Technische Universität München

Lehrstuhl für Medientechnik

Quality of Experience-driven Multi-Dimensional Video Adaptation

Fan Zhang

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.-Ing. Thomas Eibert
Prüfer der Dissertation:	1.	Univ.-Prof. Dr.-Ing. Eckehard Steinbach
	2.	Jun.-Prof. Dr.-Ing. Alexander Raake (Technische Universität Berlin)

Die Dissertation wurde am 16.09.2013 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 20.03.2014 angenommen.

To my beloved wife, for her constant support and love.

Abstract

Video adaptation is a key technology for universal video access in heterogeneous communication environments. The main challenge in this context is the selection of an optimal combination of Multi-Dimensional Adaptation (MDA) operations (such as spatial down-sampling, frame dropping and adjustment of quantization parameters) to maximize the user's Quality of Experience (QoE) under certain resource constraints. To achieve this goal, different factors that affect the perceptual quality need to be considered. The focus of this thesis is to solve this optimization problem by a QoE-driven approach.

To begin with, extensive subjective experiments are conducted to study the human preference between temporal and spatial details for different types of video content and a wide range of bit rates. A detailed analysis of the experimental data unveils how the perceived quality is influenced by the video content, available transmission rate and MDA operations.

Moreover, the impacts of SNR, temporal and spatial resolution on the perceptual video quality are modelled separately based on the observations from the subjective test and a multi-dimensional video quality metric MDVQM is proposed. Performance evaluations using subjective quality ratings show that the proposed video quality metric provides accurate quality estimation in the presence of different spatial and temporal quality impairments.

Furthermore, accurate rate adaptation based on ρ-domain analysis is studied. The proposed rate control algorithm combines ρ-domain rate model and header size estimation for H.264/AVC video. Experimental results show that the proposed algorithm achieves better rate control accuracy and video quality when compared with the original ρ-domain rate control algorithm.

Finally, this thesis ends up with a QoE-driven multi-dimensional video adaptation scheme combining both the proposed video quality metric and the rate control algorithm. The video quality metric is used to predict the resulting QoE under different adaptation modes. The optimal combination of adaptation operations is determined by considering both the resulting QoE and computational complexity. Significant QoE improvement against conventional video adaptation schemes has been confirmed by performance evaluation using various types of video contents.

Zusammenfassung

Videoanpassung ist eine Schlüsseltechnologie für universellen Video-Zugang in heterogenen Kommunikationsumgebungen. Die größte Herausforderung in diesem Kontext ist die Auswahl der optimalen Kombination von Multi-Dimensionalen Anpassungsoperationen (MDA), um die Nutzerzufriedenheit (QoE) unter begrenzten Ressourcen zu maximieren. Um dieses Ziel in heterogenen Umgebungen zu erreichen, müssen verschiedene Faktoren, welche die wahrnehmbare Qualität beeinflussen, berücksichtigt werden. Der Schwerpunkt dieser Arbeit ist, dieses Optimierungsproblem durch einen QoE-orientierten Ansatz zu lösen.

Zunächst werden umfangreiche subjektive Tests durchgeführt, um die menschliche Präferenz zwischen zeitlichen und räumlichen Details für verschiedene Videoinhalte und einen breiten Bereich von Datenraten zu studieren. Eine detaillierte Analyse der experimentellen Daten zeigt, wie die wahrgenommene Qualität von dem Videoinhalten, von verfügbaren Übertragungsraten und von MDA-Operationen beeinflusst wird.

Außerdem, basierend auf den Beobachtungen aus den subjektiven Tests, werden die Auswirkungen von SNR und zeitlicher und räumlicher Auflösung auf die wahrgenommene Videoqualität separat modelliert und eine multi-dimensionale Videoqualitätsmetrik MDVQM wird vorgeschlagen. Leistungsbewertungen mit subjektiven Qualitätsbewertungen zeigen, dass die vorgeschlagene Qualitätsmetrik in Gegenwart von unterschiedlichen räumlichen und zeitlichen Qualitätsdegradierungen eine genaue Qualitätsschätzung liefern kann.

Weiterhin wird eine genaue Datenratenanpassung basierend auf ρ-Domain-Analyse untersucht. Der vorgeschlagene Datenratensteuerungsalgorithmus kombiniert das ρ-Domain Datenratenmodell mit der Schätzung der Header-Größe für H.264/AVC Video. Experimentelle Ergebnisse zeigen, dass der vorgeschlagene Algorithmus die Datenrate genauer steuern kann und bessere Videoqualität erreicht, verglichen mit dem ursprünglichen ρ-Domain Ratenalgorithmus.

Schließlich endet diese Dissertation mit einer QoE-orientierten multi-dimensionalen Videoanpassungsmethode, welche die vorgeschlagene Videoqualitätsmetrik und den Datenratensteuerungsalgorithmus kombiniert. Die Qualitätsmetrik wird verwendet, um die resultierende

Nutzerzufriedenheit unter verschiedenen Anpassungsmodi vorherzusagen. Die optimale Kombination von Anpassungsoperationen wird durch die Berücksichtigung der resultierenden Nutzerzufriedenheit und Rechenaufwands ermittelt. Signifikante QoE-Verbesserungen im Vergleich zu herkömmliche Videoanpassungssystemen wird durch eine Leistungsbewertung mit verschiedenen Videoinhalten bestätigt.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Dr.-Ing Eckehard Steinbach, who gave me the opportunity to be involved in this interesting research topic and whose guidance, discussions and suggestions were of great help during the work. And I am thankful to my second examiner Prof. Dr.-Ing Alexander Raake for reading my thesis and giving pertinent comments. At the same time, I would like to thank Prof. Dr.-Ing ...... for chairing my thesis defence.

Moreover, I wish to thank Deutscher Akademischer Austausch Dienst for partially supporting my doctoral research.

Furthermore, I thank the current and previous members of Institute for Media Technology at TU München. They have made my life and work at our institute an enjoyable experience. I would like to give special thanks to Dr.-Ing Wei Tu, Dr.-Ing Yang Peng, Dr.-Ing Hu Chen, Jianshu Chao and Xiao Xu for the fruitful joint work and discussions.

I would like to express my deepest gratitude to my wife Cong Li who had been at my side in a supporting role giving me all the love I needed. Finally, I would like to thank my parents who, from the first day of my education, have provided tremendous support, constant encouragement, and great guidance.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description	Definition
UMA	Universal Multimedia Access	page 1
CDN	Content Delivery Network	page 1
PSTN	Public Switched Telephone Network	page 1
MDA	Multi-Dimensional Adaptation	page 3
QoE	Quality of Experience	page 3
VQM	Video Quality Metrics	page 3
MOS	Mean Opinion Score	page 7
CBR	Constant Bit Rate	page 32
CI	Confidence Interval	page 37
DMOS	Differential Mean Opinion Score	page 37
DSCQS	Double Stimulus Continuous Quality Scale	page 8
ACR	Absolute Category Rating	page 9
ACR-HR	ACR-Hidden Reference	page 10
SAMVIQ	Subjective Assessment of Multimedia VIdeo Quality	page 10
HVS	Human Visual System	page 15
MB	Macro-Block	page 62
PSNR	Peak-Signal-to-Noise Ratio	page 15
PC	Pearson Correlation	page 41
OR	Outlier Ratio	page 41
RMSE	Root Mean Square Error	page 41
FR	Full-Reference	page 16
NR	No-Reference	page 17
RR	Reduce-Reference	page 17
SI	Spatial Information	page 11
TI	Temporal Information	page 11
SA	Spatial Activity	page 29
TA	Temporal Activity	page 29
QP	Quantization Parameter	page 26
SR	Spatial Resolution	page 26

Abbreviation	Description	Definition
TR	Temporal Resolution	page 26
SCF	Spatial Correction Factor	page 52
TCF	Temporal Correction Factor	page 45
STCF	Spatial-Temporal Compensation Factor	page 56
SSIM	Structural SIMilarity	page 18
PVS	Processed Video Sequence	page 32
SRC	SouRCe video sequence	page 30
CAVLC	Context Adaptive Variable Length Coding	page 65
CBP	Coded Block Pattern	page 67
SPS	Sequence Parameter Set	page 65
PPS	Picture Parameter Set	page 65
MV	Motion Vector	page 64
MVD	Differential Motion Vector	page 67
RDO	Rate Distortion Optimization	page 61

Chapter 1

Introduction

1.1 Motivation

Over the past 20 years, internet video applications (such as video streaming, video telephony, video sharing, etc.) have gradually become an indispensable part of our daily lives. The fast development of mobile networks has inspired the idea of Universal Multimedia Access (UMA)[MSL99] which has further propelled the boom of video services in the internet. The aim of UMA is to allow the users to access the multimedia content at anytime and from anywhere.

This has raised a big challenge to the traditional video transmission system due to the different display characteristics of the video content (such as frame rate and spatial resolution), the heterogeneity of the transmission channels, the time-varying nature of the mobile networks, and the diversity of the users' end-devices.

Figure 1.1 shows a typical application scenario for video streaming. The video contents are encoded and stored on media servers, which are normally owned by content providers (such as film companies or news agencies). The media servers are connected to the core network. The core network might be a traditional best-effort IP network or a special Content Delivery Network (CDN). Then between the core network and the end-users, there might be diverse access networks (which are often referred to as "the last mile" in the delivery path). The access networks can be classified by different access rates, from low-bitrate networks such as traditional dial-up networks over PSTN or 2.5G mobile networks (GPRS), to middle bit-rate networks such as low-speed xDSL or 3G mobile networks, to broadband services such as high-speed xDSL, 4G mobile networks, Wifi/WiMax or fiber networks.

It is quite likely that the content server does not have any a priori knowledge about the device capacities or network conditions of the end users. So conventionally, the video content is often encoded with the goal of optimizing the rate-distortion performance. Therefore a

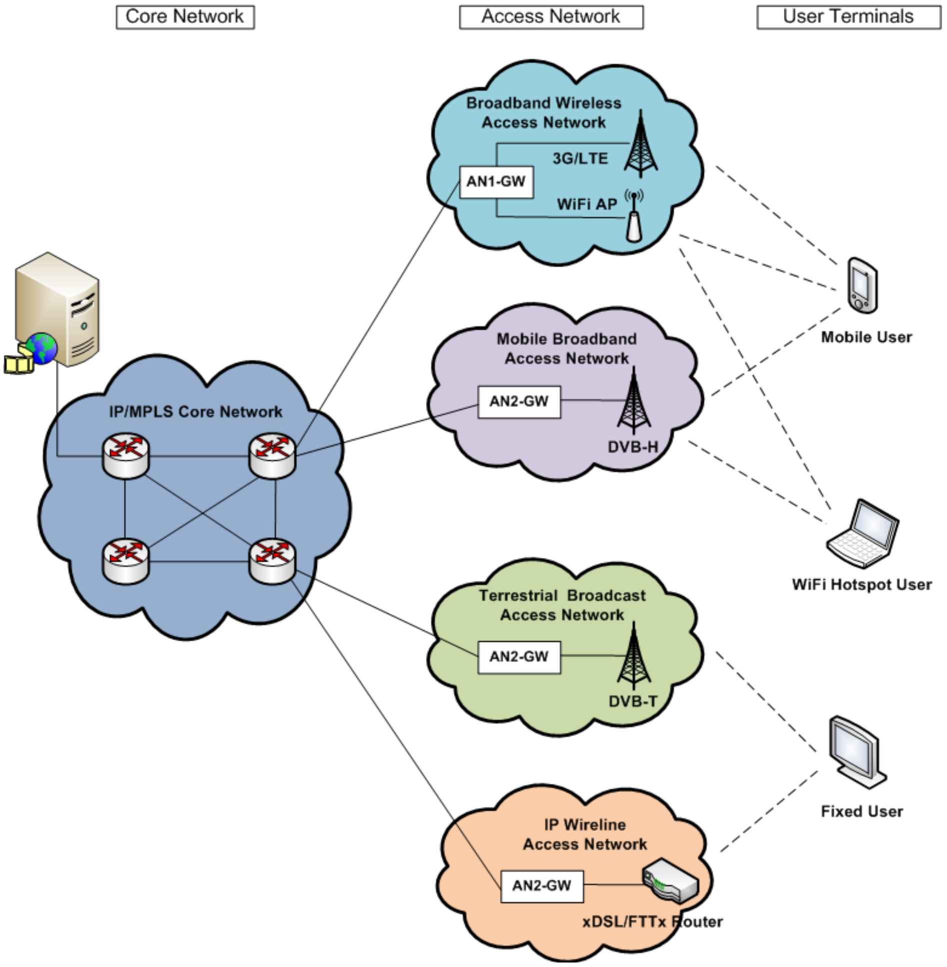


Figure 1.1: Media delivery over heterogeneous access networks (adapted from [Lab])

video stream needs to be adapted along the delivery path before it can be delivered to the end-user.

One way to manage the heterogeneity is simulcast [CAL96, MJ96], in which the same video content is encoded at several different bit-rates or even using different coding standards. Then different versions of the video content are delivered to the end users according to their specific characteristics. This strategy might work well for wired networks where the transmission capacity of the users is fixed and relatively stable. For end-users connected over mobile networks, however, the available transmission rate is time-varying and hard to predict, so the chosen bitstream may not match the user's transmission characteristics very accurately. Also, storing multiple versions of the same content consumes more storage on the content server, which is costly for content providers.

An alternative solution to simulcast is to apply video adaptation at the edge of the access network. In this solution, the video stream received by the end-user is no longer directly transmitted from the content server, but is generated at an intermediate network node by adapting the original video stream from the server to match the user's network characteristics and device capacities. This could be done, for example, on a proxy which is built on top of the gateway nodes (base stations, access points, routers etc.) in Figure 1.1. If the video stream is delivered through a CDN, normally there are also special proxy servers deployed at the edge of the core network, which can also be used to perform the video adaptation tasks. In this solution, only one high quality version of the video content needs to be stored on the content server. When a user requests to access a video content, the stored video stream is first delivered to the proxies, the proxy then performs the video adaptation in real-time to meet the user's requirement. Compared with the simulcast solution, performing the video adaptation at the edge of the core network can save valuable storage space on the content server. Furthermore, since only one version needs to be transmitted through the core network and the video stream can be cached on the proxies, this solution can also reduce the traffic in the core network.

Video adaptation can be performed by adjusting different parameters of the encoded stream such as quantization step-size, frame rate and spatial resolution of the video content. Multi-Dimensional Adaptation (MDA) refers to the schemes where the impacts of all these factors are considered jointly to meet the resource constraints and optimize the video quality. Joint optimization among different dimensions offers us more opportunities for quality optimization but also raises several new challenges, which include the assessment of video quality under different spatial/temporal resolutions and the selection of the optimal combination of adaptation operations [Wan05]. These issues in MDA can be solved by a Quality-of-Experience (QoE) driven approach.

Since the target users of most video delivery systems are human beings, the most reasonable way for quality assessment is to collect the user's opinions on the delivered video streams. The user's satisfaction level is often referred to as Quality-of-Experience of the users. The most accurate way to measure QoE is by conducting subjective tests. However, subjective tests are usually costly and time-consuming, so this approach is not practical for the evaluation of video quality in real-time. Due to the limitations of subjective quality assessment and the increasing demand for in-service quality assessment, there have been intensified studies of perceptual Video Quality Metrics (VQM) which aim to estimate the QoE of a video processing system by taking into account the characteristics of human visual perception.

QoE-driven MDA schemes utilize perceptual VQMs to assess the video quality and make optimal adaptation decisions on the fly when needed. The general diagram of a QoE-driven multi-dimensional video adaptor is shown in Figure 1.2.

