immediate reference in a randomized order, contextual effects are minimized with DSCQS, as verified by the study in [CGHS99]. However, since the procedure of DSCQS is very time-consuming, only a small number of test conditions can be tested in a test session with DSCQS. In comparison, ACR can test 4 times as many conditions as DSCQS in the same time period. Therefore, ACR is more favorable for multimedia applications, where a large number of test conditions may exist. The disadvantage of ACR is that as a single stimulus method, it is susceptible to the contextual effect and may provide unreliable subjective data. But with hidden reference and randomized presentation order, it has been shown in some studies [PW03, HTG05] that a single stimulus method can produce reliable subjective data comparable to DSCQS. As a result, VQEG decided to use ACR-HR in its MM test [VQE08a]. SAMVIQ, with its interactive interface, allows the viewer to control the pace of the test. Compared to the continuous sequential presentation structure used in DSCQS and ACR, this interactive method minimizes the error of judgment caused by the lack of concentration. Furthermore, SAMVIQ provides the viewer with the ability to review and compare the test videos against each other as well as against the explicit reference in a randomized order, adjusting the ratings at the same time as appropriate, which not only significantly reduces the contextual effect, but also helps the viewer to provide more appropriate quality ratings for contents that they find difficult to rate

on a single viewing, especially when various types of quality impairments are involved. The major drawback of SAMVIQ is that with the capability to review the test videos multiple times, the reviewer may take more time for each test video compared to ACR. Various studies have compared SAMVIQ with traditional methods for multimedia applications. [Bli06] compares SAMVIQ with DSCQS and shows that SAMVIQ provides more reliable results than DSCQS. SAMVIQ is compared to ACR in [HTBH+07], [PP08] and [RPCH10]. [HTBH+07] suggests that SAMVIQ provides results comparable to ACR except for some types of error conditions, but concludes that further investigations are necessary. Both [PP08] and [RPCH10] find that for a given number of test subjects, SAMVIQ provides more accurate subjective data than ACR, but ACR with more subjects can achieve comparable accuracy. It is also shown that the difference between SAMVIQ and ACR is more significant for videos with higher resolutions [PP08], and for videos that are difficult to differentiate in terms of perceived quality [RPCH10].

Considering the merits and drawbacks of each subjective test method, SAMVIQ is selected as the test method in the work presented in Chapter 4 for collecting subjective ratings.

2.2.1.4 Test Procedure

A subjective test should be carried out in a controlled environment with viewing conditions conforming to the general guidelines provided by ITU (e.g., in [ITU07]), such as the illumination of the test cabinet, the specifications of the display, the viewing distance, etc. Prior to the test, the test subjects should usually be screened for normal visual acuity or corrected-to-normal acuity and for normal color vision. Before starting the test, written instructions (so that all the subjects receive exactly the same information) should be provided to the test subjects about the type of the assessment, the types of impairment that may occur, the test method (including the presentation structure and the rating scale), etc. A training session with a number of representative conditions should be provided to the test subjects for them to get familiar with the test method as well as the impairment types and the quality range that are likely to occur in the test. The sequences used in the training session should be different from those used in the actual test. Questions regarding the test should only be allowed before the actual test and need to be answered with care to avoid bias.

After the test, the subjective ratings should be reported along with the details of the test setup. The test subjects are usually screened based on the reliability of their subjective ratings to exclude ratings from the subjects that may have voted randomly or inconsistently. The calculation of the subject reliability is typically based on the correlation between individual ratings and the corresponding mean values over all the subjects. Details on the screening process and the reliability measures can be found in [ITU02] for DSCQS, in [VQE08b] for ACR and in [ITU07] for SAMVIQ.

2.2.2 Objective Video Quality Assessment

Objective video quality assessment is to use objective metrics to estimate the perceptual video quality without involving human subjects. Since human perception of video quality is a very complex process, which depends on many different factors such as impairment types and levels, video content characteristics, display properties, viewing conditions, user expectations and others, finding an objective metric that provides accurate and reliable quality estimate is extremely challenging. A large amount of effort has been devoted to develop and evaluate objective quality metrics. For example, VQEG has been organizing projects[†] that collect proposals on objective quality metrics and perform validation tests for various video applications. These projects have resulted in ITU standardization of objective quality metrics for standard definition television ([ITU04a][ITU04b]) and for multimedia applications [ITU08].

Objective video quality metrics are generally classified into full-reference, no-reference and reduced-reference categories [ITU00] based on the availability of the original video that may be used as a reference for comparison to the processed video. Figure 2.7 illustrates how different types of metrics can be deployed in a typical video transmission system.

- *Full-reference (FR)* metrics require the entire reference video, pixel by pixel, to be available and evaluate the quality of the processed video by comparing it to the reference. It is generally accepted that FR metrics provide the best accuracy in estimating the perceptual video quality and have been proven to have high correlation with subjective test results (e.g., [ITU04b]). However, the requirement of the entire reference places significant limitations on the practical usability of FR metrics, as the reference is often not accessible for many practical video applications, such as in-network in-service video quality monitoring. FR metrics are most suitable for off-line applications such as codec evaluation or lab testing, and are potentially applicable to real-time in-service quality monitoring or optimization at the source. For example, in Chapter 5 of this dissertation, an FR metric is applied in a real-time wireless video transmission system for in-service video quality optimization.

- *No-reference (NR)* metrics analyze the processed video directly without comparing to a reference. This makes them much more flexible than FR metrics in terms of applicability. However, although humans are usually able to reliably evaluate the quality of a processed video without using any reference, it turns out that designing such objective metrics is a very difficult task, where the main challenge lies in distinguishing impairment from content. As a result, assumptions need to be made about the type(s) of impairment and/or the video content, which makes NR metrics subject to errors caused by video content resembling a certain type of impairment (e.g., edges at the block

[†]See http://www.vqeg.org for more information.

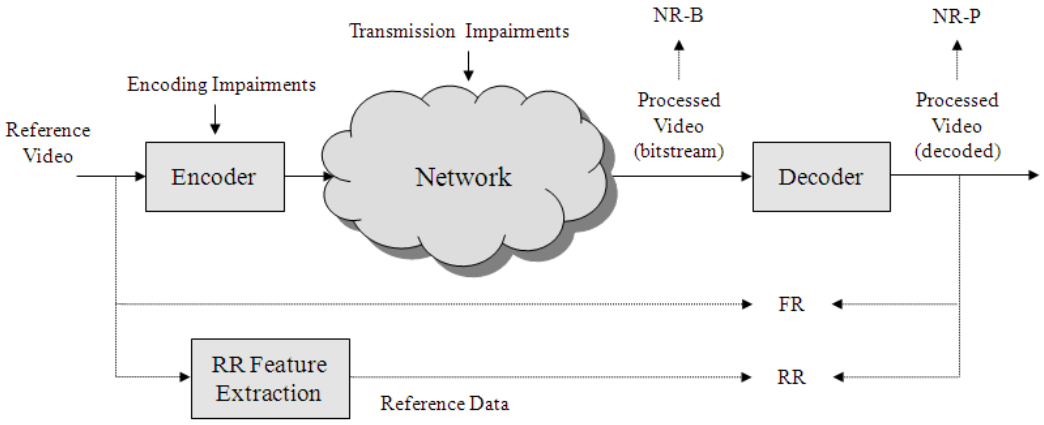


Figure 2.7: Deployment of FR, RR and NR metrics in a typical video transmission system.

boundaries in the original content could be interpreted as blocking artifacts caused by video compression). There are two types NR metrics: pixel domain (NR-P) metrics and bitstream domain (NR-B) metrics [KCRV06]. NR-P metrics analyze the decoded video, while NR-B metrics only have access to the encoded bitstream. Obviously, NR-B metrics are the best option for in-network in-service quality measurement.

- *Reduced-reference (RR)* metrics extract certain features (e.g., spatial and temporal information) from the reference video, and evaluate the quality of the processed video based only on those features. The extracted features need to be transmitted over a reliable channel to the location where the video quality is to be measured, which introduces overhead to the system. The concept of RR metrics gives the possibility to provide better quality estimate (with overhead) than NR metrics at places where the reference video is not available. Usually, the more reference information is available, the more accurately the metric can estimate the perceptual video quality [WP01], but the more overhead there will be for transmitting the reference data. The compromise between accuracy and overhead can be tailored for different applications.

The most popular objective metric for video quality assessment is PSNR, which is an FR metric defined as

$$PSNR = 10 \cdot log_{10}(\frac{I_{max}^2}{MSE}), \tag{2.3}$$

where I_{max} represents the largest possible sample value ($I_{max} = 255$ for 8-bit representation) and MSE is the mean squared error (MSE) given by

$$MSE = \frac{1}{N_R \cdot N_C} \sum_{i=1}^{N_R} \sum_{j=1}^{N_C} (f(i,j) - \hat{f}(i,j))^2. \tag{2.4}$$

Here N_R and N_C are the number of rows and columns in a video frame, and $f(i,j)$ and $\hat{f}(i,j)$ represent the sample values of the $(i,j)th$ pixel in the reference and processed video frame,

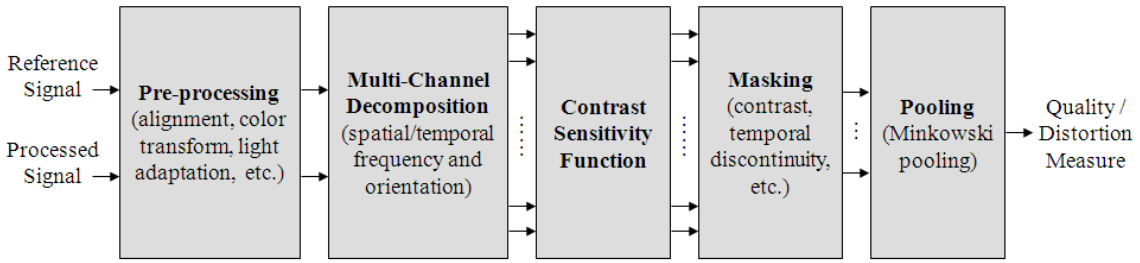


Figure 2.8: Block-diagram of HVS model-based quality assessment. "Contrast Sensitivity Function" may also be implemented as a filter before the "Multi-Channel Decomposition". Adopted from [WSB03].

respectively. Usually when applying PSNR for evaluating video quality, only the luminance component is considered ($YPSNR$), and the average PSNR, calculated over all video frames, is used as the quality measure, which can be written as

$$\overline{PSNR} = \frac{1}{N_f} \sum_{i=1}^{N_f} YPSNR_i, \tag{2.5}$$

where N_f is the number of frames. This definition of the average PSNR is assumed throughout this dissertation, unless stated otherwise.

The popularity of PSNR is mostly due to its simplicity and mathematical tractability in optimization problems. However, it neglects the properties of the HVS and therefore can have poor correlation with the quality perceived by humans [WB09, VQE03]. The same PSNR value may indicate significantly different perceptual qualities for different contents and distortion types (see [WSB03] and [WB09] for some illustrative examples). In order to improve the performance of objective quality metrics, models that account for a number of relevant psychophysical HVS features (such as multi-channel decomposition, contrast sensitivity function, masking, and others [MF06]) are developed and adopted for quality assessment. Most HVS model-based quality metrics aim to quantify the HVS sensitivity of the error between the reference and the processed signal, typically by modeling the HVS features in a sequential process as shown in Figure 2.8. Details on HVS modeling as well as reviews of HVS model-based metrics for image and video quality assessment can be found in [Win06, WSB03, Win99]. In general, HVS model-based metrics are FR metrics that may provide accurate estimates of the perceptual quality. However, the explicit modeling of HVS is usually associated with significant computational complexity. In addition, most HVS model-based metrics are developed based on psychophysical experiments focusing on the threshold of visibility (near-threshold), which may not correlate well with the perceived quality of clearly visible distortions (supra-threshold, often the case in multimedia applications).

Another group of metrics, sometimes referred to as engineering metrics [Win06], takes a more practical approach to incorporate HVS properties into objective quality assessment.

Instead of relying on sophisticated general models of HVS, these metrics estimate the overall quality based on the extraction and analysis of certain features or artifacts in the image/video, taking into account the HVS properties in a simplified or implicit manner. While such metrics are not as versatile, they generally can be computed efficiently and can perform well for specific applications. Engineering metrics cover the whole spectrum of FR/RR/NR metric categories. Many of the FR engineering metrics share the concept of extending PSNR by including certain spatial and/or temporal features of the video. For example, [TGP98] adopts a weighting function based on local gradient measures to simulate the spatial masking effect of HVS. Chapter 4 presents a PSNR-based video quality metric that estimates the overall perceptual quality in the presence of both spatial and temporal impairments. A review of PSNR-based metrics can be found in Section 4.2. [WBSS04] and [WLB04] follow a new philosophy that focuses on structural similarity (SSIM) instead of pixel-wise difference, where the SSIM metrics compare the mean, variance and covariance of small patches inside the reference and processed image/video frame, and combine the measurements into a single-value quality metric. A RR metric is presented in [PW04], which uses a number of features extracted from spatio-temporal blocks of the video. These features were selected empirically from a number of candidates so as to yield the best correlation with the subjective data. The size of the spatio-temporal blocks can be adjusted to control the trade-off between the quality estimate accuracy and the amount of overhead [WP01]. A review of NR-P metrics can be found in [Win06], most of which focus on measuring the strength of a certain type of artifact (such as blockiness, blurriness, and others) or the combination thereof. [TCC02] and [IKH$^+$06] present NR-B metrics that estimate the PSNR (quantization error) for MPEG-2 encoded video frames based on the statistical properties of the quantized DCT coefficients. NR-B metrics for estimating the impact of packet losses on the overall quality are proposed in [RVS04] and [KCRV06].

Chapter 3

Low-Delay Error-Resilient Wireless Video Transmission

In this chapter, error-resilient video transmission under stringent delay constraint is studied for point-to-point wireless communication where per-packet feedback information can be generated and transmitted instantaneously from the receiver to the transmitter. A low-delay error-resilient video transmission framework is developed, which utilizes the instantaneous feedback both in the video encoder and in the transmitter to improve video quality with no or controlled impact on the end-to-end delay.

3.1 Introduction

Due to the inherent unreliability of the wireless channel and the vulnerability of the compressed video data against errors, transmitting compressed video over wireless channels has created many technical challenges. The stringent delay requirement of interactive video applications makes these challenges even more difficult to overcome. Many those applications fit into a general point-to-point wireless transmission scenario where a live video is transmitted from one sender to one receiver over a wireless channel, and the distance between the sender and the receiver is short enough for the sender to receive near instantaneous (i.e., within a few milliseconds) feedback information from the receiver regarding the status of every transmitted video packet. Examples of such applications include, among others, conversational applications (e.g., video walkie-talkie) and teleoperated applications (e.g, navigating or controlling a teleoperator based on the video transmitted from a camera attached to it). This particularly challenging scenario with very high practical relevance, i.e., point-to-point wireless video transmission for interactive video applications with stringent delay requirement, is studied in this chapter.

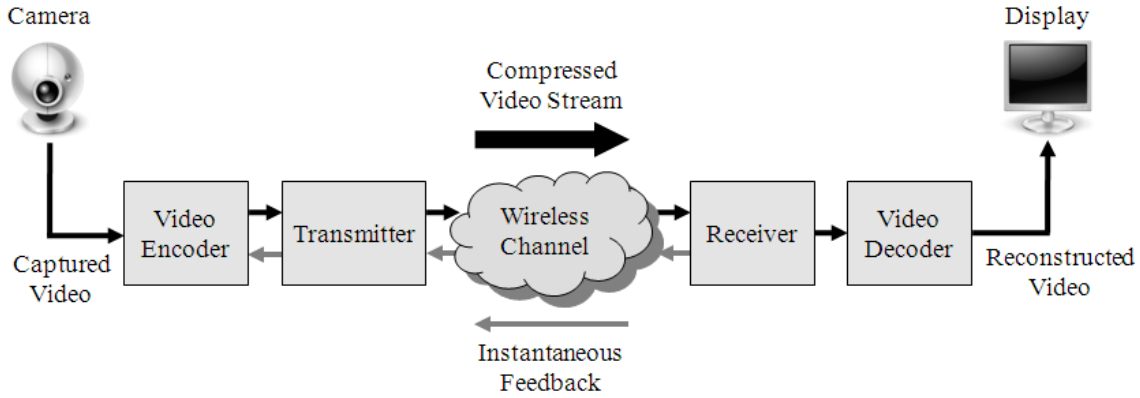


Figure 3.1: Point-to-point wireless video transmission with instantaneous feedback. The end-to-end delay in the system needs to meet the stringent delay constraint required by interactive video applications.

The general system structure abstracted from the considered scenario is illustrated in Figure 3.1. Video frames are captured live from a video camera, compressed on the fly by a video encoder, and then packetized and transmitted over a wireless channel to a receiver. During the transmission, the video packets may get corrupted in the wireless channel. The receiver receives the packets, checks their integrity (e.g., through CRC) and forwards only the useful packets to the video decoder; the packets that are corrupted during the transmission or arrive too late are discarded. The forwarded packets are then decoded and the lost image areas are concealed before the reconstructed video frames are finally displayed to a human user. In this system, the receiver is able to send per-packet feedback information back to the transmitter instantaneously, informing the transmitter about the status of every transmitted packet. The end-to-end delay in the system, i.e., the time needed for an event occurring in the camera's view to be displayed to the user, needs to meet the stringent delay constraint required by interactive video applications.

For the considered system, two major challenges exist. First, the error-prone nature of the wireless channel can cause significant quality degradations in the reconstructed video displayed to the user. Error resiliency mechanisms must be considered to improve the video quality. Second, the stringent delay constraint, generally required to be below 150ms [ITU01] and often in the range of 30-100ms [DSB+99, BO00], poses additional challenges for the system design, including the error resilience design. To address those challenges, a low-delay error-resilient video transmission framework is developed and presented in this chapter. To meet the particularly low delay constraint, the system is designed in such a way that each video frame has to be transmitted within a fixed-sized time slot, which allows the system to work without large sender and receiver buffers that would normally introduce significant delay. The available instantaneous feedback is exploited for both video coding and wireless transmission to increase the error resiliency of the system. The video encoder, before it starts to encode a

new frame, would have the necessary information from the feedbacks about which packets in the reference frame being available at the decoder. Based on this information, the encoder may exclude the corresponding lost areas from the motion compensated prediction (MCP) loop or conceal the reference frame in an encoder/decoder synchronized manner, both of which can avoid error propagation entirely without sacrificing the compression efficiency too much. At the transmitter, retransmission of the lost packets is integrated into the low-delay framework without introducing any additional delay. This is realized by dynamically allocating the fixed resource between video source coding and retransmission. In addition, different resource allocation strategies are considered so that the system can be even more adaptive to varying channel conditions. As a result of the above designs, the proposed framework achieves very low end-to-end delay and provides excellent error resilience for a wide range of channel conditions.

The rest of this chapter is organized as follows. A review of the related work is given in Section 3.2. Section 3.3 presents the low-delay design of the proposed framework. The error resilience designs, i.e., the error-resilient video coding and the delay-aware channel-adaptive retransmission, are described in Section 3.4 and Section 3.5, respectively. Experimental results are presented and discussed in Section 3.6 for performance evaluation. Several practical issues encountered during the hardware implementation are discussed in Section 3.7. Section 3.8 gives a summary of this chapter.

3.2 Related Work

One of the major challenges in wireless video transmission is that even a small number of lost packets may lead to significant degradation in the reconstructed video quality, ultimately rendering the video service unacceptable to the users. This happens because when video packets get lost, mismatch between the reference frames in the encoder and the decoder develops and leads to the infamous error propagation problem (see Section 2.1.1). Generally, this challenge can be addressed in two ways: a) by reducing the number of video packets lost during the transmission; b) by mitigating the impact of packet losses on the reconstructed video quality.

A video packet may get lost when it is corrupted in the wireless channel. There are two basic techniques that can be used to recover the packet. One is to correct the corrupted packet through channel coding (i.e., packet-level forward error correction (FEC) [Hui96]), which adds a certain amount of redundancy (parity packets) to the compressed video for error correction. The other is to retransmit the corrupted packet based on the feedback information from the receiver. For the considered scenario of point-to-point wireless transmission with feedback, retransmission-based error control is more flexible and efficient in adapting to the time-varying wireless channel conditions. However, retransmission is usually considered unsuitable for real-time video applications because of the additional delay it

introduces [ZEP+06]. Delay-constrained retransmission [WHZ00] may be useful for streaming applications where relatively large delay can be tolerated, but cannot meet the stringent delay requirement in interactive applications. In the proposed framework, however, retransmission can be integrated without introducing any additional delay, which is achieved by proactively and dynamically adjusting the video source coding rate to accommodate the potential retransmissions, keeping the total rate fixed. This can also be regarded as a joint source-channel resource allocation scheme where a fixed transmission rate is shared between video coding and retransmission. Similar ideas of considering retransmission data rate in video source coding rate control are adopted in [HOK99] and [APS01], so that the source coding rate would be reduced during poor channel conditions. But both of them consider a relatively large buffer for rate control and the rate adjustment is reacting to the retransmission data rate in a previous time slot, in which case retransmission still leads to increased end-to-end delay. The RESCU scheme proposed in [RJ00] allows retransmission for real-time interactive video transmission by changing the frame dependencies, but would not be able to fully utilize the available instantaneous feedback information. With RESCU, the compression efficiency is always reduced due to the constantly increased prediction distance, and the frame-based structure (no slice structure) is very vulnerable to the error-prone nature of the wireless channel. In this chapter, considering the low complexity requirement of most wireless applications, the joint source-channel resource allocation problem is formulated in a rather simple but effective manner based on the current channel condition. At significantly higher complexity, it can also be solved within a rate-distortion (RD) optimization framework, for example, using a formulation similar to [HCC02]. Furthermore, different resource allocation strategies are applied for different channel conditions within the low-delay framework, leading to high adaptability to a wide range of channel conditions.

Video packets may also be lost when transmitted over a CBR channel and the available transmission data rate cannot accommodate the instantaneous video source data rate generated by the video encoder (see Section 2.1.1 and Figure 2.2). This can be alleviated by buffering a certain amount of video data at the cost of increased end-to-end delay, which however may not be acceptable for interactive applications with low-delay requirement. In this case, the accuracy of the rate control scheme would have significant impact on the resulting reconstructed video quality. Rate control has been widely studied for the DCT-based hybrid coding structure adopted in all common video compression standards. Most of the proposed rate control methods determine the quantization level q from the target bit rate R based on an explicit rate-quantization (RQ) model. Some RQ models [RCL99] are deducted analytically from the information entropy theory, while others are determined empirically based on mathematical functions (of $1/q$), such as linear function [MGL05, LLS07a], quadratic function [DL96, LCZ00, WK08], and others [LOK96]. In [HKM01, HM01, HM02a, HM02b], a rate

model is constructed in the ρ-domain instead of the q-domain, where ρ represents the percentage of zeros among the quantized transform coefficients. It has been shown that compared to q-domain model-based approaches, ρ-domain rate control can achieve higher accuracy with low computational complexity. Rate control can be carried out at the frame level, the slice level or the MB level. More accurate rate control can be achieved at a smaller unit level, at the cost of slightly increased overhead due to the signaling of more quantization information. In this work, since very accurate rate control is required, an MB-level rate control scheme based on the ρ-domain rate control in [HM02a] is adopted, with some modifications that improve the accuracy at low bitrate.

Video packets may still get lost during the transmission, in spite of all the efforts to avoid it. When it happens, the reference frame in the decoder becomes erroneous compared to that in the encoder, and the error propagates in both spatial and temporal direction, causing particularly annoying artifacts in the reconstructed video. Various approaches have been proposed to mitigate the impact of error propagation. At the decoder, the lost image areas are concealed using spatially and/or temporally adjacent contents. More sophisticated error concealment methods may reduce the initial error level, resulting in less severe artifacts from the error propagation. However, those methods are usually associated with high complexity and without inserting an I-frame to stop the error propagation, the artifacts would eventually become unbearable. At the encoder, by sacrificing the compression efficiency, the video can be encoded in a more error-resilient manner. Error-resilient encoding approaches can be categorized into two classes according to the availability of feedback. Without feedback, error resilience can be improved by deliberately encoding a certain amount of MBs in INTRA mode. In order to determine how many and which MBs should be encoded in INTRA mode, error statistics of the channel, such as the average PER, need to be estimated. Those deliberately INTRA-coded MBs can be selected randomly [Sto02], heuristically based on the characteristics of the video content [LV00], or they can be added implicitly by including the expected channel distortion (i.e., distortion caused by packet losses, modeled based on the average PER) into a rate-distortion (RD) optimized INTRA/INTER mode decision process [ZRR00, WFS00, SFG00, YR07] that usually only considers source distortion (i.e., distortion caused by compression). However, due to the lack of feedback, those INTRA-update schemes have to add significant amount of redundancy to be able to effectively reduce the artifacts caused by error propagation.

When feedback is available, the performance can be significantly improved. For instance, error tracking [SFG97, GF99] utilizes feedback to track the error from the original occurrence to the current frame and stops error propagation by encoding all MBs in the affected area in INTRA mode. Simply INTRA updating the affected area may lead to significantly reduced compression efficiency. Reference picture selection schemes [FNI96, ITU96, LG06] stop the

error propagation by restricting MCP to use only the most recent error-free frame as the reference, at the cost of reduced compression efficiency due to increased prediction distance. The feedback information can also be used to improve the performance of RD-optimized mode decision through more accurate estimate of the channel distortion [ZRR00, WFS00, LC04, LG06] or even through re-decoding of the frames affected by error propagation [Wad89]. In those schemes, if the round trip time is larger than a frame interval, error propagation can not be stopped entirely as done by error tracking or by reference frame selection, and the computational complexity can be significantly higher than the non-feedback counterpart, as the expected channel distortion needs to be constantly re-computed or frames to be re-decoded. However, in this work, with instantaneous feedback available in the proposed framework, error propagation can be avoided entirely and therefore no channel distortion estimation or frame re-decoding is necessary in the mode decision. The lost areas in the reference frame can be concealed in an encocer/decoder synchronized manner to completely eliminate the mismatch between the encoder and the decoder. In this way, the system is highly resilient against transmission errors with the least possible impact on the compression efficiency. In cases where an encocer/decoder synchronized error concealment cannot be guaranteed, e.g., because the encoder cannot replicate the error concealment scheme used in the decoder, error propagation can still be entirely avoided by excluding the lost areas from the temporal prediction, at the cost of reduced compression efficiency.

3.3 Low-Delay System Design

In order to meet the stringent delay requirements of interactive applications, the video transmission system is designed in such a way that the end-to-end delay is as low as possible. Therefore, the video is encoded in the IPP...P structure and only the encoded bitstream of the current frame is buffered; no additional buffering is performed to smooth out the variations in the bitrate generated by the video encoder. It is assumed that the wireless channel is a CBR channel with constant packet size and a feedback communication channel is available.

3.3.1 End-to-End Delay Analysis

The general timeline of the proposed low-delay system is illustrated in Figure 3.2, showing how the system operates. Assuming at a certain point of time, an event occurs in the camera's view. This event could be someone talking or gesturing in a video conferencing session, or some occurrence as the result of a remote control action. Since a video camera captures at a fixed frame rate, it would take some time T_{CAP} for the event to be captured and for the captured frame (i.e., Frame i) to be accessible to the video encoder. Then, the video encoder starts encoding Frame i into a number of independently decodable slices with the

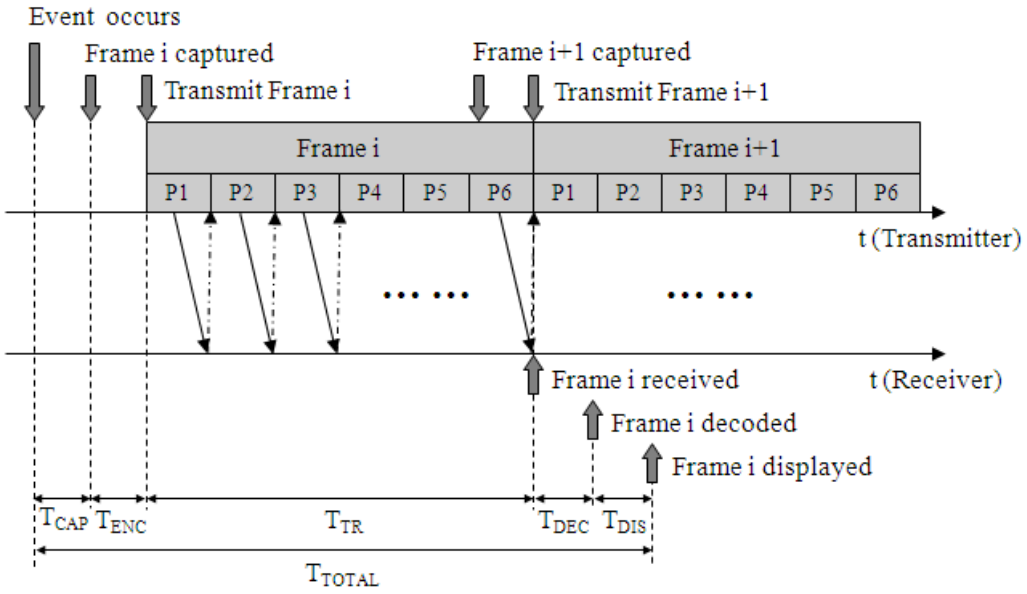


Figure 3.2: Timeline of the low-delay system design. To achieve the smallest possible end-to-end delay, the transmitter starts transmit a video frame as soon as the first packet from that frame is available; no additional buffering for rate smoothing is performed.

same size, each of which is encapsulated into a video packet for transmission. Since no additional buffering is considered, the transmitter starts transmitting as soon as the first slice is encoded and packetized (taking time T_{ENC}), assuming that the transmitter has access to the available slices during the encoding of a frame. For each received video packet, the receiver sends back the corresponding feedback information, informing the transmitter about the status of the packet. The feedback information can be negative acknowledgment (NACK) for every corrupted packet, or positive acknowledgment (ACK) for every correctly received packet, or it can be both NACK and ACK, depending on specific system conditions, especially the reliability of the feedback channel. The packets that belong to Frame i are transmitted and received only during a fixed-size time slot, referred to as the transmission delay T_{TR}; later packets would be considered lost. The reconstructed version of Frame i would be ready for display after the decoding time T_{DEC}. Finally, after a certain display delay T_{DIS}, the reconstructed Frame i, i.e., the event occurs at the camera side, is displayed to the user at the receiver side. Therefore, the end-to-end delay of the application, namely the time it takes for the event to be displayed to the user, is given by

$$T_{TOTAL} = T_{CAP} + T_{ENC} + T_{TR} + T_{DEC} + T_{DIS}. \tag{3.1}$$

Since the display device operates at a fixed frame rate, the time from a frame being captured to it being ready for display needs to be constant for a continuous playout. This constant delay, referred to as the communication delay T_{COM} here ($T_{COM} = T_{ENC} + T_{TR} +$

T_{DEC}), is typically given as a system parameter when designing a video communication system and is referred to as the end-to-end delay in most of the literature[†]. Typically, the reason for not discussing the capture and display delay is because the communication delay is usually the part that one may have control over (e.g., by controlling the amount of video data buffered) and is often much larger than the capture and display delay. However, for interactive applications that may require the end-to-end delay to be as low as 30-100ms, capture and display delay could be significant factors and therefore are briefly discussed in the following.

- The capture delay T_{CAP} consists of two elements: 1) the time from the occurrence of the event to the next capture opportunity and 2) the time from the camera starting to capture to the captured frame being accessible to the video encoder. The first element is a random variable which is related to the capture frame rate. For 30fps, it varies between 0ms and 33ms, with an average value of 17ms. This delay could be reduced by increasing the capture frame rate, but it would also generate more raw video data and require the display to match the higher frame rate. The second element is usually a constant delay in the order of several milliseconds for the entire frame to be captured. It can be reduced by enabling the encoder to start encoding as soon as a smaller part (e.g., a row of MBs) of the frame is available.

- The display delay T_{DIS}, which is the time from a reconstructed frame being moved into the display buffer to it actually being displayed to the user, also consists of two elements similarly as the capture delay. The first element, the time from a frame being ready to the next display opportunity, depends on the display frame rate and the synchronization among the receiver, the decoder and the display controller. With careful synchronization, which may be difficult to realize in practice, this delay can be kept very small. Otherwise, it can be as high as 33ms for a display frame rate of 30fps. The second element is the time the display device takes to present the frame onto the screen. It is related to the response time of the display and usually in the order of several milliseconds.

Considering that the capture delay T_{CAP} and the display delay T_{DIS} may add up to 20-80ms or even larger, there is actually very little room for the communication delay in the total delay budget. Even in cases where it is possible to buffer a few frames, the impact on the resulting video quality would be very small in the proposed system. Therefore, without loss of generality, the system is designed to operate without buffers for rate smoothing, so that the communication delay can be kept as low as possible without the usual buffering delays [BO00]. This is achieved by assigning a fixed-size time slot (i.e., bit budget) for each

[†]The constant communication delay is also referred to as the end-to-end delay in Section 2.1.1 for general discussion of the challenges in wireless video communication.

frame as shown in Figure 3.2. The encoder and the transmitter, as well as the receiver and the decoder are assumed to be able to operate in parallel, so that the transmitter can start transmitting a packet as soon as a slice is encoded, and the decoder can start decoding a slice as soon as a packet is received. In this way, the encoding delay T_{ENC} and the decoding delay T_{DEC} in Equation (3.1) are only the time needed to encode/packetize and decode one slice, respectively. Since the slice size in a wireless video transmission system is often quite small for error resiliency purpose, T_{ENC} and T_{DEC} are usually in the order of a few milliseconds. Note that T_{DEC} also includes the time for concealing the lost slices, which could be significant depending on the complexity of the error concealment method and the number of lost slices. The same is true for T_{ENC} in the case where the encoder conceals the reference frame in a encoder/decoder synchronized manner. The transmission delay T_{TR} corresponds to the fixed-size time slot assigned to a frame[†], which is related to the target frame rate of the video application. For 30fps, T_{TR} is 33ms. Here, for clarity, T_{ENC}, T_{TR} and T_{DEC} are assumed to be constant to guarantee a constant T_{COM}, which for example can be realized by a conservative delay budgeting that considers worst-case scenarios. With the proposed low-delay design, for a frame rate of 30fps, T_{COM} could be as low as about 40ms, and the end-to-end delay could be in the range of 60-120ms, which would be able to meet the requirements of most interactive applications.