In Figure 1.2, the resource allocator collects the feedback information (such as network conditions and user's preference) from the channel and the end-users. Based on the collected information, it determines the target source coding bit-rate for the adaptation operation

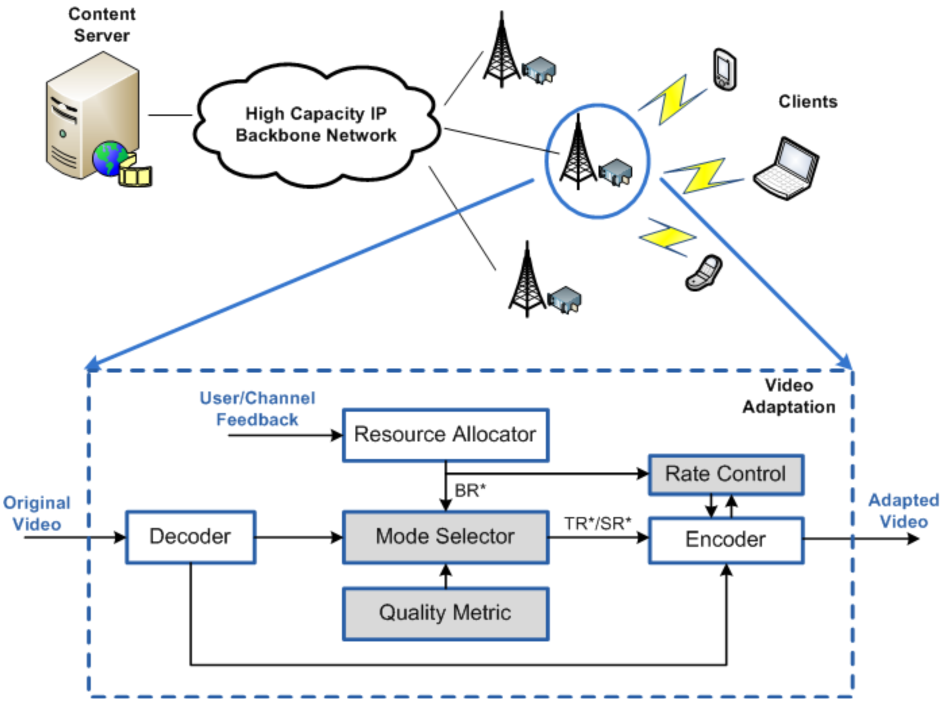


Figure 1.2: Block Diagram of a QoE-driven multi-dimensional video adaptation system

(BR^*) and passes this information to the mode selector. On the other hand, the incoming video stream is decoded by the video decoder and the decoded video frames are used by the mode selector to extract necessary video features. The perceived quality of the adapted videos under different adaptation modes can be estimated by feeding all the information to a perceptual video quality metric. Taking into account various factors (such as the computational complexity and resulting QoE of different adaptation operations), the mode selector determines the optimal parameters (e.g., spatial resolution SR^* and frame rate TR^*) for the video adaptation operations. According to the decisions of the mode selector, the encoder performs proper adaptation operations. The rate control module interacts with the encoder to guarantee the adapted video meets the rate requirements given by the resource allocator.

In this thesis, various aspects of such a QoE-driven MDA system are studied. More specifically, the focus is put on the estimation of the perceived video quality (quality metric), the selection of optimal adaptation mode (mode selector) as well as the accurate control of the bitrate (rate controller). The corresponding modules are marked in grey in Figure 1.2.

1.2 Summary of Main Contributions

The main contributions of this dissertation can be summarized as follows:

- The individual and overall impact of different video properties (such as the quantization step-size, the spatial resolution, the frame rate) on the perceived video quality are studied through specifically designed subjective tests. Prior works in the literature concerning video quality assessment mainly focus on videos with fixed frame rate. In this work, subjective tests are conducted which help us to understand how the perceived video quality is affected when the spatial and temporal resolution of the video content are changed separately or even jointly.

- An accurate no-reference VQM for evaluating the perceived video quality at different frame rates and spatial resolutions is presented and its prediction performance is analyzed. The proposed metric models the overall video quality as the product of separate items, with each of the items simulating the impact of quantization, frame dropping and spatial down-sampling, respectively. All the features used in the VQM can be easily computed from the encoded bitstream so that it is well suited for in-service video quality estimation. The performance of the proposed metric is also validated by the results of the subjective tests.

- An accurate rate control algorithm based on ρ-domain analysis is proposed. The approach uses a two-stage encoder structure to resolve the inter-dependency between RDO and ρ-domain rate control. The size of the header information is estimated using an improved rate model which considers the different components in a macroblock header. Experimental results show that the proposed algorithm achieves better rate control accuracy and video quality when compared with the original ρ-domain rate control algorithm. The proposed rate control algorithm can be used together with the video quality metric to perform accurate bitrate adaptation for QoE optimization.

- Based on the proposed VQM and rate control algorithm, a QoE-driven MDA scheme is developed for optimizing the perceived video quality. The adaptation scheme uses the proposed VQM to estimate the resulting video quality under different adaptation modes and then determines the optimal adaptation mode by taking into account the video quality as well as the computational complexity. The algorithm is evaluated and shown to provide better performance than conventional adaptation schemes.

1.3 Outline of the Thesis

The rest of the thesis is arranged as follows. Chapter 2 outlines the main aspects of video quality assessment and gives a review of the state-of-the-art video quality metrics. In Chapter 3, a multi-dimensional video quality metric is proposed and evaluated with results from extensive subjective tests. Then, an improved ρ-domain rate control algorithm for H.264/AVC video with header size estimation is presented in Chapter 4. The proposed QoE-aware video adaptation scheme is presented in Chapter 5 together with performance evaluation. The thesis concludes in Chapter 6 with a summary of the results.

Parts of this thesis have been published in [ZS11, SZPD12].

Chapter 2

Overview of Video Quality Assessment

Nowadays, video data is responsible for a considerable part of the total internet traffic due to the boom of various video related services and the remarkable evolution of network technologies and mobile devices. Video quality assessment is fundamental to monitor and guarantee the quality of these video services.

Depending on whether human observers are involved in the assessment process, video quality assessment can be performed subjectively or objectively. The purpose of subjective video quality assessment is two-fold. First, it can be used to evaluate or compare the performance of different video processing algorithms/systems. Second, it can help us to find out how the perceived video quality is affected under different conditions. On the other hand, objective video quality assessment estimates the video quality using video features which can be measured and computed objectively, thus makes it possible to monitor and optimize the video quality automatically. Both of them are indispensable parts of designing and evaluating a video system which aims to provide the best QoE to the users. In this chapter, background and related work in the field of both subjective and objective video quality assessment are discussed. Section 2.1 provides a summary of the guidelines given in the ITU standard documents for conducting subjective tests. This is followed by a review of the development of objective video quality metrics in Section 2.2.

2.1 Subjective Video Quality Assessment

In subjective video quality assessment, a set of test video sequences are presented to the human observers (also referred to as test subjects). The task of the human observers is to provide their opinions about the video quality. There are two basic forms of subjective tests: the paired comparison approach and the Mean Opinion Score (MOS) approach. In paired comparison, two test videos are displayed side by side and the human observers need to judge which one has a better quality. By the MOS approach, the videos are displayed one by one and the human observers are asked to rate the quality of each video. The MOS value is calculated as the mean value of the collected ratings.

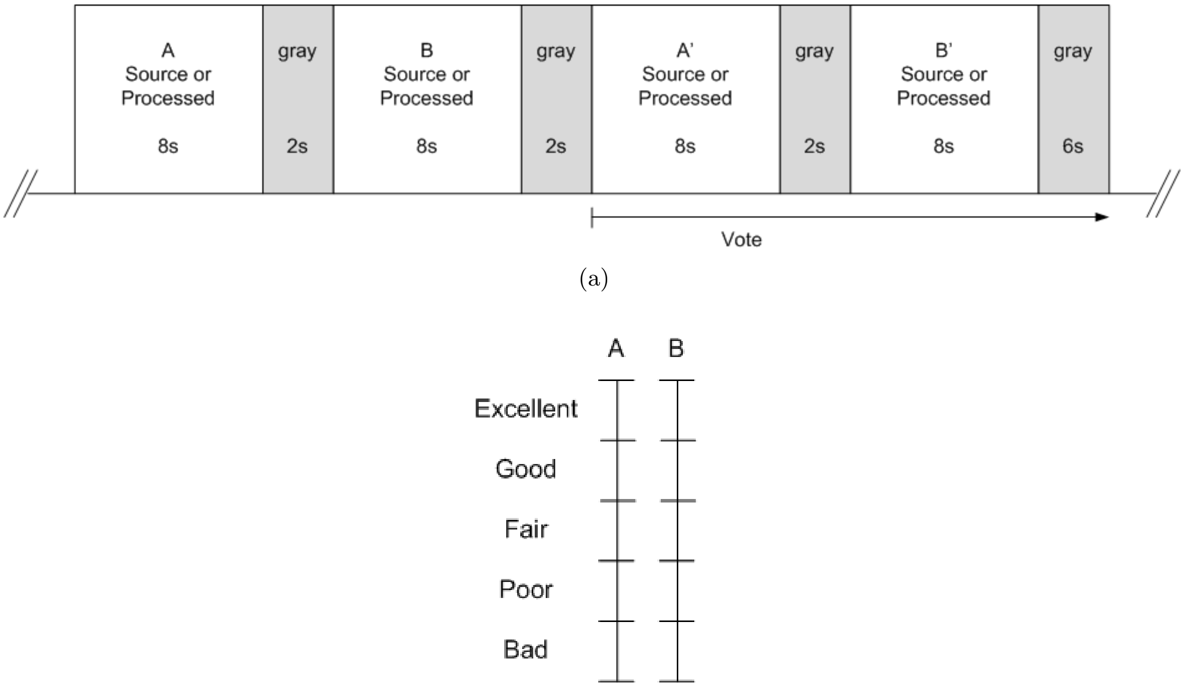
The judgement of human beings tends to be affected by many factors, such as the health situation, the mood as well as the surrounding environment. Therefore, to ensure the accuracy of the results, subjective tests must be conducted in a controlled manner. For this purpose, the International Telecommunication Union (ITU) has established a series of recommendations to standardize the design and procedure of subjective tests. The most important documents are ITU-R Rec. BT.500-11 [ITU99] (for television applications), ITU-T Rec. P.910 [ITU98] and ITU-T Rec. BT.1788 [ITU07] (for multimedia applications). The most important aspects defined in the documents regarding preparation and conduct of subjective tests are summarized in the following.

2.1.1 Test Method

When designing a subjective test, the first question one needs to answer is the purpose of the test. The standard documents recommend different test methods addressing various test scenarios. The test method should be carefully selected depending on the specific goal of the test. The following is a brief description of the most widely used test methods. Since a comparison of system performance is not the focus of this thesis, only test methods following the MOS approach are discussed. The reader can refer to [ITU98, ITU99, ITU07] for more details.

- Double-Stimulus Continuous Quality Scale (DSCQS)
  DSCQS is a Double Stimulus (DS) method defined in [ITU99], in which two videos, i.e. the original source sequence (also referred to as reference sequence) and a processed version of the same sequence, are presented twice to the test subjects. The presentation order of the two sequences is randomized, i.e., sometimes the reference sequence is presented first and sometimes the processed sequence is presented first. The test subjects are asked to give their ratings at the second presentation of each video. This voting procedure is shown in Figure 2.1a. The test subjects use a continuous grading scale as shown in Figure 2.1b for the rating. As pointed out in [ITU99], DSCQS is more resilient to contextual effects when compared with other test methods (contextual effects refer to the phenomenon that the results of the subjective tests tend to be affected by the level and ordering of the impairments that appear in the tests. For example, if an impaired test sequence is presented after several high quality test sequences, the viewers may give it a lower score than it normally deserves). This is due to the fact that the original source sequence is always available in DSCQS to serve as a reference when the test subjects rate the processed sequences. However, the use of a reference for each test sequence also

causes DSCQS to be very time-consuming and only a small number of test sequences can be evaluated during a session, which is the major disadvantage of DSCQS.



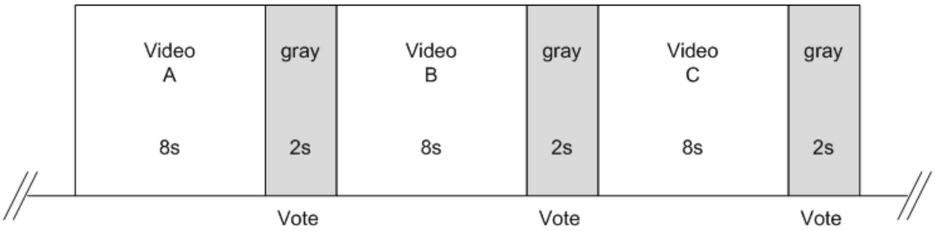
(a)

(b)

Figure 2.1: Double-Stimulus Continuous Quality Scale (DSCQS) [ITU99]: (a) Presentation structure; (b) rating scale.

- Absolute Category Rating (ACR)
  This method is a Single Stimulus (SS) method defined in [ITU98]. In the ACR method, the test sequences are presented one at a time and the test subjects are asked to rate each sequence after the presentation. The procedure of ACR is shown in Figure 2.2a. In order to alleviate the impact of contextual effect, the presentation order of the test sequences should be randomized for each individual test subject. Typically, ACR uses a five-level categorical grading scale as shown in Figure 2.2b. A nine-level scale can also be used in case a higher discriminative power is desired, as suggested in [ITU98]. Because each sequence is presented only once before being rated, ACR allows more test sequences to be evaluated during the same time interval in comparison with DSCQS. But the drawback of ACR is that it is a SS method, so it may be seriously affected by contextual effects and therefore, ACR needs more participants to achieve the same reliability as DSCQS [ITU05a]. The efficiency of ACR is partially offset by this drawback. Due to this reason, VQEG has used an enhanced version of ACR in its Multimedia Test [VQE07]. In the improved method, the original version of each video content is randomly inserted into

the test dataset to serve as a hidden reference. Therefore, this improved ACR method is also referred to as ACR-HR (ACR with Hidden Reference). In [HTG05], a comparison is performed between ACR-HR and DSCQS for low-resolution videos, the results show that ACR-HR can provide the same reliability as DSCQS while keeping the simplicity and efficiency of ACR.



(a)

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

(b)

Figure 2.2: Absolute Category Rating (ACR) [ITU99]: (a) Presentation structure; (b) rating scale.

- Subjective Assessment of Multimedia VIdeo Quality (SAMVIQ)

  SAMVIQ is a new assessment methodology defined in [ITU07]. In SAMVIQ, the evaluation is conducted scene by scene. Each scene contains all the processed test sequences of the same video content. To alleviate the contextual effects, an explicit reference and a hidden reference of the same content are also included in each scene. The hidden reference is inserted randomly into the processed test sequences.

  The major difference between SAMVIQ and conventional test methods (such as DSCQS and ACR) is that the test subjects can control the order of the presentation as well as start/stop the presentation of a test sequence at any time. There is no strict timing for the rating of each test sequence. The test subjects can freely make comparisons between a processed test sequence and the reference sequence or between two processed test sequences and then give or adjust their rating for individual test sequences accordingly. This allows SAMVIQ to produce reliable subjective ratings. Figure 2.3 shows the presentation structure and rating scale for $SAMVIQ$.

> In [BHTHB06], the performance of SAMVIQ and ACR-HR is compared using test sequences with CIF (352x288) resolution. The results suggest that the subjective ratings produced by both methods are very similar. Considering the higher efficiency of ACR-HR (In SAMVIQ, test subjects tend to spend more time to make comparisons between different sequences), ACR-HR is considered to be the preferred method. In comparison, the impact of spatial resolution on the result accuracy of SAMVIQ and ACR-HR is studied in [PP08]. The results show that for video contents of high resolutions (VGA/HD), the results from SAMVIQ are more precise than those from ACR-HR for the same number of test subjects.

According to the above discussion of different test methodologies as well as the number and spatial resolution of the test sequences, SAMVIQ is selected as the test method in the work presented in Chapter 3 for collecting subjective ratings.

2.1.2 Test Material

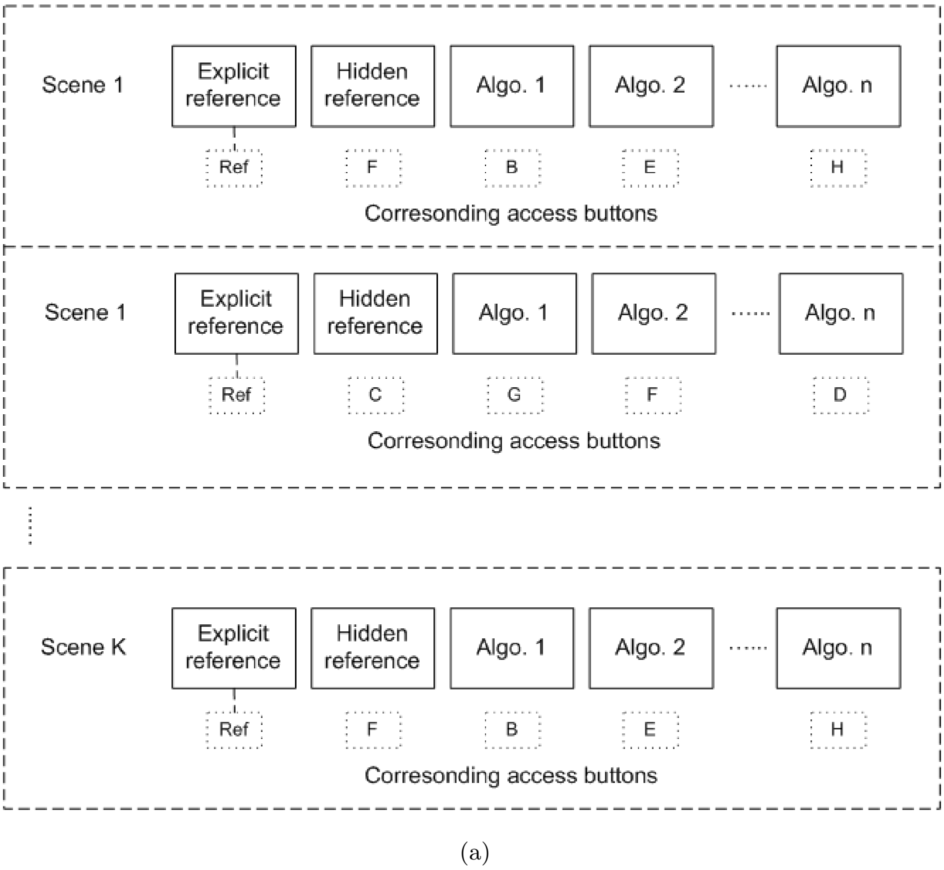
Since the purpose of video quality assessment is to evaluate the performance or help optimize the QoE of a certain video processing system, the target application of the system under consideration should be taken into account when selecting the test materials. Also, to improve the reliability of the test results, it is important that a wide variety of materials are used in the test. The variety of the test materials refers to not only the diversity of the video contents but also the quality range of the processed sequences. In the subjective tests conducted by VQEG [VQE00, VQE03, VQE08, VQE09], the Spatial perceptual Information (SI) and Temporal perceptual Information (TI) are used to determine the characteristics of the video contents. 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 $Sobel(Fn)$ is the filtered frame by the Sobel filter. Sobel filter is widely used in image processing algorithms to compute an approximation of the gradient magnitude at each point in the input image. The 2D Sobel filter uses a pair of 3x3 convolution kernels given in Eq.(2.3):

$$K_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \quad and \quad K_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{2.3}$$



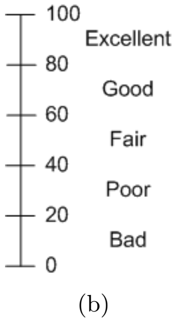


Figure 2.3: Subjective Assessment of Multimedia VIdeo Quality (SAMVIQ) [ITU07]: (a) Presentation structure; (b) rating scale.

and the filtered frame of the Sobel filter is calculated as:

$$\begin{gathered} G_x = K_x * A \quad \mathit{and} \quad G_y = K_y * A \\ G = \sqrt{{G_x}^2 + {G_y}^2} \end{gathered} \tag{2.4}$$

where A is the input image and G is the filtered image. The operator '$*$' denotes the 2D convolution operation. The selected video contents should span the full range of scene characteristics which is of interest to the system under test.

To achieve precise and reliable quality ratings, the quality range of the test materials should be as large as possible. Otherwise, if the quality range is too narrow, the test subjects tend to give quality scores which exaggerate the quality difference of two test sequences. In most cases, it is a good practice to include processed sequences with extremely high and low quality in the test material.

2.1.3 Test Subjects

When selecting the test subjects, the number and type of the viewers should be carefully considered. In most standard documents [ITU98, ITU99, ITU07], it is suggested that the number of test subjects should not be less than 15 in order to produce reliable results. In practice, the appropriate number of participants should be selected according to the reliability of the test method as well as the expected precision of the results. For example, VQEG recommends to use at least 24 test subjects for its Multimedia Test [VQE07] using the ACR-HR method while the European Broadcast Union (EBU) suggests to use at least 15 test subjects in its video codecs evaluations [KSW05] using SAMVIQ.

Two types of test subjects should be distinguished, i.e. experts and non-experts. The term "non-expert" refers to people who "are not directly concerned with picture quality as part of their normal work and are not experienced assessors" (quoted from [ITU07]). All the standard documents suggest that the test subjects in the subjective tests should be non-experts. The consideration here is that experts tend to have a fixed or preconceived way of evaluating the image/video quality which is different from that of non-experts. Since non-experts compose a much larger part of the public who consume the video contents, the results from non-experts are more representative and reliable. That does not mean, however, that the test subjects do not have any background knowledge about the tests. They need to understand the type of artifacts and quality range that are expected in the tests. This can be done through a training session before the formal test session.

2.1.4 Test Procedure

In general, a subjective test can be divided into five phases: preparation, introduction, training session, test session and post-processing.

In the preparation phase, the test environment (including the room used for the test as well as the display devices) should be set up according to the guidelines provided in [ITU99]. The visual acuity of the test subjects should also be checked.

Before the test starts, both a written and an oral introduction should be given to the test subjects. The content of the introduction should include the timing and organization of the test, how the test sequences will be presented, how the test subjects should rate the sequences and sometimes, the expected types of impairment which occur in the tests, etc.

After that, a training session should also be provided to help the test subjects get familiar with the test interface, voting process as well as the types of video contents and visual artifacts in the test. The procedure and video materials used in the training session should be similar to those of the formal test session, but the same video contents should not be included in the test session again. The ratings collected from the test session should not be considered in the final results. Any questions from the test subjects about the test can be answered at this point. After the formal test session begins, no further questions are allowed.

After the test sessions, the collected subjective data should be screened and outliers should be removed. Different screening processes are defined in the standard documents for the use of different test methodologies. In Section 3.3.6, the screening process defined for SAMVIQ is discussed in more details. For more information about different screening processes, the reader can refer to the corresponding ITU recommendations [ITU98, ITU99, ITU07].

The standardization efforts discussed above have made subjective tests the most reliable way for video quality assessment. Although subjective quality assessment is not suitable for real-time applications, they are still very important in the sense that they provide the "ground-truth" data for the design and verification of objective video quality metrics which enable real-time in-service video quality evaluations. Also, the information from the subjective tests can help us to understand the properties and limits of the human visual system.

2.2 Objective Video Quality Assessment

Objective video quality assessment can be used instead of subjective quality assessment whenever the involvement of human beings needs to be avoided. It can be useful for a number of scenarios throughout all the phases of building a video processing and communication system, such as:

- Estimation of necessary resources to deliver a certain quality level at the planning stage of a network service.
- Comparison of different processing algorithms when designing the system.
- Verification of system performance during the testing phase.

- Monitoring and optimization of the perceived video quality when the system is running.

The basic idea of objective video quality assessment is to use mathematical quality metrics to estimate the perceptual video quality in an automatic and objective manner. The most widely used objective video quality metric nowadays is perhaps the Peak Signal-to-Noise Ratio (PSNR), which can be calculated as follows:

$$MSE = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} [I_1(i,j) - I_2(i,j)]^2 \tag{2.5}$$

and

$$PSNR = 10 \cdot log \frac{255^2}{MSE} \tag{2.6}$$

where W and H denote the width and height of the picture, I_1 and I_2 are the corresponding frames in the original and processed video, respectively. Since the human eyes are more sensitive to the details of the luminance component in an image or video than those of the chrominance components, normally the PSNR value is only calculated for the luminance component. And the PSNR value for a video sequence is calculated as the average PSNR value over all the frames included in the sequence.

The popularity of PSNR is largely due to its simplicity and clear physical meaning. It does provide a good estimation of the perceived video quality as far as the video content and the type of distortion are not changed [EF95, HTG08]. However, for more complicated cases where different video contents, different frame rates and spatial resolutions are to be considered, the performance of PSNR is not satisfactory [Gir93, EF95, Win99, WB02]. The major drawback of PNSR is that it measures only the fidelity of the signal without considering the characteristics of the video content, the Human Visual System (HVS) as well as the interaction between the two. In this sense, there is no difference between a video signal and an audio/speech signal or signals of any other type. In [WM08], this kind of pure fidelity measurement is named "data metric" to differentiate it from perceptual metrics where psychophysical aspects are considered.

2.2.1 Classification of Objective Video Quality Metrics

If the HVS is considered as a processing system, then the video content is its input and the perceived video quality is its output. One straightforward way for predicting the system output is to find out the internal components of the system and to model the behavior of the fundamental functional blocks. This is the basic idea behind the so called HVS-based approach [WM08]. Over the years, several famous HVS-based VQMs have been proposed such as the Visible Difference Predictor (VDF) by Daly [Dal93], the Sarnoff model proposed by Lubin [Lub97], the Perceptual Distortion Metric (PDM) proposed by Winkler [Win98, Win99], as well as the Digital Video Quality (DVQ) proposed by Watson [WHM01]. In [WSB03a], a general framework of these HVS-based VQMs is summarized as shown in Figure 2.4. From the framework, it can be seen that several most important perceptual features (such as light adaptation, contrast sensitivity, masking and facilitation, error pooling, etc.) of the HVS are considered and integrated to imitate the visual perception process of humans. The biggest challenge for the VQMs of this category is that the human visual perception is a very complex process which involves not only the signal reception in the eyes but also the processing of the resulting signals in the human brain. Although our understanding of the whole system is much better than a decade ago, there is still a long way ahead of us until the visual perception can be modeled accurately enough. Also, the computational complexity required by the HVS-based approach has limited the scope of its possible application. Since the target of the work in this dissertation is to build a video quality metric for real-time video adaptation, HVS-based metrics are not our focus. Interested readers can refer to the overviews in the literature for more details of HVS modeling [WSB03a, UE07, WM08, CSRK11].

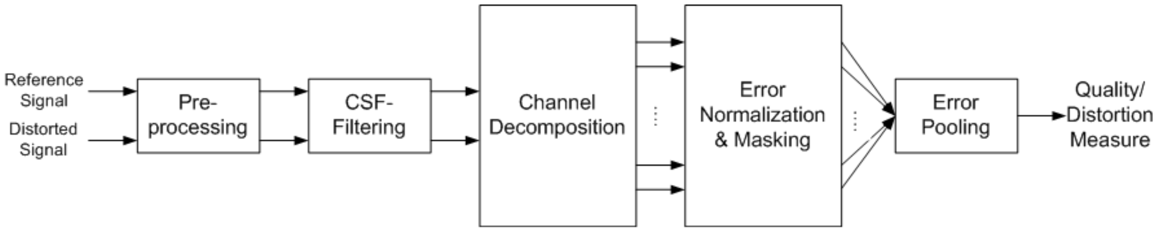


Figure 2.4: General framework of HVS-based visual quality metrics (adopted from [WSB03a]).

The second way of modeling the system is to treat it as a black box. Then the system response can be approximated by observing the relationships between the input and output signal. In [WM08], it is referred to as the Engineering Approach. In this way, complicated modeling of the building blocks of the HVS can be avoided and the problem of predicting the output signal can be solved by numerical approaches. Although the accuracy and universality of the engineering approach is not as good as the HVS-based approach, it is more suitable for real-time applications. Therefore, the engineering approach is adopted in Chapter 3 for developing the video quality metric.

Another traditional classification of video quality metrics is based on the amount of reference information available for quality estimation [ITU00]. If the metric requires the access to the whole reference video sequence for the quality estimation of a distorted video (as shown in Figure 2.5), then it is classified as a Full-Reference (FR) video quality metric. When humans determine the quality of an image/video, it is always helpful to have the original visual content as a reference for the comparison (for example, to identify the type and strength of distortions). Similarly, it is generally accepted that the use of more reference information can help to reduce the complexity and improve the accuracy of the quality metric [ITU00].

However, due to the dependency on the whole reference information, FR metrics can only be used in applications where the original video content is available at the place where quality estimation is performed, such as quality optimization at the source side or test in a laboratory scenario.

For a broader range of video systems where the quality estimation needs to be done in the middle of the network or at the end-user side, FR metrics are not feasible. This has promoted the development of No-Reference (NR) video quality metrics, where the video quality is estimated solely on the distorted video contents without any reference to the original content (as shown in Figure 2.7). NR quality metrics can be used at any place within the system, so they can be applied to a wider range of applications (such as for real-time quality estimation in a transmission scenario). However, the design of NR quality metrics faces more difficulties than FR metrics due to the lack of reference information. This is reflected by the number of established ITU-T standards for different classes of quality metrics as discussed in Section 2.2.2.

The third class of video quality metrics is the Reduced-Reference (RR) metrics, which can be seen as a compromise between FR and NR solutions. In RR metrics, normally a set of important video features are extracted at the source side and transmitted using an ancillary communication channel to the place where the video quality is estimated. The same features are also extracted from the distorted video contents and quality degradations caused by distortions is estimated by comparing the features from both sides (as shown in Figure 2.6). As discussed above, the more reference information is available, the more accurate is the quality estimation. But this also requires more transmission capacity of the ancillary channel. So the most critical issue in the design of RR metrics is the tradeoff between accuracy and the amount of overhead information.

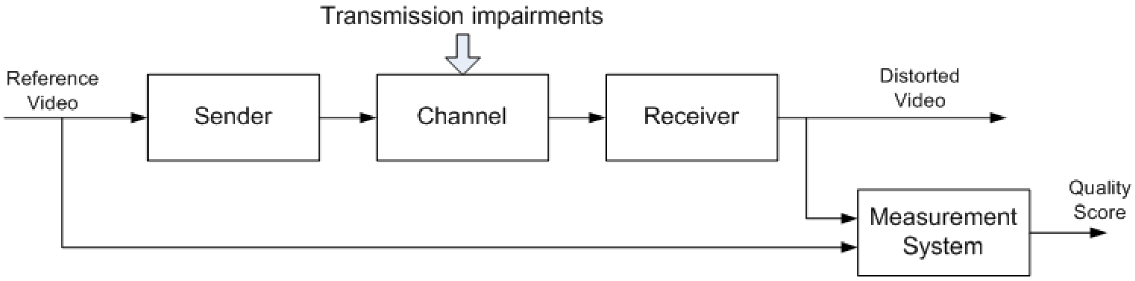


Figure 2.5: Block diagram of a full-reference video quality assessment system (adapted from [ITU00]).