3.3.2 Rate Control Algorithm

In the above low-delay system design, each video frame is to be transmitted within a fixed-size time slot, which corresponds to a fixed bit budget for every frame. However, a video encoder usually does not output a constant number of bits from frame to frame. This variation in the source bitrate, usually smoothed out by using large sender/receiver buffers, may have significant impact on the reconstructed video quality in a low-delay system that cannot afford buffering delays. Specifically, when the number of bits of an encoded frame exceeds the given bit budget, the remaining video packets in this frame would all be considered lost. In this case, in order to reduce the number of lost packets, the encoder can be configured conservatively to generate less than the available bit budget, so that the gap between the encoder's target and the real budget can absorb the variation. However, by doing this, part of the available transmission rate is wasted and the effective source coding rate is reduced, leading to stronger quantization and lower visual quality. How much the target coding rate needs to be reduced depends on the degree of the variation, which is determined by the accuracy of the rate control algorithm adopted in the video encoder. Therefore, a more accurate rate control algorithm would result in a better reconstructed video quality in the proposed low-delay design.

[†]The propagation delay is negligible in the considered point-point wireless communication scenario.

The rate control algorithm adopted in this work follows the ρ-domain rate control (ρ-RC) presented in [HM02a], which is based on a linear rate model between the bitrate and the percentage of zeros (ρ) among the quantized DCT coefficients:

$$R(\rho) = \theta \cdot (1 - \rho), \tag{3.2}$$

where θ is a content-dependent model parameter and can be adaptively estimated during the encoding. Let R_T be the target number of bits for a video frame, the ρ-RC algorithm determines the quantization parameter QP for each MB in the following steps :

Step 1. *Initialization.* Assuming that motion estimation, mode decision and DCT have been performed for the entire current frame, generate the distributions $D_0(x)$ and $D_1(x)$ for the DCT coefficients in the intra-coded and inter-coded MBs, respectively. Set N_m (the number of MBs encoded) $= R_m$ (the number of bits generated) $= \rho_m$ (the number of zeros produced) $= 0$. Set $\theta_m = 7$, which is its typical value.

Step 2. *Determine the QP for the current MB.* Let N denote the total number of MBs in a frame, the number of zeros to be produced in the remaining MBs can be calculated by

$$\rho_r = 384 \cdot (N - N_m) - \frac{R_r}{\theta_m}, \tag{3.3}$$

where $384 \cdot (N - N_m)$ is the total number of remaining DCT coefficients and $\frac{R_r}{\theta}$ is the number of non-zeros that corresponds to the remaining available $R_r = R_T - R_m$ bits according to (3.2). Determine the QP based on the one-to-one mapping between ρ and QP, which is dependent on the specific quantization scheme (refer to [HM02a] for details). Encode the current MB with this QP.

Step 3. *Update.* Let R_0 and ρ_0 be the number of bits and number of zeros produced by encoding the current MB, respectively. Set $N_m = N_m + 1$, $R_m = R_m + R_0$ and $\rho_m = \rho_m + \rho_0$. Subtract the frequencies of the DCT coefficients of the current MB from $D_0(x)$ or $D_1(x)$ according to its coding type. If $N_m \geq 10$, update θ_m based on the number of bits and the number of non-zero coefficients generated so far by

$$\theta_m = \frac{R_m}{384 \cdot N_m - \rho_m}. \tag{3.4}$$

Step 4. *Loop.* Repeat Step 2 and Step 3 for the next MB until all MBs in the current frame are encoded.

In video coding, not only DCT coefficients consume the bit budget (i.e., R_T), but so as other information such as motion vectors and header information, all of which can be considered as overhead in this context. Since the number of bits needed for encoding the overhead typically depends on the results of the quantization and therefore cannot be precisely

calculated before the rate control process starts, it needs to be estimated prior to quantization and/or updated during the rate control. The ρ-RC algorithm in [HM02a] does not explicitly separate the bits for DCT coefficients and for overhead, but the overhead bits can be considered part of R_m when computing the remaining available bit budget R_r in (3.3). However, although ρ-RC performs accurately at medium to high bitrate when the amount of overhead bits is small compared to that of the coefficient bits, its performance is significantly worsened at low bitrate when the overhead bits becomes significant. When the rate control begins, ρ-RC would overestimate the budget for coefficient bits and choose a QP that is smaller than it should be. The QP would become larger as the rate control proceeds and the number of overhead bits is updated and removed from the remaining bit budget. As a result, not only the resulting number of bits would not match well with the target, but also the QPs within a frame would vary significantly, resulting in a degraded visual quality. Therefore, in this work, the number of overhead bits is handled differently to improve the performance of the ρ-RC algorithm, especially that at low bitrate. First, before quantizing the first MB, the number of overhead bits is estimated from that in the previous frame and removed from the bit budget. Then during the rate control process, the difference between the number of overhead bits in the current frame so far and that in the previous frame at the same position is computed and the remaining budget is updated accordingly. In this way, the encoder would start with a QP quite close to the best suitable QP and adaptively refine it during the encoding process. So the remaining bit budget R_r in (3.3) is written as:

$$R_r(i) = R_T(i) - R_H(i-1) - R_C^m(i) - (R_H^m(i) - R_H^m(i-1)), \tag{3.5}$$

where $R_T(i)$ is the target number of bits for the current frame, $R_H(i-1)$ is the total number of overhead bits in the previous frame, $R_C^m(i)$ is the number of coefficient bits produced in the current frame so far, and $R_H^m(i)$ and $R_H^m(i-1)$ are the number of overhead bits produced in the current and previous frame up to the current MB, respectively. Note that Equation (3.5) is identical to the original ρ-RC algorithm for the first I-frame, and in general would not work as well for an I-frame or a P-frame directly following an I-frame. For those frames, the number of overhead bits could be estimated by a rate model for the overhead bits, such as the one presented in [KSK07].

3.4 Error-Resilient Video Coding

As mentioned in Section 3.3, a video frame in this work is encoded into several slices. Each slice can be decoded independently from the other slices in the same frame, so that losing one packet during the transmission would not result in the loss of an entire frame. The lost packets are concealed at the decoder based on the spatially and/or temporally adjacent pixels. However, because of the error propagation, even a few packet losses can still cause significant degradation

in the video quality, which is one of the major challenges of transmitting compressed video data over error-prone channels, especially over wireless channels. In the proposed framework, the available instantaneous feedback from the receiver is utilized in the encoder in such a way that the error propagation is entirely avoided. With the proposed error resilience design, the image error caused by concealment is constrained within the current frame and the system is highly resilient against transmission errors with minimal sacrifice in compression efficiency. Two error resilient coding schemes are discussed in this section, each suitable for a different set of specific system situations. In this work, it is assumed that the encoder uses only one previous frame as the reference.

3.4.1 Error-Resilient Motion Estimation

Error propagation happens when the decoder uses an erroneous version (compared to the one used in the encoder) of the previous frame as the reference in the MCP to reconstruct the current frame. Typically, the encoder references the normally decoded version, which would also be available at the decoder if there is no packet loss during the transmission. But with packet losses, the decoder has to conceal the lost areas in the previous frame, resulting in errors that propagate in both spatial and temporal directions. However, in the proposed framework, when the encoder starts to encode a new video frame, all[†] the acknowledgments for the packets in the previous frame (i.e., the reference frame) are already available at the encoder (see Figure 3.2), which allows the encoder to avoid error propagation entirely by excluding the lost packets from the motion estimation process. In other words, the encoder uses only the areas that are correctly received at the decoder as the reference so that no error would occur when decoding the current frame at the decoder. Although the decoder may still have an erroneous version of the previous frame, the concealment errors would not propagate because the erroneous parts are not used as predictions. In this work, this scheme is referred to as error resilient motion estimation (ERME), which is illustrated in Figure 3.3. Excluding the lost MBs in the reference frame from the motion estimation would result in a reduced compression efficiency, as some MBs in the current frame may not be able to find a good prediction and therefore need to be encoded in the INTRA-mode.

3.4.2 Synchronized Error Concealment

Although ERME can avoid error propagation, the decrease in the compression efficiency may be quite significant. If the encoder also knows how the lost image areas are concealed at the decoder, it can perform exactly the same error concealment on the same areas in the reference

[†]The acknowledgments for the last few packets may not be available yet when the encoding starts if there is some feedback delay. However, since the motion estimation has a limited search range, as long as the feedback delay is not too large (assumed in this work), those acknowledgments can still be incorporated later without any influence on the performance of the error resilient coding schemes here.

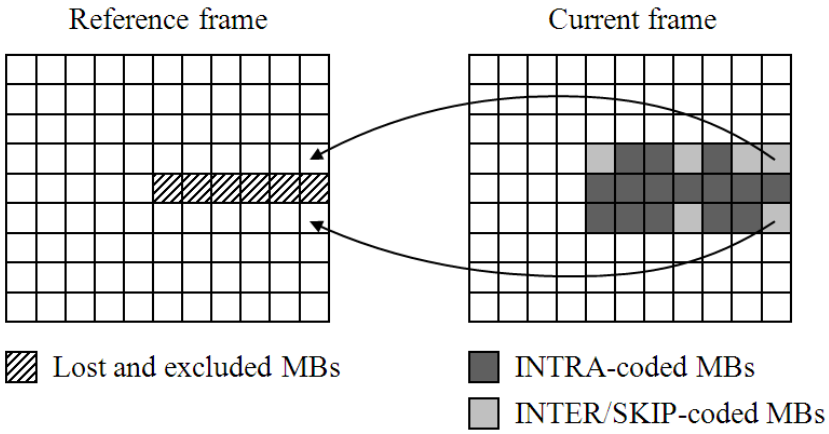


Figure 3.3: Error resilient motion estimation at the encoder. The lost MBs in the reference frame are excluded from the motion compensated prediction at the encoder.

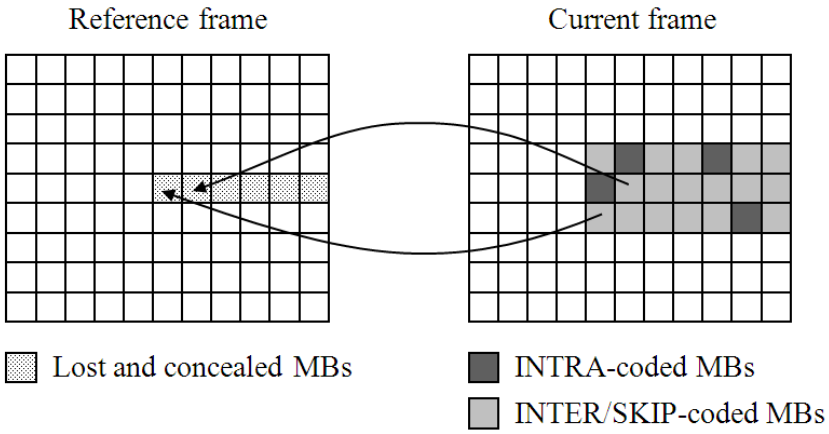


Figure 3.4: Synchronized error concealment at the encoder. The lost MBs in the reference frame are concealed at the encoder in a encoder/decoder synchronized manner.

frame as if they were not available either at the encoder. By doing this, the encoder and the decoder are resynchronized in the sense that they have identical reference frames again. Therefore, this scheme is referred to as the synchronized error concealment (SYNEC). As illustrated in Figure 3.4, with SYNEC, the lost MBs in the reference frame are concealed at the encoder before the encoding starts and then the usual motion estimation is performed based on the concealed reference frame, as would be done at the decoder.

Since the lost areas may still be used as predictions after concealment, SYNEC has higher compression efficiency than ERME, especially when the lost areas can be well concealed. This is illustrated in Figure 3.4 as most of the MBs in the potentially affected area of the current frames being coded in the INTER/SKIP-mode by using the concealed MBs in the reference frame as predictions. When comparing ERME and SYNEC in terms of compression efficiency, ERME can be understood as an extreme case of SYNEC, where the encoder applies an extremely bad concealment scheme which guarantees that the lost areas would never be

used as predictions. The other extreme case, generally speaking, would be applying a perfect concealment scheme that can recover the lost MBs with 100% accuracy. So the compression efficiency of SYNEC is between ERME and the error-free case, depending on how well the lost areas can be concealed, which is in turn determined by the video content, the slice size as well as the concealment method applied.

On the other hand, SYNEC has certain requirements on the system for it to work. First, the encoder needs to have exact knowledge of the error concealment scheme applied at the decoder, which sometimes may not be possible. Second, the acknowledgments have to be error-free to make sure the reference frames used in the encoder and the decoder are identical. With SYNEC, lost acknowledgments would lead to error propagation, which is not always the case with ERME. If only the correctly received packets are acknowledged, losing some of those positive acknowledgments would just cause ERME to exclude some unnecessary areas from the motion estimation, which would leads to reduced compression efficiency but no error propagation. If both positive and negative acknowledgments are used in the system, SYNEC and ERME can be applied for different packets. The packets with negative acknowledgment can be concealed using SYNEC and the packets without any acknowledgment (i.e., status undetermined) can be excluded from the motion estimation.

3.4.2.1 Error Concealment Methods

The error concealment method applied in the proposed system may have significant impact on the system performance. In general, concealment methods can be classified into three categories: spatial, temporal and spatio-temporal concealment. Four representative methods with low computational complexity are considered in this work, which are briefly introduced in the following.

Spatial concealment conceals a lost MB based on the available spatially adjacent pixel values (e.g., [AF95, KS93]) and is typically adopted for I-frames. One of the most commonly adopted spatial concealment methods is considered in this work, which interpolates the pixel values in a lost MB from its four one-pixel-wide boundaries using weighted average based on distance [KS93]. This considered spatial concealment method is referred to as SEC.

Temporal concealment conceals a lost MB based on the MBs in the temporally adjacent frames, usually in the previous frame (e.g., [AF95, LRL93, ZAF00]), and is typically adopted for P- and B-frames. Two temporal methods are considered in this work. The first method (referred to as CPB) copies the MB at the same spatial position in the previous frame as the concealment, which is the most commonly adopted temporal concealment method. However, with CPB, there will be clearly visible artifacts in the areas with motion. Various methods are proposed to improve the performance of temporal concealment by estimating the motion vector of the lost MB. The missing motion vector can be estimated from the available motion

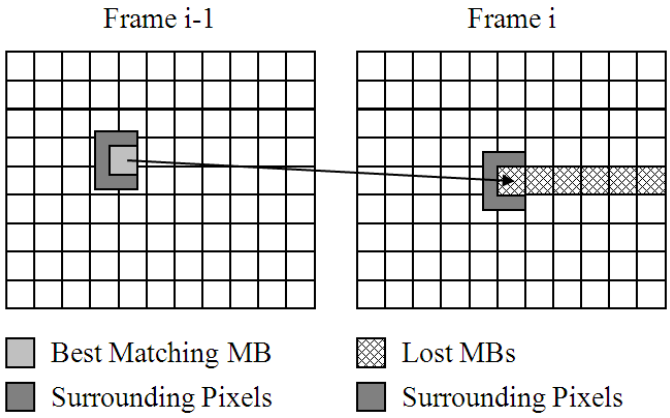


Figure 3.5: Illustration of the DMVE method. Minimizing the SAD between the surrounding pixels is used as the criteria to search for the best matching MB in Frame i-1 for concealing the lost MB in Frame i. The surrounding pixels could be correctly received or concealed pixels. The width of the surrounding pixels is typically two to eight pixels.

vectors of the spatially adjacent MBs [AF95], or by searching the best matching MB in the previous frame based on the surrounding pixel values of the lost MB [LRL93, ZAF00]. The DMVE method in [ZAF00] is considered in this work as a representative of the improved temporal concealment methods. As illustrated in Figure 3.5, DMVE searches in the previous frame for the MB whose surrounding pixel values best match that of the lost MB and uses that MB as the concealment. The best matching MB would have the least Sum of Absolute Differences (SAD) between the surrounding pixels, which typically have a width of two to eight pixels (eight adopted in this work). A full search over some area could be performed to find the best match, but it is associated with high computational complexity. In this work, the optional candidate search is performed, which considers only a few candidate motion vectors and therefore adds much less computational overhead.

Spatio-temporal concealment combines spatial and temporal concealment in an content-adaptive manner (e.g., [PPIES04, FK07]). In this work, a spatio-temporal method (referred to as STEC) is considered, which combines the results from SEC and DMVE in the following steps. First, perform DMVE on the lost MB. If the matching error (i.e., SAD between the surrounding pixels) is smaller than a threshold T_e (determined empirically), use the results from DMVE as the concealment. Otherwise, perform SEC and then conceal the lost MB by

$$\widetilde{M}(i,j) = (1 - w(i,j)) \cdot \widetilde{M}_t(i,j) + w(i,j) \cdot \widetilde{M}_s(i,j), \tag{3.6}$$

where a concealed pixel in the lost MB $\widetilde{M}(i,j)$ is the weighted average of the result from SEC $\widetilde{M}_s(i,j)$ and DMVE $\widetilde{M}_t(i,j)$. The pixel-wise weighting function $w(i,j)$ follows that in [FK07], which depends on the matching error around the lost MB in DMVE.

3.5 Delay-Aware Channel-Adaptive Retransmission

Retransmission is applied in the proposed framework to reduce the number of lost packets during the transmission, which is integrated into the low-delay system design with no or acceptable impact on the end-to-end delay.

3.5.1 Delay-Aware Retransmission

As shown in Section 3.3.1, in order to keep the end-to-end delay as low as possible, each video frame is to be transmitted within a fixed-size time slot, i.e., a fixed bit budget is assigned for transmitting each frame. If some of the packets of a frame are lost and need to be retransmitted, the retransmissions also have to be performed within the given time slot, so that no additional delay would be introduced. This requires part of the bit budget to be reserved in advance (i.e., before the encoding) for retransmissions that may be necessary later, which leads to a joint source-channel resource allocation problem where the bit budget for a video frame is shared between video source coding and retransmission. Four resource allocation strategies, referred to as retransmission schemes, are considered in this work, which are illustrated in Figure 3.6 and introduced in the following.

- **Retransmission Scheme 0 (RS0)** does not consider retransmission and assigns the entire bit budget for video source coding. Let R_{TS} be the bit budget corresponding to one time slot and R_S be the bit budget for source coding (i.e., the target bitrate for the rate control algorithm). Then R_S is given by

$$R_S = R_{TS}. \tag{3.7}$$

  RS0 is suitable for situations where the channel condition is very good and the error concealment alone may already provide satisfactory image quality.

- **Retransmission Scheme 1 (RS1)** reserves some bit budget for retransmission by encoding the video frame with a coarser quantization compared to RS0. The total budget R_{TS} is allocated in such a way that all the packets in the current frame would be successfully received by the end of the current time slot. In this case, R_S can be written as

$$R_S = R_{TS} \cdot (1 - EPER), \tag{3.8}$$

  where $EPER$ denotes the estimated packet error rate for the current time slot. In this work, the actual packet error rate in the previous time slot is used as $EPER$. RS1 is suitable for situations where the packet losses may lead to visible concealment artifacts.

- **Retransmission Scheme 2 (RS2)** considers not only quantization but also frame skipping regarding the source coding. In RS2, the following frame is skipped and the

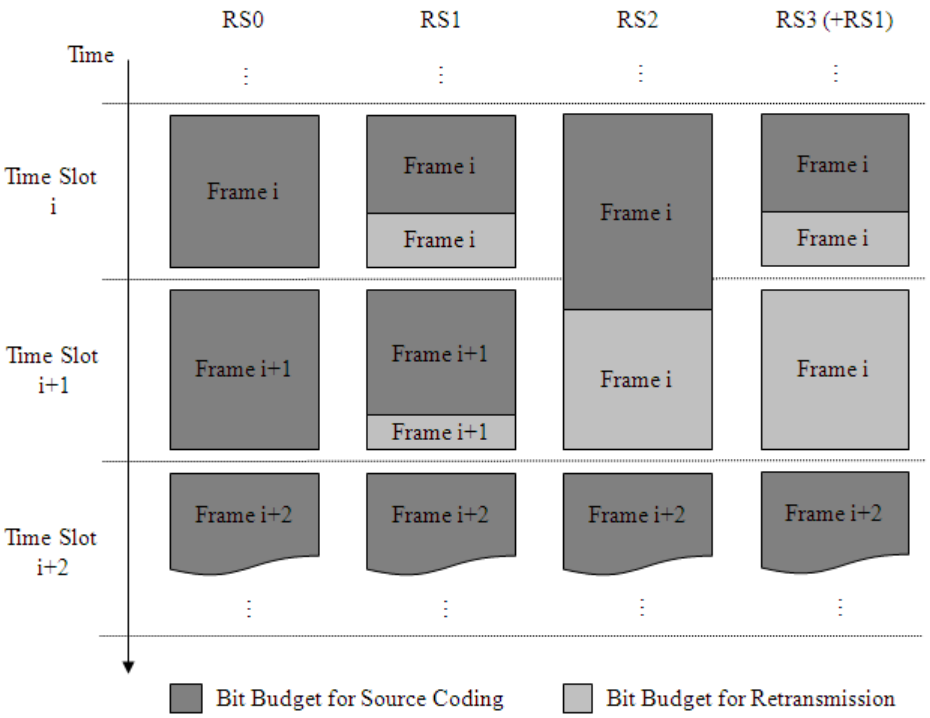


Figure 3.6: Illustration of the retransmission schemes considered in this work. Note that RS3 can be combined with any one of RS0–RS2. The combination of RS3 and RS1 is given here as an example.

total bit budget for the current frame corresponds to two time slots. This doubled bit budget is then allocated in the same way as in RS1, which yields

$$R_S = 2 \cdot R_{TS} \cdot (1 - EPER). \tag{3.9}$$

RS2 is suitable for situations where one may prefer reducing the frame rate over a coarser quantization that would reduce the image quality too much. Note that skipping a frame would also increase the end-to-end delay for the current frame by one frame interval, which may not be acceptable for certain applications.

- **Retransmission Scheme 3 (RS3)** is quite different from the previous three retransmission schemes, which are proactive measures adopted before the encoding to reserve bit budget for retransmission (i.e., determine bit budget for source coding). Those proactive schemes are based on certain estimations (e.g., the estimation of the packet error rate) that may deviate significantly from the actual situation, especially given the rapidly time-varying nature of the wireless channel. For example, the actual packet error rate may be significantly larger than the estimated one, which may lead to unacceptable degradation of the resulting image quality. Therefore, RS3 is designed as a reactive measure that is based on actual measurements to mitigate the impact of these

mis-estimation situations. It is adopted after the transmission of the current frame, where one can decide to skip the next frame and use the extra time slot for retransmission if the remaining packet losses would degrade the image quality too much. RS3 can be combined with any one of RS0–RS2. In the example illustrated in Figure 3.6, RS3 is combined with RS1, where at the end of the time slot i, RS3 observes that a significant number of video packets are still lost and decides to skip the following frame and keep retransmitting the lost packets in Frame i. If necessary, the frame skipping may continue as long as the increased delay is acceptable to the application.

3.5.2 Channel Adaptation

The above four retransmission schemes are combined in a channel-adaptive manner to form a delay-aware channel-adaptive retransmission (DACAR) scheme. Different retransmission schemes are applied for different channel conditions that are characterized by the packet error rate. The proposed scheme operates as follows.

Before encoding a new video frame, the PER during the last time slot is measured and used as an estimate of the channel condition during the transmission of the current frame. Based on this estimated PER (EPER), one of RS0, RS1 and RS2 is applied. If the EPER is below a certain threshold $s1$, it is considered that the channel condition would be very good and the error concealment alone could already provide satisfactory quality. Therefore, RS0 is applied, i.e., no retransmission is considered. If $s1 \leq \text{EPER} \leq s2$, the channel is considered to be in fair condition and retransmission becomes necessary. In this case, RS1 is applied, which assigns only part of the total bit budget for encoding the frame according to Equation (3.8). If the channel condition is bad, i.e., EPER $> s2$, applying RS1 may decrease the source coding budget R_S too much to keep a satisfactory image quality. So RS2 is applied, which skips the next frame and uses two time slots for transmitting the current frame. The doubled total bit budget would keep the image quality at a decent level despite the bad channel condition, but at the cost of reduced frame rate. Since skipping a frame introduces additional delay, RS2 is only considered when the resulting end-to-end delay T_{TOTAL} does not exceed the application requirement T_{MAX}.

Once it is decided which of the retransmission schemes is to be applied, the rate control algorithm in the video encoder is configured to meet the corresponding R_S given in Equation (3.7)–(3.9). The current frame is then encoded, packetized and transmitted. After all the video packets are transmitted once, the transmitter starts retransmitting the lost packets, until the available transmission rate for the current frame is used up or all the video packets are received correctly.

After the transmission of the current frame finishes, the residual PER (RPER) in the transmitted frame is measured. If the RPER is high, i.e., larger than a threshold $s3$, which

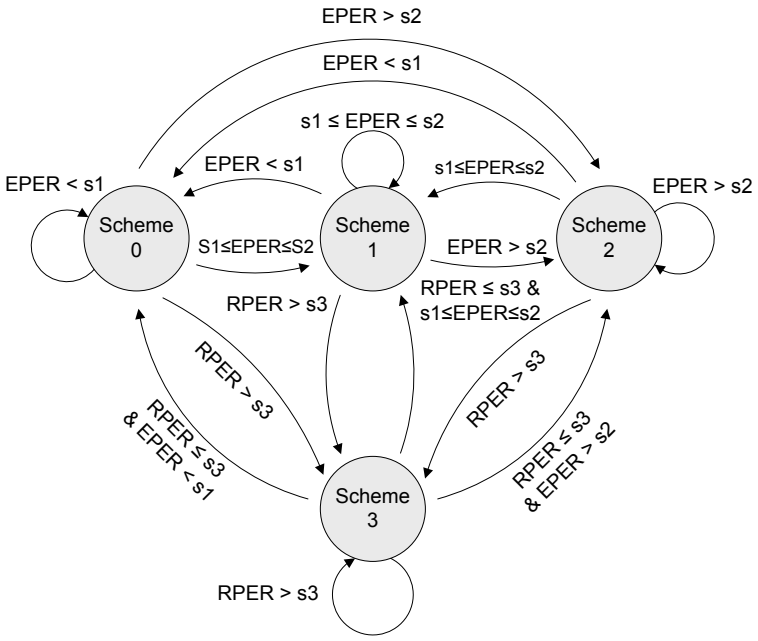


Figure 3.7: State diagram of the channel adaptation.

for example could be caused by a sudden change of the channel condition, the reconstructed frame may have unsatisfactory image quality. In this case, RS3 is applied, which skips the next frame and uses the extra time slot to retransmit the lost packets in the current frame. This would keep the image quality at a decent level even at drastic channel condition changes where the actual PER is much higher than the EPER. The measured RPER is used here as an estimate of the quality of the reconstructed current frame, which could also be estimated by some image quality metric, such as PSNR. Similar to RS2, the resulting end-to-end delay should meet the application requirement for RS3 to be applied.

The state diagram of the proposed DACAR scheme is shown in Figure 3.7, which illustrates how the four retransmission schemes are combined. Note that the delay constraint is not shown for clarity. In the proposed scheme, the choice of the thresholds (i.e., $s1$, $s2$ and $s3$) depends on various factors, including video content, concealment method, slice size, human perception of different types of quality degradation (e.g., quantization, frame rate reduction, concealment), specific application requirements, etc. For example, $s1$ is determined by the trade-off between quantization distortion and concealment distortion, and could be larger when the lost packets can be well concealed. To determine $s2$, a trade-off between spatial quality (i.e., image quality) and temporal quality (i.e., frame rate) needs to be considered, which relates to how a human user perceives the degradation of these qualities. This trade-off is also relevant when determining $s3$. For certain applications, $s3$ could also be tailored to control the minimum image quality. In this chapter, all three thresholds are selected empirically to generate the experimental results.

3.6 Experimental Results

3.6.1 Rate Control Performance

The improved ρ-domain rate control algorithm (referred to as ρ-RC-2) presented in Section 3.3.2 is implemented and evaluated in an MPEG-4 video encoder (Xvid [Xvi]) with the H.263-type quantization. The experimental results for four widely-used test video sequences in CIF (352x288) resolution at 30fps are presented: Mother&Daughter (MD), Foreman (FM), Coastguard (CG) and Cheerleaders (CL). For each video, the first 300 frames are considered and all frames are encoded as P-frames except for the first I-frame. The target number of bits of the first I-frame is twice as much as that of a P-frame so that the I-frame would have similar quality as the P-frames. Each frame is encoded as one slice for evaluating the rate control performance.

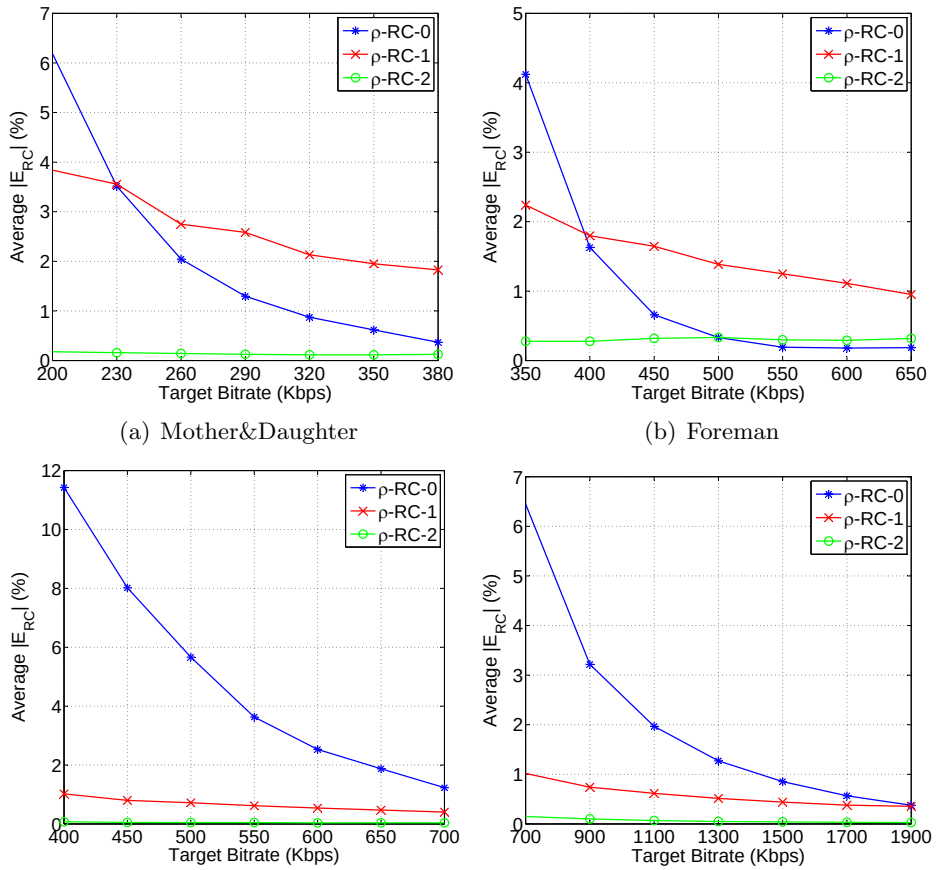
The ρ-RC-2 algorithm is compared with two other algorithms based on the ρ-domain rate control. One is the original ρ-RC algorithm (ρ-RC-0), which does not estimate the overhead before encoding but update it during the encoding process. The other one, referred to as ρ-RC-1, only estimates the overhead from the previous frame without the adaptive update during the encoding. For all three algorithms, ρ-RC-0 is applied for the first I-frame and the following P-frame; these two frames are excluded from the performance comparison. The relative control error is used to measure the accuracy of a rate control algorithm, which is defined for each frame as:

$$E_{RC} = \frac{R - R_T}{R_T} \times 100\%, \tag{3.10}$$

where R and R_T are the actual and target number of bits of the frame, respectively.

The average accuracy of the three rate control algorithms are evaluated and compared in Figure 3.8 for all four test videos. For each video, the average of the absolute value of the E_{RC} is plotted against the target bitrate. It can be seen that ρ-RC-0 always has difficulties producing an accurate output at low bitrate, but it performs better as the target bitrate increases, finally reaching a high level of accuracy. ρ-RC-1 performs differently for different videos. For Mother&Daughter and Foreman, ρ-RC-1 has better performance than ρ-RC-0 at very low bitrate, but as the bitrate increases, ρ-RC-1 does not improve as quickly as ρ-RC-0 and its accuracy is still rather limited at relatively high bitrate. For the other two test videos, namely Coastguard and Cheerleaders, simply estimating the overhead from the previous frame (i.e., ρ-RC-1) already leads to a very low control error, which suggests that the number of overhead bits does not change much from frame to frame for these two videos. With both overhead estimation and adaptive update, ρ-RC-2 has significantly higher average accuracy than the comparing algorithms.