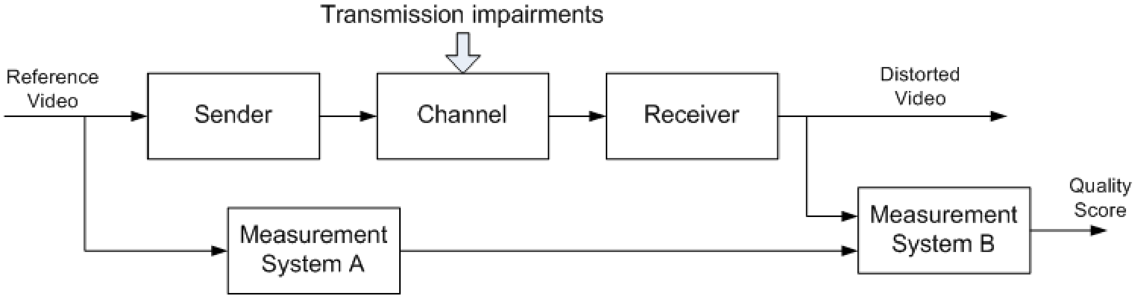


Figure 2.6: Block diagram of a reduced-reference video quality assessment system (adapted from [ITU00]).

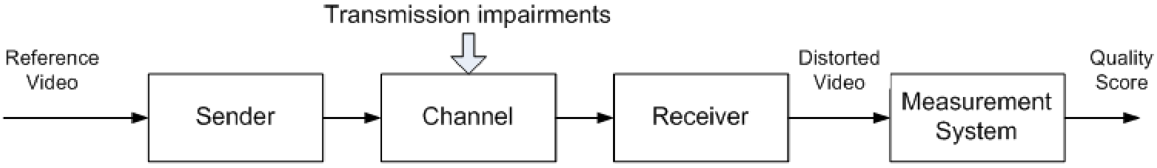


Figure 2.7: Block diagram of a no-reference video quality assessment system (adapted from [ITU00]).

2.2.2 Advances of Objective Video Quality Metrics

Due to the increasing demand of reliable and accurate video quality metrics, there has been a large amount of effort devoted to this topic from both industry and academia. The most remarkable work has been done by VQEG from ITU. From 1997, VQEG has conducted a number of validation tests to evaluate the performance of various proposed VQMs. Based on the test results, VQEG has also established a series of standards which give recommendations for the choice of objective video quality metrics for different applications. A summary of the work by VQEG is given in Table 2.1. Apart from the standardization efforts from VQEG, there are also contributions in other literatures. In the following, a review of several most important works in the field of perceptual video quality metrics will be given. As mentioned previously, the focus is put on metrics following the engineering approach.

Full-reference video quality metrics

The Structure SIMilarity (SSIM) index is proposed by Wang et al. in [WBSS04]. Similar to PSNR, SSIM does not make any assumption about the type of artifacts in the video. But different from PSNR, which calculates the picture quality based on pixel-to-pixel errors, SSIM estimates the quality by measuring how well the structural information contained in the picture is preserved. Since it is observed that human perception is more sensitive to distortions in

Table 2.1: ITU Recommendations for objective video quality metrics

ITU Standard	Metric Type	Target Application	Validation Test
ITU-T J.144[ITU04b] ITU-R BT.1683[ITU04a]	FR	SDTV	FR-TV1/FR-TV2 (1997-2003)
ITU-T J.249[ITU10]	RR	SDTV	RRNR-TV (2000-2008)
ITU-T J.247[ITU08b]	FR	Multimedia	MM-I (2003-2008)
ITU-T J.246[ITU08a]	RR	Multimedia	MM-I (2003-2008)
ITU-T J.341[ITU11a]	FR	HDTV	HDTV-I (2004-2010)
ITU-T J.342[ITU11b]	RR	HDTV	HDTV-I (2004-2010)

structural information [WBSS04], SSIM provides much better quality predictions than PSNR. To measure the structure similarity between the original content x and distorted content y, SSIM calculates the following three components [WBSS04]:

$$l(x,y) = \frac{2\cdot \overline{x}\cdot \overline{y}}{\overline{x}^2 + \overline{y}^2} \tag{2.7}$$

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

$$s(x,y) = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{2.9}$$

where $\overline{x}$ and $\overline{y}$ are the mean pixel values of x and y, respectively. σ denotes the standard deviation. The first two components, i.e. $l(x,y)$ and $c(x,y)$, can be seen roughly as measures of similarity of brightness and contrast between x and y, respectively. The third components $s(x,y)$ is the linear correlation of the two signals, which is a indication of how well the structural information is preserved. The SSIM index is then calculated as the product of the three components:

$$SSIM(x,y) = l(x,y)\cdot c(x,y)\cdot s(x,y) \tag{2.10}$$

The range of SSIM index is $[0,1]$, with a higher value indicating better perceptual quality.

The SSIM index has been originally proposed for still image quality assessment. In [WLB04], it is adapted for video quality assessment by calculating the weighted sum of the SSIM indices of the Y, Cb and Cr components. Other extensions of the SSIM index include the MultiScale-SSIM in [WSB03b] and the Speed SSIM proposed in [WL07].

The Psytechnics full-reference video quality metric is one of the four metrics suggested by ITU-T J.247 [ITU08b] for multimedia applications with spatial resolutions from QCIF to VGA. It performed best in VQEG's Phase-I Multimedia Test [VQE08]. After the spatial and temporal alignment process between the reference and distorted video, seven features are extracted from the videos. The spatial distortion is measured by decomposing the frames into sub-bands using a pyramid transform (similar to the concept of wavelet transform) and then calculating the PSNR values for selected sub-bands according to the spatial resolution of the frames. The temporal distortion is calculated based on the frequency and duration of dropped/frozen frames. These features, together with other features measuring the edge distortion, the blocking artifacts, the blurring artifacts and spatial complexity, are combined by a linear integration function to produce a final estimation of the video quality. Different integration functions are used for different spatial resolutions (QCIF/CIF/VGA).

VQuad-HD developed by SwissQual is the only full-reference video quality metric suggested by ITU-T J.341 [ITU11a] for HDTV applications. A jerkiness feature is calculated based on the local and global motion intensity to indicate temporal degradation. A blockiness measure is used to measure the spatial degradation. The basic idea of the blockiness detection algorithm is that if the video is processed using a block structure of size n, then average edge strength values calculated at a step-size of n could be very different for different offsets. The third component is calculated as the distribution of the local similarity and difference features. Finally, a logistic function is used to integrate these three features together for the quality estimation.

Other important full-reference video quality metrics include the Picture Quality Scale (PQS) [YMM00], the Perceptual Evaluation of Video Quality (PEVQ) by Opticom [ITU08b], the Motion-based Video Integrity Evaluation (MOVIE) index by Seshadrinathan and Bovik [SB10].

Reduced-reference video quality metrics

In [WJP+93], Webster et al. proposed a reduced-reference quality metric based on localized TI and SI values (refer to Eq.(2.2)(2.1)). The TI and SI are calculated for a certain Spatial-Temporal region (S-T region) in the video sequence. The values from the original video are transmitted to the quality estimator and compared with the values calculated from the distorted video. The outputs of the comparison are three measurements indicating the level of spatial and temporal distortions. A weighted sum of these three measurements is used as the quality estimation. The size of the overhead information can be controlled by selecting a suitable size of the S-T region.

The Yonsei reduced-reference quality metric (proposed by Yonsei University, Korea) is the only metric which is included in all the three RR video quality standards by ITU-T (J.246 [ITU08a] /J.249 [ITU10]/J.342 [ITU11b]). In this scheme, an edge map is generated by applying edge enhancement filters to the original video frames. The position and value of a set of edge pixels are transmitted to the quality estimator. The quality estimator calculates again an edge map based on the distorted video frames. The pixel values in the distorted edge map are then compared with the corresponding transmitted values to calculate the Edge PSNR (EPSNR). EPSNR is then adjusted according to the strength of different artifacts (blocking/blurring/jerkiness). A piecewise linear function is finally applied to form the quality estimation based on EPSNR. By adjusting the number of position/value pairs transmitted to the estimator, a compromise between prediction accuracy and side information is achieved.

Perhaps the most widely used RR metric is the General Model of the Video Quality Model (GMVQM) from National Telecommunications and Information Administration (NTIA) [WP99, WL07]. It is included in ITU-T J.144 [ITU04b] and ITU-T J.249 [ITU10] for SDTV application (although ITU-T J.144 is a standard for full-reference quality metrics, the techniques used in GMVQM are actually reduced-reference). It was the best-performing metric in VQEG's FR-TV tests [VQE00, VQE03] and also presented good performance in VQEG's RRNR-TV test [VQE09]. In GMVQM, a number of video features, such as SI/TI (as shown in Eq.(2.1)(2.2)), ratio between the strength of Horizontal/Vertical edges and diagonal edges, mean chrominance pixel values, standard deviation of luminance component, are calculated for both the reference and distorted videos. Then the strength of various artifacts, such as blocking, blurring, noise, jerkiness and color distortion, are measured according to the gain or loss of these features. These measurements are then combined by a linear function to provide an estimate of the overall video quality. Similar to the Webster metric, the rate required to transmit the features from the reference video can be controlled by adjusting the size of the S-T region for which the video features are calculated.

Another technique that can be used to implement reduced-reference video quality metrics is data hiding (such as watermarking) [FCM05, NCA06, CMB02]. Although the schemes based on data hiding are often classified as no-reference quality metrics, the quality estimator does need certain shared information from the reference video (for example, in the case of watermarking, the undistorted version of the watermark needs to be available). From this point of view, it is more suitable to consider them as reduced-reference metrics.

No-reference video quality metrics

The main focus of previous works in the field of objective video quality assessment has been put on FR and RR metrics. Due to the ever increasing demand of in-service quality monitoring, more and more research efforts have been devoted to the development of NR metrics in the recent years. According to the video features used in the metrics, no-reference metrics can be further divided into 3 categories [THB08]: bitstream layer metrics, media layer metrics and hybrid metrics.

Bitstream layer metrics utilize information from the encoded video bitstreams as well as information related to network performance (such as packet-loss rate). The metrics require no or only partial decoding of the encoded bitstream, so they can be used in lightweight solutions for quality estimation. But since they do not fully exploit content characteristics, they are often less accurate than media layer metrics. ITU-T G.1070 [ITU12] describes quality assessment metrics for videophone applications over IP networks. The content contains metrics for speech, video and overall multimedia quality. The video metric in ITU-T G.1070 is based on the work of Yamagishi et al. in [YH06, HYTT07]. The metric models the quality degradation caused by coding distortion and transmission errors separately. The coding quality of the video is modelled as the product of a power function of bitrate and an exponential function using both the bitrate and frame rate. The transmission quality of the video is modelled based on an exponential function of the packet loss ratio, which also considering the impact of frame rate and bitrate. The overall video quality is estimated using the product of the two items. The metrics include 12 model parameters which need to be trained for different video codecs. Suggested parameter values for different video resolution and codecs are also given in the recommendation. The metric above considers any random packet loss and in [BM10], it is extended by considering the duration and strength of burst packet loss.

In [RGSM+08], a bitstream layer metric for SD and HD IPTV applications is proposed. Similar to the metric in ITU-T G.1070, the coding distortion and transmission distortion are modelled separately. The coding distortion is based on an exponential function of bitrate and the transmission distortion is modelled using bitrate and packet loss rate. To take into account the video content, the same authors proposed in [GSR10] a new model for the coding distortion using information from the encoding process such as motion vectors and quantization parameters. Other bitstream layer models include [RCNR07, KSI09, KKHD11].

Media layer metrics assess the video quality based on the decoded pixel values. Most NR metrics in this category try to estimate the video quality by measuring the physical strength of different types of artifacts and their psychophysical impact on human perception. The main artifacts considered are blocking, blurring, ringing and motion jerkiness. For a complete review of models for different types of artifacts, the readers can refer to [HR10, Cha13]. Usually, a distorted video stream contains more than one artifact, so metrics considering the impacts of multiple artifacts are more robust and practical. In [FM05], such a metric is proposed by accounting for three artifacts. The blocking artifact is measured by comparing the correlations between adjacent pixels within and across the borders of the block structure used in the codec. The blurring is estimated by examining the spread of edges in the frame. To measure the noisiness, the frames are first filtered to remove its nature structure (such as edges and textures) and keep only the noise. Then the noise variance is calculated to estimate the strength of the noise. The overall frame quality VQ is modelled by using a weighted p-Minkowski metric to combine the three measurements (see Eq(2.11)).

$$VQ_p = (\alpha \cdot Blockiness^p + \beta \cdot Blurriness^p + \gamma \cdot Noisiness^p)^{1/p} \tag{2.11}$$

where p, α, β and γ are parameters which are determined by least-squares fitting.

Although humans can easily identify the type and strength of visual distortions in a video without the reference to the original content, it is not an easy job for NR quality metrics.

To address this issue, a extensive framework has been proposed in [MB11] for blind image quality assessment based on Nature Scene Statistics (NSS). The basic assumption is that natural scenes hold certain statistical properties which tend to be destroyed by distortions, so the abnormality of picture statistics is a good indication for the type and strength of different distortions. The proposed algorithm first extracts 88 statistical features from the content, then two vectors are calculated based on these extracted features. The first vector tells the probabilities of the content suffering from different types of artifacts and the second vector estimates the resulting picture quality when the content is affected by a certain artifact. The overall quality is computed as the inner product of the two vectors.

Hybrid metrics aim to combine the merits from media layer metrics and bitstream layer metrics by combining all the available information. On one hand, the decoded pixel information can help to improve the accuracy of the estimation. On the other hand, information from the network and the bitstream can be used to extract video features more efficiently and thus avoid unnecessary computation.

In [KHD12], Keimel et al. propose a NR hybrid video quality metric for HDTV content coded with H.264/AVC. The metric utilizes features extracted from both the bitstream (such as slice type, average quantization parameter for each slice, motion information, percentage of different MB type, etc.) and the pixel domain (such as blocking/blurring measurements, motion continuity, edge continuity, etc.). The weighted sum of these features is then used in a sigmoid function for the estimation of the overall quality. In comparison to a previous metric using bitstream layer information [KKHD11], the hybrid metric provides a better prediction accuracy. The hybrid metric also outperforms FR metrics such as PSNR, SSIM and GMVQM according to the evaluation.

VFactor is a patent-protected hybrid video quality metric which is used for many commercial applications for quality monitoring [Che]. According to the introduction in [WM08], VFactor uses not only information from the decoded pixel domain and the video coding layer of the bitstream, but also those calculated from the Packetized Elementary Stream (PES) layer (such as timing information) and Transport Stream (TS) layer (packer loss, delay and delay jitter, etc.).

The development of hybrid video quality metrics is also a main focus of the VQEG. One of the initial work focuses of the Joint Efforts Group (JEG) newly formed by VQEG is to develop a no-reference hybrid video quality metric for H.264/AVC.