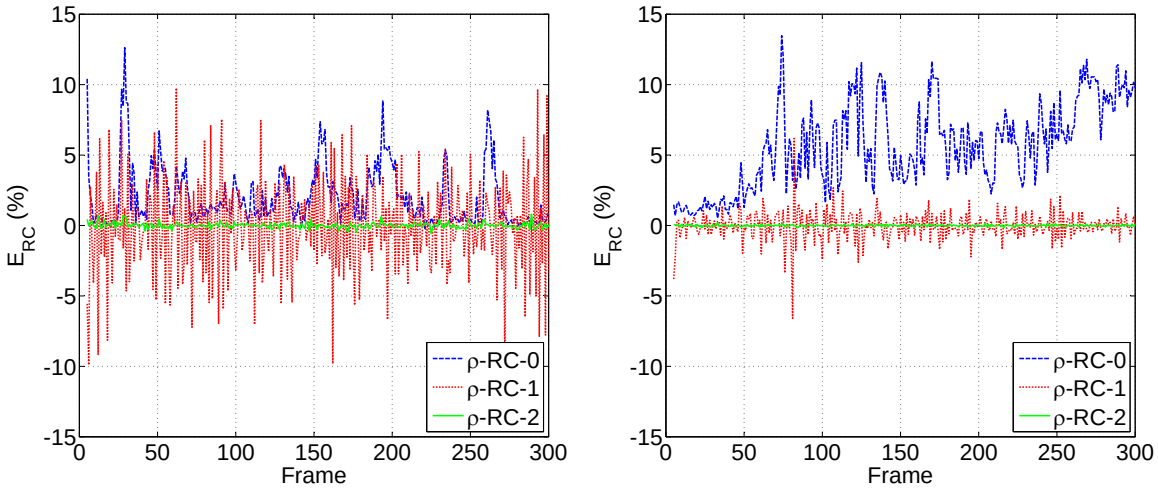
For the proposed low-delay system design, not only the average of the control error, but also the variation may have significant impact on the reconstructed video quality, as for each



(a) Mother&Daughter (b) Foreman

(c) Coastguard (d) Cheerleaders

Figure 3.8: Average of the absolute relative control error versus target bitrate for all test videos. Three algorithms based on the ρ-domain rate control are compared.



(a) Mother&Daughter, R_T = 260Kbps (b) Coastguard, R_T = 500Kbps

Figure 3.9: Relative rate control error for each frame in (a) Mother&Daughter at 260Kbps and (b) Coastguard at 500Kbps. Three algorithms based on the ρ-domain rate control are compared.

Table 3.1: Performance of the rate control algorithms

Video	R_T *(Kbps)*	*Average* $\lvert E_{RC} \rvert$ *(%)*			$E_{RC}^{Max} \mid E_{RC}^{Min}$ *(%)*		
		ρ-RC-0	*ρ-RC-1*	*ρ-RC-2*	*ρ-RC-0*	*ρ-RC-1*	*ρ-RC-2*
MD	200	6.2	3.8	0.2	33.5 | -0.03	15.1 | -15.5	1.3 | -0.4
	290	1.3	2.6	0.1	9.3 | -0.4	9.5 | -7.8	0.6 | -0.6
FM	350	4.1	2.2	0.3	20.1 | -0.7	12.1 | -11.0	0.3 | -1.2
	500	0.3	1.4	0.3	6.3 | -0.7	6.4 | -5.7	0.6 | -1.4
CG	400	11.4	1.0	0.07	19.3 | -2.5	6.6 | -6.4	0.9 | -0.2
	550	3.6	0.6	0.05	11.0 | -0.3	3.8 | -4.2	0.2 | -0.2
CL	700	6.4	1.0	0.2	10.9 | -2.5	3.6 | -4.0	1.4 | -0.1
	1300	1.3	0.5	0.05	2.8 | -0.3	1.6 | -1.5	0.5 | -0.1

frame, the level of overshoot would indicate the amount of packet losses due to inaccurate rate control. In order to avoid those packet losses, the system would need to consider the maximum error level and reduce the target encoding bitrate accordingly. In such a system, the maximum control error or the level of variation would determine how much of the available transmission bitrate would be wasted for accommodating the rate control error. In general, the variation level has similar behavior as the average value for the three algorithms shown in Figure 3.8. As examples, the E_{RC} of each frame is plotted for Mother&Daughter at 260Kbps and for Coastguard at 500Kbps in Figure 3.9, which shows that the maximum control error can be higher than 10% for both ρ-RC-0 and ρ-RC-1 even at relatively low average error level. In comparison, ρ-RC-2 has a much smaller variation level. Representative results of both the average error and the error variation are also summarized in Table 3.1 for all four test videos. It can be seen that with ρ-RC-2, the resulting bitrate is very close to the target bitrate for a wide range of bitrate and video content. The relative control error is always within a range of 1.5%, and in average below 0.3%. This high level of accuracy of the rate control algorithm allows the proposed low-delay design to fully utilize the available transmission data rate.

3.6.2 System Performance

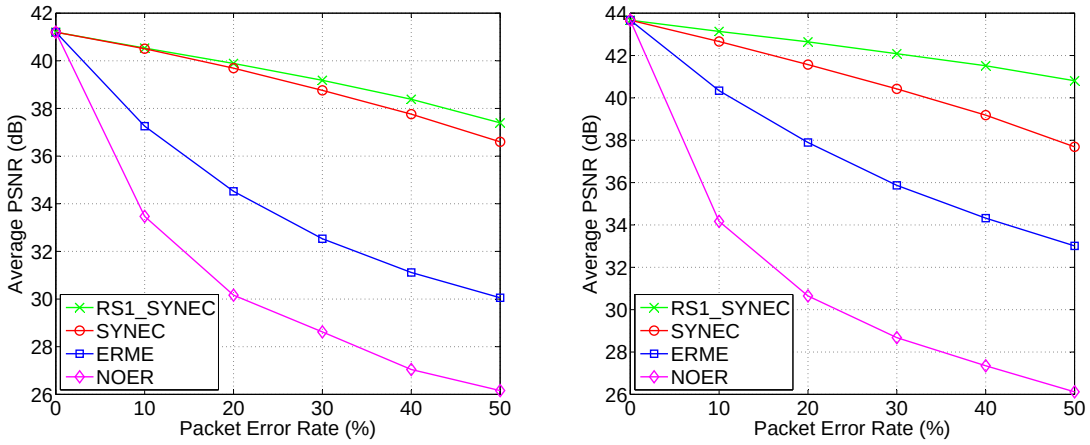
In order to evaluate the performance of the proposed framework, a real-time video transmission system is implemented. The Xvid [Xvi] codec, an open-source MPEG-4[†] video codec that is capable of real-time encoding, is modified to encode and decode the video. The four error concealment methods introduced in Section 3.4.2.1 are implemented in both the encoder and the decoder for synchronized error concealment. The wireless channel is modeled as a packet erasure channel with constant transmission data rate and random packet losses. A software channel emulator is implemented to emulate the wireless channel characteristics, such as the

[†]In this dissertation, an MPEG-4 video codec refers to an MPEG-4 Part 2 video codec.

available transmission data rate and the packet error rate. A feedback channel with no delay and no loss is also implemented in the channel emulator.

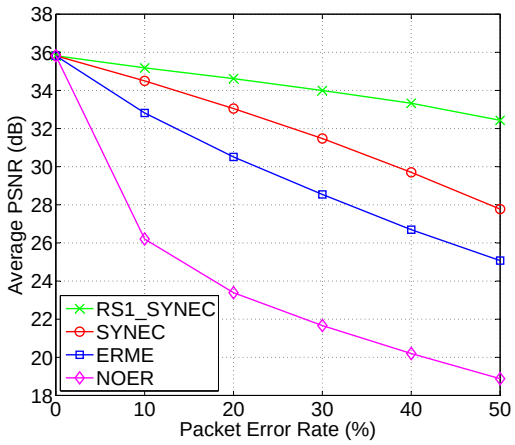
Three widely-used test video sequences with different content characteristics are adopted as the input video source: Mother&Daughter, Foreman and Football. The test videos are in CIF (352x288) resolution and have a frame rate of 30fps. The first 300 frames of each test video are encoded in the IPPP...P structure with only the first frame being an I-frame, which is designed to be transmitted using two time slots so that the image quality can be kept at a similar level to the P-frames. The following P-frame is skipped to make the extra time slot available without introducing additional delay to the rest of the frames. In the experiments, this first I-frame is considered as free of losses and not included in the results. Each video frame is encoded into slices with the same number of bytes and each slice is encapsulated into one video packet. The slice/packet size affects the overall system performance in several aspects. For example, smaller slice/packet size would decrease the compression efficiency due to the limitation on the spatial prediction and cause increased overhead, but it would also reduce the negative impact of a lost source packet on the image quality (better concealment), decrease the packet loss probabilities for a given wireless channel and increase the granularity of the retransmission/resource allocation. The optimal size would depend on various factors, including the channel conditions, the video content, the concealment method, and others. The slice/packet size used in the experiments is 125 bytes, which is determined empirically and would provide a good compromise among the related factors.

In the first experiment, the system performance is evaluated in an averaged manner for a wide range of packet error rates and various transmission data rates. Four different systems are compared: 1) NOER, where none of the proposed error resilience features are applied and DMVE is performed at the decoder for error concealment; 2) ERME, where the error resilient motion estimation is applied at the encoder and DMVE is performed at the decoder; 3) SYNEC, where DMVE is applied at both the encoder and the decoder for synchronized error concealment; 4) RS1_SYNEC, where both retransmission and SYNEC are applied and RS1 is always used as the retransmission scheme. Note that all of these four systems have the lowest possible end-to-end delay as none of the considered error resilience features introduces any additional delay. The reconstructed video quality with different systems, measured as the average PSNR against the original video, is compared in Figure 3.10. The PSNR values are computed over the entire video (the first I-frame excluded) and 10 different channel realizations. It can be seen that the systems utilizing instantaneous feedback to avoid error propagation (i.e., ERME, SYNEC and RS1_SYNEC) perform significantly better than the NOER system where error propagation sharply decreases the video quality. With improved compression efficiency, SYNEC outperforms ERME by a significant margin and the difference varies from video to video. For Mother&Daughter, a video with little motion, the improvement

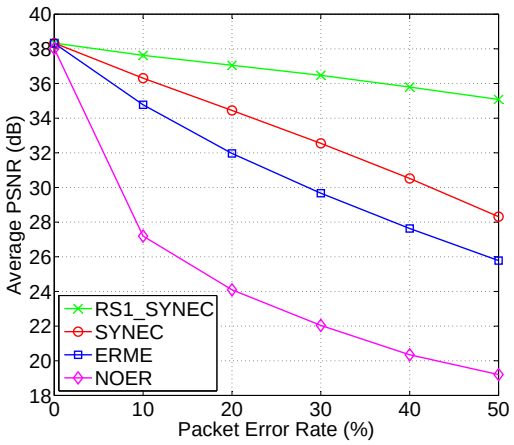


(a) Mother&Daughter, 500Kbps

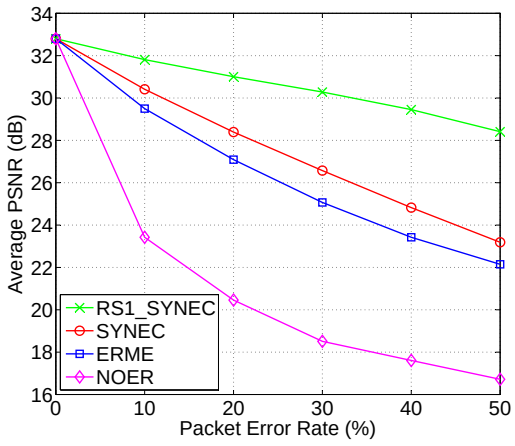
(b) Mother&Daughter, 1Mbps



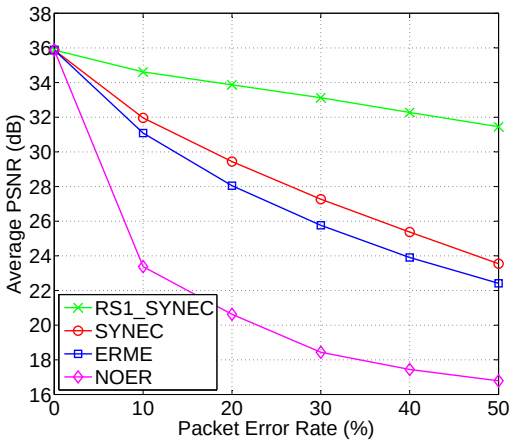
(c) Foreman, 1Mbps



(d) Foreman, 2Mbps



(e) Football, 1Mbps

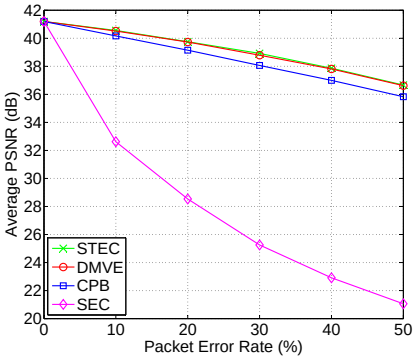


(f) Football, 2Mbps

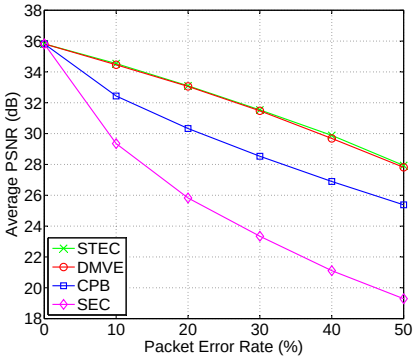
Figure 3.10: Average PSNR comparison between various systems (with an MPEG-4 video codec). The results are generated for various videos at different transmission data rates and packet error rates. The PSNR values are averaged over 10 random channel realizations.

of SYNEC over ERME can be as high as up to 7dB, with larger difference at higher packet loss rates and at lower transmission data rates. This is due to the fact that the applied temporal concealment method is able to conceal the lost MBs very well for low motion videos. When the amount of motion in the video increases, the performance difference between SYNEC and ERME decreases, as the concealment does not perform as well for videos with higher motion. For the Foreman video with medium motion, the difference is about 2-4dB, and for the high motion video Football, it reduces to about 1dB. Note that the performance of ERME can be seen as the upper bound for the approaches based on INTRA-update, as with ERME, an MB is encoded in INTRA-mode only when the INTRA-mode gives better rate-distortion performance. The integration of retransmission (i.e., RS1) leads to another significant gain on top of SYNEC, which is larger at higher packet loss rates and for videos with higher motion. This shows that for high motion videos, the system should try to avoid packet losses (e.g., by being conservative when reserving resources for retransmission), as the negative impact would be significant even without error propagation.

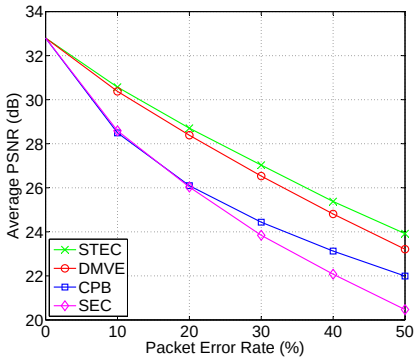
The impact of the applied error concealment method is studied by comparing four representative concealment methods (see Section 3.4.2.1) in the SYNEC system. The results averaged over 10 channel realizations for different videos are shown in Figure 3.11. For the low motion video Mother&Daughter, concealment methods that utilize temporal redundancies (i.e., CPB, DMVE and STEC) clearly outperform the spatial concealment method SEC and the difference is already more than 7dB at 10% packet loss rate. DMVE and STEC perform only slightly better than CPB in this case. For the medium motion video Foreman, there is a clear difference between DMVE and CPB, and between CPB and SEC, while DMVE and STEC have similar performances. When it comes to the high motion video Football, CPB is only better than SEC at high packet error rate levels and STEC becomes slightly better than DMVE, which indicates that for a significant number of MBs, DMVE does not perform very well and STEC chooses to combine SEC and DMVE. Figure 3.12 shows representative images from the Foreman sequence that illustrate the typical artifacts and the overall visual quality of the different concealment methods. It is clear in this example that both spatial (SEC) and temporal (CPB and DMVE) concealment methods have their respective strengths and weaknesses. SEC introduces strong artifacts in areas with details (e.g, the building facade) but performs well in smooth areas (e.g., the hand). On the contrary, the temporal methods, especially DMVE, conceal the detailed areas with low motion very well, but have problem with the waving hand (i.e., high motion). The STEC method utilizes the respective strengths of both SEC and DMVE in combination and provides significantly more visually pleasing results. Therefore, although the difference between DMVE and STEC is very small in terms of average PSNR, the visual difference may be more significant, especially in the presence of strong and/or irregular motion.



(a) Mother&Daughter, 500Kbps



(b) Foreman, 1Mbps



(c) Football, 1Mbps

Figure 3.11: Performance comparison between different error concealment methods with SYNEC and an MPEG-4 video codec.

(a) SEC

(b) CPB

(c) DMVE

(d) STEC

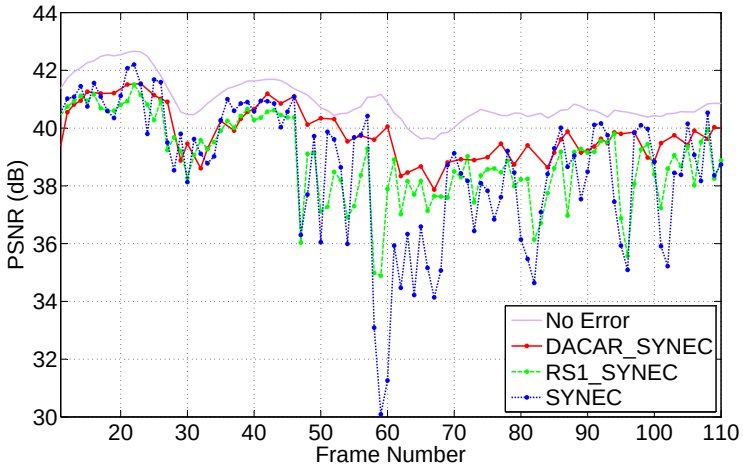
Figure 3.12: Example images for different concealment methods.

To evaluate the performance of the DACAR scheme, a second experiment is carried out for a particular channel realization with time-varying packet error rate. The channel packet error rate changes every two frames, varying between 4, 20 and 50% with a probability of 0.2, 0.6 and 0.2, respectively. Three different systems, each applying a different retransmission scheme, are compared: 1) SYNEC, where no retransmission is considered and the entire available transmission data rate is used for video source coding; 2) RS1_SYNEC, where the retransmission scheme RS1 is always adopted; 3) DACAR_SYNEC, where different retransmission schemes are considered for different channel conditions. The synchronized error concealment with DMVE is adopted in all three systems. In DACAR_SYNEC, the thresholds in the DACAR scheme, i.e., $s1$, $s2$ and $s3$, are set to be 5, 30 and 15%, respectively. The maximum end-to-end delay is assumed to allow skipping two frames, which would make the application of both RS2 and RS3 possible. For the three test video sequences, the PSNR values of Frame 11 to Frame 110 are plotted in Figure 3.13, where the performance of an error-free system (No Error) is also included as a reference. Note that since frame skipping is involved in DACAR_SYNEC, only the PSNR values of the displayed frames are shown in the corresponding results. It can be seen from the results that even for such rapidly and drastically changing channel conditions, DACAR_SYNEC keeps the image qualities at a high and relatively constant level that is close (within 1-2dB) to the error-free case. In comparison, some of the frames in RS1_SYNEC have significantly lower PSNR values than the error-free reference. The difference can be up to 5dB for low/medium motion videos and up to 8dB for the high motion video. The performance is even worse in the SYNEC system where up to 10dB differences can be observed for all three videos. Those drastically varying image qualities in RS1_SYNEC and SYNEC cause clear visual artifacts that would significantly degrade the perceived quality of the reconstructed video. The average PSNR values over the entire sequence are summarized in Table 3.2 for all three videos. For the displayed images, DACAR_SYNEC provides an additional average quality improvement of 1-3dB compared to RS1_SYNEC. Although the frame rate decreases sometimes with DACAR_SYNEC, the overall perceived video quality is still improved significantly compared to the other systems.

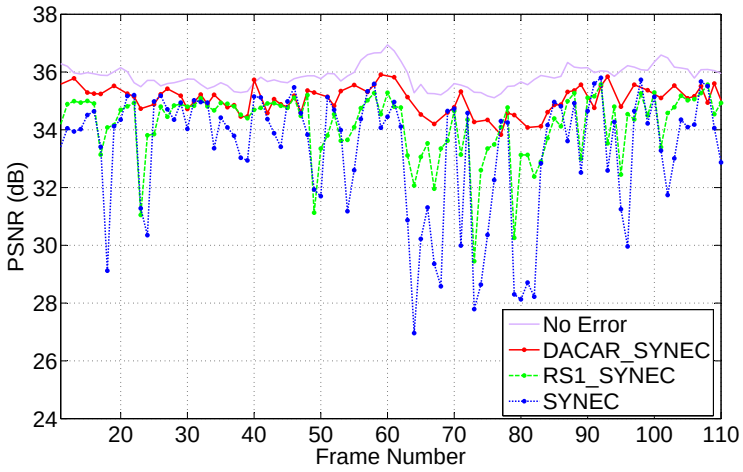
Table 3.2: Average PSNR comparison between various systems (with an MPEG-4 video codec) for a particular channel realization with time-varying packet error rate.

System	*Average PSNR (dB)*		
	Mother&Daughter	*Foreman*	*Football*
No Error	41.2	35.8	32.7
DACAR_SYNEC	40.2*	34.9*	31.5*
RS1_SYNEC	39.5	34.1	30.3
SYNEC	39.4	32.8	28.2

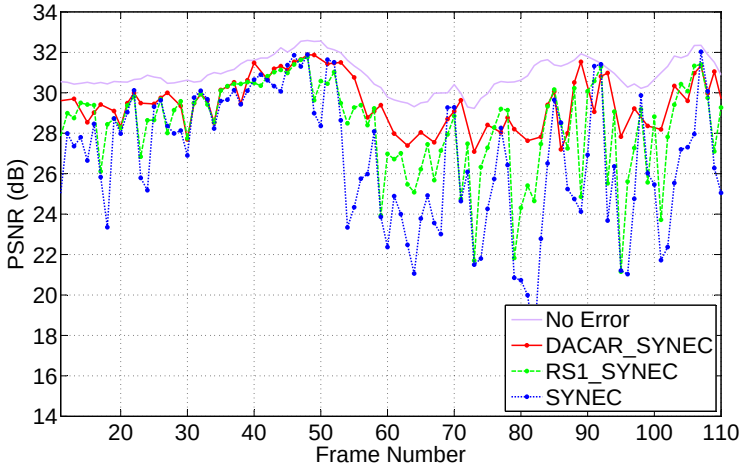
*Only displayed frames are considered here.



(a) Mother&Daughter, 500Kbps



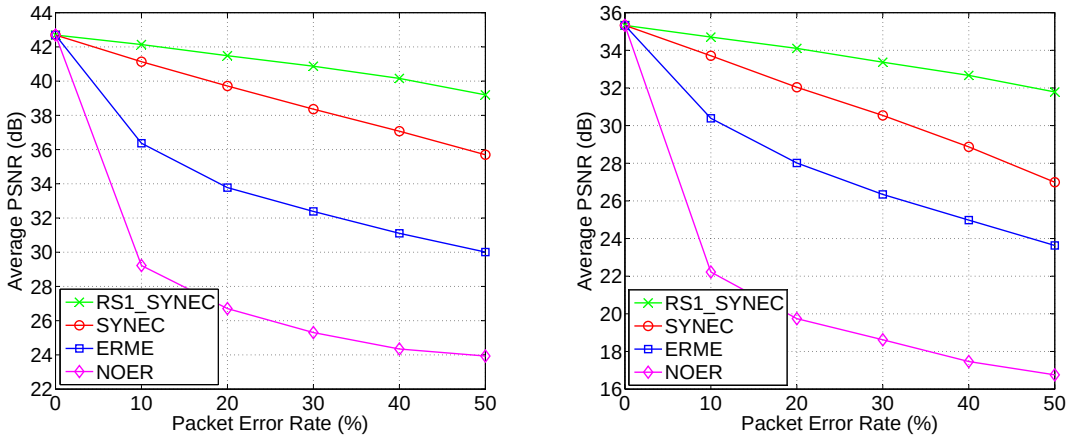
(b) Foreman, 1Mbps



(c) Football, 1Mbps

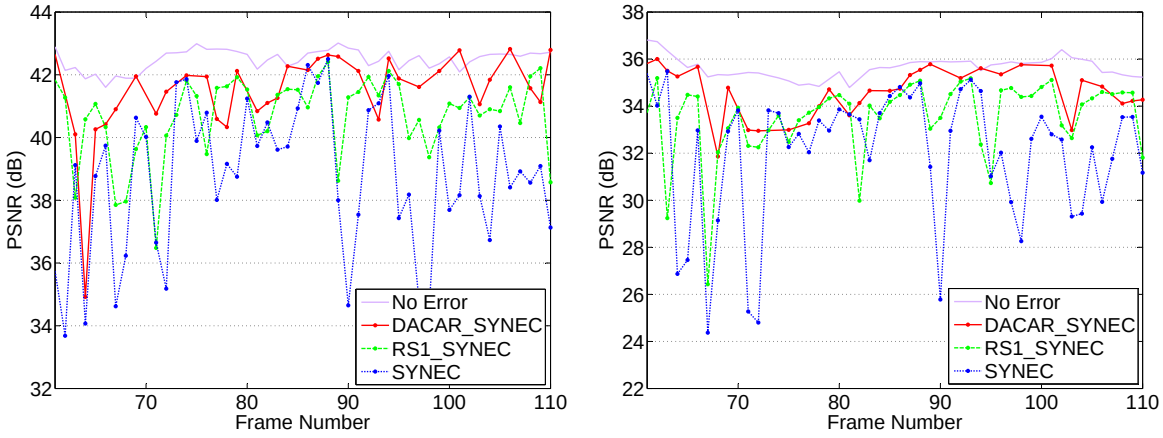
Figure 3.13: Frame PSNR comparison between various systems (with an MPEG-4 video codec) for a particular channel realization with time-varying packet error rate. The packet error rate changes every two frames, varying between 4, 20 and 50% with a probability of 0.2, 0.6 and 0.2, respectively.

The performance of the proposed framework is also evaluated in an H.264/AVC-based real-time video transmission system, where the x264 [x26] encoder and the H.264/AVC decoder in the FFmpeg [FFm] project are integrated. Similarly to the MPEG-4 based case, the reconstructed video quality is evaluated both in an average manner (see Figure 3.14) and in a particular channel with time-varying packet error rate (see Figure 3.15 and Table 3.3). As expected, the experimental results show similar system behaviors with the H.264/AVC video codec as with the MPEG-4 codec.



(a) Mother&Daughter, 500Kbps (b) Foreman, 500Kbps

Figure 3.14: Average PSNR comparison between various systems (with an H.264/AVC video codec). The results are generated for various videos at different packet error rates. The PSNR values are averaged over 10 random channel realizations.



(a) Mother&Daughter, 500Kbps (b) Foreman, 500Kbps

Figure 3.15: Frame PSNR comparison between various systems (with an H.264/AVC video codec) for a particular channel realization with time-varying packet error rate. The packet error rate changes every two frames, varying between 4, 20 and 50% with a probability of 0.2, 0.6 and 0.2, respectively.

Table 3.3: Average PSNR comparison between various systems (with an H.264/AVC video codec) for a particular channel realization with time-varying packet error rate.

System	*Average PSNR (dB)*	
	Mother&Daughter	*Foreman*
No Error	42.7	35.3
DACAR_SYNEC	41.9*	34.4*
RS1_SYNEC	41.0	33.3
SYNEC	39.7	31.8

*Only displayed frames are considered here.

3.7 Practical Issues in Hardware Implementation

Based on the proposed framework, a real-time FPGA-based software/hardware testbed with a hardware video codec (see Figure 3.16) is implemented, which is used to find the best hardware configuration for specific system requirements. During the testbed implementation, several practical issues appeared that may have significant impact on the system design and the overall performance. Some of those practical issues are briefly discussed in the following.

It has been assumed in this chapter that the video encoder and the transmitter are able to operate simultaneously at the slice/packet level. This enables the transmitter to start transmitting a frame as soon as the first slice is encoded and packetized, reducing the overall end-to-end delay. And more importantly, this makes it possible for the encoder to perform ERME/SYNEC as presented in Section 3.4, as the necessary acknowledgments for the packets in the previous frame would be available before the encoder starts encoding a new frame. However, in a practical system based on a hardware video codec, it may happen that the transmitter would not have access to the encoded bitstream until an entire frame is encoded. In this case, the end-to-end delay would be larger and ERME/SYNEC needs to be modified to adapt to this situation, which can be done by extending the prediction distance of the MCP. More specifically, a video frame now does not use the previous frame, but the frame earlier as the reference, for which all necessary feedback information would be available. This can also be seen as a case where the feedback delay is one frame interval.

A hardware video codec typically would not have an accurate MB-level rate control available and therefore the resulting bitrate variation would be quite significant. As discussed in Section 3.6.1, such variation could degrade the reconstructed video quality significantly in the presented low-delay system design. One way to alleviate this problem is to configure the target source coding rate lower than the available bitrate, at the cost of not fully utilizing the transmission capacity. Another alternative would be to configure the hardware encoder at a higher bitrate than the target, and then reduce the bitrate to the target level by software "re-quantization" with an accurate rate control operating at the MB-level. This could be a

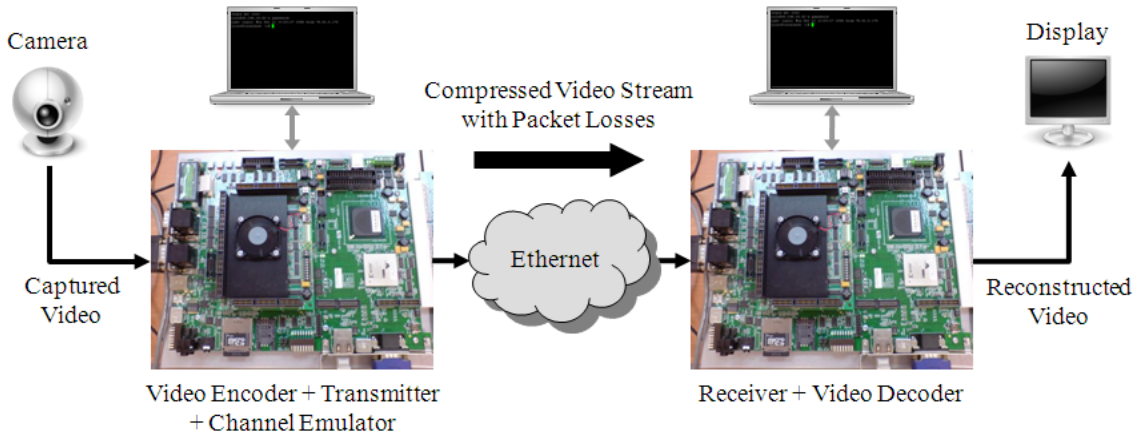


Figure 3.16: A real-time FPGA-based software/hardware testbed with a hardware video codec based on the proposed framework.

feasible way to increase the accuracy of the rate control if the entropy coding is done in the software or partial decoding/re-encoding is acceptable in terms of computational overhead.

3.8 Summary

In this chapter, a low-delay error-resilient video transmission framework for point-to-point wireless communication with instantaneous feedback is presented. The end-to-end delay is designed to be as small as possible, which is realized by transmitting each video frame within a given fixed-sized time slot. The available per-packet instantaneous feedback is integrated into the encoder for error-resilient video coding as well as into the transmitter for retransmission, both are designed under the low-delay constraint. The retransmission is integrated without introducing any additional delay, which is realized by dynamically allocating the fixed channel resource between video source coding and retransmission. The channel adaptability is further improved by adopting different retransmission schemes for different channel conditions, with a controlled impact on the end-to-end delay that is acceptable to the target video application. Experimental results have shown that the presented framework provides significantly improved video quality for a wide range of channel conditions and is highly adaptive to channel dynamics. In addition, the presented framework has been designed with very low complexity and high flexibility, and therefore has high practical significance.

Chapter 4

Perceptual Video Quality Modeling

In this chapter, the respective impact of spatial and temporal impairments[†] on the overall perceptual video quality and their interaction are investigated. Based on the quality evaluations from subjective tests, a full-reference objective video quality metric is developed, which captures the trade-off between the picture quality and the temporal resolution of a compressed video sequence. The proposed metric is based entirely on parameters that can be easily computed from the video, making it useful for dynamic adaptation in video communication systems, such as the system design presented in Chapter 3.

4.1 Introduction

Video transmission over wireless channels faces many challenges, such as limited transmission capacity, time-varying error-prone channel conditions, stringent delay requirements of video applications, and others. In order to improve the QoE of wireless video applications, video coding and scheduling parameters are often adapted to wireless channel characteristics (e.g., transmission data rate, packet error rate, delay jitter, etc.), where multiple video coding parameters, including the quantization parameter as well as the spatial and temporal resolution of the video, may be adjusted at the same time so as to achieve the best perceptual video quality. For example, for live video streaming applications (e.g., video surveillance), multi-dimensional rate control [SK01, RL02, LK05] involving multiple coding parameters as aforementioned may be performed by the video encoder to achieve improved video quality for a given encoding bitrate. Chapter 3 presents a video transmission system where the trade-off between picture quality and temporal resolution is exploited. For pre-encoded video

[†]In this chapter, unless otherwise stated, the term spatial impairment refers to the quantization, and spatial quality refers to the video quality subject to quantization only. The term temporal impairment refers to the frame rate reduction, and temporal quality refers to the perceived motion smoothness at a certain frame rate.

streaming applications (e.g., Video-on-Demand), transcoding [VCS03] may be necessary for bitrate reduction, where again the adjustment of multiple coding parameters may lead to a better video quality [JR04]. The emergence of the Scalable Video Coding (SVC) extension [SMW07, SSW07] of the H.264/AVC standard, which allows efficient scalability of quality, spatial and temporal resolution of a compressed video bitstream, further expands the scope of the applications of such multi-dimensionanl video adaptation.