2.3 Summary

In this chapter, both objective and subjective methods for video quality assessment are introduced. For the discussion of subjective video quality assessment, the guidelines provided in the ITU standards are summarized. Also, the advantages and disadvantages of different test methodologies are analyzed. This is followed by a review of the previous works on objective video quality assessment. Different approaches for designing objective video quality metrics are discussed and the metrics are classified according to the utilized information. From the review, it can be seen that most of the achievements so far are in the field of full-reference and reduced-reference quality metrics. Most no-reference metrics proposed so far are distortion and application specific due to the lack of reference information. It is still very difficult to build generic no-reference metrics without a deeper understanding of the HVS. In Chapter 3, the guidelines presented in this chapter are followed to conduct extensive subjective quality assessments, and the results are used to develop a no-reference objective video quality metric for QoE-driven multi-dimensional video adaptation.

Chapter 3

Perceptual Video Quality Modeling

In this chapter, the impact of frame size, frame rate and quantization on the perceived quality of a video is explored and a Multi-Dimensional Video Quality Metric (MDVQM) is proposed to estimate the video quality in the presence of quantization, frame dropping and spatial down-sampling. The SNR video quality is captured by a logistic function whereas the impact of frame rate reduction and spatial down-sampling are modelled separately as temporal/spatial correction factors. The overall video quality metric is then calculated as the product of these components. The proposed metric uses only several features that can be easily extracted from the bitstream or decoded frames and thus is practical for real-time video adaptation applications.

3.1 Introduction

The remarkable evolution of communication networks has enabled video content delivery over mobile networks. The increased power of end-devices and the user's ever-increasing demand for video content further boosted the popularity of video applications. The video quality perceived by the end users is the most crucial factor for the success of video services. Therefore, in-service monitoring and optimization of the video quality is becoming more and more important for service providers. As discussed in Chapter 1, subjective quality assessment is not feasible in this scenario due to the involvement of human observers and it can only be achieved by employing objective video quality metrics which can estimate the perceived video quality automatically and accurately. Many quality metrics have been proposed so far in the literature and some have already been used in commercial solutions. However, the heterogeneity of the end-users brings new challenges for quality estimation. Most prior video quality metrics deal with a fixed spatial and temporal resolution. Meanwhile, as mentioned in Chapter 1, transmitted videos often need to be adapted to a different display size and frame rate. Hence, it is important to develop new video quality metrics which consider the impacts of different adaptation schemes on the perceived video quality.

Typically, the video adaptation can be performed by changing either the Quantization Parameter (QP), the frame rate/Temporal Resolution (TR), or the frame size/Spatial Resolution (SR). Using a larger QP results in stronger coding artifacts (e.g. blocking artifacts for block-based hybrid video coder). Reducing TR by dropping frames affects the smoothness of motion and reducing the SR by spatial down-sampling introduces blurring artifacts if the video is later up-sampled and displayed in the original resolution. In the following, unless otherwise stated, the term SNR Video Quality ($SNRVQ$) is used to denote the video quality resulting from quantization only. The term Spatial Video Quality (SVQ) and Temporal Video Quality (TVQ) are used to refer to the perceived video quality when a reduction of SR and TR is performed, respectively. The term Spatial-Temporal Video Quality ($STVQ$) is used to denote the video quality when both TR and SR are reduced.

The remainder of this chapter is structured as follows. Section 3.2 reviews the related work on video quality assessment involving quantization, frame rate and frame size, separately and jointly. Section 3.3 gives a description of the conducted subjective tests. In Section 3.4 the results of the subjective tests are analyzed and a novel no-reference video quality metric is introduced. The performance of the proposed quality metric is evaluated and compared with the related metrics in the literature. In Section 3.5, the work presented in this chapter is summarized.

3.2 Related Work

Several subjective studies have been performed and reported in the literature to analyze the impact of frame rate and spatial resolution on the subjective quality.

In [WCL04], the authors study the preference of frame rate by performing subjective tests using CIF (352x288) sequences encoded at different bit-rates (50-1000kbps) and frame rates (30fps/15fps/7.5fps). The results show a general trend that the preferred frame rate reduces when the encoding bit-rate decreases. The sequences are further divided into three categories according to their content complexity and the analysis shows that for videos of different categories, the switching bit-rates of the optimal frame rate vary significantly, which indicates the content dependency of the user preference - the higher the content complexity, the higher the switching bit-rates.

In [CT07], the results from a number of previous studies are summarized to study the effects of different frame rates on human perception for various scenarios. The finding is that although the results vary slightly according to the task, the viewing condition and the viewers' characteristics, the minimum frame rate should be kept between 10-15fps to achieve an acceptable performance.

The study in [WSV$^+$03] investigates the impact on subjective quality of QP, spatial resolution and frame rate for H.263 encoded videos. Two subjective tests are conducted using five source sequences with an original resolution of 320x192 at 30 fps. The first one studies how the subjective quality is affected when jointly adjusting QP and the spatial resolution while the second one focuses on jointly adjusting QP and frame rate. The overall conclusion is that human vision is more sensitive to quantization artifacts than blur and motion jerkiness, especially at middle and low bit-rates. The authors suggest that when the QP used for encoding reaches a certain threshold, the frame rate and/or spatial resolution should be changed in order to achieve a better subjective quality and the QP threshold depends highly on the spatial/temporal activity of the content. A similar study of the joint impact of the same parameters (QP, SR and TR) for low bit-rate cases is conducted in [ZCL$^+$08]. Both H.263 and H.264/AVC codecs are used to encode the test sequences (CIF@50fps) at a constant bit-rate (in comparison to [WSV$^+$03], where constant QP values are used). The test results confirm the conclusions in [WSV$^+$03].

In [LSR$^+$10], an extensive study is performed for HD (1280x720@50fps) resolution videos encoded with two scalable video codecs - H.264/SVC and a wavelet-based scalable video codec (W-SVC). The sequences are encoded for a wide range of bit-rates (from 300kbps to 4Mbps) using 3 spatial layers (HD/640x360/320x180) and 4 temporal layers (from 50fps to 6.25fps). The conclusion is that when the bit-rate is small, it is preferable to reduce the spatial resolution from HD to 640x360 to prevent strong blocking artifacts. But further spatial-downsampling (to 320x180) should be avoided due to the strong blurring artifacts caused by up-sampling back to HD. While for relatively high bit-rate cases, since a certain level of spatial quality is already guaranteed, a higher frame rate is more desirable than a higher spatial resolution. It is also found that although the choice of codec type does have influence on the test results, the overall tendency is consistent across the two codecs.

The above works do not propose any concrete video quality metrics for different spatial and temporal resolutions.

In [LLS$^+$05], the authors propose a metric based on an expo-logarithm function of the frame rate to estimate the negative impact of frame dropping on the perceived video quality. The average of every frame's maximal motion vector magnitude is used in the metric as a representation of the motion intensity to consider the impact of the video content. Another work in [QG08] considers the jitter and jerkiness effects. A subjective study is conducted, in which the video quality is deteriorated by frame dropping with varying strength, burst length and frequency. An interesting finding is that jitter is more annoying than jerkiness, therefore the change of frame rate should not be performed too frequently. Unfortunately, only the jerkiness effect is modelled with a sigmoid function of the frame rate. A problem in the above metrics is that only the temporal quality of the video is considered which has limited their application in practice.

A video quality metric QM is proposed by Feghali et al. in [FWSV07]. The metric considers both the SNR and the temporal quality of the video. The video quality is estimated simply by the average PSNR value when no frame rate reduction is conducted (FR=30fps). In case of frame dropping, the PSNR value is significantly affected due to the difference between the repeated frames and the original frames. To address this issue, the authors propose to add a compensation term to the PSNR value which depends on the frame rate and motion intensity for a more accurate estimation of the overall quality. The motion intensity is estimated by the average magnitude of the top 25% of the largest motion vectors in each frame. In [KJSR08, SYN$^+$10], the above metric is extended by considering also the impact of spatial characteristics of the video. The SNR quality is still estimated by PSNR and temporal quality is modeled similarly except that the motion activity measure is calculated as the standard deviation of the motion vector magnitudes. Spatial quality is modelled as a sigmoid function of the height of the frame in [KJSR08] and in [SYN$^+$10] it is modeled using an exponential function of the height and a spatial activity measure. The overall quality is computed as the weighted sum of the three quality values. However, the accuracy of motion vectors is strongly affected by the chosen motion estimation algorithm and sometimes also the bit-rate (which affects the quality of the reference frames), therefore the above metrics sometimes suffer large estimation errors.

In [OMW09, OMLW11], Ou et al. propose the metric VQMTQ which models the impact of frame dropping and quantization on the perceived video quality. The overall video quality is estimated as:

$$SQF = \hat{Q}_{max} \cdot \left(1 - \frac{1}{1 + e^{p(SPSNR - s)}}\right) \tag{3.1}$$

$$TCF = \frac{1 - e^{-\alpha_t \frac{f}{f_{max}}}}{1 - e^{-\alpha_t}} \tag{3.2}$$

$$VQMTQ = SQF \cdot TCF \tag{3.3}$$

where $\hat{Q}_{max}$ is the subjective rating for the highest quality video (which is empirically set to 90 for a 0-100 MOS scale). f and f_{max} are the frame rate after and before frame dropping, respectively. α_t, p and s are parameters depending on the video content. SQF estimates the SNR quality of the video and TCF is a correction factor modeling the negative impact of frame dropping on the video quality.

In [XOMW10, OXMW11], VQMTQ is extended to QSTAR, where the impact of frame size is also considered by introducing a spatial correction factor:

$$SCF = \frac{1 - e^{-\alpha_s \left(\frac{s}{s_{max}}\right)^{\beta_s}}}{1 - e^{-\alpha_s}} \tag{3.4}$$

$$QSTAR = VQMTQ \cdot SCF \tag{3.5}$$

where α_s and β_s are content dependent parameters.

In [PS11], Peng et al. propose a full-reference video quality metric STVQM for the estimation of SNR and temporal video quality:

$$SVQM = \frac{100}{1 + e^{-(SPSNR + w_s \cdot SA + w_t \cdot TA - \mu)/s}} \tag{3.6}$$

$$TVQM = \frac{1 + a \cdot TA^b}{1 + a \cdot TA^b \cdot \dfrac{30}{FR}} \tag{3.7}$$

$$STVQM = SVQM \cdot TVQM \tag{3.8}$$

where SVQM and TVQM model the SNR video quality and quality degradation caused by frame dropping, respectively. SPSNR is the spatial PSNR (which is computed by averaging the PSNR values over the non-dropped frames). w_s, w_t, μ, s, a and b are parameters that need to be trained from the quality ratings collected from the subjective tests. TA and SA are measures of spatial and temporal activity of the video content, respectively. TA and SA are calculated by the following equations:

$$SA = mean_{time}\{std_{space}[Sobel(F_n)]\} \tag{3.9}$$

$$TA = mean_{time}\{std_{space}[F_n - F_{n-1}]\} \tag{3.10}$$

It can be seen that the calculations of TA and SA are very similar to that of TI and SI in Eqs.(2.2)(2.1), except that the average value over time is calculated instead of the maximum value.

According to the evaluations in [PS11], VQMTQ and STVQM provide significantly better estimation performance than QM, while the performance difference between VQMTQ and STVQM is not statistically significant. More concretely, for both VQMTQ and STVQM, the Pearson Correlation (PC) with the ratings from subjective tests is higher than 0.95 and the Root-Mean-Square Error (RMSE) is less than 10 on a 0-100 MOS scale.

Summarizing the above results, almost all the current quality metrics which deal with the problem of multi-dimensional optimization of perceived video quality are FR metrics and designed based on PSNR. Although they can provide accurate prediction of video quality, it is not feasible to use them for real-time video adaptation inside the network due to the absence of the original video which is requested for the PSNR, SA and TA calculation. In this chapter, a no-reference video quality model named MDVQM is proposed to address the demand for multi-dimensional video adaptation. The impacts of quantization, frame rate and frame size are modelled separately and the overall video quality is determined as the product of theses different factors. The metric uses only two activity measures from the video content and thus is computationally efficient. Validation tests show that the quality predictions of the metric correlate very well with subjective ratings obtained in subjective tests.

3.3 Details of the Subjective Study

In order to understand how different factors (i.e. quantization, frame rate and frame size) affect the perceptual video quality, two separate subjective tests are conducted. The first test (Test I) focuses on the impact of individual impairments such as those caused by frame dropping or spatial down-sampling. The second test (Test II) aims to evaluate the impact on video quality when TR and SR are changed at the same time.

Since the current 3G mobile networks employ powerful error correction techniques at the physical and link layer, it is assumed in this work that the channel impairments such as bit-error and packet loss are hidden from the application layer, so that from the perspective of the video applications, changing channel conditions are only reflected by varying transmission rates, which define the target rate for the video adaptation. Therefore, network errors are not explicitly considered during the design of the subjective tests and the development of objective quality metrics.

Furthermore, this work focuses on the non-scalable version of H.264/AVC video, because it covers the lion share of the video traffic in today's internet. All the test materials are encoded using H.264/AVC video codecs and the proposed metric is trained based on the corresponding subjective data. Although the choice of video codec might affect the results, the analysis and evaluation in this work are general and can easily be extended to other codec types.

3.3.1 Source Sequences

In Test I, eight source video sequences (SRC) with a wide range of spatial and temporal content characteristics are used. Six of them are well-known standard test sequences available from [Xip]: CREW (CR), HARBOUR (HA), SOCCER (SC), PEDESTRIAN AREA (PA), PARK JOY (PJ), FOOTBALL (FB). Two of them are internet videos from Youtube: OBAMA (OB)[Youb] and KOBE (KO)[Youa].

In Test II, three standard test sequences (PA, FB, and Rush Hour (RH)) are used.

A clip of 10 seconds from each SRC is selected in order to maintain a high concentration of the subjects. All the standard test sequences are in 4CIF(704x576) resolution. The original spatial resolution of the two Youtube sequences is 1024x768 and the sequences are center-cropped to 4CIF resolution. The frame rates of the SRCs are either 60fps or 30fps. Figure 3.1 shows example frames of the SRCs and their original frame rates are given in the titles.

Their spatial information (SI) and temporal information (TI) indices [ITU07] are shown in Figure 3.2. It can be observed that they span a wide range in the SI-TI space.