While the involvement of multiple video coding parameters opens up additional possibilities for improving the video quality, it also introduces new challenges in making the right choices among these parameters that affect the video quality in different manners. In order to apply multi-dimensional video adaptation effectively, it is essential to understand the respective impact of each adjustable parameter on the overall perceived quality as well as their interactions, if any, with each other, and to be able to accurately estimate the resulting quality when the parameters are adjusted. Although many MDA schemes (e.g., [JR04, LK05]) adopt PSNR as the video quality metric and adjust the parameters so as to improve or optimize the PSNR, it has been shown that PSNR has poor correlation with subjective quality ratings in such context [FWSV07]. As a result, these adaptation schemes generally do not provide the best possible performance in terms of perceptual video quality. Therefore, developing an objective metric that can accurately estimate the perceptual quality when multiple video coding parameters are adjusted is crucial to the success of a MDA scheme, such as the one presented in Chapter 3. Based on this motivation, a full-reference objective video quality metric is developed in this chapter that considers both quantization (spatial quality) and frame rate (temporal quality), the two most often adjusted encoding parameters in video adaptation schemes. In addition, the metric is designed in such a way that the impact of other typical spatial quality impairments (e.g., error concealment, spatial resolution change) can be easily incorporated in a weighted average manner. The proposed metric has very high correlation with the subjective ratings, and contains only parameters that can be calculated from the video directly, including PSNR, frame rate as well as spatial and temporal activity measures. The high accuracy, the content-independency, as well as the low computational complexity of the metric make it highly applicable and preferable for dynamic adaptation and optimization in practical video transmission systems.

The rest of this chapter is organized as follows. Section 4.2 reviews the related work on video quality assessment involving quantization and frame rate reduction, respectively and jointly. Section 4.3 presents the subjective test and the test results. Section 4.4 analyzes the respective impact of quantization and frame rate reduction on the perceptual video quality as well as their interaction, and presents the proposed video quality metric. The performance of the proposed metric is evaluated and compared to other related metrics in Section 4.5. Section 4.6 summarizes the chapter.

4.2 Related Work

4.2.1 Spatial Quality Assessment

Most of the prior work on objective video quality assessment† has been concerned with the impact of quantization alone on the perceptual quality. It has been shown that even in this case, PSNR has poor correlation with the perceptual video quality. A great deal of effort has been made to develop better objective image and video quality metrics, among which some metrics share the same philosophy - to extend MSE/PSNR by taking into account the HVS properties (spatial and/or temporal), which MSE/PSNR alone is completely ignorant of. The motivation of developing new metrics based on MSE/PSNR is two-fold. On one hand, MSE/PSNR is easy to calculate, has clear physical meaning, is well understood for optimization purposes, and the video research community is very familiar with MSE/PSNR. On the other hand, MSE/PSNR does have higher correlation with the perceptual quality for one video content quantized at different levels (i.e., encoded at different bitrate). The general approach of these MSE/PSNR-based metrics is to incorporate certain perceptual weighting factors to the MSE/PSNR calculation that account for the HVS properties, so that the new metric would have better accuracy for quality estimation across video contents. Several weighted-MSE metrics are presented in [Mar86] for image quality assessment in general, considering luminance masking (weighted by local mean of the luminance level), texture masking (weighted by local standard deviation), as well as other HVS properties. In [TGP98], a spatial masking function based on local luminance gradients is proposed for video quality assessment to simulate the spatial masking effect. [OYL+05] presents a video quality metric based on the absolute pixel difference, where luminance masking, textural masking and temporal masking are estimated by using local mean luminance, local gradients and inter-frame difference, respectively, which are combined to yield a distortion visibility threshold. In comparison to the above metrics that mimic certain low-level HVS properties (visibility/sensibility), several metrics focus on high-level HVS properties (perception/attention). [TMJ00] discusses several energy measures and proposes an information measure of local interest for image quality assessment, where the MSE is weighted more strongly in areas of interest. The Edge-PSNR metric standardized in [ITU04b], motivated by the observation that the HVS is more attracted to edges, performs edge detection and computes the MSE only in the detected edge areas.

The weighting has also been applied on the sequence level instead of for each pixel. This sequence level weighting is easy to compute and circumvents the tricky problem of combining both low-level and high-level HVS properties by estimating the overall HVS behavior instead of treating each individually. [WP02] presents a PSNR-based metric which uses a logistic function to model the relationship between the average PSNR and the perceptual video quality,

†A review of objective video quality assessment and metric classifications can be found in Section 2.2.2.

but ignores the content-dependency of PSNR. [ODZ07] adopts a liner model and computes the model parameters for each video content by assuming that the perceptual qualities of a low quality and a high quality version of the same video content are available. The assumption is too strong for the metric to have any practical relevance. [BRK09] estimates the linear model parameters on the sequence level based on a spatial activity measure (average edge strength), but performs the weighting on the macroblock level. The temporal aspect of HVS properties is ignored in [BRK09]. The work presented in this chapter also follows the sequence level approach for modeling the spatial quality. The proposed model uses a logistic function to account for the non-linear relationship between perceptual quality and PSNR, and considers both spatial and temporal aspects of the HVS properties.

A new philosophy is followed in [WBSS04] for designing a full-reference engineering metric, which focuses on structural similarity instead of pixel-wise difference. [WBSS04] compares the mean, variance and covariance of small patches and combines them into a single quality metric SSIM. Similar to PSNR, weighted versions of SSIM have also been proposed to take into account the HVS properties. In [WLB04], a local luminance based weighting (assuming that dark regions do not attract attention) as well as a motion vector magnitude based weighting (for temporal masking) are performed. [Sha06], in the same spirit of [TMJ00], proposes to use energy/information measures of local interest as the weighting factors to account for high-level HVS properties. [WL07] uses an information measure of motion to estimate local interest level, considering that moving objects attract attention.

4.2.2 Temporal Quality Assessment

All of the above metrics assume a fixed frame rate of the processed video. When frame rate reduction is involved, those metrics are often calculated by comparing the temporally upsampled (e.g., by repetition) version of the processed video with the original video, in which case the quality estimation performance is much lower than in the full frame rate case. Therefore, the impact of frame rate reduction on the perceptual quality needs to be measured differently.

Several studies have investigated how video quality is affected by frame rate reduction. It has been shown that the temporal perceptual quality decreases non-linearly with frame rate reduction [HTG08], and the impact is content-dependent [WSV+03, HTG08]: in general, high motion video content is more negatively affected by the frame rate reduction than low motion content. As a result, the temporal quality is often modeled as a non-linear function of the frame rate (reduction) and parameters that measure the degree of motion in the video content [LLS+07b, YGEMD07, HTG08, OLZ+08, ZCL+08]. [LLS+07b] proposes to model the non-linearity based on a logarithmic function, and uses the average of the maximal motion vector magnitude of each frame (obtained via a optical flow algorithm) as the motion

measure. A power function and a exponential based function is adopted in [YGEMD07] and [HTG08], respectively, but no motion measure is proposed in either work. [OLZ+08] adopts an exponential decay function and suggests to use a combination of normalized frame difference and the average of the top 10% motion vector magnitude (obtained via full-search block matching) as the motion measure. [ZCL+08] proposes a jerkiness measure which is the product of the inverse of frame rate and the average frame difference.

4.2.3 Spatio-Temporal Quality Assessment

Only a few prior work considers the presence of both spatial and temporal impairment, and all of them model the overall quality as a linear combination of spatial and temporal quality, implying that the impact of each type of impairment is independent from each other. [VW93] linearly combines a spatial and a temporal quality impairment measure to estimate the overall quality. [FWSV07] focuses on the fact that the average PSNR of all frames (including the repeated frames in case of frame rate reduction) underestimates the perceptual quality at reduced frame rate, and proposes to add a compensation term based on the frame rate. A motion measure, defined as the average of the top 25% motion vector magnitude, is used as the weighting factor between PSNR and frame rate. The proposed metric is simple to calculate, but inherits the content-dependency of the PSNR and therefore can not provide accurate quality estimate across different video contents. In addition, the two terms do not have clear physical meanings. The PSNR term measures the effect of both spatial and temporal impairment (therefore referred to as STPSNR hereafter), while the frame rate term is designed empirically to compensate for STPSNR's underestimation of the overall quality. Although the frame rate term is a linear function of frame rate reduction, the true non-linear relationship is actually hidden in the STPSNR term. Another drawback of this metric is that a measure based on the motion vector amplitude may not be an accurate or reliable estimate of the motion level of the video, and is highly dependent on the motion estimation scheme adopted in the video codec (see Section 4.4.2.2 for illustrative examples and more details). [JKSR07] extends the model in [FWSV07] by using the standard deviation of the motion vector magnitude instead of the mean value as the motion measure.

However, the underlying assumption that the respective impact of spatial and temporal impairment are independent may not be valid. The results reported in [MSM04] and [SHH+08] indicate that when the spatial quality is low, changing the frame rate does not affect the overall quality much even with high motion video content, which seems to contradict with aforementioned findings with videos subject to frame rate reduction only. This discrepancy suggests that an interaction may exist between the impacts of the two impairment types, i.e., the impact of temporal impairment may depend on the level of the spatial impairment and vice versa. In this chapter, the existence of the interaction is confirmed by graphical analy-

sis as well as formal statistical analysis (see Section 4.4.2) of the subjective quality ratings from a carefully designed subjective test (see Section 4.3). As a result, the overall quality is modeled as the product (instead of a linear combination) of two terms. The first term is a logistic function of the average PSNR over non-repeated frames (SPSNR) that estimates the spatial quality only. The logistic function is used to model the non-linear relationship between SPSNR and perceptual quality, with two content activity measures (one spatial and one temporal) accounting for the HVS properties, both spatially and temporally, on the sequence level. The second term estimates the temporal quality and uses a non-linear function to model the relationship between frame rate reduction and perceptual quality, with a temporal activity measure (the same as the one used for spatial quality modeling) representing the content dependency here. Therefore, the proposed metric resolves the non-linearity and content-dependency of the respective impact of spatial and temporal impairments and takes into account the interaction between the two types of impairments, leading to a very high correlation with the subjective ratings for a wide range of video contents and quality levels. The computational complexity of the metric is still kept low, making the metric highly relevant for practical applications.

In parallel to the work presented in this chapter, two other PSNR-based metrics were developed for perceptual video quality estimation in the presence of both quantization and frame rate reduction. [SYN+10] extends the metric in [FWSV07] by taking into account the content-dependency of the STPSNR, but still shares the other drawbacks, i.e., ignoring the interaction, having no clear physical meanings and using a motion vector magnitude measure. [OMW09] presents a metric consisting of SPSNR and frame rate reduction. The impacts of the two impairment types are assumed to be independent from each other, but the metric is designed as the product of the two terms, implying a dependency. There are two content-dependent parameters in this metric that still need to be determined for each video individually. [OMW11] resolves the content-dependency of the metric in [OMW09] by estimating the content-dependent parameters from several spatial and temporal activity measures, which results in a metric that is similar to the metric proposed in this chapter. However, the metric in [OMW11] contains four content activity measures (including a motion vector based measure) with high computational complexity, limiting its usability in practical applications. The included motion vector based measure (retrieved from the video encoder directly) also renders the metric dependent on encoder's configurations (e.g., motion estimation scheme, encoding bitrate, mode decision, etc.). In comparison, the proposed metric has only two standard content activity measures, which can be easily computed from the source video and independent of encoder's configuration. In addition, the interaction of the two impairment types is confirmed in this work by graphical and formal statistical analysis of the subjective data; modeling the overall quality as the product of the two terms is well founded.

(a) Mother&Daughter

(b) Foreman

(c) Football

Figure 4.1: Sample images of the source video sequences used to generate the test videos in the subjective test.

4.3 Subjective Quality Assessment

In order to understand how quantization and frame rate reduction affect the perceptual video quality, a subjective test is carried out to collect subjective quality ratings. The subjective test is designed in such a way that the obtained subjective ratings allow formal statistical analysis of the impact of video content, spatial quality (SPSNR) and temporal quality (frame rate) as well as their interactions. The subjective ratings are also used for designing and evaluating the proposed objective quality metric.

4.3.1 Test Settings

Several important settings of the subjective test are reported in this section, including the test material, the test subject, the test method as well as the test procedure.

Test Material

Three well-known source video sequences (SRC) with a wide range of spatial and temporal content characteristics are used to generate the test videos: Mother&Daughter (MD), Foreman (FM) and Football (FB). Representive sample images of the three SRCs are shown in Figure 4.1. The SRCs are 10 seconds long in CIF (352x288) resolution and have an original frame rate (FR) of 30fps. Each SRC is temporally downsampled to 15, 10 and 7.5fps to generate four different temporal quality levels. Then for each temporal quality level, the videos are encoded and decoded using an MPEG-4 video codec (Xvid [Xvi]) to generate three different spatial quality levels, i.e., SPSNR at about 38dB, 34dB and 31dB. The videos are encoded in the IPPP...P structure with a constant quantization parameter that results in one of the aforementioned spatial quality levels. The combination of four temporal and three spatial quality levels results in 12 processed video sequences (PVS) for each SRC, 36 in total for the three SRCs, covering a wide range of perceptual quality levels. After the encoding/decoding,

frame repetition is performed on the videos with reduced frame rate so that each PVS has the same duration. The first two seconds of each PVS are removed to avoid the potential quality instability in the video frames close to the first I-frame. This is also done to the SRCs to match the video duration. The resulting 8-second videos are stored for the subjective test and the quality modeling. The total number of test videos are selected so that the duration of the entire test (including a introduction and a trial test) would be about 30 minutes.

Test Subject

A total of 27 non-expert subjects participated in the subjective test. The subjects are university students between the age of 20 and 26, including both males and females, with a male majority. All subjects reported to have normal or correct-to-normal visual acuity and normal color vision. Most of the subjects had never participated in a subjective test before; none of them had participated in a subjective test in the previous 12 months. The number of subjects is in line with the requirement of ITU-R recommendation BT.1788 [ITU07] and is similar to the number of subjects participated in the VQEG multimedia test [VQE08a].

Test Method

The SAMVIQ method [ITU07], which is specifically designed for subjective quality assessment of multimedia contents, is adopted in this work to collect subjective ratings for the test videos. It has been shown that SAMVIQ, with its interactive interface and review/compare ability, can provide more accurate and reliable subjective data than conventional test methods such as DSCQS (adopted in [SYN+10]) and ACR (adopted in [OMW11]), especially when it is difficult to rate or differentiate the test videos in terms of perceived quality (see Section 2.2.2 for details), which is indeed the case when both spatial and temporal impairments are involved. As to this work, preliminary tests (as well as the real tests) indicate that the test subjects find it difficult to give appropriate ratings to videos with different dominant types of impairments or differentiate them without comparing them to each other. Therefore, SAMVIQ should provide the most accurate subjective ratings in the context of this work.

Generally, SAMVIQ is a single stimulus method with random access where an explicit reference is always accessible. A graphical user interface (the central part shown in Figure 4.2) is developed for the subjective test implementing the SAMVIQ method. The video is displayed at the original resolution at the center of the screen on a background with mid-level grey color. The test is carried out scene by scene[†]; each scene corresponds to a video content. For each scene, the test subject can use the 14 access buttons placed under the video window to access the test videos. The unprocessed source video is included as an explicit reference and is always accessible through the "Ref" button. An implicit reference and the 12 processed

[†]A general example of the test organization for the SAMVIQ method can be found in Figure 2.6(a).

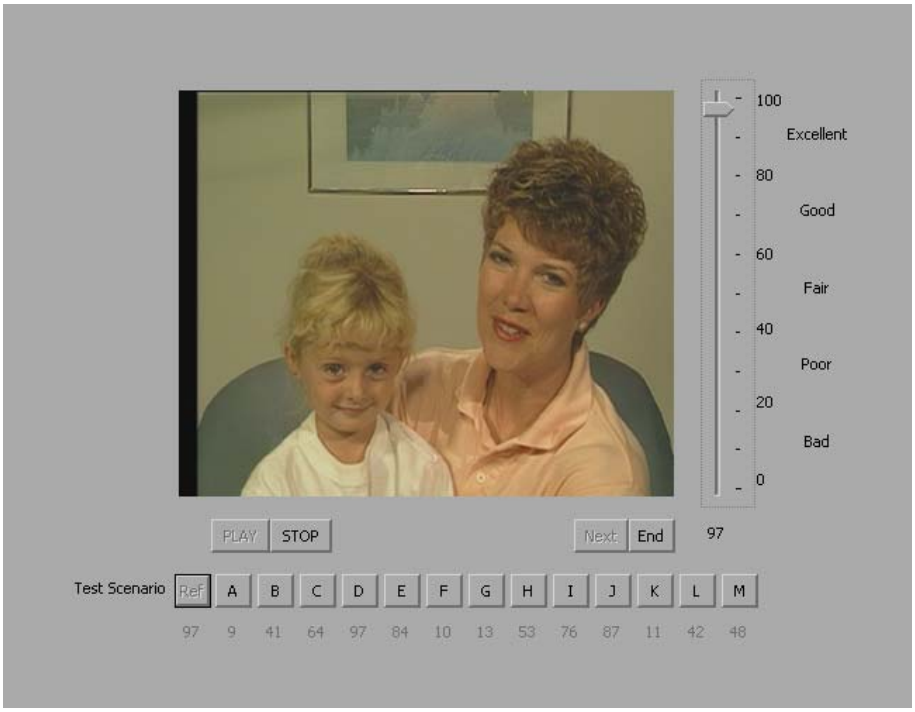


Figure 4.2: The graphical user interface implementing the SAMVIQ method.

videos are associated with the "A" to "M" buttons in a randomized order from scene to scene and for each test subject. This randomized order is to prevent the test subject to vote according to an established order and to reduce the "contextual effect", and the inclusion of the reference videos is to obtain more reliable test results. In this implementation (in comparison to the reference implementation in [ITU07]), when a button is selected, the associated video is played automatically, which makes the interface more user-friendly (especially when comparing videos) and reduces the overall test time. Therefore, the "PLAY" button is only for indicating the playing status. The "STOP" button may be used to stop the playing if desired. The "NEXT" button allows the subject to proceed to the next scene, but only when all the test videos of the current scene are rated. The "End" button ends the subjective test, but only when all the test videos are rated. The slider on the right of the video window is used to rate the video quality on a continuous quality scale graded from 0 to 100. The quality scale is also divided into five equal intervals and annotated by five adjectival quality terms (Excellent, Good, Fair, Poor, Bad) for general guidance.

The test subject may play and rate any video in any order and for multiple times. When a video is accessed for the first time, it has to be viewed entirely; all the control items are disabled during the viewing. After that, the video can be rated using the slider and the rating is shown under the corresponding access button. The subject may also review the video, compare it against the reference video and other test videos that have already been rated, and adjust the ratings as appropriate, during which the videos may be started or stopped

immediately. Once all the videos of the current scene are rated, the test can proceed to the next scene. The test ends when all the scenes are tested.

Test Procedure

The subjective test was carried out in a carefully controlled room with viewing conditions conforming to the general guidelines given in [ITU07]. A group of 5 or 6 test subjects was tested in one test session, each sitting in front of an identically configured LCD display. The configuration of the LCD displays is summarized in Table 4.1. Before the test, the subjects were requested to report whether they had normal or correct-to-normal visual acuity and normal color vision, as well as whether and when they had participated in a subjective test before. Non-suitable subjects were excluded from the test.

A test session consists of three phases: the introduction, the training session and finally the real test. During the introduction phase, written instructions were provided to the test subjects, carefully introducing the types of impairment likely to occur in the test, the test organization and the user interface. The subjects were informed that both spatial and temporal impairments may occur, and they should judge the overall quality and rate accordingly. Questions from the subjects were allowed during the introduction phase.

After the introduction, a training session is provided to the test subjects for them to learn how to use the user interface, as well as to get familiar with the types of artifacts and quality ranges that would likely to occur in the real test. The Carphone source video sequence, with a number of representative test conditions, is used in the training session. First, the test procedure is demonstrated to the test subjects, during which the subjects were shown how to access videos, how to use the rating scale, how to review and compare, etc. The subjects were encouraged to use the full range of the rating scale, but as they find appropriate. After the demonstration, the subjects were required to finish a mock test, during which the subjects were instructed to adjust the chair and the viewing distance according to their preference. Questions were allowed during the mock test and the test administrator also observed whether there was any unexpected behavior. The real test took place directly after all the subjects finished the mock test. The average duration of the real test was 16 (range 10-24) minutes, with a standard deviation of 4 minutes. No question was allowed during the real test.

Table 4.1: Configuration of the displays used in the subjective test

Parameter	*Specification*
Type of display	LCD
Display size	17 inch
Manufacturer	FUJITSU SIEMENS
Model	SCENICVIEW B17-2 CI

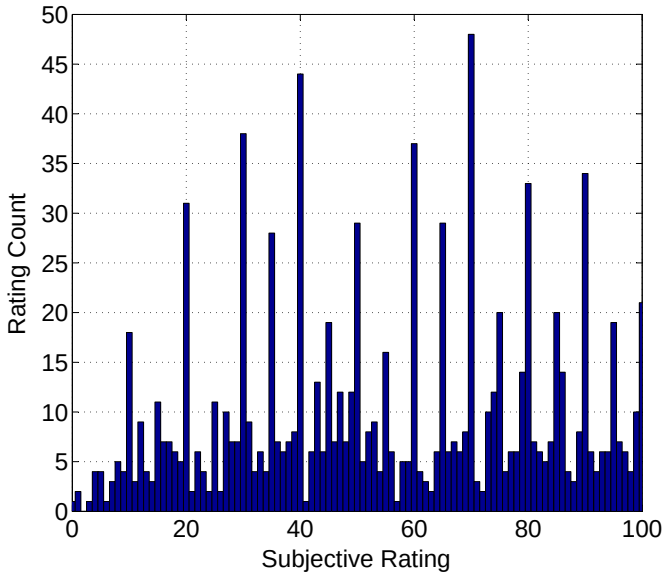


Figure 4.3: Histogram of the valid raw subjective ratings collected from the subjective test.

4.3.2 Subjective Data

After the subjective ratings are collected from the test, a screening process is carried out to ensure that only ratings from subjects who have rated in a stable and coherent manner are used for further analysis; subjects who may have rated randomly are rejected. Typically, the rejection is done by verifying the level of consistency of the ratings of one subject according to the mean ratings of all subjects. In this work, the screening procedure recommended by ITU-R BT.1788 [ITU07] for the SAMVIQ method is adopted to screen the collected data.

The adopted screening procedure first calculates the Pearson correlation coefficient r_p and the Spearman's rank correlation coefficient r_s for the ratings of each subject against the mean ratings of all subjects. Then for each subject, the following rejection criteria is applied:

IF $r(\text{subject}) \leq RejectionThreshold$, **THEN** the subject is rejected,

where $r = min(r_p, r_s)$ and the $RejectionThreshold$ is 0.85 for the collected data.

Subjective ratings of 25 subjects are verified to be valid in the screening process; 2 subjects are rejected. The histogram of the valid ratings (explicit reference excluded) is shown in Figure 4.3, which demonstrates that the subjective ratings span the entire quality range and the distribution over different quality ranges (i.e., "Bad", "Poor", etc.) is quite uniform. This verifies that the subjects were indeed using the full range of the scale for rating the quality and the quality levels of the test videos were well selected. Notice also how subjects tend to quantize the ratings to the ticks on the scale and the mid-points between the ticks (i.e., ratings like 5, 10, 15, etc.). 54% of the total ratings are at those positions.

The valid raw ratings are further processed to form a differential mean opinion score

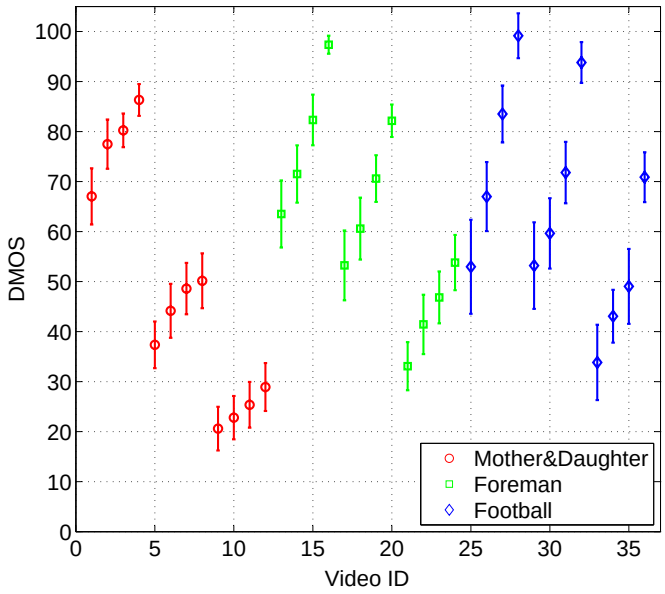


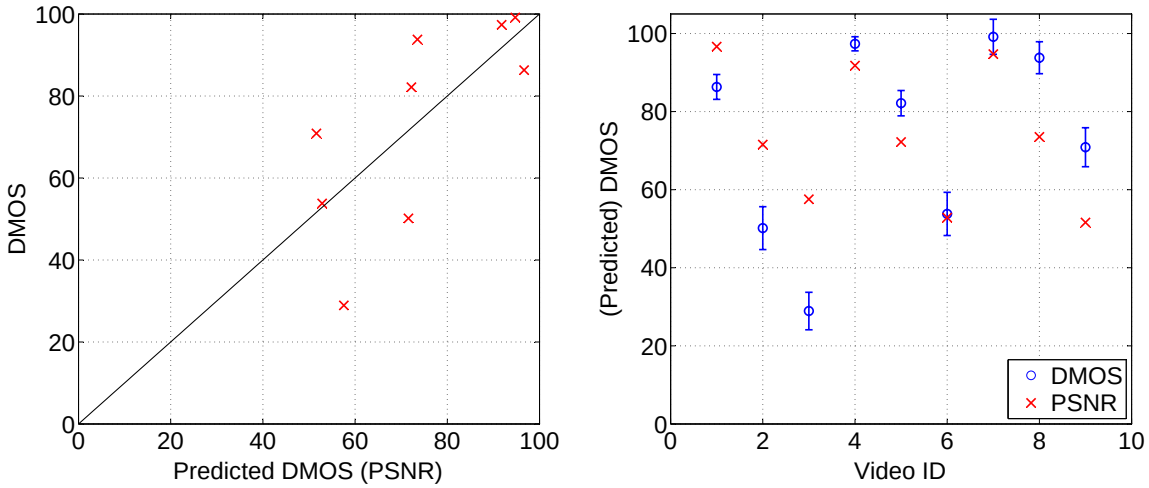
Figure 4.4: DMOS values of all test videos. The vertical bar indicates the corresponding 95% confidence interval.

(DMOS), which is the average of the difference between the ratings of a processed video and those of the corresponding hidden reference. Specifically, let s_i^j denote the rating given by subject j to the processed video i, and s_{ref}^j denote the rating given by the same subject to the corresponding hidden reference. Then the DMOS value of video i is calculated as

$$DMOS_i = \frac{1}{N_s}\sum_{j=1}^{N_s}(s_i^j - s_{ref}^j + 100), \tag{4.1}$$

where N_s is the number of subjects. By subtracting the rating of the reference video, DMOS removes the bias in the quality ratings caused by individual's preference of the video content. This process is also referred to as hidden reference removal.

DMOS is used as the subjective quality measure for each processed video. Since each raw rating is in the range [0, 100], DMOS is in the range [0, 200], with higher DMOS values indicating better quality. DMOS values greater than 100 (indicating better quality than the reference) are considered valid and included in the data analysis. The DMOS values of all test videos, along with the corresponding 95% confidence interval (calculated using the Student's t-distribution), are presented in Figure 4.4. Notice that the 95% confidence intervals are generally larger for video contents with higher motion, especially for videos with reduced frame rate, which indicates that the subjects are less consistent when rating high motion video contents with reduced frame rate. The average of the 95% confidence intervals for Mother&Daughter (low motion), Foreman (medium motion), and Football (high motion) are ± 4.7, ± 5.2 and ± 6.5, respectively.



(a) Correlation between PSNR and DMOS

(b) Accuracy of PSNR's prediction

Figure 4.5: Performance of PSNR for predicting DMOS. The vertical bar indicates the corresponding 95% confidence interval of each DMOS.

4.4 Spatio-Temporal Perceptual Quality Modeling

The modeling of the perceptual video quality in the presence of both spatial and temporal impairment is carried out in two steps. First, the impact of the spatial impairment (i.e., quantization) alone is analyzed and a spatial quality model is developed, which is presented in Section 4.4.1. Then, in Section 4.4.2, the impact of the temporal impairment (i.e., frame rate reduction) at different spatial quality levels is studied and a temporal quality model as well as an overall quality model are developed.

4.4.1 Spatial Quality Analysis and Modeling

4.4.1.1 Spatial Quality Analysis

The most popular objective quality metric for measuring the spatial quality is the average PSNR defined in Equation (2.5). But as discussed in Section 2.2.2, PSNR ignores the HVS properties and has poor correlation with the perceptual quality, which is confirmed by the subjective data collected from this work as shown in Figure 4.5. Figure 4.5(a) plots the DMOS against the predicted DMOS from PSNR (by a least-squares linear fitting to the DMOS values) for all videos with full frame rate, where a low correlation between DMOS and PSNR is evident. The accuracy (or the lack of it) of PSNR's prediction is depicted in Figure 4.5(b).

Considering the merits of PSNR and the potential to improve its performance, this work extends PSNR to form a spatial quality model, which models the overall HVS behavior on the sequence level (refer to Section 4.2.1). There are two major problems with PSNR as a

spatial quality metric. First, PSNR is content-dependent, which means that similar PSNR values may indicate significantly different subjective quality levels for different video contents, as can be seen in Figure 4.5(a). In addition, the relation between PSNR and the subjective quality is not linear, as shown in [WP02]. Therefore, in oder to improve the performance, content features should be included to reduce the content-dependency of the new model and a non-linear function should be considered.

A two-way ANOVA with repeated measures is performed on the subjective data to show how the subjective quality relates to different factors (i.e., *PSNR* and *Content* here). The results of the ANOVA test[†] are reported in Table 4.2, which show that both *PSNR* and *Content* have significant impact ($p < 0.0001$) on the subjective quality. This confirms that including content features in the model is necessary. The interaction between the two factors is also found to be significant ($p < 0.0001$), indicating that how PSNR impacts the subjective quality is content-dependent. The significant interaction term suggests that the model should not simply be a linear combination of PSNR and content features.

Table 4.2: Two-way ANOVA results for the full frame rate videos

Source of Variation	*df*	*Mean Square*	*F-value*	*p-value*
PSNR	2	34971	364.1	<0.0001
Content	2	21160	220.3	<0.0001
PSNR : Content	4	1973	20.5	<0.0001

4.4.1.2 Spatial Quality Modeling

To model the spatial quality based on PSNR, two issues need to be addressed. First, the form of the non-linear relation between DMOS and PSNR needs to be determined. Following the recommendation in [WP02], the relation is modeled as a form of the logistic function. Also taking into consideration of the impact of content on the subjective quality as well as the interaction between PSNR and content, the spatial video quality is modeled as

$$SVQM = \frac{100}{1 + e^{-(PSNR + w \cdot VC_s - \mu)/s}}, \tag{4.2}$$

where VC_s is a model parameter related to the video content and w is a weighting factor between $PSNR$ and VC. μ and s are model coefficients that can be determined by fitting to the subjective data.

The second issue is to find suitable content features to include into the model, i.e., to determine VC in Equation (4.2). As discussed in Section 4.2.1, both spatial and temporal aspects of the HVS properties have impacts on the perceptual quality, suggesting that both

[†]The ANOVA results in this dissertation are generated using R version 2.12.1 [Com].

Table 4.3: SA and TA values of the source videos

Source Video	*SA*	*TA*
Mother&Daughter	54.6	3.8
Foreman	85.9	15.2
Football	65.6	27.6

spatial and temporal features of the video content should be considered. Therefore, VC is designed as a weighted sum of two parameters:

$$VC_S = SA + k \cdot TA, \tag{4.3}$$

where SA and TA represent the spatial and temporal activity level of the video content, respectively. The spatial and temporal perceptual information measures recommended by ITU [ITU99] for selecting source videos for a subjective test (see Section 2.2.1) are adopted here, slightly modified, as the TA and SA, which are defined as

$$SA = mean_{time}\{std_{space}[Sobel(F_n)]\}, \tag{4.4}$$

$$TA = mean_{time}\{std_{space}[F_n - F_{n-1}]\}. \tag{4.5}$$

The calculation of SA and TA is based on the luminance component of each video frame. For calculating SA, each frame (F_n) is first processed by the Sobel operator ($Sobel()$), which computes the gradient at each pixel. The standard deviation of the gradient magnitudes over the pixels ($std_{space}()$) is then computed. It is averaged over all frames ($mean_{time}()$) to generate a single-valued SA. Larger SA value indicates higher spatial activity level in the video content. Similarly, TA is computed based on the standard deviation of the difference between consecutive frames, with higher TA value indicating higher temporal activity level. The SA and TA values of the source videos used in this work are given in Table 4.3, which correspond well to the perceived spatial and temporal activity level in these videos.