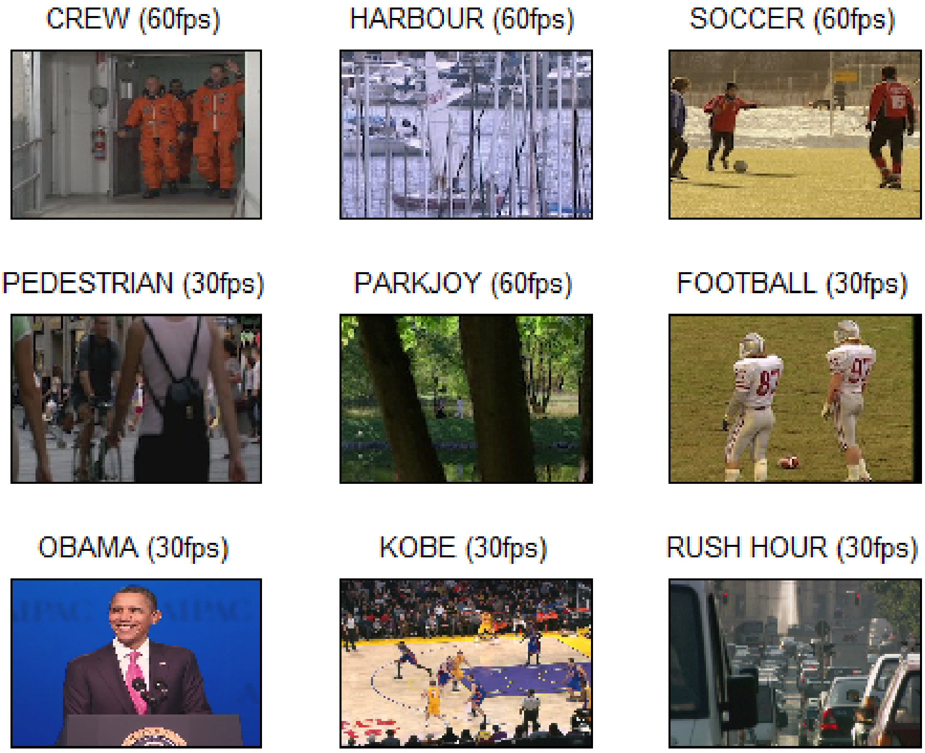


Figure 3.1: Example frames of the source videos used for the subjective tests

3.3.2 Test Sequences

Table 3.1: Bit-rates, frame rates and spatial resolutions of the processed video sequences for Test I

SRC	BR (kbps)				FR(fps)xSR		
CR	5200	2800	1600	1200	60x4CIF	15x4CIF	60xCIF
HA	8000	3500	2000	1300	60x4CIF	15x4CIF	60xCIF
SC	3600	2400	1400	1000	60x4CIF	15x4CIF	60xCIF
PJ	8000	6000	4000	2000	60x4CIF	20x4CIF	60xCIF
PA	2000	1500	1000	500	30x4CIF	10x4CIF	30xCIF
OB	640	384	256	192	30x4CIF	15x4CIF	30xCIF
KO	1500	1000	800	640	30x4CIF	15x4CIF	30xCIF
FB	1500	1000	800	640	30x4CIF	15x4CIF	30xCIF

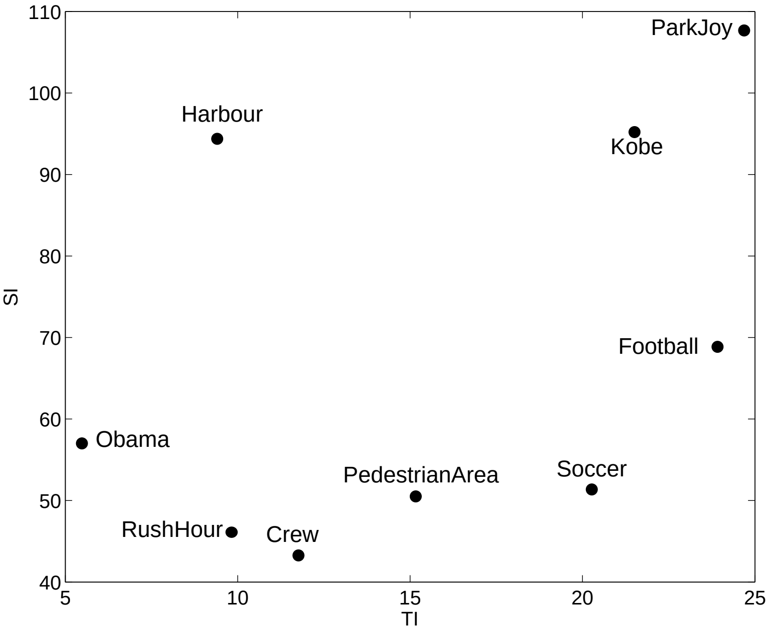


Figure 3.2: Spatial Information vs. Temporal Information indices of the source videos

Table 3.2: Bit-rates, frame rates and spatial resolutions of the processed video sequences for Test II

SRC	BR (kbps)				FR(fps)xSR	
PA	1500	1000	800	500	30x4CIF	15xCIF
FB	1500	1000	800	640	30x4CIF	15xCIF
RH	2000	1000	640	512	30x4CIF	15xCIF

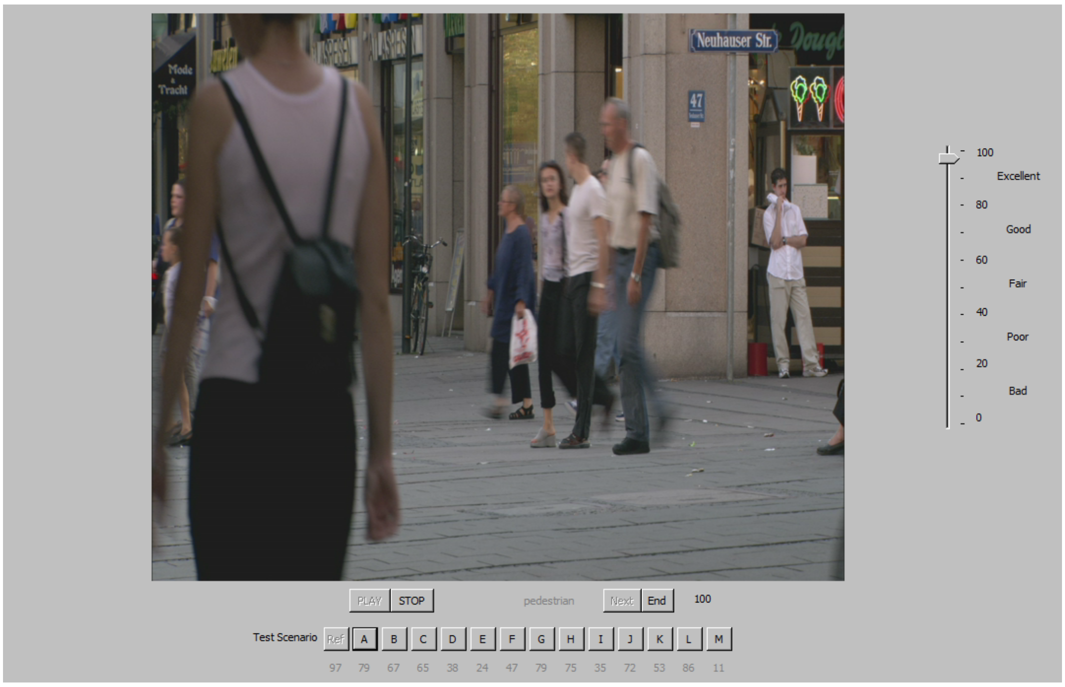
In Test I, 12 processed video sequences (PVS) (96 in total for the 8 SRCs) are generated for each SRC. The PVSs are encoded using the open source X264 encoder in an IPPP...P structure with Constant Bit-Rate (CBR). The original rate control algorithm in X264 is replaced by a new algorithm based on ρ-domain analysis, which will be discussed in more detail in Chapter 4. The PVSs for each SRC are divided into 3 groups (4 for each group). For the first group, the original SR and TR are kept unchanged and it is referred to as the *SNR group*. For the second group, the PVSs are spatially down-sampled to CIF resolution, referred to as the *SR group*. And for the last group (referred to as *TR group*), the PVSs are temporally down-sampled by a factor of 2-4. For each group, the PVSs are encoded at 4 different bit-rates. A description of the encoding bit-rates, frame rates and spatial resolutions of the PVSs is given in Table 3.1. One thing to note here is that in the test a display window of fixed size (4CIF) is used, so all the spatially down-sampled sequences are resampled back to their original resolution for playback. Details of how the subjective data is split for model training and subsequent validations are given in Section 3.4.1.

In Test II, 8 PVSs are generated for each SRC, among which 4 are encoded in full-resolution and 4 are down-sampled both spatially and temporally before encoding. Within each group, the PVSs are encoded with 4 different bit-rates in a CBR manner. Detailed information of PVSs in Test II is given in Table 3.2.

To avoid fatigue of the test subjects, Test I is divided into 3 subtests. The first subtest includes all the PVSs from CR, HA and SC. The second subtest includes all the PVSs from HA, PA and PJ. The third subtest includes all the PVSs from HA, FB, OB, KB. The PVSs from HA are included in all three subtests so that this common set can be later used to combine the scores from different subtests into a super dataset as will be discussed in Section 3.3.6. Similarly, in Test II, a set of common sequences from FB and PA are included for calibration purposes.

3.3.3 Test Methodology

As mentioned in Section 2.1, the SAMVIQ method [ITU07] is adopted in this work to collect subjective ratings for the test videos. A graphical software interface is developed which implements SAMVIQ for the subjective test. The central part of the interface is shown in Figure 3.3. The video is displayed at the original resolution at the center of the screen and the background is set to mid-level grey color. The test sequences are accessed through the access buttons ("REF" buttons corresponds to the reference sequence and button "A"-"M" correspond to the processed sequences and the hidden reference). After viewing each sequence, the test subject can use the slider bar on the right hand side to score the sequence. The score is displayed under the corresponding access button. The slider uses a continuous quality scale from 0 to 100 and is divided into five equal intervals with annotation by five adjectival quality terms (Excellent, Good, Fair, Poor, Bad) for general guidance according to [ITU07]. If a test subject is viewing a sequence for the first time, the whole sequence should be watched and no jump to other sequences is allowed during the play (the access buttons of all other sequences are disabled during the first play). If the test subject is viewing a sequence to which a score has already been given, the playout process can be stopped and resumed (through the "STOP" and "PLAY" button, respectively). In this case, the test subject can also switch to other sequences at any time (through the access buttons). Once all the sequences in a test scene have been scored, the "NEXT" button can be used to proceed to the next scene (a test scene contains all the test sequences from the same source sequence). After the test subject has finished all the test scenes, the subjective test can be ended using the "END" button.



Test Scenario	Ref	A	B	C	D	E	F	G	H	I	J	K	L	M
	97	79	67	65	38	24	47	79	75	35	72	53	86	11

Figure 3.3: The graphical user interface implementing the SAMVIQ method.

3.3.4 Test Subjects

A total of 56 test subjects have participated in the tests. The number of test subjects in each test is summarized in Table 3.3 (the number in the bracket is the number of subjects that are rejected by the screening process as discussed in Section 3.3.6). Note that the participants in the tests are overlapping. All the participants are non-experts, which means that they were not professionally involved in image/video quality assessment at their work. The subjects are all with normal or correct-to-normal visual acuity and between 21 and 38 years old, including both males and females.

3.3.5 Test Environment and Procedure

The general viewing conditions in the subjective tests were arranged as specified by ITU-T Rec. BT.1788 [ITU07] for a laboratory environment. The room for the experiments was equipped with 17-inch LCD monitors of type FUJITSU SIEMENS SCENICVIEW B17-2 CI. The ratio of inactive screen luminance to peak luminance was kept below a value of 0.02. The viewing distance is about 4 times the height of the video stimulus.

A test session is divided into three phases: instruction, the training session and the formal test session. During the introduction phase, a written instruction was distributed to the participants, explaining the tasks to be performed in the tests. The training session was conducted prior to the test session to get the participants familiar with the test mechanism and to demonstrate the range of artifacts to be expected during the actual test session. The scores obtained during the training session were not considered in the final results. Questions from the subjects were allowed during the training session. The test session begins after the training session, the average duration of the test session was about 20 minutes. No question was allowed during the test session.

Table 3.3: Number of subjects in the tests. The numbers in the bracket indicate the number of test subjects rejected by the screening process in each subtest as discussed in Section 3.3.6

Test	Test I			Test II
	Sub. I	Sub. II	Sub. III	
#Subj.	18 (2)	18 (1)	24 (3)	24 (2)

3.3.6 Subjective Data Post-Processing

The screening process defined in [ITU07] is adopted to reject test subjects who may have rated randomly or inconsistently. More specifically, the Pearson correlation coefficient r_p and the Spearman's rank correlation coefficient r_s between the ratings of each viewer and the mean ratings of all viewers are calculated using Eq.(3.11) and Eq.(3.12), respectively:

$$r_p = \frac{\sum\limits_{i=1}^{N_v}(x_i - \overline{x})\cdot (y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{N_v}(x_i - \overline{x})^2\cdot}\sqrt{\sum\limits_{i=1}^{N_v}(y_i - \overline{y})^2}} \tag{3.11}$$

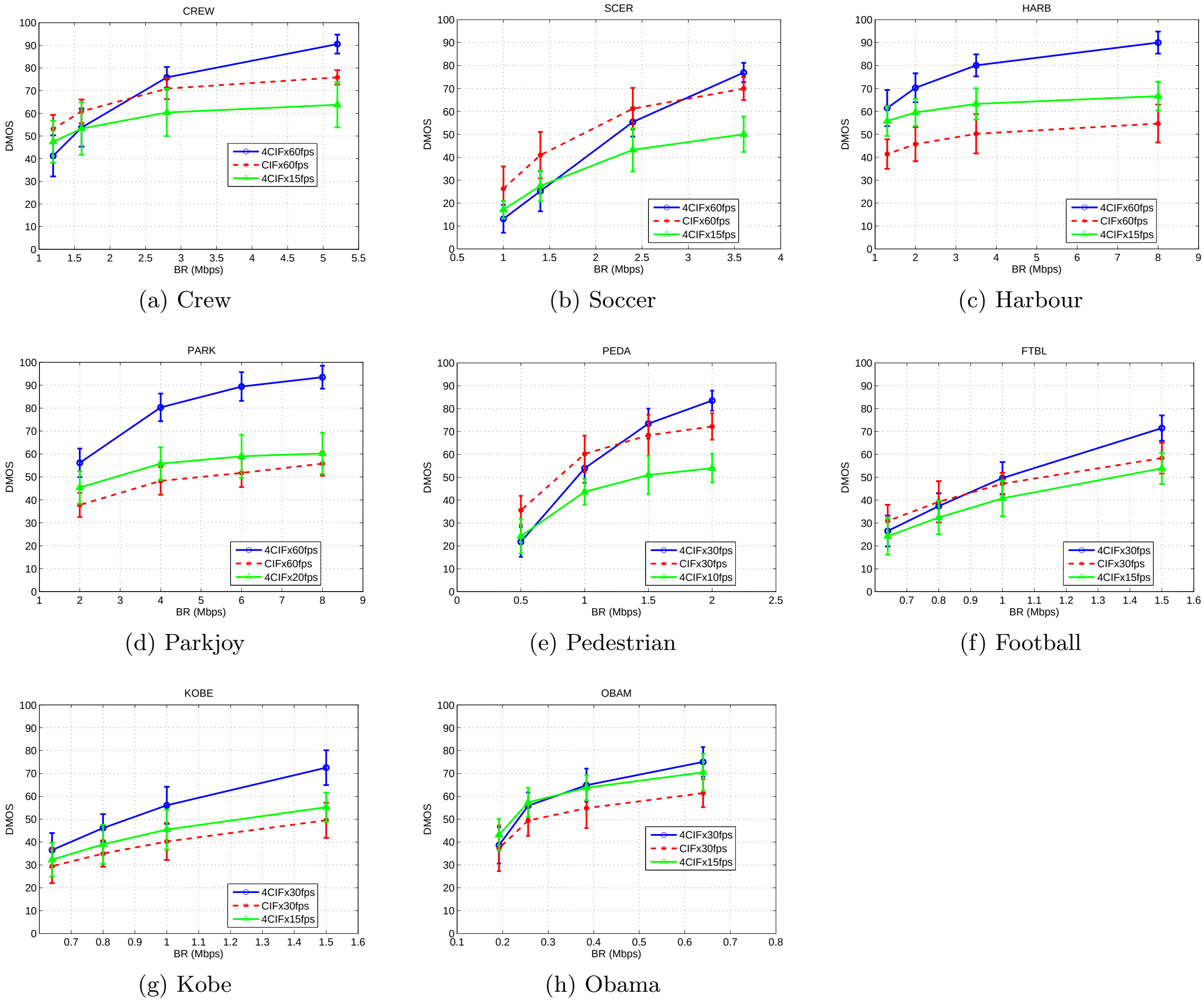
$$r_s = 1 - \frac{6\cdot \sum\limits_{i=1}^{N_v}[R(x_i) - R(y_i)]^2}{N_v^3 - N_v} \tag{3.12}$$

where x_i is the individual score of the viewer for video i and y_i is the mean score of all the viewers for video i. N_v is the number of test sequences. $\overline{x}$ and $\overline{y}$ are the mean value of $\{x_i|i = 1...N_v\}$ and $\{y_i|i = 1...N_v\}$, respectively. $R(x_i)$ is the ranking order of the score x_i in $\{x_i|i = 1...N_v\}$. Then, the correlation of individual scores from viewer j against corresponding mean scores from all the viewers is r_j calculated by:

$$r_j = min(r_{pj}, r_{sj}) \tag{3.13}$$

The rejection threshold is determined by:

$$th_{reject} = \begin{cases} 0.85, & \text{if } [mean(r) - std(r)] > 0.85 \\ mean(r) - sdt(r), & \text{otherwise} \end{cases} \tag{3.14}$$