Finally, substituting Equation (4.3) into Equation (4.2), the spatial video quality model (SVQM) is given by

$$SVQM = \frac{100}{1 + e^{-(PSNR + w_s \cdot SA + w_t \cdot TA - \mu)/s}}, \tag{4.6}$$

where w_s and w_t are weighting factors for SA and TA, respectively. The values of the constants w_s, w_t, μ and s are determined by a non-linear least-squares fitting using the subjective data, which leads to $w_s = 0.0356$, $w_t = 0.236$, $\mu = 36.9$ and $s = 2.59$. SVQM has the range of 0 to 100, higher value indicating better quality. The positive values of w_s and w_t indicate that at the same PSNR level, contents with higher spatial and temporal activity levels have better perceived quality. DMOS and SVQM are plotted against PSNR for the full frame rate videos in Figure 4.6, which shows that SVQM fits the subjective data (DMOS) very well. The non-linearity and the content-dependency are well resolved by the logistic function and the inclusion of content features (i.e., SA and TA), respectively.

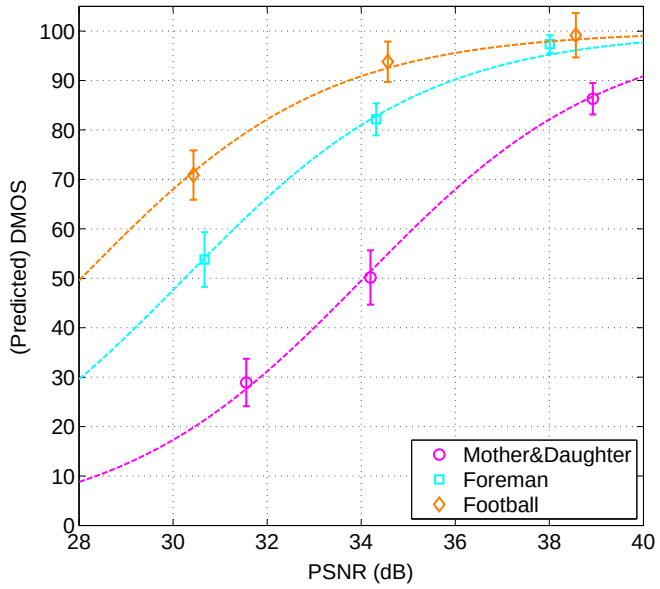


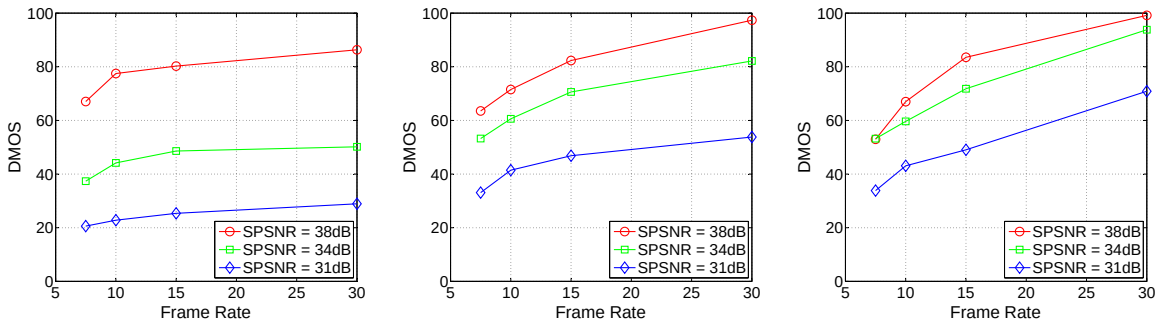
Figure 4.6: DMOS (points) and SVQM (curves) versus PSNR for the full frame rate videos. The vertical bar indicates the corresponding 95% confidence interval of each DMOS. Notice that the non-linearity and the content-dependency are well resolved in the SVQM model.

4.4.2 Temporal Quality Analysis and Overall Quality Modeling

4.4.2.1 Temporal Quality Analysis

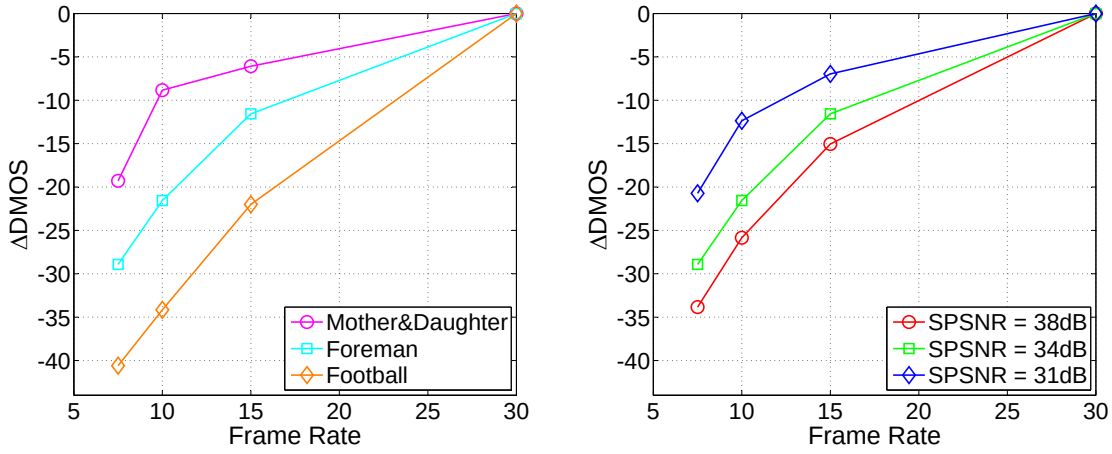
To understand how temporal impairment (i.e., frame rate reduction) affects the overall perceptual quality, three questions need to be addressed. First, what is the relationship between perceptual quality and frame rate? Is it linear or non-linear? Second, is the relationship content-dependent, i.e., different for different video contents? Third, is the relationship dependent on the spatial quality level (i.e., is there an interaction between spatial quality and temporal quality perception)? All three questions are addressed in this section.

In Figure 4.7, the DMOS values are plotted against the corresponding frame rate at different spatial quality (SPSNR) levels for each source video separately. Recall that while SPSNR is determined solely by the spatial impairment (i.e., quantization) level, similar SPSNR values may represent different spatial quality levels for different video contents. As expected, DMOS becomes lower as frame rate decreases. Generally, the relationship appears to be non-linear, which is in accordance to the results obtained by studies on uncompressed videos (see Section 4.2.2). Figure 4.8(a) compares the curves at a similar spatial quality level for different source videos (i.e., SPSNR = 38dB for Mother&Daughter, SPSNR = 34dB for Foreman and Football), where the DMOS reduction (ΔDMOS) compared to the corresponding full frame rate video is plotted. The comparison shows that DMOS decreases slower for low-motion video content like Mother&Daughter and faster for high-motion content like Football, indicating that the impact of frame rate is content-dependent and reducing frame rate has



(a) Mother&Daughter (b) Foreman (c) Football

Figure 4.7: DMOS versus frame rate at different SPSNR levels for each source video.



(a) For different source videos. (b) At different spatial quality levels (Foreman).

Figure 4.8: ΔDMOS versus frame rate a) for different source videos at a similar spatial quality level and b) for a specific source video (Foreman) at different spatial quality levels.

stronger negative impact for videos with higher temporal activity. This suggests that a temporal activity measure of the video content should be considered when modeling the temporal perceptual quality. Also, as depicted in Figure 4.8(b), the curves for a specific source video have different slops for different spatial quality levels; DMOS decreases faster at higher spatial quality level. This observation indicates that the negative impact of frame rate reduction is more perceivable or annoying at high spatial quality levels than at low spatial quality levels. Therefore, the impact of frame rate is also dependent on the spatial quality level (i.e., there is an interaction between spatial and temporal quality perception).

The above observations are confirmed by a three-way ANOVA test with repeated measures, which considers three factors that may have impact on the overall perceptual quality: *SPSNR*, *Frame Rate (FR)* and *Content*. The results of the ANOVA test are reported in Table 4.4. As expected, the main effects of the three factors are significant ($p < 0.0001$). The two-way interaction between *SPSNR* and *Content* is significant ($p < 0.0001$), in accordance to the

findings in Section 4.4.1.1 for spatial quality modeling. Both the interaction between FR and *Content* ($p < 0.0001$) and that between *SPSNR* and FR ($p = 0.0016$) are significant, which confirms that the impact of frame rate on the overall perceptual quality depends on the video content as well as on the spatial quality level.

Table 4.4: Three-way ANOVA results for all videos

Source of Variation	*df*	*Mean Square*	*F-value*	*p-value*
SPSNR	2	110094	783.0	<0.0001
Frame Rate (FR)	3	30945	220.1	<0.0001
Content	2	22282	158.5	<0.0001
SPSNR : Content	4	6444	45.8	<0.0001
FR : Content	6	2892	20.6	<0.0001
SPSNR : FR	6	505	3.6	0.0016
SPSNR : Content : FR	12	114	0.8	0.6367

4.4.2.2 Overall Quality Modeling

Based on the observation that an interaction exists between spatial quality and temporal quality perception, the overall quality is modeled as the product of the two quality terms:

$$STVQM = SVQM \cdot TVQM. \tag{4.7}$$

$SVQM$ models the spatial quality as given in Equation (4.6), where the $PSNR$ is averaged over non-repeated video frames for videos with reduced frame rate, i.e., SPSNR. $TVQM$ models the temporal quality and is given in the following.

Similar to the spatial quality modeling, modeling the temporal quality requires to resolve the non-linear relationship between subjective quality and frame rate as well as the content-dependency. Since there is no evidence that spatial characteristics of the video content have impact on the temporal quality perception (i.e., motion smoothness), only a temporal activity measure (i.e., the TA defined in Equation (4.5)) is considered in modeling the temporal quality. The proposed temporal video quality model (TVQM) is given as

$$TVQM = \frac{1 + a \cdot TA^b}{1 + a \cdot TA^b \cdot \frac{30}{FR}}, \tag{4.8}$$

where FR denotes the frame rate and $FR \leq 30$fps. The two constants a and b are determined by fitting to the subjective data, which leads to $a = 0.028, b = 0.764$. TVQM has the range of 0 to 1, higher value indicating better temporal quality. When $FR = 30$ or $TA = 0$, TVQM reaches 1, which is in accordance to the fact that at full frame or for static scene (still image), the overall quality equals the spatial quality. At the same reduced frame rate level, TVQM is lower for larger TA, i.e., temporal quality is lower for video contents with higher temporal

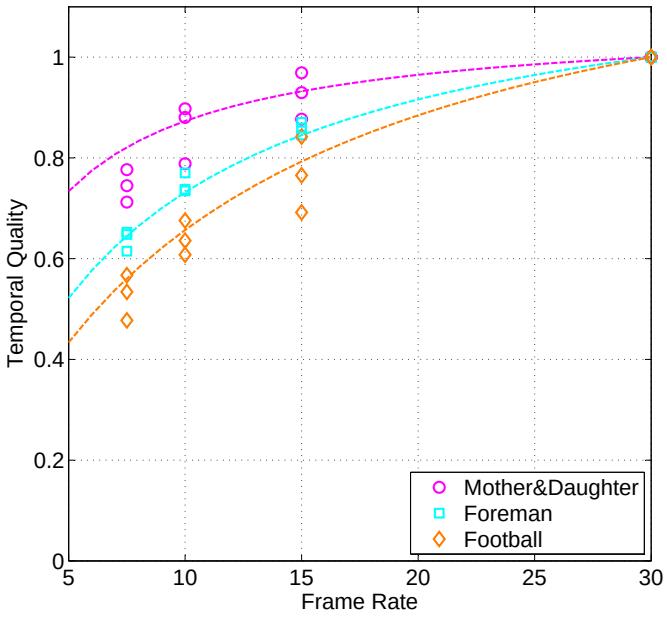


Figure 4.9: Subjective temporal quality (points) and TVQM (curves) versus frame rate for all test videos. For each source video and each frame rate, three subjective temporal quality points are available, which correspond to the three different spatial quality levels. Notice that the non-linearity and the content-dependency are well resolved in the TVQM model.

activity levels. The subjective temporal quality, which is calculated by dividing the DMOS value of the test video by that of the full frame rate video with the same spatial quality level, and TVQM are plotted against frame rate in Figure 4.9. TVQM fits the subjective data very well; both the non-linearity and the content-dependency are well resolved.

Many previous studies use temporal activity measures based on motion vectors, which are often obtained using block-matching methods that are adopted by video encoders for motion estimation. However, since a typical block-matching method does not intent to find the real motion level, the resulting motion vectors may not accurately or reliably reflect the temporal activity level of the video. As an example, a full-search full-pel block-matching is performed on the source videos used in this work. Sample images along with the resulting motion vectors of 8x8 blocks are shown in Figure 4.10. One issue that can be clearly seen here is that the block-matching algorithm generates large motion vectors for non-textured areas (such as the background wall in Mother&Daughter and the helmet in Foreman), which significantly undermine the ability of motion vectors for measuring the temporal activity level. The mean and standard deviation of the motion vector magnitudes (averaged over all frames) are summarized in Table 4.5, indicating that those measures do not accurately or reliably reflect the perceived temporal activity level. In general, motion vectors are highly dependent on the block-matching method, the search range, and if they are generated by a video encoder, on the quantization level. This indicates that if some of these parameters

(a) Mother&Daughter (b) Foreman (c) Football

Figure 4.10: Sample images and the corresponding motion vectors of 8x8 blocks (arrows). The motion vectors are obtained using a full-search full-pel block-matching algorithm with a search range of ±16. Note that the arrows are scaled for each video to fit within the display grid, so no comparison should be made between videos.

Table 4.5: Mean and standard deviation of motion vector magnitudes using the full-search full-pel block-matching method with different search ranges

Video	*Mean*			*STD*		
	±8	*±16*	*±32*	*±8*	*±16*	*±32*
Mother&Daughter	1.9	2.9	4.5	2.7	4.8	8.1
Foreman	2.9	3.9	5.4	2.2	3.9	7.0
Football	6.0	9.3	13.0	2.5	4.4	8.1

change, a quality model based on motion vector would have different performance and different model coefficients. Therefore, the TA defined in Equation (4.5), which is independent from motion vectors, is used in this work as the temporal activity measure, for both spatial and temporal quality modeling.

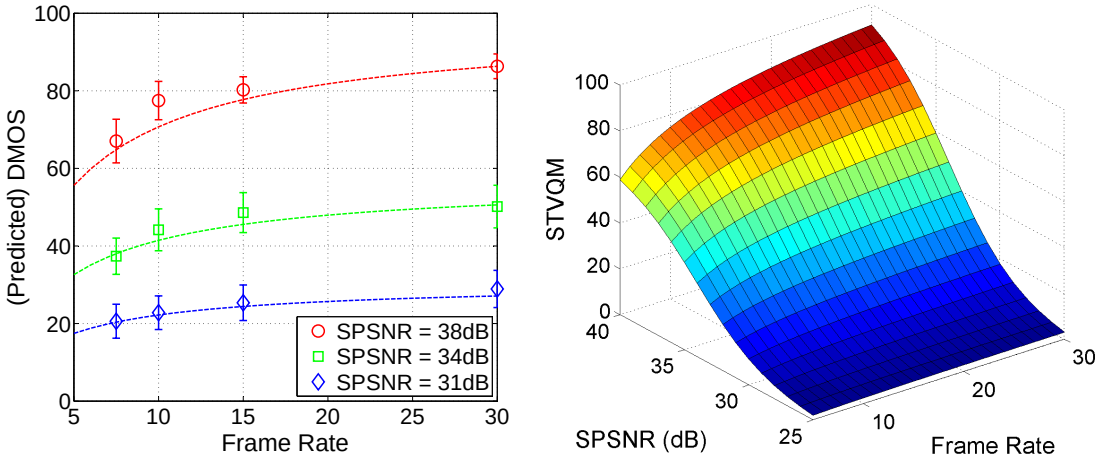
Combining Equation (4.6), (4.7) and (4.8) yields the spatio-temporal video quality model (STVQM), which is written as:

$$STVQM = \frac{100}{1+e^{-(SPSNR+w_s \cdot SA+w_t \cdot TA-\mu)/s}} \cdot \frac{1+a \cdot TA^b}{1+a \cdot TA^b \cdot \frac{30}{FR}}. \tag{4.9}$$

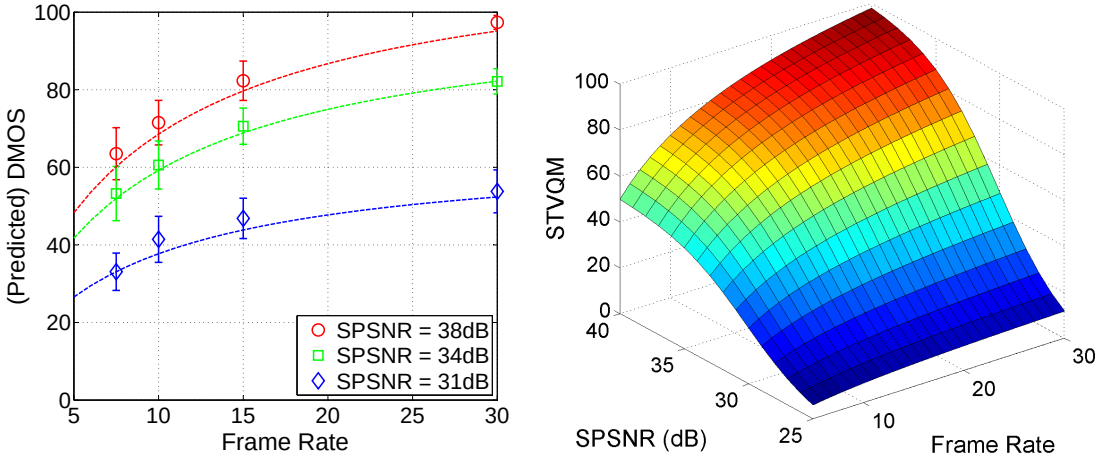
The STVQM model coefficients are summarized in Table 4.6. Figure 4.11, on the left side, depicts DMOS and STVQM versus frame rate at different SPSNR levels for each source video separately, showing that the curves generated from STVQM fit the DMOS values very well. On the right side, STVQM is depicted as a function of SPSNR and frame rate in a three-dimensional view.

Table 4.6: Summary of the STVQM model coefficients

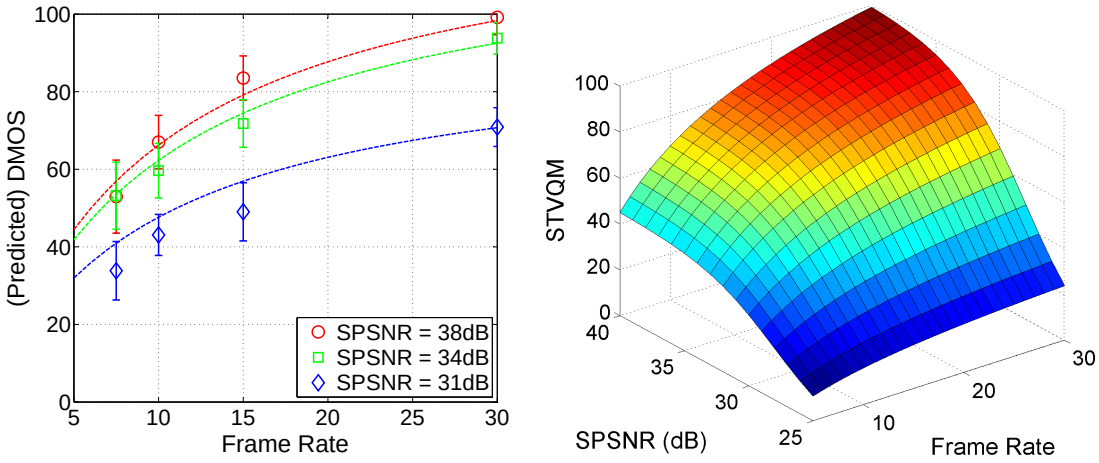
w_s	w_t	μ	s	a	b
0.0356	0.236	36.9	2.59	0.0280	0.764



(a) Mother&Daughter



(b) Foreman



(c) Football

Figure 4.11: The STVQM model. Left: DMOS (points) and STVQM (curves) versus frame rate at different SPSNR levels. The vertical bar indicates the corresponding 95% confidence interval of each DMOS. Right: STVQM model as a function of SPSNR and frame rate in a three-dimensional view.

4.5 Performance Evaluation

In this section, the spatial quality model SVQM and the spatial-temporal quality model STVQM are evaluated against several state-of-the-art video quality models for their performance in predicting the subjective quality ratings. The evaluation is carried out using three statistical metrics that characterize various aspects of a model's performance, including accuracy, monotonicity and consistency. Significance tests are performed for each evaluation metric to determine whether the differences between models are statistically significant.

4.5.1 Evaluation Metrics

The performance of an objective video quality model is evaluated by comparing its predictions with the corresponding subjective quality ratings (i.e., DMOS values). Following the performance evaluation process adopted by VQEG in its multimedia test [VQE08a], three statistical metrics: Pearson correlation coefficient (PCC), root mean square error (RMSE) and outlier ratio (OR), along with their 95% confidence intervals, are used to quantify the model performance in this work. The definition of the evaluation metrics is briefly reviewed in the following. The calculation of the 95% confidence interval and the details about the significance test for each metric can be found in [VQE08a].

The PCC measures the linearity between the model predictions and the subjective quality ratings. It is an intuitive indicator of a model's overall performance in various aspects and is given by

$$PCC = \frac{\sum_{i=1}^{N_v}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N_v}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{N_v}(y_i - \overline{y})^2}}, \tag{4.10}$$

where x_i and y_i denote the video quality model output for video i (VQM_i) and the corresponding subjective rating ($DMOS_i$), respectively. N_v is the number of videos considered in the analysis. The PCC value of a video quality model is normally between 0 and 1, and the closer the PCC is to 1, the better the model is. The exact goodness-of-fit is measured by the square of the PCC, typically denoted as R^2, which can be interpreted as the proportion of variance in the subjective ratings that is explained by the objective model.

The root mean square error measures the accuracy of a model (i.e., how close or far the model predictions are from the corresponding subjective ratings) and is defined as

$$RMSE = \sqrt{\frac{1}{N_v - d} \cdot \sum_{i=1}^{N_v}(VQM_i - DMOS_i)^2}, \tag{4.11}$$

where N_v is the number of videos considered in the analysis, and d is the number of model coefficients that need to be determined by fitting to the subjective ratings, which is included

to penalize for model complexity (i.e., overparameterization). Smaller RMSE values indicate better model accuracy.

The consistency of a model is measured by the outlier ratio. A model prediction is defined as an outlier if it significantly deviates from the corresponding subjective rating. Let I_{outl} denote the set of outliers, then

$$VQM_i \in I_{outl}, \textbf{IF } |VQM_i - DMOS_i| > k \cdot \frac{s_i(DMOS)}{\sqrt{N_s}}, \tag{4.12}$$

where $s_i(DMOS)$ represents the standard deviation of the individual subjective ratings for video i, and N_s is the number of subjects ($N_s = 25$). k is the critical value for the 95% confidence interval and $k = 2.064$ in this work. The outlier ratio is then defined as the ratio of outliers to all videos considered in the analysis:

$$OR = \frac{|I_{outl}|}{N_v}. \tag{4.13}$$

4.5.2 Spatial Quality Model Evaluation

The performance of SVQM is evaluated and compared with three other objective models: PSNR, SSIM [WBSS04] and a weighted MSE model presented in [BRK09] (referred to as WMSE hereafter) using the subjective ratings of the full frame rate test videos.

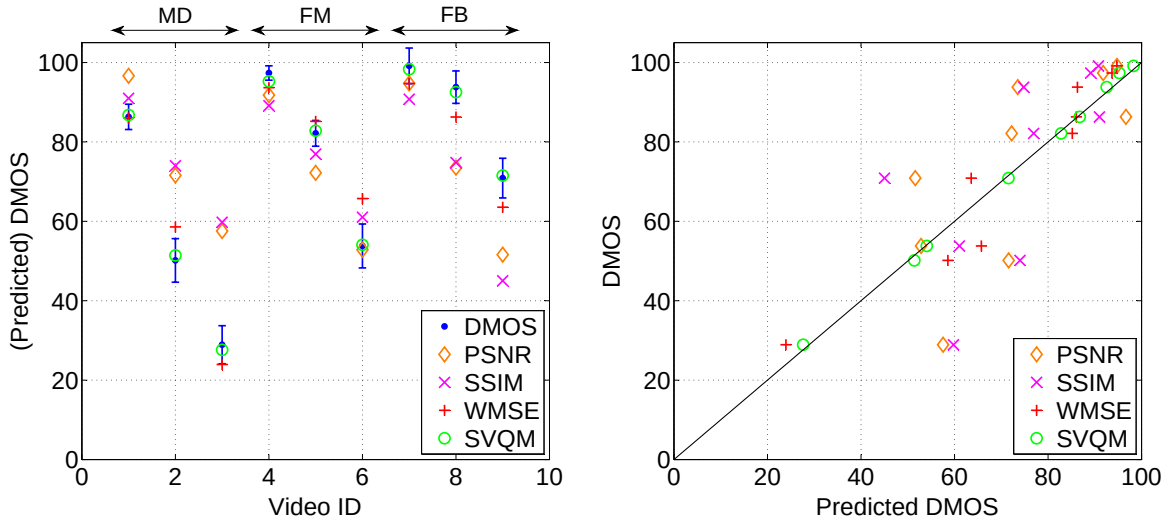
PSNR and SSIM are two very popular models for video quality assessment. While PSNR measures the pixel-wise difference, SSIM measures the structural similarity between small windows inside video frames. For each pair of windows, an SSIM index compares the local mean, variance and covariance, which can be written as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{4.14}$$

where x and y represents a local window in a test video and in its corresponding reference, respectively. C_1 and C_2 are constants. In this work, the local SSIM indices are computed by the MATLAB implementation available online at [Wan] using the default settings. A single SSIM measure is calculated for each test video by averaging the local SSIM indices over all windows within a frame and then over all frames. A VSSIM model is proposed in [WLB04] that includes local weighting factors in the averaging process to account for HVS properties, but the performance is similar to the non-weighted average. Therefore, for simplicity, the non-weighted average is used for comparison. For both PSNR and SSIM, the predicted DMOS values are obtained by fitting the model to the measured DMOS values of all considered videos using a linear function with two coefficients (i.e., the slope and the intercept), which are determined by least-squares fitting.

The WMSE model in [BRK09] is an extension to MSE/PSNR, which adopts sequence-level perceptual weighting to take HVS properties into account. WMSE is given by

$$WMSE = 100 \cdot (1 - \alpha_1 \cdot e^{\alpha_2 \cdot ES} \cdot MSE), \tag{4.15}$$



(a) Accuracy of model predictions (b) Correlation between model predictions and DMOS

Figure 4.12: Performance evaluation and comparison for the SVQM model. The vertical bar indicates the corresponding 95% confidence interval of each DMOS. Only the test videos with full frame rate are included in the analysis.

where MSE is the mean square error averaged over all frames, and ES measures the average edge strength based on Sobel filtered frames. The two model coefficients α_1 and α_2 are determined by a non-linear least-squares fitting to the measured DMOS values. In [BRK09], the model coefficients are determined on the sequence level, but the actual weighting is performed for each macroblock in order to incorporate the model into mode decision algorithms, which would compromise the performance of the model for predicting the overall quality. Therefore, in this work, both the determination of the model coefficients and the weighting process are carried out on the sequence level based on the subjective data, which should lead to the best possible performance of the WMSE model.

The predicted DMOS values from all four models, i.e., PSNR, SSIM, WMSE and SVQM, are compared with the measured DMOS values in Figure 4.12(a), which illustrates how close or far the model predictions are from the subjective ratings. It is clear that the predicted DMOS values from both PSNR and SSIM are very inaccurate due to the strong content-dependency of these models. WMSE considers the spatial masking effect of HVS by including a spatial content activity measure into the model, leading to a significant improvement compared to PSNR and SSIM. The proposed SVQM considers both spatial and temporal aspects of the HVS properties and as a result, the predicted DMOS values from SVQM are very close to the measured DMOS values. Figure 4.12(b) illustrates the linear correlation between the model predictions and the DMOS values for all four models, where it can be seen that the SVQM model predictions are more linearly correlated with the subjective ratings than the comparing models. The statistical evaluation metrics that quantify the performance of the models, along

Table 4.7: Pearson correlation coefficients of the spatial quality models

Model	*PCC*	*LB PCC*	*UB PCC*	*Sig. Level*	R^2
PSNR	0.719	-0.017	0.959	1.00	0.518
SSIM	0.648	-0.151	0.935	1.00	0.420
WMSE	0.959	0.767	0.993	0.99	0.920
SVQM	0.999	0.993	1.000	-	0.998

Table 4.8: RMSE values of the spatial quality models

Model	*RMSE*	*LB RMSE*	*UB RMSE*	*Sig. Level*
PSNR	18.20	12.03	37.04	1.00
SSIM	19.96	13.20	40.63	1.00
WMSE	7.47	4.94	15.21	1.00
SVQM	1.50	0.94	3.67	-

Table 4.9: Outlier ratios of the spatial quality models

Model	*OR*	*CI*	*Sig. Level*
PSNR	0.78	±0.32	0.98
SSIM	1.00	±0.00	1.00
WMSE	0.67	±0.37	0.96
SVQM	0.11	±0.24	-

with the corresponding 95% confidence intervals and significance test results, are summarized in Table 4.7, 4.8 and 4.9, where lower bound (LB) and upper bound (UB) represents the limits of the 95% confidence intervals (CI). The results show that in every aspect of the model performance, the proposed SVQM model provides better results than the comparing models with smaller confidence intervals. The statistical significance of the difference between SVQM and the comparing models is well above the typical 95% significance level for all three metrics.

4.5.3 Overall Quality Model Evaluation

The performance of the spatio-temporal quality model STVQM is evaluated and compared with four other objective models that consider both quantization and frame rate reduction: STPSNR, the QM model presented in [FWSV07], an STPSNR-based extension of the SVQM model (referred to as SVQM$^+$ hereafter) and the VQMTQ model presented in [OMW11]. The subjective ratings of all test videos are used for the performance evaluation.

Recall that STPSNR is calculated by averaging over all frames in the test videos, including the repeated frames for videos with reduced frame rate. With STPSNR, the impact of frame rate reduction is taken into consideration in the way that the PSNR values of the repeated frames are smaller than those of the non-repeated frames. This is the typical way of using PSNR for video quality assessment when frame rate reduction is involved. The predicted

DMOS values from STPSNR are obtained by a linear least-squares fitting to the measured DMOS values. Note that STSSIM, the spatio-temporal version of SSIM, is found to have similar performance as STPSNR here as in the spatial quality case. Therefore, for clarity results for STSSIM are not reported.

The QM model in [FWSV07] extends STPSNR by adding a temporal compensation term based on the frame rate reduction. A motion vector magnitude measure is included in the temporal compensation term, considering that the frame rate reduction has different impact on videos with different temporal activity levels. The QM model is given by

$$QM = STPSNR + \beta_1 \cdot MA^{\beta_2} \cdot (30 - FR), \tag{4.16}$$

where MA is the average magnitude of the top 25% largest motion vectors normalized by the frame width. β_1 and β_2 are model coefficients. In this work, the motion vectors are obtained using the full-search full-pel block matching method with ± 32 search range on the uncompressed source videos. Many pairs of β_1 and β_2 are tested and the one that leads to the best mean correlation coefficient averaged over all source videos is selected, similar as done in [FWSV07]. To obtain the predicted DMOS values, the QM model is further fitted to the measure DMOS values by linear least-squares fitting.

Since STPSNR and QM are identical to PSNR at full frame rate, both of them have strong content-dependency as PSNR does in the spatial quality case. It can be expected that their performance across different video contents would be very poor. Therefore, with the intention to reduce the content-dependency of STPSNR, the SVQM model is extended to predict the spatio-temporal quality by replacing PSNR by STPSNR, which is written as

$$SVQM^{+} = \frac{100}{1 + e^{-(STPSNR + w_s' \cdot SA + w_t' \cdot TA - \mu')/s'}}. \tag{4.17}$$

The model coefficients here are determined by a non-linear least-squares fitting to the DMOS values of all test videos, which leads to $w_s' = 0.0925$, $w_t' = 0.384$, $\mu' = 40.3$ and $s' = 3.55$. Note that since the content-dependency as well as the non-linearity of the impact of both spatial and temporal impairment are considered by SVQM^{+}, its performance would be much better than STPSNR and QM. The comparison between STVQM and SVQM^{+} is to show how much improvement is achieved due to the distinct physical meaning and clear structure of the STVQM model.

Figure 4.13 compares the predicted DMOS values from four models: STPSNR, QM, SVQM^{+} and STVQM, with the measured DMOS values, and Figure 4.14 illustrates the linear correlation between the model predictions and the DMOS values. The results for the VQMTQ model are close to that of the STVQM model and therefore are not included in the figures for clarity. As expected, STPSNR and QM have poor performance predicting the DMOS, both in terms of accuracy and linear correlation. SVQM^{+} indeed performs much better than STPSNR and QM because of its reduced content-dependency. The performance of the STVQM

model is clearly better than all three comparing models. Its predictions are very close to and quite linearly correlated to the measured DMOS values. These observations are confirmed and quantified by the statistical metrics summarized in Table 4.10, 4.11 and 4.12, where the results for the VQMTQ model are also included. The significance tests show that the difference between the STVQM model and the first three comparing models are statistically significant at the 100% level. Comparing the performance of STVQM and SVQM$^+$ shows that the distinct physical meaning and clear structure of STVQM lead to a significant improvement. This is also illustrated by comparing the Pearson correlation coefficients of the models for individual source video (summarized in Table 4.13), where the content-dependency issue is entirely excluded. In this case, SVQM$^+$ performs similarly as STPSNR and QM, whereas STVQM still outperforms all three comparing models by a significant margin. The statistical analysis also show that STVQM and VQMTQ have similar performance and the differences between the two models are not statistically significant. On the other hand, the proposed STVQM model has several advantages over VQMTQ, as discussed in Section 4.2.3, which make STVQM more independent and more suitable for real-time applications.