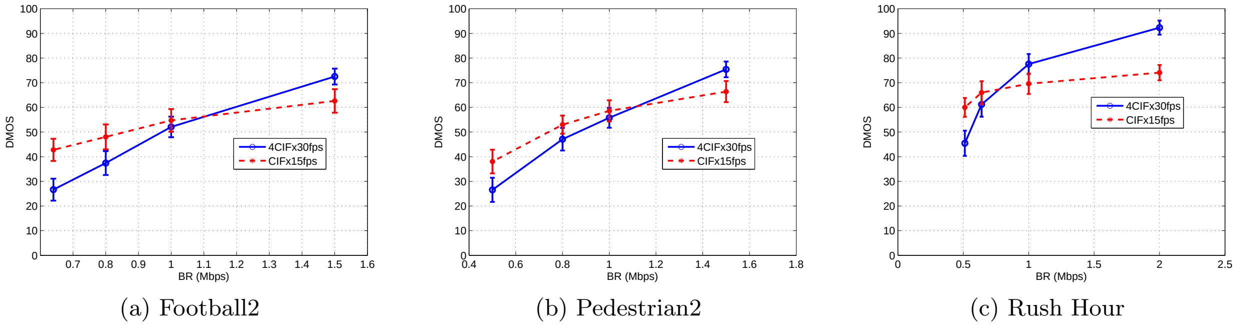


(a) Crew (b) Soccer (c) Harbour

(d) Parkjoy (e) Pedestrian (f) Football

(g) Kobe (h) Obama

Figure 3.4: DMOS values of all test videos in Test I. The vertical bar indicates the corresponding 95% confidence interval. Data are calibrated and merged as described in Section 3.3.6.



(a) Football2 (b) Pedestrian2 (c) Rush Hour

Figure 3.5: DMOS values of all test videos in Test II. The vertical bar indicates the corresponding 95% confidence interval. Data are calibrated and merged as described in Section 3.3.6.

where $r = [r_1, r_2, ...r_j..., r_{Ns}]$ is the vector of correlation values of all the viewers. Finally, the following rejection criteria is applied:

$$\begin{cases} \text{observer j is rejected,} & \text{if } r_j < th_{reject} \\ \text{observer j is not rejected,} & \text{otherwise} \end{cases} \tag{3.15}$$

The number of rejected subjects in each test is given in Table 3.3.

After screening, the Mean Opinion Score (MOS) is calculated from the subjective ratings. Let x_i^j denote the rating of test sequence i given by subject j, and x_{ref}^j be the rating of the corresponding hidden reference given by the same subject. The Differential Mean Opinion Score (DMOS) value of test sequence i (denoted as μ_i) is calculated as:

$$\mu_i = \frac{1}{N_s} \sum_{j=1}^{N_s} (x_i^j - x_{ref}^j + 100) \tag{3.16}$$

where N_s is the number of test subjects. The DMOS value is used as the subjective quality measure for the PVSs. Note that since the raw subjective rating is in the range [0,100], it is possible that DMOS values are greater than 100, and these values are considered valid and included in the analysis.

With the subjective ratings from the common set as mentioned in Section 3.3.2, the method proposed in [PW08] is used to generate a super dataset for the development of our video quality metric. Briefly speaking, an overall average (over all subtest) of the DMOS values is first calculated for each of the videos in the common set. These overall average values are considered as the most accurate measurements of the video quality. By fitting the average values of the common videos from each subtest to the overall average values, a linear mapping function is determined. This linear mapping function is used to convert the DMOS values from the subtests to form a super dataset for our later analysis.

The Confidence Interval (CI) associated with the DMOS value of test sequence i is given by:

$$[\mu_i - \delta_i, \mu_i + \delta_i] \tag{3.17}$$

The term δ_i in Eq.(3.17) can be derived from the standard deviation σ_i and the number of test subjects N_s. For example, a 95% CI is calculated as:

$$\delta_i = 1.96 \frac{\sigma_i}{\sqrt{N_s}} \tag{3.18}$$

where the standard deviation σ_i for test sequence i is defined as:

$$\sigma_i^2 = \frac{1}{N_s - 1} \sum_{j=1}^{N_s} (x_i^j - \mu_i)^2 \tag{3.19}$$

The derived DMOS values of all test videos, along with the corresponding 95% confidence interval are shown in Figures 3.4 and 3.5. In the figures, the blue curves always correspond to the videos with full resolution. In Figure 3.4, the red dotted curves correspond to the videos with reduced SR and the green curves correspond to the videos with reduced TR. In Figure 3.5, the red dotted curves correspond to the videos whose TR and SR are reduced at the same time.

From the results, it can be seen that at high bit rate, the blue curves are always above the other curves, which indicates full spatial/temporal resolution is preferred. With the decrease of bit rate, some blue curves intersect with the red or green curves, indicating that at a lower bit rate, reducing the spatial/temporal resolution is the better choice for transcoding. The different appearance of the curves also suggests that the characteristics of the video content have a strong impact on the perceived video quality.

3.4 Design of the NR Video Quality Metric

As mentioned in the previous section, in order to understand the impact of changing different parameters, subjective tests are conducted to collect subjective quality ratings. These data serve as the “ground-truth” quality ratings for the design, development and validation of objective quality metrics. As mentioned above, the three considered parameters that affect the perceived video quality are QP, TR and SR. It has been shown in [OXMW11][PS11] that the impact of quantization (QP) is separable from that of TR and SR, so they are studied and modelled separately in the following.

3.4.1 SNR Quality Metric

3.4.1.1 Design of the SNR Quality Metric

Many objective quality metrics for measuring the SNR quality of video sequences have been proposed in the literature. According to our application scenario, video adaptation is usually performed at an intermediate network node (e.g. a proxy server) at the edge of the core and access network, where the reference video is not available. Therefore, a no-reference video quality metric is best suited for this situation.

To model the SNR video quality, the first step is to select an appropriate functional form. Many PSNR-based full-reference video quality metrics, such as the PSNR-VQM in [PW02], the PEVQ in [ITU08b] and VQuad-HD in [ITU11a], choose to use the sigmoid function as the basic function form. The popularity of the sigmoid function is due to the finding from the subjective results [VQE00, VQE03] that PSNR usually only correlates linearly with the MOS values in the middle of the quality range, while saturation of MOS values appears towards the two extremes of the quality range. This phenomenon accords with the fact that human observers tend to have difficulties to identify quality difference between two videos with extremely good or bad quality. The typical form of a sigmoid function can be written as:

$$P(t) = \frac{1}{1 + e^{c(t-d)}} \tag{3.20}$$

where c and d are parameters which can be used to adjust the shape of the sigmoid function. Figure 3.6 shows several sigmoid functions with different parameters.

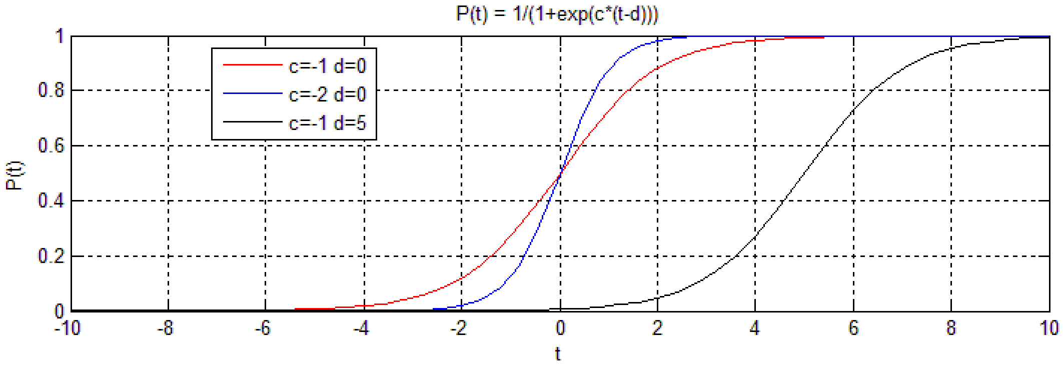


Figure 3.6: sigmoid functions with different parameters

From the figure, it can be seen that the parameter c controls the dropping rate of the middle range of the curve while the parameter d can be used to control the position of the saturation point. In practice, c and d can be modelled as functions of the spatial-temporal characteristics of the video content.

The full-reference metric STVQM in Eq.(3.1) also uses the sigmoid function for the estimation of SNR video quality and has been shown to provide good quality prediction. In the following, it is used as a starting point to derive a no-reference video quality metric. To change the metric into a no-reference quality model, the features SPSNR, SA and TA in Eqs.(3.9)(3.10) need to be estimated from the decoded video frames instead of from the reference video.

To observe the difference of TA and SA values between the original video and the encoded video subjected to quantization artifact, experiments are conducted in which several typical test video sequences at CIF resolution (including Foreman, Mother&Daughter, etc.) are encoded with different QPs and TA and SA values are extracted from the encoded sequences. The obtained TA and SA values are shown in Figure 3.7. It can be seen that although the TA and SA values do change as a function of the QP values, the extent of change is quite limited. In our experiments, the change of TA and SA from high QP (low bit-rate) to low QP (high bit-rate) for most test sequences is no larger than 8%. This observation indicates that TA and SA extracted from decoded sequences can be seen as a good approximation to those extracted from the original sequences.

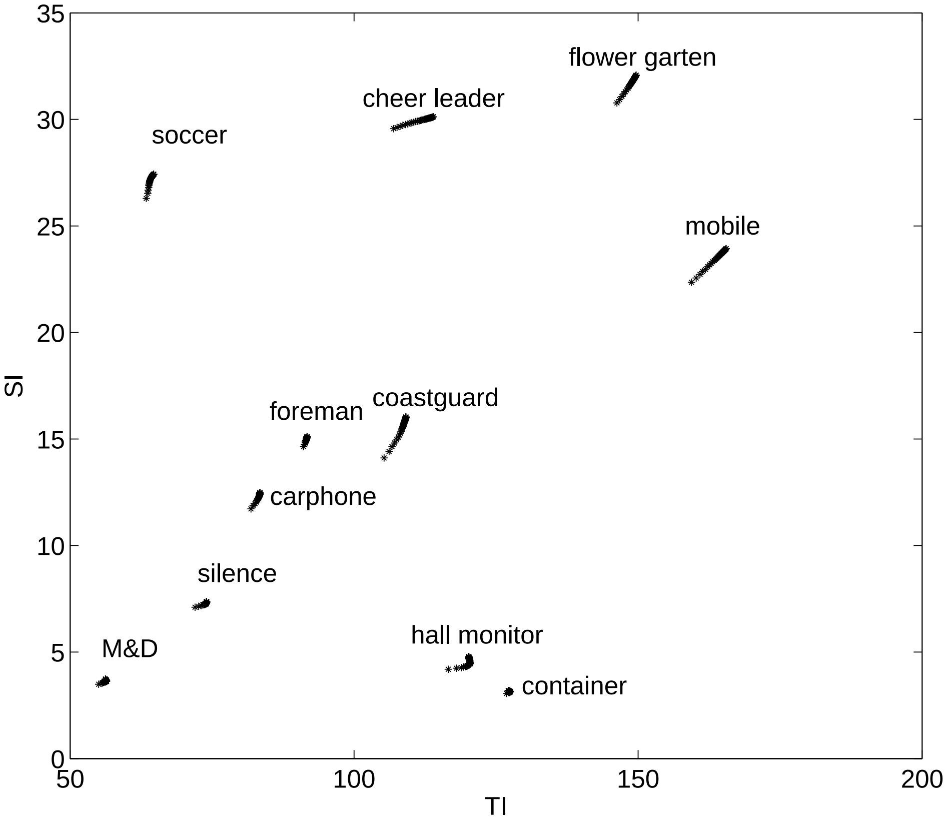


Figure 3.7: Spatial Activity (SA) and Temporal Activity (TA) variation against bit-rate for typical test sequences

It is known from rate-distortion theory that the relationship between the bit-rate and PSNR can be approximately modeled using a logarithmic function. Since our test videos have different frame rates and frame sizes, here the pixel bit-rate (bit-per-pixel) is used instead of normal bit-rate in bit-per-second:

$$bpp = \frac{BR}{FR\cdot FS} \tag{3.21}$$

where FR and FS are the frame rate and frame size respectively. Then the SPSNR in Eq.(3.6) can be estimated by:

$$SPSNR = m\cdot ln(bpp) + n \tag{3.22}$$

where bpp is the pixel bit-rate and $ln(x)$ is the natural logarithm of x. m and n are content-dependent parameters. For simplicity, the parameter n is modelled as a linear combination of SA and TA, so that it can be merged with the other items in Eq.(3.6). For the parameter m, different types of functions are examined and the power function seems to provide the best performance. Finally, the SNR quality of a video is modeled as:

$$m = TA^{a_0}\cdot SA^{a_1}\cdot a_2 \tag{3.23}$$

$$SNRVQ = \frac{100}{1+e^{-(m\cdot ln(bpp)+a_3\cdot SA+a_4\cdot TA+a_5)}} \tag{3.24}$$

where $a_0, \ldots, a_5$ are model parameters which need to be trained using subjective data. Compared to the FR quality metric in STVQM, our NR quality metric has two more parameters. But all the features can be extracted from the decoded frames, which makes this metric applicable for video adaptation in the absence of the original content.

3.4.1.2 Performance Analysis of the SNR Quality Metric

In this section, the performance of the proposed SNR quality model is evaluated against several state-of-the-art video quality metrics.

According to the criteria used by VQEG in its multimedia test [VQE08], the performance of a video quality metric can be measured by the accuracy and consistency of the predictions. In [VQE08], accuracy is defined as "the ability to predict the subjective quality ratings with low error" while consistency is defined as "the degree to which the model maintains prediction accuracy over the range of video test sequences". The Pearson Correlation (PC) and the Root Mean Square Error (RMSE) are used to measure the accuracy of a metric and the consistency is measured by the Outlier Ratio (OR).

The formula to calculate the PC value has already been given in Eq.(3.11), but x_i in the formula now denotes the quality prediction from the metric for video sequence i. The PC values are within the range [0,1], with 1 indicating the highest linear relationship between the model predictions and the subjective quality ratings.

The RMSE value is defined as:

$$RMSE = \sqrt{\frac{1}{N_v - d} \sum_{k=1}^{N_v} [DMOS(k) - PQ(k)]^2} \tag{3.25}$$

where N_v denotes the number of videos considered in the analysis, and d denotes the number of metric parameters which need to be trained from the subjective data. $DMOS$ is the obtained quality rating from the subjective test and PQ is the predicted quality from the metrics.