4.6 Summary

In this chapter, a full-reference objective video quality metric is presented, which considers the impact of both spatial (i.e., quantization) and temporal (i.e., frame rate reduction) quality impairment on the overall perceptual video quality. The metric is based on PSNR, frame rate as well as spatial and temporal video content activity measures that can be easily computed from the original source video. Unlike most existing metrics, the presented metric is content-independent, which is suitable for autonomous adaptation in a system where the trade-off between spatial and temporal quality may be exploited to improve the overall perceptual video quality. Statistical analysis with the data collected from subjective tests shows that the presented metric is very accurate in predicting the perceptual quality. The performance is either significantly better than or as good as (but with other advantages) that of the related metrics.

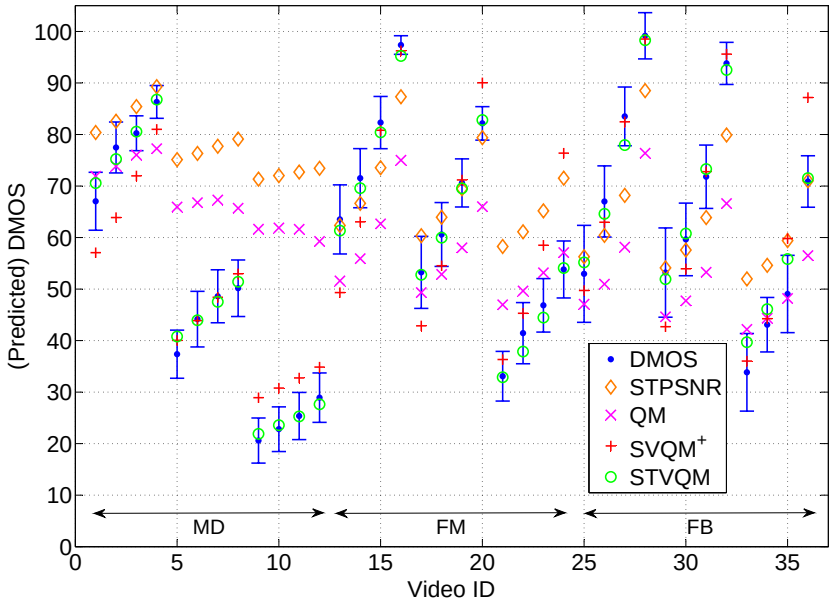
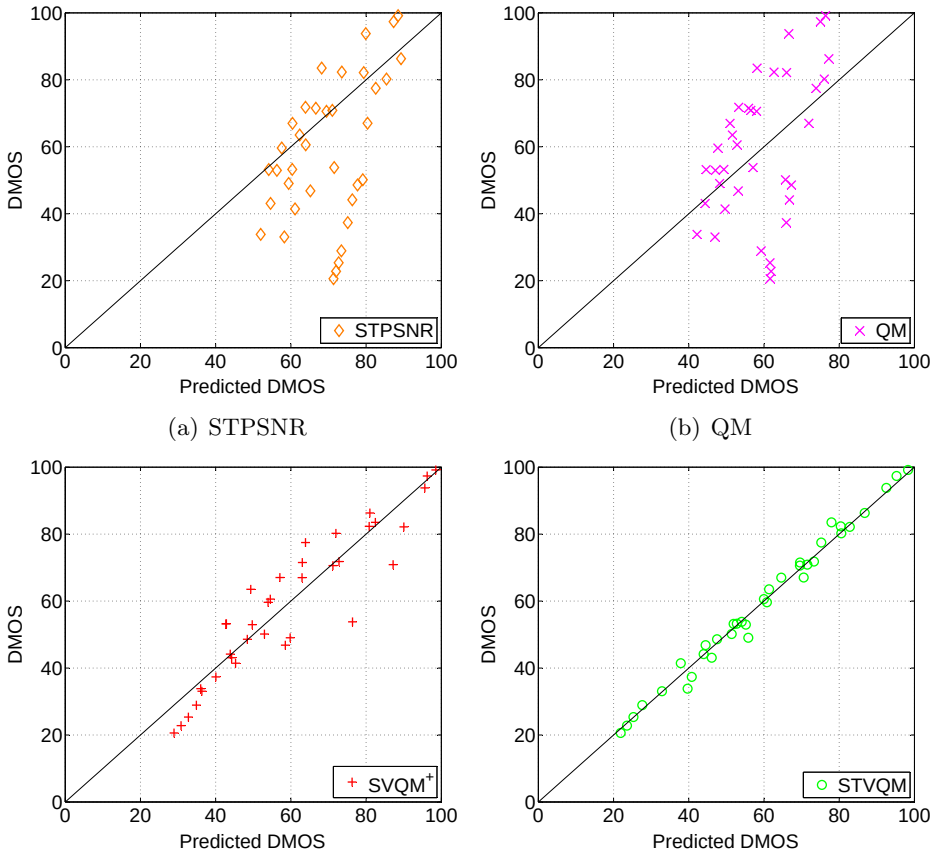


Figure 4.13: Performance evaluation and comparison for the STVQM model – accuracy. The vertical bar indicates the corresponding 95% confidence interval of each DMOS.



(a) STPSNR (b) QM

(c) SVQM$^+$ (d) STVQM

Figure 4.14: Performance evaluation and comparison for the STVQM model – linear correlation with DMOS.

Table 4.10: Pearson correlation coefficients of the spatio-temporal quality models

Model	*PCC*	*LB PCC*	*UB PCC*	*Sig. Level*	R^2
STPSNR	0.484	0.185	0.701	1.00	0.234
QM	0.471	0.169	0.692	1.00	0.222
SVQM$^+$	0.925	0.858	0.962	1.00	0.856
VQMTQ	0.995	0.988	0.997	0.23	0.988
STVQM	0.994	0.989	0.997	-	0.989

Table 4.11: RMSE values of the spatio-temporal quality models

Model	*RMSE*	*LB RMSE*	*UB RMSE*	*Sig. Level*
STPSNR	19.10	15.45	25.03	1.00
QM	19.55	15.77	25.73	1.00
SVQM$^+$	8.55	6.87	11.31	1.00
VQMTQ	2.82	2.25	3.77	0.62
STVQM	2.66	2.13	3.56	-

Table 4.12: Outlier ratios of the spatio-temporal quality models

Model	*OR*	*CI*	*Sig. Level*
STPSNR	0.83	±0.12	1.00
QM	0.78	±0.14	1.00
SVQM$^+$	0.47	±0.16	1.00
VQMTQ	0.00	±0.00	0.69
STVQM	0.03	±0.05	-

Table 4.13: Pearson correlation coefficients for individual source video

Model	*MD*		*FM*		*FB*		*Average*	
	PCC	R^2	*PCC*	R^2	*PCC*	R^2	*PCC*	R^2
STPSNR	0.978	0.956	0.863	0.745	0.929	0.862	0.923	0.855
QM	0.974	0.949	0.906	0.821	0.943	0.889	0.941	0.887
SVQM$^+$	0.985	0.970	0.856	0.733	0.943	0.890	0.928	0.864
VQMTQ	0.999	0.998	0.997	0.995	0.985	0.971	0.994	0.988
STVQM	0.997	0.995	0.998	0.996	0.990	0.981	0.995	0.990

Chapter 5

QoE-Driven Multi-Dimensional Adaptation

In this chapter, a QoE-driven MDA scheme is developed and integrated into the low-delay error-resilient video transmission framework presented in Chapter 3. Instead of using the PER-based heuristic approach, the decision of which retransmission scheme to apply is based on the resulting QoE, which is estimated based on the objective video quality metric STVQM presented in Chapter 4. With the QoE-driven MDA, the system can deliver significantly improved QoE with high adaptability to both channel conditions and video content characteristics.

5.1 Introduction

In a wireless video transmission system with limited transmission capacity and time-varying channel characteristics, an MDA scheme that may adjust multiple spatial and temporal video coding parameters at the same time would be able to deliver better QoE and provide the system with higher adaptability. An example of such an MDA scheme is the delay-aware channel-adaptive retransmission scheme presented in Section 3.5. Other related examples include multi-dimensional rate control schemes (e.g., [LK05]), multi-dimensional transcoding schemes (e.g., [JR04]), adaptive scheduling schemes based on scalable video coding (e.g., [SSW07]) or frame dropping (e.g., [TCS08]), and others. Most of the MDA schemes either adopt heuristic approaches [BW07], or formulate the problem in such a way to optimize certain Quality of Service (QoS) parameters (e.g., throughput [SSW07]) or the MSE/PSNR of the reconstructed video [LK05]. A heuristic approach based on packet error rate is also applied in Chapter 3 for the channel adaptive retransmission. However, although heuristic approaches generally can improve the system performance compared to schemes without MDA, they perform well

in some situations, less well in others, and may even perform worse than without in certain situations. Also, the parameters in these heuristic approaches (if any) have significant impact on the overall performance and selecting suitable parameters is rather challenging, which typically involves empirically testing many parameter sets in many different situations.

The problem formulation of optimizing QoS parameters or MSE/PSNR would lead to a solution that achieves the best possible QoS or MSE/PSNR, but typically not the best QoE, especially not in the MDA scenario. QoS parameters, such as transmission bitrate or packet error rate, are good quality measures for general data transmission, but have only grossly approximate relationship with the perceptual video quality. MSE/PSNR has long been used for image and video quality assessment, but in general does not correlate well with perceptual quality either (see Section 2.2.2). One significant problem with MSE/PSNR in an MDA scenario is that the resulting temporal aspect of the perceptual quality is poorly modeled by MSE/PSNR when the temporal resolution (i.e., frame rate) is adjusted, which has been shown by various previous works (e.g., [FWSV07]) as well as by the work presented in Chapter 4 based on subjective test results. The multi-dimensional rate control scheme in [RL02] uses the sum of absolute error (SAE) instead of the MSE as the optimization objective, intending to place less emphasis (no squaring) on the large differences associated with the skipped frames. But as shown in [WSV+03], SAE still has problems in predicting the perceptual quality in the context of multi-dimensional rate control. Since MSE/PSNR and SAE place too much weight on the frame rate, an MDA scheme using them as the optimization objective would favor spatial adjustment (e.g., quantization) too strongly over temporal adjustment (i.e, frame rate reduction) and select the temporal option much more rarely than it should be selected. Therefore, both heuristic approaches and QoS or MSE/PSNR optimization based schemes are rather far from being able to fully exploit the potential gain of MDA.

To address the drawbacks of the existing MDA schemes discussed above, a QoE-driven approach is proposed and integrated into the adaptive retransmission scheme presented in Chapter 3. The presented QoE-driven MDA decides between spatial and temporal adjustment based on their respective resulting QoE. The QoE is estimated based on the objective quality metric STVQM presented in Chapter 4, which has been shown to have high accuracy in predicting the perceptual quality in the presence of both spatial and temporal quality impairments. The resulting adaptation scheme automatically takes into account various factors that may affect the QoE, including available bitrate, packet error rate and error pattern, video content, slice size, concealment method, etc., and therefore can provide significantly improved QoE for arbitrary videos over a wide range of channel conditions. Thus, the meaning of "multi-dimensional" here is double-edged. One one hand, it refers to that multiple spatial and temporal video encoding parameters may be adjusted. One the other hand, it also implies that the scheme is adapted to multiple factors, including channel conditions, video content

characteristics, and others.

The rest of this chapter is organized as follows. Section 5.2 presents the problem formulation of the proposed QoE-driven MDA scheme. The related quality estimation issues, for both source coding and channel introduced quality degradations, are addressed in Section 5.3. Experimental results with various channel models are presented and discussed in Section 5.4, showing consistent QoE improvement for a wide range of system conditions. Section 5.5 summarizes the chapter.

5.2 Problem Formulation

In the DACAR scheme presented in Section 3.5, four retransmission schemes are combined in a heuristic manner based on PER thresholds that are selected empirically. As discussed in Section 3.5.2, the choice of the thresholds depends on various factors, which makes it impossible to find a set of thresholds that perform the best in all situations. For example, the best thresholds for a low-motion video would be different from those for a high-motion video, as the frame rate reduction would affect the perceptual quality of those videos differently. In this section, the combination of the retransmission schemes is revisited and a QoE-driven MDA (QMDA) scheme is proposed, which always chooses the retransmission scheme that results in the best estimated QoE.

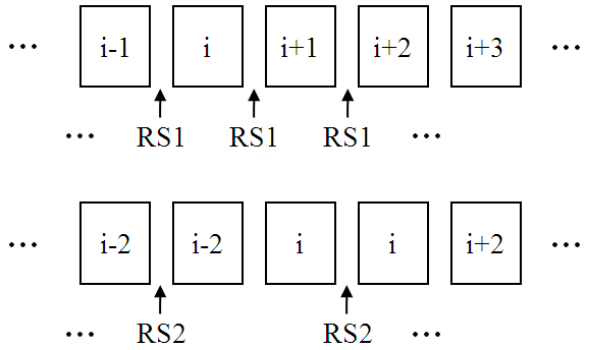
5.2.1 QoE-Driven Decisions

As in the DACAR scheme, the trade-off between spatial and temporal quality is exploited in the QMDA scheme at two places where decisions are to be made. By making these decisions, the system is automatically adapted to multiple related factors.

Decision-I

The first decision to make is that, before encoding a new video frame, which one of the retransmission schemes[†] RS0, RS1 and RS2 should be applied. Two trade-offs exist behind this decision. First, to decide between RS0 and RS1, the trade-off between the impact of concealment (RS0) and that of quantization (RS1) on the perceptual video quality needs to be considered. Applying RS0 would result in a larger number of lost MBs that need to be concealed, while applying RS1 would result in a reduced source coding bit budget, thus a coarser quantization. This choice mainly depends on how well the lost MBs can be concealed, which is determined by the video content, the concealment scheme, the slice size as well as the packet error rate and error pattern in the wireless channel. The second trade-off is between the spatial quality (RS1) and the temporal quality (RS2), which needs to be

[†]Refer to Section 3.5.1 for the definition of the retransmission schemes.



(a) Example of displayed frames structure. Notice how RS2 reduces the frame rate and increases the delay.

(b) Example of image quality. The video Mother&Daughter is encoded at 150kbps/30fps with RS1 (left) and at 150kbps/15fps with RS2 (right).

Figure 5.1: Differences between RS1 and RS2.

considered when deciding between RS1 and RS2. The difference between RS1 and RS2 is illustrated in Figure 5.1 using examples of displayed frames structure and image quality. It can be seen that compared to RS1, RS2 reduces the frame rate (see the frame repetition in Figure 5.1(a)), but the resulting doubled bit budget could significantly improve the image quality (see Figure 5.1(b)). In addition, the end-to-end delay is increased when applying RS2 (see the frame position shift in Figure 5.1(a)), which also needs to be taken into consideration when making the decision.

Given the video content (VC), the channel statistics (CS) and the codec configuration (CC), the problem of Decision-I is formulated as to find the retransmission scheme γ^* that maximizes the resulting QoE, i.e.,

$$\gamma^* = \arg\max_{\gamma \in \mathbf{\Gamma}} Q(\gamma, VC, CS, CC), \tag{5.1}$$

where $\mathbf{\Gamma}$ denotes the set of candidate retransmission schemes {RS0, RS1, RS2} and $Q(\cdot)$ denotes the resulting QoE. The resulting QoE of a particular retransmission scheme depends

on the perceptual quality of the reconstructed video (VQ) and the end-to-end delay (T), which can be written as

$$Q(\gamma, VC, CS, CC) = f(VQ(\gamma), T(\gamma)). \tag{5.2}$$

The impact of the end-to-end delay on the QoE is application-specific and may vary significantly for different video applications. In this work, this impact is modeled as a simple on-off function. In other words, it is assumed that as long as the delay does not exceed the application requirement T_{MAX}, its impact can be ignored, and a delay larger than T_{MAX} is not acceptable. Thus, Equation (5.2) becomes

$$Q(\gamma, VC, CS, CC) = \begin{cases} VQ(\gamma), & T \leq T_{MAX} \\ 0, & T > T_{MAX} \end{cases}. \tag{5.3}$$

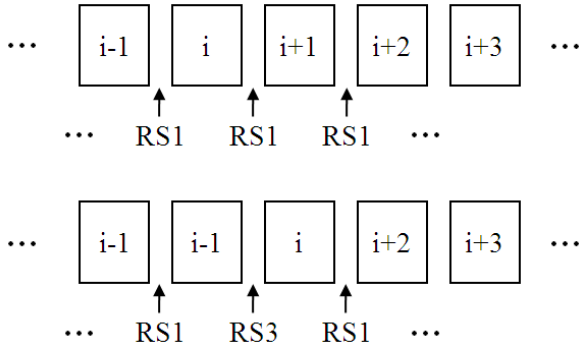
In addition, it is assumed that error concealment and quantization have the same impact on the perceptual video quality if they introduce the same MSE, as done in various previous works (e.g., [HCC02]). Then the video quality can be measured by the PSNR-based STVQM model presented in Chapter 4 (see Equation (4.9)), which has shown to be very accurate in estimating perceptual quality. Extensive experimental results with natural test video sequences show that most of the time, RS1 leads to better quality than RS0. Even when RS0 outperforms RS1, typically when PER is below 3%, the quality difference is very small. Therefore, RS0 is no longer considered in the rest of this chapter and the focus of Decision-I is on exploiting the trade-off between spatial and temporal quality, which is shown to be able to provide significant gain. As a result, Equation (5.1) is reformulated as

$$\begin{aligned} &\gamma^* = \arg \max_{\gamma \in \{RS1, RS2\}} STVQM(\gamma) \\ &\text{subject to } T(\gamma^*) \leq T_{MAX}. \end{aligned} \tag{5.4}$$

Since Decision-I needs to be made before encoding/transmission, the SPSNR term in the STVQM metric needs to be estimated, which is addressed in Section 5.3.1.

Decision-II

The second place where the trade-off between spatial and temporal quality is exploited is after transmitting a frame in its allocated time slot(s). It needs to be decided whether the residual lost packets should just be concealed (sacrificing the spatial quality) or the next frame should be skipped (sacrificing the temporal quality) so that the lost packets can be further retransmitted (RS3). The main idea is to adapt to situations where the actual PER during the transmission is significantly larger than estimated before encoding. Applying RS3 would reduce the frame rate and introduce additional delay (see Figure 5.2(a)), but the retransmissions could improve the image quality significantly, as illustrated in Figure 5.2(b).



(a) Example of displayed frames structure. Notice how RS3 reduces the frame rate and increases the delay.

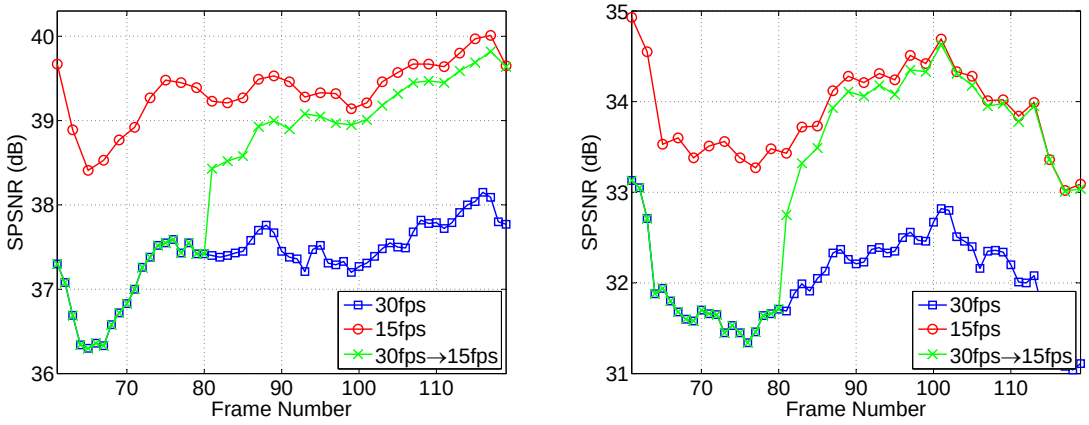
(b) Example of image quality. The video Mother&Daughter is encoded/transmitted at 500kbps/30fps without RS3 (left) and with RS3 (right). The PER is 30% and DMVE is applied for error concealment.

Figure 5.2: Differences between without and with RS3.

Similar as in Decision-I, the impact of the end-to-end delay on the QoE is modeled as an on-off function and the resulting QoE is as given by Equation (5.3). Also, when both quantization and error concealment are involved, it is assumed that their impacts on the video quality would be the same if they introduce the same MSE and the overall quality can be measured by the PSNR-based STVQM model. Thus, the solution of Decision-II can be formulated as

$$\begin{aligned} &\text{Apply RS3, as long as } STVQM(RS3) > STVQM(\text{w/o } RS3) \\ &\text{and } T(RS3) \leq T_{MAX}. \end{aligned} \tag{5.5}$$

With this formulation, RS3 would only be applied when it leads to improved QoE under the constraint on the end-to-end delay. Since Decision-II is made after the encoding, the SPSNR term in $STVQM(\text{w/o } RS3)$ can be computed. But the SPSNR term in $STVQM(RS3)$ need to be estimated if all the lost packets cannot be successfully retransmitted within the extra time slot (i.e., the current PER is very high). This quality estimation is addressed in Section 5.3.2.



(a) Mother&Daughter, 250Kbps (b) Foreman, 350Kbps

Figure 5.3: Examples of the gradual spatial quality (measured by SPSNR) improvement when frame rate changes from 30fps to 15fps at the same bitrate. The frame rate change occurs after Frame 80. Notice that the transition process may take up to 10 frames and is longer for the low-motion video Mother&Daughter.

5.2.2 Channel Adaptation

Decision-I and Decision-II constitute the final QMDA scheme. Unlike in the DACAR scheme where both Decision-I and Decision-II are made for each frame, the channel adaptation time-scale for Decision-I in the QMDA scheme is considered to be a (fixed) number of frames, which is referred to as an adaption group of pictures (AGOP), while Decision-II is considered to be per-frame adaptation. Various aspects are taken into consideration when determining the adaptation time-scales, which are discussed in the following.

The choice of the adaptation time-scale depends on the time-scale of the channel variation. Generally, with a small time-scale, the adaptation scheme can quickly follow even rapidly varying channel conditions. However, in the considered system, a small adaptation time-scale, such as the length of one video frame, may not be the best choice in terms of maximizing the QoE, especially when associated with Decision-I. There are two issues if Decision-I is to be carried out with a small time-scale for following rapid short-term channel variations. First, the potential of the frame rate reduction may not be fully utilized. More specifically, when Decision-I decides for reducing the frame rate, it is sacrificing the temporal quality to achieve an improved spatial quality so that the overall quality would be maximized. This improvement of the spatial quality is a gradual transition process that would take several frames to reach the plateau. Examples of this process are given in Figure 5.3, which shows that the transition may take up to 10 frames and the transition is longer for a video with lower motion level (i.e., Mother&Daughter). In this case, if Decision-I is made for every frame, and assuming the SPSNR is accurately estimated for the current frame, the system would underestimate the potential spatial quality improvement (that could be achieved by

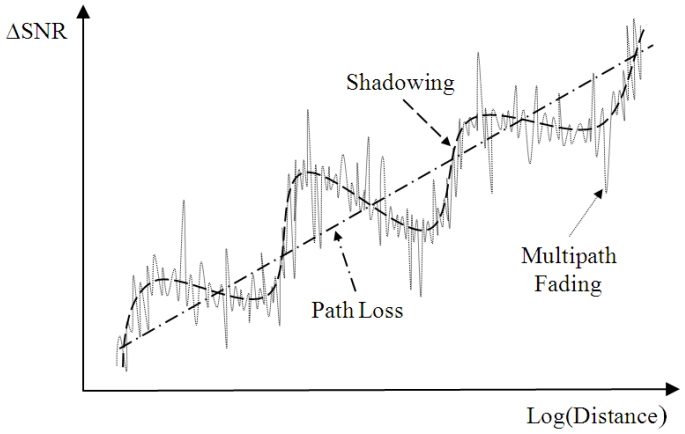


Figure 5.4: Illustration of typical wireless channel variations consisting of large-scale (long-term) variation (caused by path loss and shadowing) and small-scale (short-term) variation (caused by multipath propagation).

a longer term of frame rate reduction) and therefore may often decide against frame rate reduction. The second issue here is that a small adaptation time-scale may cause undesired short-term video quality variations (both spatially and temporally), especially as Decision-I is a proactive approach that depends on estimations of the channel condition from most recent observations. In case of rapid short-term channel variations, these channel estimations could be quite different from the actual conditions and therefore lead to adjustments that may not be desirable. In comparison, although Decision-II with small-scale adaptation would also cause quality variations (temporal only), since the decisions are based on the actual channel conditions and the resulting picture quality, the adjustments, if carefully controlled, would generally improve the overall QoE. Therefore, from the video application perspective, adopting Decision-I on a relatively large time-scale (i.e., a relatively large AGOP size) is desired.

From the channel perspective, typical wireless channels may experience two different types of variations: large-scale (long-term) variation and small-scale (short-term) variation [Skl97]. Large-scale variation is caused by path loss and shadowing effects with a time-scale of seconds, while small-scale variation is caused by multipath fading with a time-scale of milliseconds [ZCK01]. As illustrated in Figure 5.4, these two types of variations are considered superimposed to form the overall channel variation. Thus, considering the discussions above from the video application perspective, the QMDA scheme applies Decision-I to follow the large-scale variation and Decision-II to adapt to the small-scale variation. In this way, the resulting system can be highly adaptive to the wireless channel, while at the same time being relatively insensitive to the issues from the video application perspective. The final QMDA scheme operates as follows.

Before encoding a new AGOP, the average PER during the transmission of the last AGOP is measured and used to estimate the current average channel condition. Based on this average

channel quality estimate, Decision-I is made according to Equation (5.4), where the calculation of the STVQM metric for the current AGOP (to be encoded) is based on parameters estimated from the previous AGOP, such as the spatial and temporal content activity measures. From this perspective, the size of an AGOP should be kept relatively small so that video content changes can be quickly tracked. In this work, the AGOP size is chosen to be 30 frames (i.e., one second), which should provide a good balance among several factors that the AGOP size may have impact on. The decided retransmission scheme will then be applied to all the video frames in the new AGOP and the video encoder will be configured to generate the corresponding source rate R_S according to Equation (3.8)–(3.9). After the transmission of each frame, Decision-II is made according to Equation (5.5) to decide whether RS3 should be applied.

5.3 Video Quality Estimation

In the proposed QMDA scheme, since the decisions may need to be made before the actual encoding or transmission of the video frames, the resulting video quality would need to be estimated. First, at the beginning of each AGOP where Decision-I is to be made before encoding, the $STVQM(\gamma)$ term in Equation (5.4) needs to be estimated. This is mainly a problem of estimating the source coding quality, which is discussed in Section 5.3.1. Second, after the transmission of each frame where Decision-II is to be made, the $STVQM(RS3)$ term in Equation (5.5) needs to be estimated if the RPER is expected not to be zero after the retransmissions in the extra time slot. This is mainly a problem of estimating the distortion caused by packet errors in the channel, which is discussed in Section 5.3.2.

5.3.1 Quality Estimation for Decision-I

To make Decision-I according to Equation (5.4), three parameters in the STVQM model need to be estimated: $SPSNR$, SA and TA, for both $STVQM(RS1)$ and $STVQM(RS2)$. The content activity measures SA and TA are estimated from the measured content statistics of the previous AGOP, which is formulated as

$$\widetilde{SA}_i = \overline{SA}_{i-1}, \tag{5.6}$$

$$\widetilde{TA}_i = \overline{TA}_{i-1}, \tag{5.7}$$

where $\widetilde{SA}_i$ and $\widetilde{TA}_i$ represent the estimated content activity measures of the current AGOP, while $\overline{SA}_{i-1}$ and $\overline{TA}_{i-1}$ represent the average values measured from the previous AGOP. Note that $\overline{SA}_{i-1}$ and $\overline{TA}_{i-1}$ are averaged over all video frames in the previous AGOP, including the skipped frames when RS2 is applied.

The estimation of $SPSNR$ can be formulated as a rate-distortion modeling problem for video source coding that has been studied extensively for the commonly used hybrid video coding structure. In this work, a widely used RD model (e.g., in [SW98, HM02b]) is adopted, which is given by

$$D(R) = \sigma^2 \cdot e^{-\alpha \cdot R}, \tag{5.8}$$

where D denotes the MSE, R denotes the source coding rate, σ^2 is the variance of the source data and α is a content-dependent parameter to be determined. Based on this RD model, the PSNR can be written as a function of the rate R as

$$PSNR(R) = 10 \cdot log_{10}(\frac{255^2}{D(R)}) = PSNR0 + \gamma \cdot R, \tag{5.9}$$

where $PSNR0$ is the logarithmic form of σ^2 and γ is a content-dependent parameter to be determined.

Since both RS1 and RS2 are involved, two cases are considered when estimating the SPSNR values for the current AGOP. If the target retransmission scheme (for which the SPSNR is to be estimated) for the current AGOP is the same as the one applied for the previous AGOP, both $PSNR0$ and γ are estimated from the encoding results of the previous AGOP. It is formulated as

$$\widetilde{SPSNR}_i = \overline{PSNR0}_{i-1} + \gamma_{i-1} \cdot R_i. \tag{5.10}$$

Here, the term $\widetilde{SPSNR}_i$ represents the estimated SPSNR of the current AGOP, $\overline{PSNR0}_{i-1}$ represents the average $PSNR0$ of the previous AGOP and R_i is the target source coding rate for the current AGOP. γ_{i-1} is determined by

$$\gamma_{i-1} = \frac{1}{\overline{R}_{i-1}} \cdot (\overline{SPSNR}_{i-1} - \overline{PSNR0}_{i-1}), \tag{5.11}$$

where $\overline{R}_{i-1}$ represents the actual source coding rate of the previous AGOP, and the two PSNR terms, $\overline{SPSNR}_{i-1}$ and $\overline{PSNR0}_{i-1}$ represent the average SPSNR and PSNR0 of the previous AGOP, respectively.

For the other retransmission scheme, it would not be proper to estimate $PSNR0$ and γ from the previous AGOP, as the prediction distance is different. In this case, the problem is formulated so as to estimate the difference between $SPSNR(RS1)$ and $SPSNR(RS2)$ at the same per-second bitrate, so that once one SPSNR is known, the other SPSNR can be calculated. This can be written as

$$SPSNR(RS2) = SPSNR(RS1) + \Delta PSNR. \tag{5.12}$$

Two factors have impacts on the $\Delta SPSNR$ here. On one hand, RS2 leads to a doubled per-pixel bitrate that would increase the SPSNR. On the other hand, the increased prediction distance in RS2 would decrease the SPSNR. The overall difference results from the combination of the two impacting factors, which is found to be similar for different video sequences.

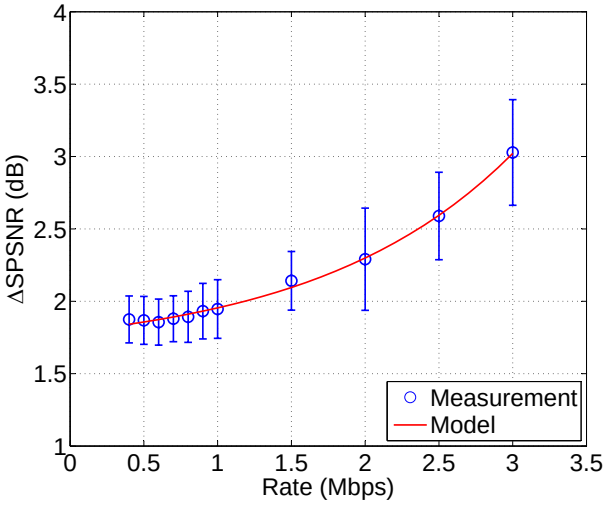


Figure 5.5: Modeling the SPSNR difference between RS1 and RS2 at the same per-second bitrate. The measurement points are averaged over several (at least 8, mostly $\geq$ 15) test sequences. The vertical bar indicates the corresponding 95% confidence interval.

Therefore, the overall difference $\Delta PSNR$ is modeled empirically based on the average encoding results from 20 test video sequences with different content characteristics (all with CIF (352x288) resolution and an original frame rate of 30fps), which are illustrated in Figure 5.5 with the average values and the corresponding 95% confidence intervals. Since each video is only encoded within a proper rate range and the test video sequences cover a wide range of content complexities, not all statistics in Figure 5.5 are computed from 20 samples, but each rate point involves at least 8 samples, with more samples ($\geq$ 15) for the low to medium rate range. The average $\Delta SPSNR$ is modeled by an exponential function as

$$\Delta SPSNR = a + b \cdot e^{c \cdot R}, \tag{5.13}$$

where R is the source coding bitrate in Mbps and the constants a, b, and c are determined by a non-linear least-squares fitting to the average measurement data, which leads to $a = 1.64$, $b = 0.15$ and $c = 0.74$. Figure 5.5 shows that this model approximates the measurements very accurately.

In summary, the estimation of $STVQM(RS1)$ and $STVQM(RS2)$ works as follows. If RS1 is applied for the previous AGOP, $SPSNR(RS1)$ is estimated from the coding statistics of the previous AGOP using Equation (5.10) and (5.11). Then $SPSNR(RS2)$ is estimated from the estimated $SPSNR(RS1)$ using Equation (5.12) and (5.13). If RS2 is applied for the previous AGOP, $SPSNR(RS2)$ is estimated first from the previous AGOP before estimating $SPSNR(RS1)$ from the estimated $SPSNR(RS2)$. With the content activity measures estimated from the previous AGOP using Equation (5.6) and (5.7), the two $STVQM$s can be computed using Equation (4.7).

5.3.2 Quality Estimation for Decision-II

To make Decision-II according to Equation 5.5, since the current video frame has been encoded and transmitted, the term $STVQM(\text{w/o } RS3)$ can be readily computed from the available data (i.e., SA and TA from the original frame and $SPSNR$ from the decoded and concealed frame). As for the term $STVQM(RS3)$, depending on the current channel status, there are two possible situations regarding the parameter $SPSNR$. If the current PER is not very high so that all the lost video packets are expected to be successfully retransmitted during the extra time slot with RS3, the resulting $SPSNR$ can be computed from the decoded frame. Otherwise (i.e., $RPER > 0$), the $SPSNR$ needs to be estimated, as which video packets would still be lost is unknown. The estimation of the $SPSNR$ in this situation is addressed in the following.

Since both source coding distortion and transmission errors are involved, the estimation problem here is often referred to as end-to-end distortion estimation [WHL+00, HCC02, ZGL+07]. Let f_n^i represent the original value of pixel i in frame n, $\hat{f}_n^i$ represent the corresponding decoded value in the encoder (no concealment), and $\tilde{f}_n^i$ represent the final reconstructed value (same in the encoder and decoder, with possible concealment). In this work, since there is no error propagation, $\tilde{f}_n^i$ can be written as

$$\tilde{f}_n^i = \begin{cases} f_n^{i(ec)}, & pixel\ i\ is\ lost \quad w.p. \quad p \\ \hat{f}_n^i, & otherwise \quad\quad w.p. \quad 1-p \end{cases}, \tag{5.14}$$

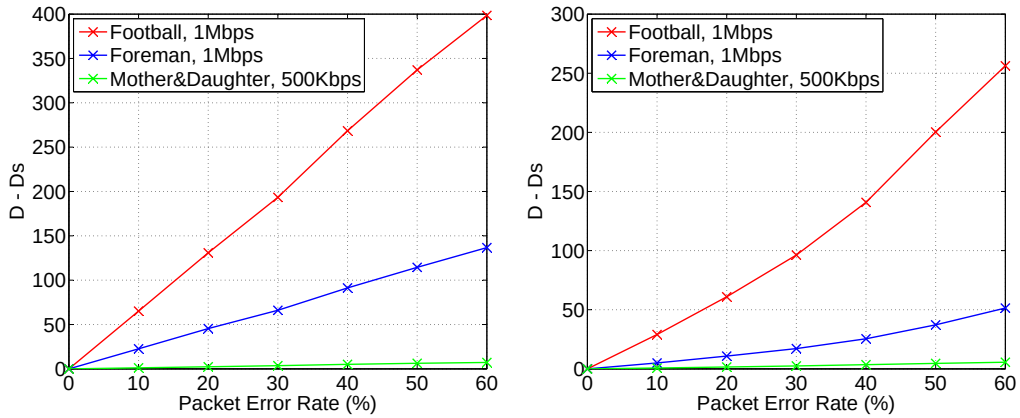
where p represents the packet error rate and $f_n^{i(ec)}$ represents the concealed value when pixel i is lost. Then, the expected end-to-end distortion of frame n measured by MSE is given by

$$\begin{aligned} D &= E\left\{\left(f_n^i - \tilde{f}_n^i\right)^2\right\} \\ &= (1-p)\cdot E\left\{\left(f_n^i - \hat{f}_n^i\right)^2\right\} + p\cdot E\left\{\left(f_n^i - f_n^{i(ec)}\right)^2\right\} \\ &= (1-p)\cdot D_s + p\cdot D_{ec}, \end{aligned} \tag{5.15}$$

where D_s denotes the source coding distortion and D_{ec} denotes the error concealment distortion. For Decision-II, D_s can be readily computed from the decoded frame in the encoder. Assuming CPB is applied as the error concealment method, D_{ec} can be written as

$$\begin{aligned} D_{ec} &= E\left\{\left(f_n^i - f_n^{i(ec)}\right)^2\right\} \\ &= E\left\{\left(f_n^i - f_{ref}^i\right)^2\right\}, \end{aligned} \tag{5.16}$$

where f_{ref}^i is the value of pixel i in the reference frame. In this case, D_{ec} is the average distortion between frame n and its reference and is constant when p changes. Thus, the



(a) CPB (b) DMVE

Figure 5.6: Simulation results for end-to-end distortion estimation.

end-to-end distortion can be computed by

$$\begin{aligned} D(p_1) &= (1-p_1)\cdot D_s + p_1 \cdot D_{ec} \\ &= (1-p_1)\cdot D_s + p_1 \cdot \frac{1}{p_0}\cdot (D(p_0) - (1-p_0)\cdot D_s) \\ &= D_s + (D(p_0) - D_s)\cdot \frac{p_1}{p_0}, \end{aligned} \tag{5.17}$$

where p_1 is the expected RPER after the retransmissions in the extra time slot and p_0 is the current RPER. The to-be-estimated $SPSNR$ is simply a logarithmic form of the $D(p_1)$ here.

If the applied error concealment method is not CPB, D_{ec} may be dependent on p. For example, for DMVE (or any other improved temporal concealment method) , D_{ec} is given by

$$\begin{aligned} D_{ec} &= E\left\{\left(f_n^i - f_n^{i(ec)}\right)^2\right\} \\ &= E\left\{\left(f_n^i - f_{ref}^j\right)^2\right\}, \end{aligned} \tag{5.18}$$

where f_{ref}^j is the pixel value in the reference frame used to conceal pixel i. Ideally, D_{ec} would be the average distortion between frame n and the motion-compensated reference frame (independent from p). However, since the motion vectors need to be estimated, the larger the packet error rate p is, the further away the actual D_{ec} would be from the ideal case (i.e., larger D_{ec}). Thus, the end-to-end distortion would become a non-linear function of p, instead of the linear function with CPB. This is illustrated by simulation results shown in Figure 5.6 for both CPB and DMVE, where $D - D_s$ is plotted against p for three different videos; each point is generated by averaging over 20 sample frames and 10 channel realizations. A detailed analysis and the determination of the actual non-linear model are left for future work. Note that the quality estimation here is only to be carried out in rare situations and therefore the model accuracy would not have significant impact on the overall performance of the proposed system. Thus, a linear model should provide sufficient accuracy for other concealment schemes.

All the experimental results in Section 5.4 are generated using CPB as the error concealment method and Equation (5.17) for quality estimation.

5.4 Experimental Results

The performance of the proposed QMDA scheme is evaluated in a real-time video transmission system with an MPEG-4 video codec (i.e., the Xvid codec [Xvi]). Synchronized error concealment (see Section 3.4.2) is integrated, where CPB is adopted as the error concealment method. As in Section 3.6.2, the results for three source video sequences with different content characteristics are analyzed: Mother&Daughter, Foreman and Football, all in CIF (352x288) resolution and with an original frame rate of 30fps. The encoder settings adopted in Section 3.6.2 are also applied here, including coding structure, slice structure and size, etc. Four different systems are compared: 1) SP, where Decision-I is RS1 (i.e., quantization adjustment) and Decision-II is no RS3 (i.e., concealment only); 2) TP, where Decision is RS2 (i.e., frame rate reduction) and Decision-II is no RS3; 3) ADP: Decision-I is made adaptively based on the QoE metric STVQM following Equation (5.4) and Decision-II is no RS3; 4) ADP+: both Decision-I and Decision-II are made adaptively based on STVQM following Equation (5.4) and (5.5), respectively. The T_{MAX} in Equation (5.4) and (5.5) is set so that two consecutive frames may be skipped. Note that adaptive decisions based on PSNR would almost always avoid frame skipping due to the excess weight PSNR places on the skipped frames, resulting in a performance very close to that of the system SP.

The wireless channel is modeled as a packet erasure channel with constant transmission data rate. Different packet error patterns are tested when evaluating the system performance with QMDA. In Section 5.4.1, results for a wide range of channel conditions (represented by static channel models) are presented, where both independent packet errors and bursty packet errors are considered. In Section 5.4.2, results for highly dynamic channel conditions characterized by a user mobility model are presented.

5.4.1 Static Channel Model

Two widely adopted channel models are considered in this section for modeling the packet error behavior in a wireless channel. One is the independent model, which assumes that the packet errors are independently and identically distributed (i.i.d.). The other one is the Gilbert-Elliot model [Gil60, Ell63], which has been shown to model the bursty nature of the Rayleigh fading channel with adequate accuracy [WC96, ZR97]. To generate the experimental results, the parameters of the channel model are kept constant during the entire duration of the video; it is hence referred to as "static channel model". Note that the resulting packet error rate may be quite different from frame to frame, especially with the Gilbert-Elliot model.

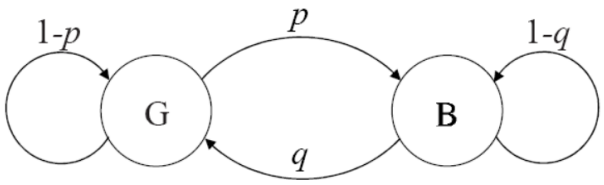


Figure 5.7: The state diagram of the Gilbert-Elliot channel model.

The Gilbert-Elliot channel model is a two-state Markov model as illustrated in Figure 5.7, where the two states of the model are denoted as G (good) and B (bad). In the state G, packets are considered to be lost with (low) probability P_G, while in the state B, packets are considered to be lost with (high) probability P_B. Let p denote the transition probability from the state G to the state B, and q denote the transition probability from the state B to the state G, then the Gilbert-Elliot model can be described by the transition matrix

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}. \tag{5.19}$$

In this work, it is assumed that $P_G = 0$ and $P_B = 1$ (i.e., packets are received successfully in the state G and are lost in the state B). This specific version of the Gilbert-Elliot model is also referred to as the simplified Gilbert model [YW95]. In this case, the mean packet error rate P_e and the mean burst length L_e can be written as [YW95]

$$P_e = \frac{p}{p+q}, \tag{5.20}$$

$$L_e = \frac{1}{q}. \tag{5.21}$$

The experimental results with static channel models are shown in Figure 5.9 – 5.10, where the resulting video quality (measured by the STVQM metric, averaged over 10 random channel realizations) of different systems for a wide range of packet error rates is compared. The results for the Gilbert-Elliot channel model are generated with the average burst length $L_e = 16$. It can be seen from the results that in general, the system ADP is highly adaptive, both to the channel conditions as well as to the video content characteristics, as the curves of ADP always follow the option with better perceptual video quality (estimated by STVQM), no matter at what rate, for which content or at what packet error rate/pattern. The system ADP+ adds another level of adaptability to channel variations by reacting to the actual channel conditions. This leads to further performance gains that are higher for channels with higher packet error rate and for videos with larger motion. Recall that the performance of the system SP also represents the performance of an adaptive system which applies PSNR as the video quality metric. So the gap between ADP+ and SP also shows how much can be improved by adopting

the more accurate quality metric STVQM in such an MDA scenario in comparison to PSNR-based adaptation/optimization. Both ADP and ADP+ run automatically in real-time; there is no parameter that needs to be determined manually to achieve the best performance for a particular condition.

The results also show that for the low motion video Mother&Daughter (see Figure 5.9), the system TP (frame skipping) always delivers better perceptual video quality than the system SP (quantization adjustment). The gain of choosing frame skipping over quantization adjustment becomes smaller as the rate/spatial quality level increases, until it saturates at high spatial quality level. For videos with higher motion (i.e., Foreman and Football, see Figure 5.11 and Figure 5.10, respectively), TP delivers better perceptual quality than SP at low rate/spatial quality level, indicating that for high motion contents, frame skipping is preferred over quantization adjustment when the rate/spatial quality is low. As the rate/spatial quality level increases, the two curves move towards each other, meet at medium rate, and then separate at high rate, where SP delivers better perceptual quality than TP. This indicates that at high rate/spatial quality level, quantization adjustment is preferred over frame skipping.

Note that the case with zero packet error rate is equivalent to a multi-dimensional rate control scheme that jointly adjust quantization and frame rate. Thus, it can also be seen from the results here how effective it would be to integrate the STVQM metric in applications involving such a rate control scheme.

5.4.2 User Mobility Model

In this section, the performance of the QMDA scheme is evaluated with highly dynamic channel conditions. A particular user mobility model is considered, which represents the wireless channel characteristics between a fixed transmitter and a moving receiver in an environment where both large-scale and small-scale variations exist. The receiver is assumed to be moving away from the transmitter at a constant speed v.

Large-scale variation is caused by path loss and shadowing effects. It has been shown that the mean path loss increases exponentially with the transmitter-receiver separation distance and the power law relationship can be expressed as [SR92]

$$\overline{PL}(d) = \overline{PL}(d_0) + 10 \cdot n \cdot \log_{10}\left(\frac{d}{d_0}\right), \tag{5.22}$$

where $\overline{PL}(d)$ denotes the mean path loss (in decibels) at the distance d, $\overline{PL}(d_0)$ denotes the mean path loss at a reference distance d_0, and n is the path loss exponent that indicates how fast the mean path loss increases with respect to distance. The value of the path loss exponent depends on the specific propagation environment and may range from 1 to 7 [Mol05] (e.g., $n = 2$ for free space).

Shadowing effects lead to slow fluctuations around the mean path loss (often referred to as shadow fading), which can be expressed as a log-normally distributed random variable [SR92]. Let X denote the random variable in decibels that represents the shadow fading, the path loss at distance d is then given by

$$PL(d) = \overline{PL}(d) + X = \overline{PL}(d_0) + 10 \cdot n \cdot \log_{10}\left(\frac{d}{d_0}\right) + X. \tag{5.23}$$

Here X is normally distributed (in decibels) with zero-mean and standard deviation σ. The value of σ depends on the specific propagation environment.

With the path loss model in Equation (5.23), the received signal-to-noise ratio (SNR, in decibels) can be expressed as a function of the distance d

$$SNR(d) = SNR(d_0) - 10 \cdot n \cdot \log_{10}\left(\frac{d}{d_0}\right) - X \tag{5.24}$$

where $SNR(d)$ and $SNR(d_0)$ represents the received SNR at the distance d and the reference distance d_0, respectively.

The received SNR given in Equation (5.24) is an average quantity. The actual instantaneous SNR varies much more rapidly due to multi-path propagation. In this work, this small-scale variation is modeled by a Rayleigh random process (often referred to as Rayleigh fading). The packet error behavior in such a Rayleigh fading channel is then represented by the packet-level simplified Gilbert model, whose transition probabilities are given by [BL00]

$$p = f_D \cdot T \cdot \sqrt{2\pi} \cdot \sqrt{-\log(1-\varepsilon)}, \tag{5.25}$$

$$q = f_D \cdot T \cdot \sqrt{2\pi} \cdot \frac{1-\varepsilon}{\varepsilon} \cdot \sqrt{-\log(1-\varepsilon)}, \tag{5.26}$$

where f_D represents the maximum Doppler shift, T represents the packet duration and ε represents the mean packet error rate. The maximum Doppler shift is given by

$$f_D = \frac{v}{c} \cdot f, \tag{5.27}$$

where v is the receiver speed, c is the speed of light and f is the carrier frequency. As shown in [EL77] and [BL95], for very slow Rayleigh fading, where the signal strength can be considered constant during the transmission of a packet (which is true with the selected parameters for the experiments in this work), the mean packet error rate ε (assuming non-coherent FSK and no FEC) can be expressed as a function of the average received SNR

$$\epsilon = 1 - 2^{-n}\left[1 + \sum_{i=1}^{n}\prod_{j=1}^{i}\frac{n+1-j}{j+2/\gamma}\right], \tag{5.28}$$

where n is the packet size in bits and γ represents the average received SNR that can be computed using Equation (5.24).

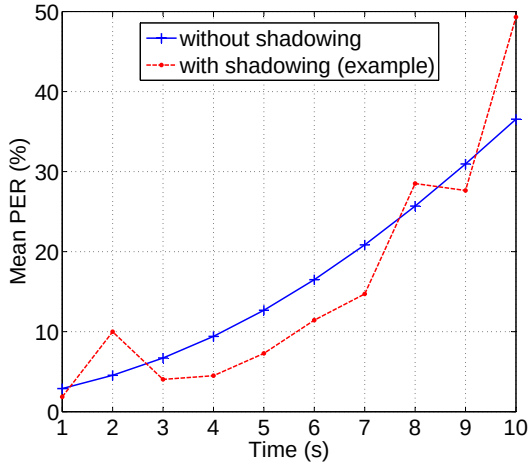


Figure 5.8: The mean packet error rate for every second with the user mobility model.

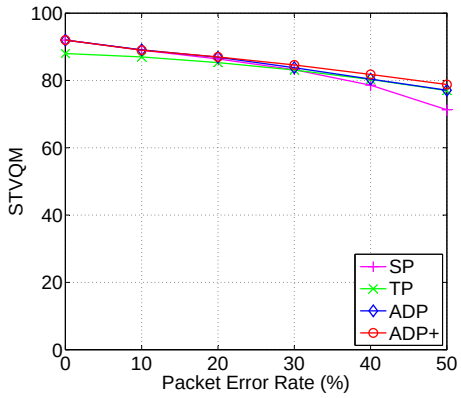
For the user mobility model, an indoor propagation environment is assumed. The receiver starts at a distance of 6 meters and moves away from the transmitter at a constant speed of $v = 1m/s$ (walking speed). The reference distance d_0 is set to be 1 meter. The path loss exponent n and the standard deviation σ of X are 3.0 and 3.0, respectively, matching typical values for indoor non-line-of-sight (NLOS) propagation environments [KWX+04]. $SNR(d_0)$ is set to be 50dB, so that the resulting mean packet error rate without shadow fading varies between 0% and 40% for a 10-second video, which is shown in Figure 5.8. The average received SNR is assumed to change every second according to Equation (5.24) and for each packet within the second, the packet status is generated by the simplified Gilbert model with the parameters calculated by Equation (5.25) and (5.26), where the carrier frequency f is set to be 3.8GHz.

The experimental results with the user mobility model are shown in Figure 5.12, where for each test video, the resulting video quality (measured by the STVQM metric, averaged over 10 random channel realizations) is plotted against the transmission data rate for different systems. As can be seen from the results, even for such dynamic channel conditions, the curves of the system ADP are still able to follow the option with better perceptual video quality at any rate and for any content. With ADP+, further significant performance gains are achieved with the improved channel adaptability. Other typical values of the user mobility model parameters (e.g., carrier frequency and the standard deviation of the shadow fading) are also tested and the results show similar behavior.

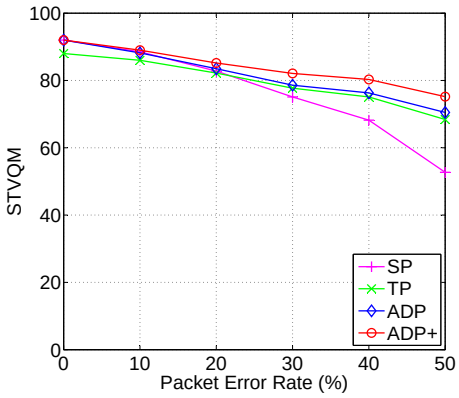
5.5 Summary

In this chapter, based on the objective video quality metric STVQM presented in Chapter 4, a QoE-driven MDA scheme is formulated and integrated into the low-delay error-resilient system

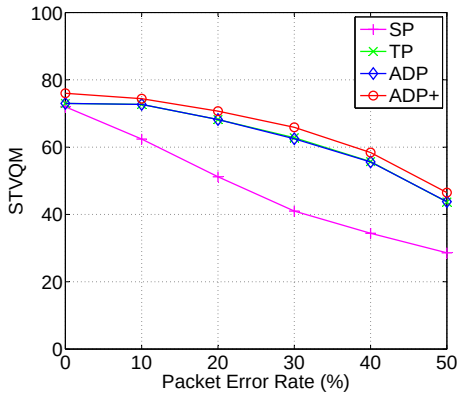
design presented in Chapter 3. Related quality estimation problems are addressed, so that the QMDA scheme is able to make decisions before the actual encoding, transmitting or processing starts. Extensive experimental results have shown that with the ability to accurately predict the perceptual video quality for different video contents with different types and amounts of quality impairment, the system with QMDA can deliver significantly improved QoE for a wide range of situations (e.g., channel conditions, video contents) comparing to non-adaptive systems or non-QoE-driven adaptive systems (i.e., systems adopting PSNR as the optimization objective).



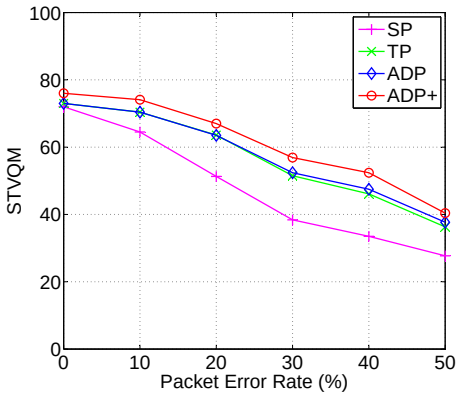
(a) High rate (400Kbps), i.i.d. error



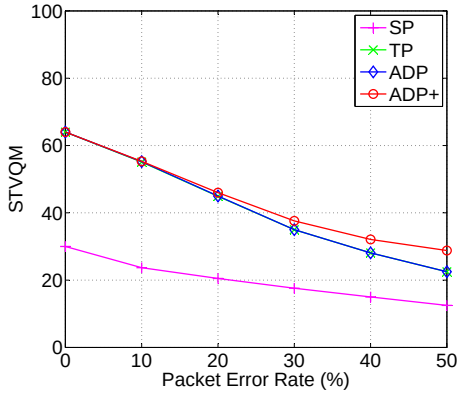
(b) High rate (400Kbps), burst error



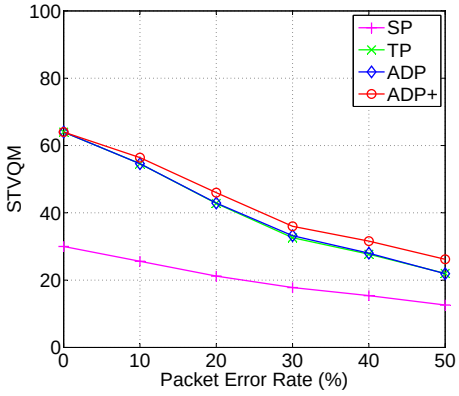
(c) Medium rate (200Kbps), i.i.d. error



(d) Medium rate (200Kbps), burst error

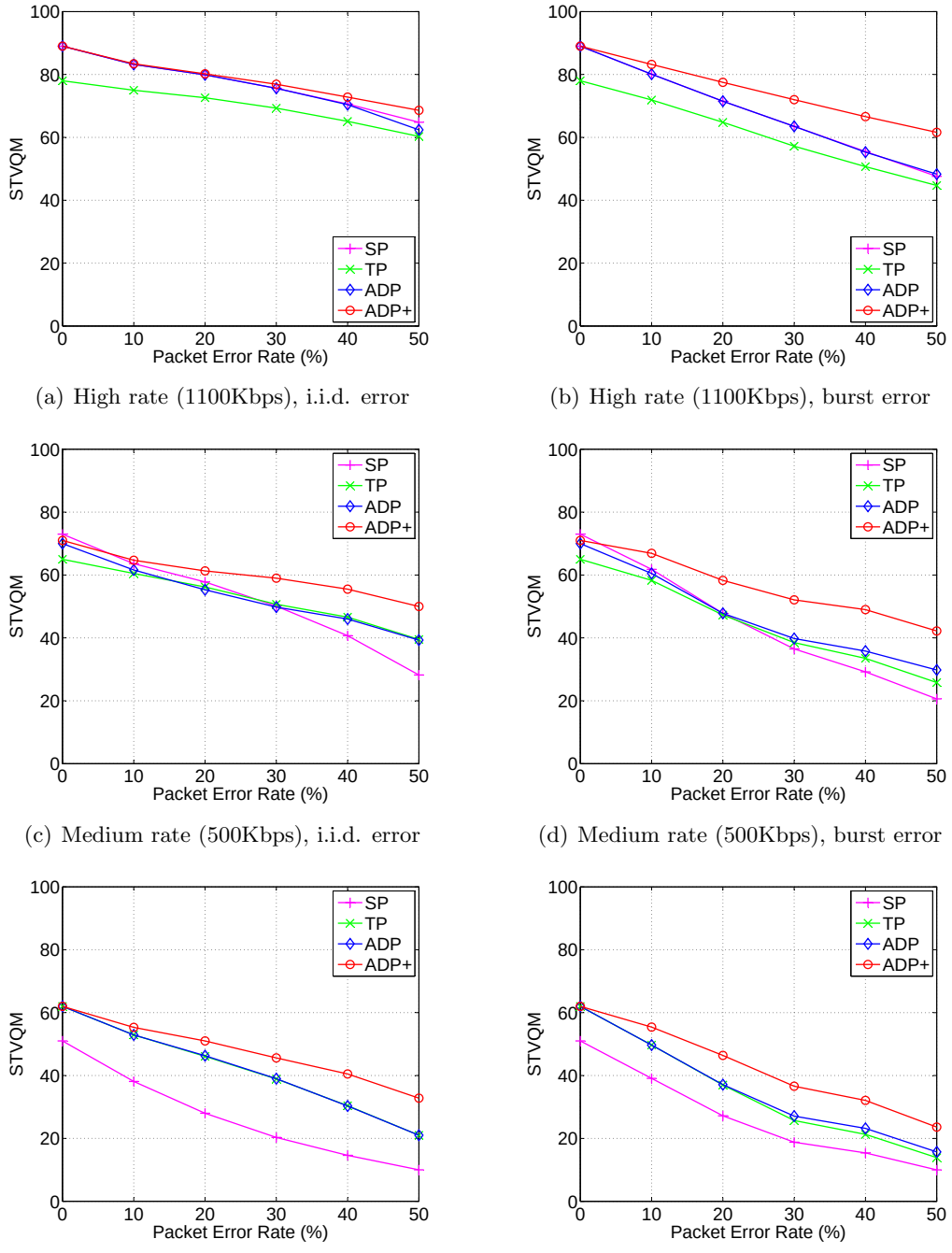


(e) Low rate (120Kbps), i.i.d. error



(f) Low rate (120Kbps), burst error

Figure 5.9: Performance comparison between various systems for Mother&Daughter. The results are generated for both i.i.d error (left) and burst error (right) at different transmission data rates and packet error rates. The STVQM values are averaged over 10 random channel realizations.



(a) High rate (1100Kbps), i.i.d. error

(b) High rate (1100Kbps), burst error

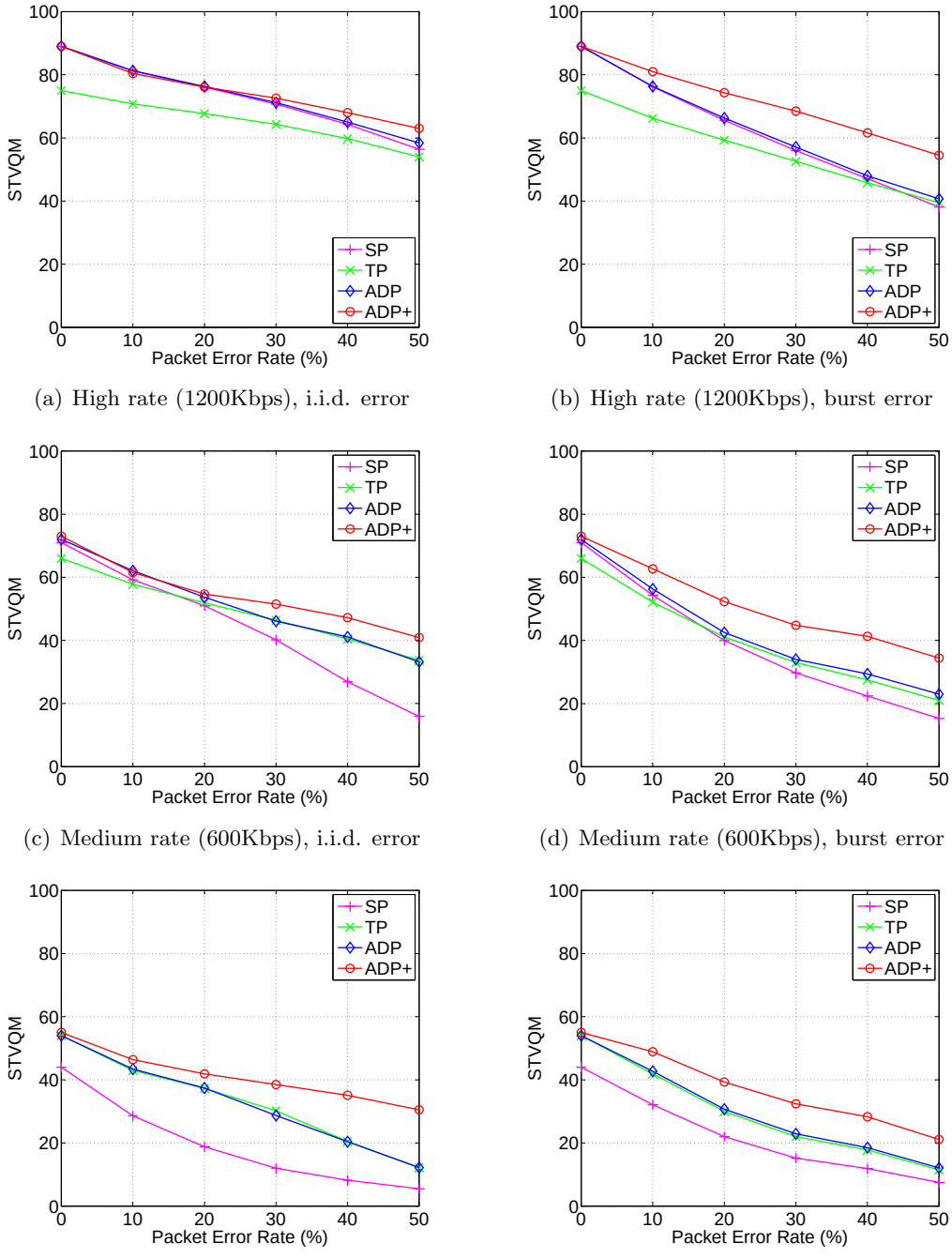
(c) Medium rate (500Kbps), i.i.d. error

(d) Medium rate (500Kbps), burst error

(e) Low rate (300Kbps), i.i.d. error

(f) Low rate (300Kbps), burst error

Figure 5.10: Performance comparison between various systems for Foreman. The results are generated for both i.i.d error (left) and burst error (right) at different transmission data rates and packet error rates. The STVQM values are averaged over 10 random channel realizations.



(a) High rate (1200Kbps), i.i.d. error

(b) High rate (1200Kbps), burst error

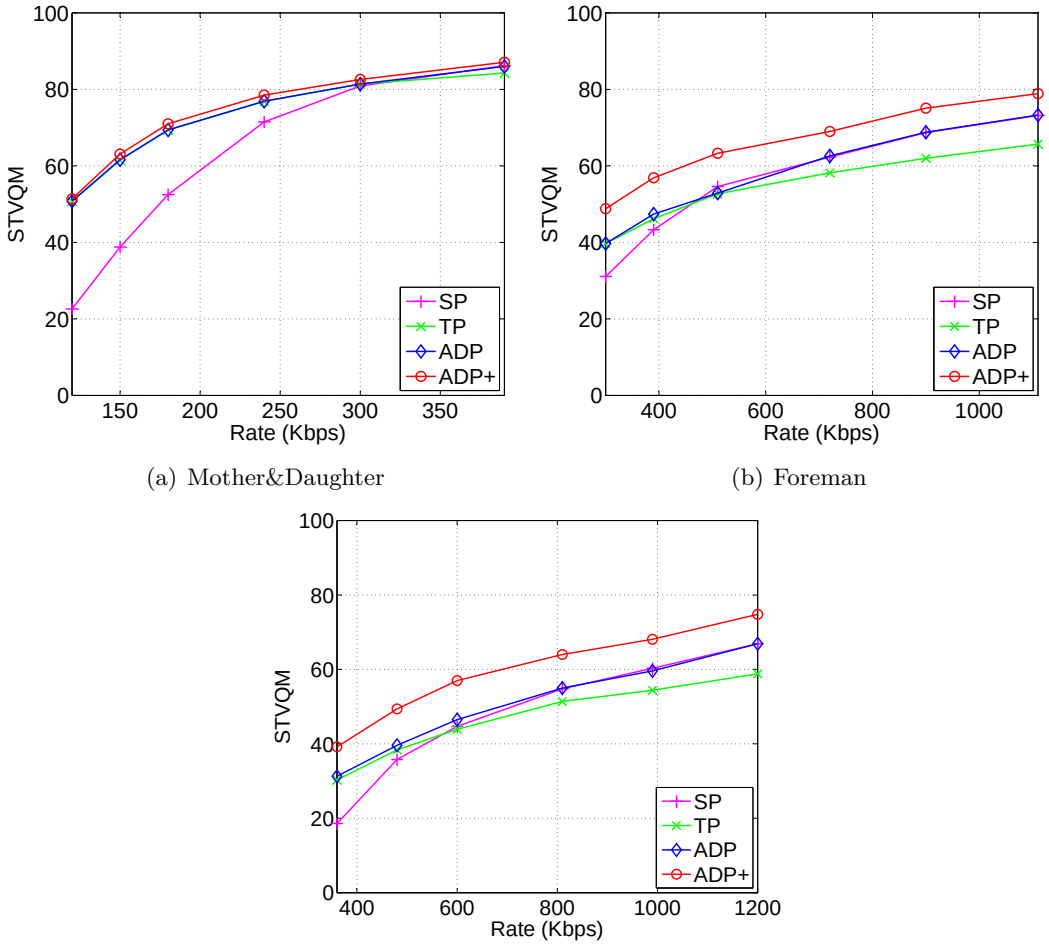
(c) Medium rate (600Kbps), i.i.d. error

(d) Medium rate (600Kbps), burst error

(e) Low rate (400Kbps), i.i.d. error

(f) Low rate (400Kbps), burst error

Figure 5.11: Performance comparison between various systems for Football. The results are generated for both i.i.d error (left) and burst error (right) at different transmission data rates and packet error rates. The STVQM values are averaged over 10 random channel realizations.



(a) Mother&Daughter

(b) Foreman

(c) Football

Figure 5.12: Performance comparison between various systems with the user mobility model. The results are generated for various videos at different transmission data rates. The STVQM values are averaged over 10 random channel realizations.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this dissertation, QoE improvement for wireless video transmission with highly dynamic error-prone channels and stringent delay constraints is investigated.

First, a complete low-delay error-resilient video transmission framework for point-to-point wireless communication with instantaneous feedback is developed. The framework is architectured in such a way that user-perceived video quality is significantly improved for a wide range of channel conditions, while the end-to-end delay of the video application is kept as low as possible. The available instantaneous feedback is integrated into both video coding and transmission, leading to highly error-resilient coding schemes and a delay-aware retransmission scheme that improve the perceptual video quality with no or controlled impact on the end-to-end delay. With the integration of multi-dimensional video adaptation into the framework, further improvement on channel adaptability and perceptual video quality is achieved. The trade-off between spatial and temporal video quality is exploited using a PER-based heuristic approach, which is although not optimal, but simple and effective. Extensive experimental results show that the proposed framework provides significantly improved video quality for a wide variety of system settings and a wide range of channel conditions. Furthermore, the framework has low computational complexity and therefore high applicability in practical real-time applications.

Second, a full-reference objective video quality metric STVQM is developed for QoE estimation with MDA, where both spatial and temporal quality impairments may exist. Based on the results from specifically designed subjective tests, the spatial quality perception is first investigated and modeled. Then the temporal quality perception and its interaction with the spatial quality perception is analyzed, based on which the overall spatio-temporal quality is modeled. The proposed metric consists of PSNR, frame rate as well as spatial and temporal

video content activity measures. The content activity measures are included to resolve the content-dependencies that most of the existing video quality metrics have. Several statistical metrics are used to evaluate various aspects of STVQM's performance, showing that the proposed metric is very accurate in estimating the perceptual video quality and performs significantly better than or as well as (but with other advantages) related metrics in the literature.

Finally, with STVQM's ability to accurately estimate the perceptual video quality in the presence of both spatial and temporal quality impairments, a QoE-driven solution is formulated for the multi-dimensional video adaptation integrated into the proposed video transmission framework. This QoE-driven MDA scheme adjusts the trade-off between spatial and temporal qualities in such a way that the best possible QoE (estimated by STVQM) is achieved. Since decisions for the QoE-driven MDA need to be made before actual encoding/transmission, estimation of the source coding distortion as well as the channel introduced distortion is addressed. Extensive experimental results have shown that the integration of the QoE-driven MDA leads to significantly improved QoE with high adaptability to both video content characteristics and channel conditions.

6.2 Future Work

The potential extensions and applications of the work presented in this dissertation can be summarized into two areas.

QoE-Driven Video Transmission

The design of the low-delay error-resilient video transmission framework has focused on the most effective techniques on the application layer. Other error-resilient coding techniques, such as data partitioning and flexible macroblock ordering, can be easily integrated into the framework to further improve the overall performance. Furthermore, error control and channel adaption tools on the lower layers (e.g., adaptive modulation, FEC) may also be considered, which would require an appropriate cross-layer design.

Although point-to-point wireless communication has been the considered scenario, the general design can be extended to a point-to-multipoint scenario where one sender is multicasting one video to several receivers. For example, packet-level FEC [Hui96] can be applied to generate the retransmission packets so that the same packet can be used by every receiver despite every receiver may have lost different original video packets.

A few more degrees of freedom or dimensions may also be considered in the QoE-driven MDA scheme, such as the spatial resolution and the end-to-end delay, so that the overall QoE

can be further improved. This would require the corresponding extension of the video quality metric and integration of the extended metric.

Video Quality Metric

A full-reference video quality metric that considers both spatial and temporal quality impairments is developed in this work. Suggestions on the extensions of the metric are as follows.

- The spatial resolution has been fixed when investigating the QoE impact of quantization and frame rate reduction. If the spatial resolution is to be considered as another dimension in the MDA, proper extension of the video quality metric to include this new dimension needs to be investigated. A possible solution would be to include the distortion caused by spatial resolution change into the PSNR calculation (i.e., weighted average of quantization and spatial resolution change). Related studies in this direction can be found in [ZCL+08, SHH+08, SYN+10].

- The impact of the end-to-end delay on the QoE has been modeled as a simple on-off function. Further studies and more sophisticated modeling would be interesting as future work, especially for applications where adding delay as another adjustable parameter may lead to significant overall performance improvement.

- Both the spatial and the temporal video quality have been modeled in an averaged manner. Further studies on the impact of local quality variations on the QoE, both spatially and temporally, and the integration of the corresponding results into the video quality metric would be of great interest and importance.

- Although the proposed FR metric is not applicable in situations where the unprocessed reference video is not available, this work provides results and analysis that would guide the design of related NR and RR metrics. In fact, in the proposed QoE-driven MDA scheme, the FR metric STVQM is used in the RR manner, where the spatial and temporal activity measures can be considered as reduced reference information. Estimating the spatial and temporal activity measures from the processed video would be an interesting direction towards an NR metric.

In addition, there are many potential application areas of the proposed video quality metric, such as multi-dimensional rate control, transcoding and resource allocation among multiple users with different required videos and channel conditions, where conventional MSE/PSNR-based approaches would not be able to provide the best possible QoE.

Bibliography

[AF95] S. Aign and K. Fazel. Temporal and spatial error concealment techniques for hierarchical MPEG-2 video codec. In *Proc. IEEE International Conference on Communications (ICC'95)*, Seattle, WA, June 1995. [cited at p. 36, 37]

[APS01] S. Aramvith, I. Pao, and M. Sun. A rate-control scheme for video transport over wireless channels. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(5):569–580, May 2001. [cited at p. 26]

[BEK03] A. M. Bruckstein, M. Elad, and R. Kimmel. Down-scaling for better transform compression. *IEEE Transactions on Image Processing*, 12(9):1132–1145, September 2003. [cited at p. 11]

[BL95] H. Bischl and E. Lutz. Packet error rate in the noninterleaved Rayleigh channel. *IEEE Transactions on Communications*, 43(2/3/4):1375–1382, February/March/April 1995. [cited at p. 101]

[BL00] F. Babich and G. Lombardi. A Markov model for the mobile propagation channel. *IEEE Transactions on Vehicular Technology*, 49(1):63–73, January 2000. [cited at p. 101]

[Bli06] Jean-Louis Blin. New quality evaluation method suited to multimedia context: SAMVIQ. In *Proc. International Workshop on Video Processing and Quality Metrics (VPQM'06)*, Scottsdale, USA, January 2006. [cited at p. 17]

[BO00] M. Baldi and Y. Ofek. End-to-end delay analysis of videoconferencing over packet switched networks. *IEEE/ACM Transactions on Networking*, 8(4):479–492, August 2000. [cited at p. 24, 30]

[BRK09] A. Bhat, I. Richardson, and S. Kannangara. A new perceptual quality metric for compressed video. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, pages 933–936, April 2009. [cited at p. 58, 77, 78]

[BW07] J. Brandt and L. Wolf. Multidimensional transcoding for adaptive video streaming. In *Proc. International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'07)*, Urbana, IL, USA, June 2007. [cited at p. 85]

[CGHS99] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach. All subjective scales are not created equal: the effects of context on different scales. *Signal Processing*, 77:1–9, August 1999. [cited at p. 14, 16]

[Com] The R Project For Statistical Computing. `http://www.r-project.org/`. [cited at p. 68]

[DL96] W. Ding and B. Liu. Rate control of MPEG video coding and recording by rate-quantization modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(1):12–20, February 1996. [cited at p. 26]

[DSB+99] T. DeFanti, D. Sandin, M. Brown, D. Pape, J. Anstey, M. Bogucki, G. Dawe, and A. Johnson. Technologies for virtual reality/tele-immersion applications: Issues of research in image display and global networking. In *Proc. of the EC/NSF Workshop on Research Frontiers in Virtual Environments and Human-Centered Computing*, Chateau de Bonas, France, June 1999. [cited at p. 24]

[EL77] E. R. Eaves and A. H. Levesque. Probability of block error for very slow rayleigh fading in Gaussian noise. *IEEE Transactions on Communications*, 25(3):368–374, March 1977. [cited at p. 101]

[Ell63] E. O. Elliott. Estimates of error rates for codes on burst-noise channels. *The Bell System Technical Journal*, 42:1977–1997, September 1963. [cited at p. 98]

[EY05] M. Etoh and T. Yoshimura. Advances in wireless video delivery. *Proceedings of the IEEE*, 93(1):111–122, January 2005. [cited at p. 11]

[FFm] FFmpeg. `http://www.ffmpeg.org`. [cited at p. 51]

[FK07] M. Friebe and A. Kaup. Fading techniques for error concealment in block-based video decoding systems. *IEEE Transactions on Broadcasting*, 53(1):286–296, February 2007. [cited at p. 37]

[FNI96] S. Fukunaga, T. Nakai, and H. Inoue. Error resilient video coding by dynamic replacing of reference pictures. In *Proc. IEEE GLOBECOM'96*, London, UK, November 1996. [cited at p. 27]

[FWSV07] R. Feghali, D. Wang, F. Speranza, and A. Vincent. Video quality metric for bit rate control via joint adjustment of quantization and frame rate. *IEEE Transactions on Broadcasting*, 53(1):441–446, March 2007. [cited at p. 56, 59, 60, 79, 80, 86]

[GF99] B. Girod and N. Färber. Feedback-based error control for mobile video transmission. *Proceedings of the IEEE*, 87(10):1707–1723, October 1999. [cited at p. 11, 27]

[Gil60] E. N. Gilbert. Capacity of a burst-noise channel. *The Bell System Technical Journal*, 39:1253–1265, September 1960. [cited at p. 98]

[HCC02] Z. He, J. Cai, and C. W. Chen. Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(6):511–523, June 2002. [cited at p. 26, 89, 96]

[HKM01] Z. He, Y. K. Kim, and S. K. Mitra. Low-delay rate control for DCT video coding via ρ-domain source modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(8):928–940, August 2001. [cited at p. 26]

[HM01] Z. He and S. K. Mitra. A unified rate-distortion analysis framework for transform coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(12):1221–1236, December 2001. [cited at p. 26]

[HM02a] Z. He and S. K. Mitra. A linear source model and a unified rate control algorithm for dct video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(11):970–982, November 2002. [cited at p. 26, 27, 32, 33]

[HM02b] Z. He and S. K. Mitra. Optimum bit allocation and accurate rate control for video coding via ρ-domain source modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(10):840–849, October 2002. [cited at p. 26, 94]

[HOK99] C.-Y. Hsu, A. Ortega, and M. Khansari. Rate control for robust video transmission over burst-error wireless channels. *IEEE Journal on Selected Areas in Communications*, 17(5):756–773, May 1999. [cited at p. 26]

[HTBH+07] Q. Huynh-Thu, M. Brotherton, D. Hands, K. Brunnström, and M. Ghanbari. Examination of the samviq methodology for the subjective assessment of multimedia quality. In *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM'07)*, Scottsdale, USA, January 2007. [cited at p. 17]

[HTG05] Q. Huynh-Thu and M. Ghanbari. A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video. In *Proc. IASTED International Conference on Signal and Image Processing*, volume 479, pages 70–76, August 2005. [cited at p. 16]

[HTG08] Q. Huynh-Thu and M. Ghanbari. Temporal aspect of perceived quality in mobile video broadcasting. *IEEE Transactions on Broadcasting*, 54(3):641–651, September 2008. [cited at p. 58, 59]

[Hui96] C. Huitema. The case for packet level FEC. In *Proc. IFIP 5th International Workshop on Protocols for High Speed Networks*, pages 109–120, Sophia Antipolis, France, October 1996. [cited at p. 25, 110]

[IKH+06] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu. A method of estimating coding psnr using quantized dct coefficients. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(2):251–259, February 2006. [cited at p. 21]

[ISO04] ISO/IEC. Information technology - coding of audio-visual objects - part 2: Visual. ISO/IEC 14496-2, 2004. [cited at p. 8]

[ITU96] ITU. *ITU-T/SG15/LBC-96-033 An error resilience method based on back channel signalling and FEC*. San Jose: Telenor R&D, January 1996. [cited at p. 27]

[ITU99] ITU. Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910, 1999. [cited at p. v, 13, 14, 15, 69]

[ITU00] ITU. User requirements for objective perceptual video quality measurements in digital cable television. ITU-T Recommendation J.143, May 2000. [cited at p. 18]

[ITU01] ITU. End-user multimedia QoS categories. ITU-T Recommendation G.1010, 2001. [cited at p. 9, 24]

[ITU02] ITU. Methodology for the subjective assessment of the quality of television pictures. ITU-R Recommendation BT.500-11, 2002. [cited at p. v, 13, 14, 15, 17]

[ITU04a] ITU. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. ITU-T Recommendation J.144, March 2004. [cited at p. 18]

[ITU04b] ITU. Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference. ITU-R Recommendation BT.1683, 2004. [cited at p. 18, 57]

[ITU05] ITU. Video coding for low bit rate communication. ITU-T Recommendation H.263, 2005. [cited at p. 8, 10]

[ITU07] ITU. Methodology for the subjective assessment of video quality in multimedia applications. ITU-R Recommendation BT.1788, 2007. [cited at p. v, 13, 14, 15, 16, 17, 62, 63, 64, 65]

[ITU08] ITU. Objective perceptual multimedia video quality measurement in the presence of a full reference. ITU-T Recommendation J.247, August 2008. [cited at p. 18]

[JKSR07] S. H. Jin, C. S. Kim, D. J. Seo, and Y. M. Ro. Quality measurement modeling on scalable video applications. In *Proc. IEEE Workshop on Multimedia Signal Processing (MMSP'07)*, pages 131–134, October 2007. [cited at p. 59]

[JR04] Y. J. Jung and Y. M. Ro. Joint control for hybrid transcoding using multidimensional rate distortion modeling. In *Proc. International Conference on Image Processing (ICIP'04)*, pages 2789–2792 vol.4, October 2004. [cited at p. 56, 85]

[JVT03] JVT. Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC)). Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, 2003. [cited at p. 8, 10]

[KCRV06] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan. Modeling packet-loss visibility in MPEG-2 video. *IEEE Transactions on Multimedia*, 8(2):341–355, April 2006. [cited at p. 19, 21]

[KS93] W. Kwok and H. Sun. Multi-directional interpolation for spatial error concealment. *IEEE Transactions on Consumer Electronics*, 39(3):455–460, June 1993. [cited at p. 36]

[KSK07] D.-K. Kwon, M.-Y. Shen, and C.-C. J. Kuo. Rate control for H.264 video with enhanced rate and distortion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(5):517–529, May 2007. [cited at p. 33]

[KWX+04] B. Kannan, K. C. Wee, S. Xu, C. L. Chuan, F. Chin, C. Y. Huat, C. C. Choy, T. T. Thiang, X. Peng, M. Ong, and S. Krishnan. UWB channel characterization in indoor office environments,. Technical Report IEEE 802.15-04-0439-00-004a, IEEE, August 2004. [cited at p. 102]

[KXMP06] S. Kumar, L. Xu, M. K. Mandal, and S. Panchanathan. Error resiliency schemes in H.264/AVC standard. *Elsevier Journal of Visual Communication & Image Representation*, 17(2):425–450, April 2006. [cited at p. 11]

[LC04] A. Leontaris and P. C. Cosman. Video compression for lossy packet networks with mode switching and a dual frame buffer. *IEEE Transactions on Image Processing*, 13(7):885–897, July 2004. [cited at p. 28]

[LCZ00] H.-J. Lee, T. Chiang, and Y.-Q. Zhang. Scalable rate control for mpeg-4 video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(6):878–894, September 2000. [cited at p. 26]

[LG06] Y. Liang and B. Girod. Network-adaptive low-latency video communication over best-effort networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(1):72–81, January 2006. [cited at p. 27, 28]

[LK05] S. Liu and C.-C. J. Kuo. Joint temporal-spatial bit rate control for video coding with dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):15–26, January 2005. [cited at p. 11, 55, 56, 85]

[LLS07a] Y. Liu, Z. G. Li, and Y. C. Soh. A novel rate control scheme for low delay video communication of H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(1):68–78, January 2007. [cited at p. 26]

[LLS+07b] Z. Lu, W. Lin, B. C. Seng, S. Kato, E. Ong, and S. Yao. Perceptual quality evaluation on periodic frame-dropping video. In *Proc. International Conference on Image Processing (ICIP'07)*, pages III–433–III–436 vol.3, September 2007. [cited at p. 58]

[LOK96] L.-J. Lin, A. Ortega, and C.-C. J. Kuo. Rate control using spline-interpolated R-D characteristics. In *Proc. SPIE Visual Communications and Image Processing (VCIP'96)*, Orlando, FL, March 1996. [cited at p. 26]

[LOR98] T. V. Lakshman, A. Ortega, and A. R. Reibman. VBR video: tradeoffs and potentials. *Proceedings of the IEEE*, 86(5):952–973, May 1998. [cited at p. 9]

[LRL93] W. M. Lam, A. R. Reibman, and B. Liu. Recovery of lost or erroneously received motion vectors. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP'93)*, pages 417–420, April 1993. [cited at p. 36, 37]

[LV00] J. Y. Liao and J. Villasenor. Adaptive intra block update for robust transmission of H.263. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1):30–35, February 2000. [cited at p. 27]

[Mar86] Hans Marmolin. Subjective MSE measures. *IEEE Transactions on Systems, Man, and Cybernetics*, 16(3):486–489, May 1986. [cited at p. 57]

[MF06] E. D. Montag and M. D. Fairchild. Fundamentals of human vision and vision modeling. In H. R. Wu and K. R. Rao, editors, *Digital video image quality and perceptual coding*, chapter 5. CRC Press, Boca Raton, FL, 2006. [cited at p. 20]

[MGL05] S. Ma, W. Gao, and Y. Lu. Rate-distortion analysis for H.264/AVC video coding and its application to rate control. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(12):1533–1544, December 2005. [cited at p. 26]

[Mol05] A. F. Molisch. Ultrawideband propagation channels-theory, measurement, and modeling. *IEEE Transactions on Vehicular Technology*, 54(5):1528–1545, September 2005. [cited at p. 100]

[MSM04] J. D. McCarthy, M. A. Sasse, and D. Miras. Sharp or smooth?: comparing the effects of quantization vs. frame rate for streamed video. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'04)*, pages 535–542, April 2004. [cited at p. 59]

[ODZ07] T. Oelbaum, K. Diepold, and W. Zia. A generic method to increase the prediction accuracy of visual quality metrics. In *Proc. Picture Coding Symp. (PCS'07)*, pages CD–ROM, November 2007. [cited at p. 58]

[OLZ+08] Y.-F. Ou, T. Liu, Z. Zhao, Z. Ma, and Y. Wang. Modeling the impact of frame rate on perceptual quality of video. In *Proc. International Conference on Image Processing (ICIP'08)*, pages 689–692, October 2008. [cited at p. 58, 59]

[OMW09] Y. Ou, Z. Ma, and Y. Wang. A novel quality metric for compressed video considering both frame rate and quantization artifacts. In *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM'09)*, Scottsdale, Arizona, USA, January 2009. [cited at p. 60]

[OMW11] Y. Ou, Z. Ma, and Y. Wang. Perceptual quality assessment of video considering both frame rate and quantization artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(3):286–298, March 2011. [cited at p. 60, 62, 79]

[OYL+05] E. P. Ong, X. Yang, W. Lin, Z. Lu, and S. Yao. Perceptual quality metric for compressed videos. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, pages 581–584, March 2005. [cited at p. 57]

[PP08] S. Péchard and R. Pépion. Suitable methodology in subjective video quality assessment: A resolution dependent paradigm. In *Proc. International Workshop on Image Media Quality and its Applications (IMQA'08)*, Kyoto, Japan, September 2008. [cited at p. 17]

[PPIES04] J. Park, D.-C. Park, R. J. Marks II, and M. A. El-Sharkawi. Content-based adaptive spatio-temporal methods for MPEG repair. *IEEE Transactions on Image Processing*, 13(8):1066–1077, August 2004. [cited at p. 37]

[PS11] Y. Peng and E. Steinbach. A novel full-reference video quality metric and its application to wireless video transmission. In *Proc. International Conference on Image Processing (ICIP'11)*, Brussels, Belgium, September 2011. [cited at p. 5]

[PW03] M. H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. In *Proc. SPIE Visual Communications and Image Processing*, volume 5150, pages 573–582, July 2003. [cited at p. 16]

[PW04] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, 50(3):312–322, September 2004. [cited at p. 21]

[PZS10] Y. Peng, F. Zhang, and E. Steinbach. Error-resilient video transmission for short-range point-to-point wireless communication. In *Proc. International Conference on Computer Communications and Networks (ICCCN'10)*, Zurich, Switzerland, September 2010. [cited at p. 5]

[RCL99] J. Ribas-Corbera and S. Lei. Rate control in dct video coding for low-delay communications. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(1):172–185, February 1999. [cited at p. 26]

[RJ00] I. Rhee and S. Joshi. Error recovery for interactive video transmission over the internet. *IEEE Journal on Selected Areas in Communications*, 18(6):1033–1049, June 2000. [cited at p. 26]

[RL02] E. C. Reed and J. S. Lim. Optimal multidimensional bit-rate control for video communication. *IEEE Transactions on Image Processing*, 11(8):873–885, August 2002. [cited at p. 55, 86]

[RPCH10] D. M. Rouse, R. Pépion, P. Le Callet, and S. S. Hemami. Tradeoffs in subjective testing methods for image and video quality assessment. In *Proc. SPIE Advances in Image Quality*, volume 7527, San Jose, California, USA, January 2010. [cited at p. 17]

[RVS04] A. R. Reibman, V. A. Vaishampayan, and Y. Sermadevi. Quality monitoring of video over a packet network. *IEEE Transactions on Multimedia*, 6(2):327–334, April 2004. [cited at p. 21]

[SFG97] E. Steinbach, N. Färber, and B. Girod. Standard compatible extension of H.263 for robust video transmission in mobile environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(6):872–881, December 1997. [cited at p. 27]

[SFG00] K. Stuhlmüller, N. Färber, and B. Girod. Adaptive optimal intra-update for lossy video transmission. In *Proc. SPIE Visual Communications and Image Processing (VCIP'00)*, Perth, Australia, June 2000. [cited at p. 27]

[Sha06] Z. Wang; X. Shang. Spatial pooling strategies for perceptual image quality assessment. In *Proc. International Conference on Image Processing (ICIP'06)*, pages 2945–2948, October 2006. [cited at p. 58]

[SHH+08] R. Shmueli, O. Hadar, R. Huber, M. Maltz, and M. Huber. Effects of an encoding scheme on perceived video quality transmitted over lossy Internet protocol networks. *IEEE Transactions on Broadcasting*, 54(3):628–640, September 2008. [cited at p. 59, 111]

[SHW03] T. Stockhammer, M. M. Hannuksela, and T. Wiegand. H.264/AVC in wireless environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):657–673, July 2003. [cited at p. 11]

[SK98] M.-Y. Shen and C.-C. J. Kuo. Review of postprocessing techniques for compression artifact removal. *Journal of Visual Communication and Image Representation*, 9(1):2–14, March 1998. [cited at p. 11]

[SK01] H. Song and C.-C. J. Kuo. Rate control for low-bit-rate video via variable-encoding frame rates. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(4):512–521, April 2001. [cited at p. 55]

[Skl97] B. Sklar. Rayleigh fading channels in mobile digital communication systems part I: Characterization. *IEEE Communications Magazine*, 35(7):90–100, July 1997. [cited at p. 92]

[SMW07] H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, September 2007. [cited at p. 56]

[SR92] S. Y. Seidel and T. S. Rappaport. 914 mhz path loss prediction models for indoor wireless communications in multifloored buildings. *IEEE Transactions on Antennas and Propagation*, 40(2):207–217, February 1992. [cited at p. 100, 101]

[SSW07] T. Schierl, T. Stockhammer, and T. Wiegand. Mobile video transmission using scalable video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1204–1217, September 2007. [cited at p. 56, 85]

[Sto02] T. Stockhammer. Error robust macroblock mode and reference frame selection. In *Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG JVT-B102*, Geneva, Switzerland, January 2002. [cited at p. 27]

[SW98] G. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15(6):74–90, November 1998. [cited at p. 94]

[SYN+10] H. Sohn, H. Yoo, W. D. Neve, C. S. Kim, and Y. M. Ro. Full-reference video quality metric for fully scalable and mobile svc content. *IEEE Transactions on Broadcasting*, 56(3):269–280, September 2010. [cited at p. 60, 62, 111]

[TCC02] D. S. Turaga, Y. Chen, and J. Caviedes. No reference PSNR estimation for compressed pictures. In *Proc. International Conference on Image Processing (ICIP'02)*, pages III–61–III–64 vol.3, September 2002. [cited at p. 21]

[TCS08] W. Tu, J. Chakareski, and E. Steinbach. Rate-distortion optimized frame dropping for multiuser streaming and conversational videos. *Advances in Multimedia, Article ID 628970*, January 2008. [cited at p. 85]

[TGP98] K. T. Tan, M. Ghanbari, and D. E. Pearson. An objective measurement tool for MPEG video quality. *Signal Processing*, 70:279–294, November 1998. [cited at p. 21, 57]

[TMJ00] D. Tompa, J. Morton, and E. Jernigan. Perceptually based image comparison. In *Proc. International Conference on Image Processing (ICIP'00)*, pages 489–492 vol.1, September 2000. [cited at p. 57, 58]

[VCS03] A. Vetro, C. Christopoulos, and H. Sun. Video transcoding architectures and techniques: a overview. *IEEE Signal Processing Magazine*, 20(2):18–29, March 2003. [cited at p. 56]

[VQE00] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, June 2000. [cited at p. 16]

[VQE03] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II, August 2003. [cited at p. 16, 20]

[VQE08a] VQEG. Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I, September 2008. [cited at p. 14, 16, 62, 76]

[VQE08b] VQEG. Multimedia group test plan, version 1.21, March 2008. [cited at p. 17]

[VW93] S. Voran and S. Wolf. An objective technique for assessing video impairments. In *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 161–165 vol.1, May 1993. [cited at p. 59]

[Wad89] M. Wada. Selective recovery of video packet loss using error concealment. *IEEE Journal on Selected Areas in Communications*, 7(5):807–814, June 1989. [cited at p. 28]

[Wan] Z. Wang. The ssim index for image quality assessment. `http://www.cns.nyu.edu/~lcv/ssim/`. [cited at p. 77]

[WB09] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, January 2009. [cited at p. 20]

[WBSS04] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. [cited at p. 21, 58, 77]

[WC96] H. S. Wang and P.-C. Chang. On verifying the first-order markovian assumption for a rayleigh fading channel model. *IEEE Transactions on Vehicular Technology*, 45(2):353–357, May 1996. [cited at p. 98]

[WFS00] T. Wiegand, N. Färber, and K. Stuhlmüller. Error-resilient video transmission using long-term memory motion-compensated prediction. *IEEE Journal on Selected Areas in Communications*, 18(6):1050–1062, June 2000. [cited at p. 27, 28]

[WHL+00] D. Wu, Y. T. Hou, B. Li, W. Zhu, Y.-Q. Zhang, and H. J. Chao. An end-to-end approach for optimal mode selection in Internet video communication: Theory and application. *IEEE Journal on Selected Areas in Communications*, 18(6):977–995, June 2000. [cited at p. 96]

[WHZ00] D. Wu, Y. T. Hou, and Y.-Q. Zhang. Transporting real-time video over the Internet: challenges and approaches. *Proceedings of the IEEE*, 88(12):1855–1877, December 2000. [cited at p. 7, 11, 26]

[WHZ+01] D. Wu, Y. T. Hou, W. Zhu, Y.-Q. Zhang, and J. M. Peha. Streaming video over the Internet: approaches and directions. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(3):282–300, March 2001. [cited at p. 11]

[Win99] Stefan Winkler. Issues in vision modeling for perceptual video quality assessment. *Signal Processing*, 78(2):231–252, October 1999. [cited at p. 20]

[Win06] Stefan Winkler. Perceptual video quality metrics - a review. In H. R. Wu and K. R. Rao, editors, *Digital video image quality and perceptual coding*, chapter 5. CRC Press, Boca Raton, FL, 2006. [cited at p. 20, 21]

[WK08] H. Wang and S. Kwong. Rate-distortion optimization of rate control for H.264 with adaptive initial quantization parameter determination. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(1):140–144, January 2008. [cited at p. 26]

[WL07] Zhou Wang and Qiang Li. Video quality assessment using a statistical model of human visual speed perception. *J. Opt. Soc. Am. A*, 24(12):B61–B69, September 2007. [cited at p. 58]

[WLB04] Z. Wang, L. Lu, and A. C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing:Image Communication*, 19(2):121–132, February 2004. [cited at p. 21, 58, 77]

[WOZ02] Y. Wang, J. Ostermann, and Y.-Q. Zhang. *Video Processing and Communications*. Prentice Hall, 2002. [cited at p. v, 8]

[WP01] S. Wolf and M. Pinson. The relationship between performance and spatial-temporal region size for reduced-reference, in-service video quality monitoring systems. In *Proc. Systematics, Cybernetics, and Informatics/Information Systems Analysis and Synthesis (SCI/ISAS'01)*, Orlando, USA, July 2001. [cited at p. 19, 21]

[WP02] S. Wolf and M. Pinson. Video quality measurement techniques. Technical Report 02-392, National Telecommunications and Information Administration (NTIA), June 2002. [cited at p. 57, 68]

[WSB03] Z. Wang, H. R. Sheikh, and A. C. Bovik. Objective video quality assessment. In B. Furht and O. Marqure, editors, *The handbook of video database: Design and applications*, chapter 41. CRC Press, Boca Raton, FL, September 2003. [cited at p. v, 20]

[WSV+03] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield. Towards optimal rate control: A study of the impact of spatial resolution, frame rate and quantization on subjective quality and bitrate. In *Proc. Visual Communications and Image Processing (VCIP'03)*, pages 198–209, June 2003. [cited at p. 58, 86]

[WWWK00] Y. Wang, S. Wenger, J. T. Wen, and A. K. Katsaggelos. Review of error resilient coding techniques for real-time video communications. *IEEE Signal Processing Magazine*, 17(4):61–82, July 2000. [cited at p. 7, 11]

[WZ98] Y. Wang and Q. Zhu. Error control and concealment for video communications: A review. *Proceedings of the IEEE*, 86(5):974–997, May 1998. [cited at p. 7, 11]

[x26] x264. `http://www.videolan.org/developers/x264.html`. [cited at p. 51]

[Xvi] Xvid. `http://www.xvid.org`. [cited at p. 42, 44, 61, 98]

[YGEMD07] K. C. Yang, C. C. Guest, K. El-Maleh, and P. K. Das. Perceptual temporal quality metric for compressed video. *IEEE Transactions on Broadcasting*, 53(1):441–446, March 2007. [cited at p. 58, 59]

[YR07] H. Yang and K. Rose. Advances in recursive per-pixel end-to-end distortion estimation for robust video coding in H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(7):845–856, July 2007. [cited at p. 27]

[YW95] J. Yee and E. Weldon. Evaluation of the performance of error-correcting codes on a Gilbert channel. *IEEE Transactions on Communications*, 43(8):2316–2323, August 1995. [cited at p. 99]

[YW98] M. Yuen and H. R. Wu. A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing*, 70(3):247–278, November 1998. [cited at p. 11]

[ZAF00] J. Zhang, J. F. Arnold, and M. R. Frater. A cell-loss concealment technique for MPEG-2 coded video. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(4):659–665, June 2000. [cited at p. 36, 37]

[ZCK01] J. Zhang, E. K. P. Chong, and I. Kontoyiannis. Unified spatial diversity combining and power allocation for cdma systems in multiple time-scale fading channels. *IEEE Journal on Selected Areas in Communications*, 19(7):1276–1288, July 2001. [cited at p. 92]

[ZCL+08] G. Zhai, J. Cai, W. Lin, X. Yang, and W. Zhang. Three dimensional scalable video adaptation via user-end perceptual quality assessment. *IEEE Transactions on Broadcasting*, 54(3):719–727, September 2008. [cited at p. 58, 59, 111]

[ZEP+06] F. Zhai, Y. Eisenberg, T. N. Pappas, R. Berry, and A. K. Katsaggelos. Rate-distortion optimized hybrid error control for real-time packetized video transmission. *IEEE Transactions on Image Processing*, 15(1):40–53, January 2006. [cited at p. 26]

[ZGL+07] Y. Zhang, W. Gao, Y. Lu, Q. Huang, and D. Zhao. Joint source-channel rate-distortion optimization for H.264 video coding over error-prone networks. *IEEE Transactions on Multimedia*, 9(3):445–454, April 2007. [cited at p. 96]

[ZR97] M. Zorzi and R. R. Rao. On the statistics of block errors in bursty channels. *IEEE Transactions on Communications*, 45(6):660–667, June 1997. [cited at p. 98]

[ZRR00] R. Zhang, S. L. Regunathan, and K. Rose. Video coding with optimal inter/intra-mode switching for packet loss resilience. *IEEE Journal on Selected Areas in Communications*, 18(6):966–976, June 2000. [cited at p. 27, 28]