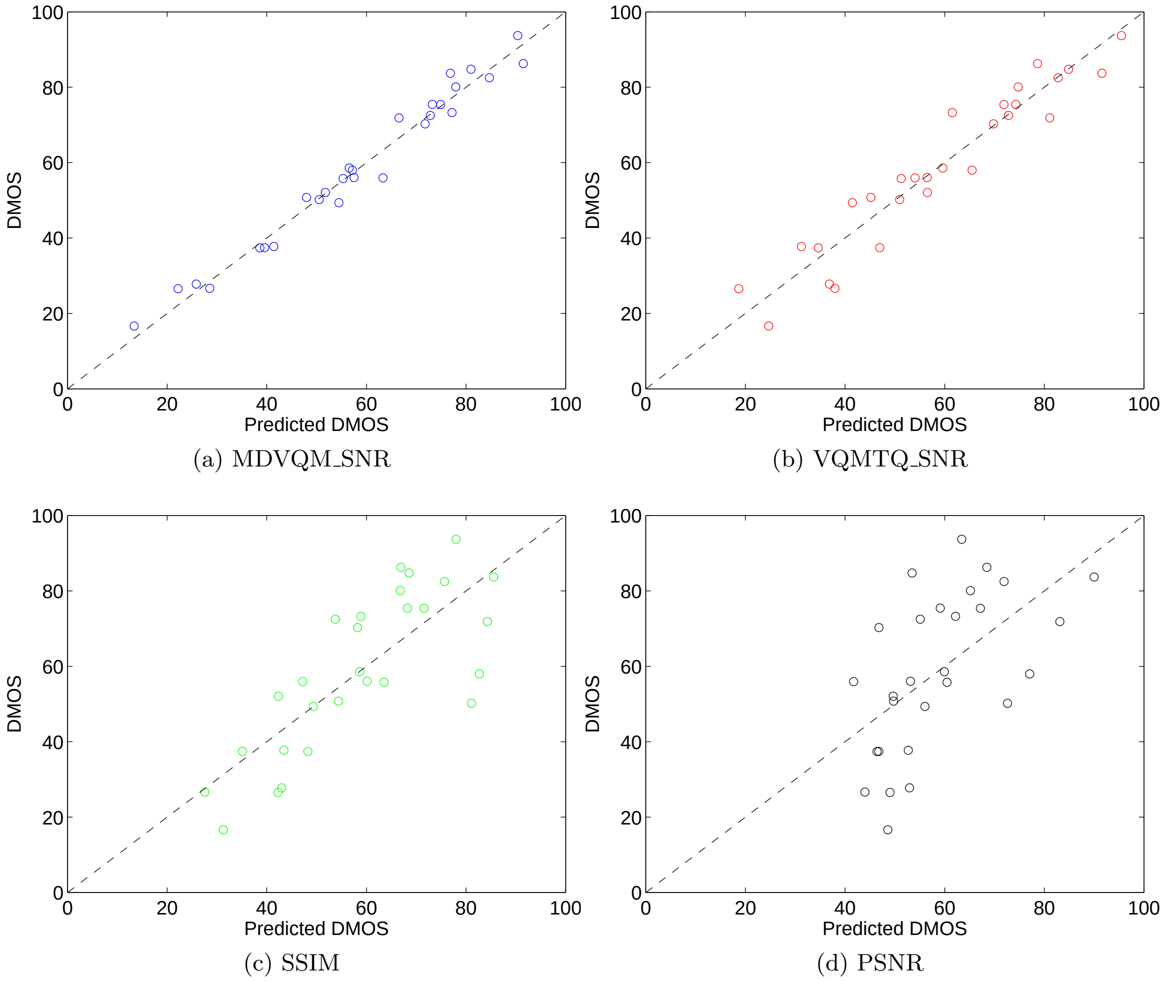
If the prediction of a video quality metric (PQ) deviates too far from the subjective data ($DMOS$), then it is considered as an outlier:

$$|DMOS(k) - PQ(k)| > 2 \cdot \frac{\sigma_{DMOS}(k)}{\sqrt{N_s}} \tag{3.26}$$

where N_s is the number of test subjects, and σ_{DMOS} denotes the standard deviation of the DMOS value over all N_s subjects. The OR is then calculated as the ratio of number of outliers R_0 to the total number of test videos in the analysis:

$$OR = \frac{R_0}{N_v} \tag{3.27}$$

The performance of the proposed SNR quality metric (noted as MDVQM_SNR) is evaluated and compared with three other objective metrics: PSNR, SSIM [WBSS04, WLB04] and



(a) MDVQM_SNR

(b) VQMTQ_SNR

(c) SSIM

(d) PSNR

Figure 3.8: Actutal DMOS vs. predicted DMOS from the SNR quality models

the SNR quality metric in VQMTQ as given in Eq.(3.1) (referred to as VQMTQ_SNR). For a fair comparison, the PSNR and SSIM values are first fitted to the DMOS values measured from the subjective tests by the use of first order least-squares fitting. The linear relationship between the actual DMOS values and predicted quality values from all the four metrics are given in Figure 3.8. It can be seen that the predicted quality values from PSNR are very inaccurate due to the neglect of content characteristics. The performance of SSIM is much better by considering the structural information of the video content, but is still not very satisfactory. In comparison, the predictions from MDVQM_SNR and VQMTQ_SNR are more linearly correlated with the subjective ratings. The statistical metrics for performance evaluation along with the corresponding 95% confidence intervals are given in Tables 3.4-3.6. The limits of the 95% confidence intervals are represented by the lower bound (LB) and upper bound (UB). The results show that in every aspect of the metric performance, MDVQM_SNR provides better results than the comparison metrics.

To determine whether the performance of the metrics is significantly different from a statistical point of view, significance tests based on F-test are performed. For example, if the result from the significance test between two metrics is 0.95, then it can be concluded with 95% confidence that the performance difference of the two comparison metrics is statistically significant. For more information about significance tests for video quality metrics, the readers can refer to [VQE00, PW08]. The results of the significance tests are also given together with the corresponding performance metrics in Tables 3.4-3.6. From the results, it can be seen that the statistical significance of the performance difference between MDVQM_SNR and the comparison metrics is well above the 95% significance level for all the three performance metrics.

Table 3.4: Pearson correlation values of the SNR quality metrics

Metric	*PC*	*LB PC*	*UB PC*	*Sig. Level*
PSNR	0.5753	0.2575	0.7808	1
SSIM	0.7795	0.5731	0.8929	1
VQMTQ_SNR	0.955	0.9039	0.9792	1
MDVQM_SNR	0.987	0.9717	0.994	-

Table 3.5: RMSE values of the SNR quality metrics

Metric	*RMSE*	*LB RMSE*	*UB RMSE*	*Sig. Level*
PSNR	16.936	13.4401	22.9051	1
SSIM	12.9703	10.2929	17.5417	1
VQMTQ_SNR	6.1614	4.8896	8.333	0.9989
MDVQM_SNR	3.3377	2.6488	4.5141	-

Table 3.6: Outlier ratios of the SNR quality metrics

Metric	*OR*	*CI*	*Sig. Level*
PSNR	0.8214	0.1419	0.9996
SSIM	0.75	0.1604	0.9978
VQMTQ_SNR	0.5	0.1852	0.8199
MDVQM_SNR	0.3214	0.173	-

Another concern when evaluating a quality metric is the performance on unknown data. In this case, the data sets for training and validation should be separate. This is the way VQEG proceeds in their tests [VQE00, VQE03]. Compared with the subjective tests conducted by VQEG, a relatively small set of test sequences are used in our subjective tests. To overcome the problem of limited subjective data for training and verification of the quality metrics, cross validation [Gei93, DK82] is used to evaluate the proposed metric for unknown data. There are different forms of cross validation, and the most widely used variant is the K-fold cross validation. In K-fold cross validation, the entire data set is divided into K subsets of equal size. From the K subsets, one is selected as the validation set and the other $K - 1$ subsets are used for training the metrics. This process is repeated for K times, with each subset being used once as the validation data. In our validation, the leave-one-out cross validation (LOOCV) [Sto74], which is the simplest case of K-fold cross validation with K equals to the size of the entire dataset, is used. This means that each time one source sequence out of the training data set is excluded from the training data set and the metric parameters are trained using data from other sequences. Afterwards, the excluded data are used for validation purpose. If the proposed metric works well for all the verification sequences, it is stable and accurate.

To simplify the cross validation, 5 source sequences (HA, CR, PJ, OB, FB) with different characteristics in terms of motion and spatial details are first selected. They are always kept in the training data set. For the remaining 3 sequences (SC, PA, KO), one sequence is used each time for validation and the other two are used for the training together with the other 5 sequences above. The results of the cross validation are shown in Table 3.7. Here for comparison purposes, the validation result for VQMTQ_SNR is also included. Note that TA and SA values from the processed sequences are used for MDVQM_SNR, while for VQMTQ_SNR, the features are extracted from the reference sequences.

Table 3.7: Cross validation result for the SNR metrics

Test	Veri.Seq.	MDVQM_SNR		VQMTQ_SNR	
		PC	RMSE	PC	RMSE
Test1	Soccer	0.9989	4.4839	0.9970	8.2911
Test2	Kobe	0.9914	2.0296	0.9902	5.6786
Test3	Peda	0.9988	2.3611	0.9998	3.3248

From the results, it can be seen that both metrics provide very stable and accurate predictions for the unknown data sets, with all PC values higher than 0.99. The proposed MDVQM_SNR metric performs a little better with smaller RMSE values. This is in accordance with the performance evaluation results in Table 3.4-3.6.

The better performance of the proposed metric is due to the fact that in VQMTQ_SNR, the drop rate of video quality against PSNR (which is the multiplier to the PSNR value) is modelled as a content-independent constant, which ignores the characteristics of the underlying videos. But in fact, the video content has a masking effect on the perceived video quality. Video with high spatial or temporal details can hide the negative impact of encoding noise

(PSNR drop) to some extent, so that the perceived quality of such videos drops more slowly when the PSNR decreases. In the proposed MDVQM_SNR metric, the drop rate of video quality against the pixel bit-rate is modelled as a function of TA and SA, so that it adapts to the characteristics of the video content. The cost of this is that MDVQM_SNR needs two more model parameters than VQMTQ_SNR. But the better prediction performance justifies the increased complexity of the metric.

When subjective ratings from all the SRCs are used to train MDVQM_SNR, the obtained model parameters are given in Table 3.8a. For reference, the obtained model parameters for VQMTQ_SNR are also given in Table 3.8b. They are used in the following sections for the estimation of temporal and spatial quality.

Table 3.8: Model parameters trained with all the subjective ratings

(a) Model parameters for MDVQM_SNR (Eq.(3.24))

a_0	a_1	a_2	a_3	a_4	a_5
-0.632	0.6591	4.1797	-0.0352	0.1133	5.6086

(b) Model parameters for VQMTQ_SNR (Eq.(3.1))

p	b_0	b_1	b_2
0.4023	40.7058	2.0562	10.6558

3.4.2 Temporal Quality Metric

Apart from changing the QP values, another option for video adaptation is frame rate reduction. On one hand, reducing the frame rate allows us to use a smaller QP value for the compression, so that the SNR quality of encoded pictures could be improved. On the other hand, the resulting jitter/jerkiness artifacts also impair the user experience. How frame rate reduction affects the overall perceived quality is hence a very practical question to be answered. In this section, the impact of frame rate reduction is studied and a temporal video quality metric is proposed to simulate the impact.

3.4.2.1 Design of the Temporal Quality Metric

As proposed in [OMLW11, PS11], the overall video quality in the presence of frame rate reduction is modelled as the product of two terms:

$$TVQM = SNRVQ \cdot TCF \tag{3.28}$$

where $SNRVQ$ models the SNR video quality without considering the impact of jerkiness introduced by frame rate reduction. The second item, Temporal Compensation Factor (TCF), models the negative impact of frame rate reduction. Normally, the value of TCF is in the range [0,1].

For the estimation of SNRVQ, the metric in Eq.(3.24) can be used. One thing to note here is that after frame rate reduction, the TA values of the adapted sequences (noted as TA_r) needed in Eq.(3.24) are different from those of the full-resolution sequences (noted as TA_f) due to the larger temporal distance of the successive frames. Since our purpose is to use this metric for video adaptation and before any real adaptation operations only TA_f can be measured, TA_f is used in the estimation of SNRVQ here for the temporal quality model. Of course, this may affect the accuracy of the prediction and a correction process is introduced later in this subsection.

The bpp value in Eq.(3.24) can be calculated as:

$$bpp = bpp_0 \cdot (\frac{FR_{max}}{FR}) \tag{3.29}$$

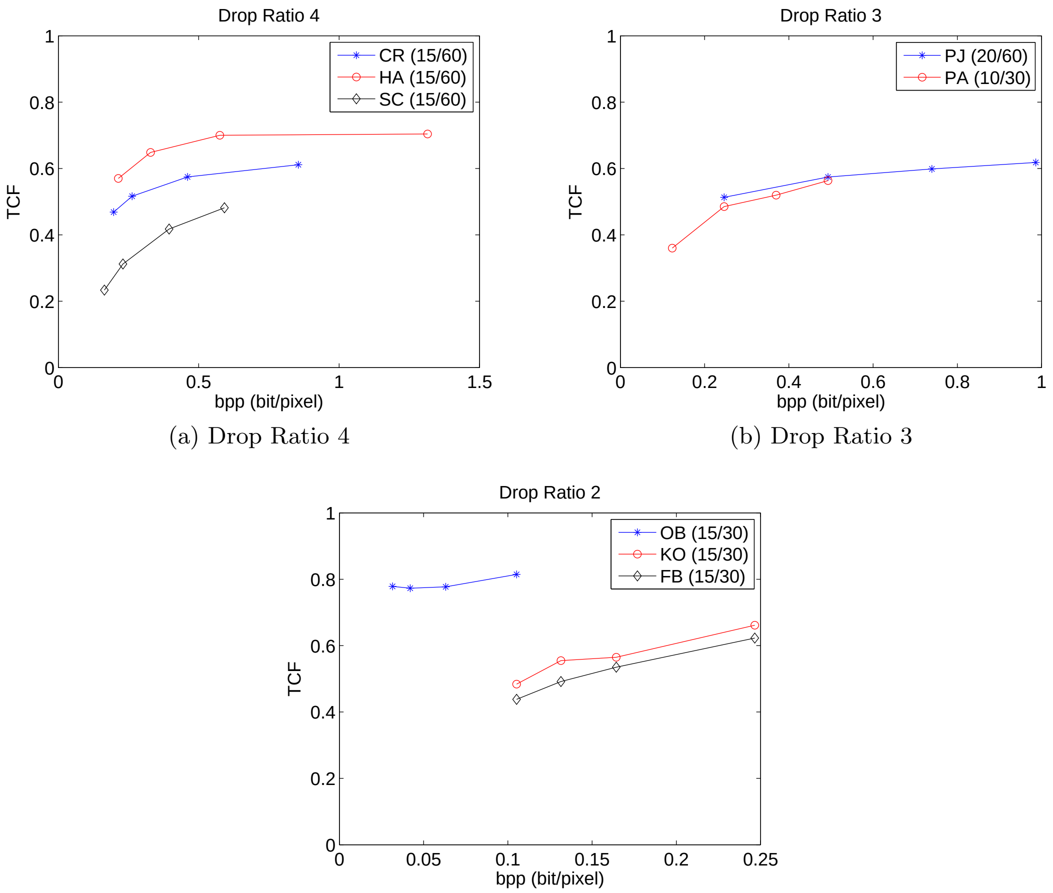
where bpp_0 is the pixel bit-rate when the sequence is encoded at the target bit-rate but with full temporal resolution, FR_{max} is the original frame rate and FR is the actual frame rate after frame rate reduction.

The actual video qualities of the test sequences are known from the results of the subjective test, and SNRVQ can be estimated using Eq.(3.24). In this way, the TCF is derived as:

$$TCF = DMOS/SNRVQ \tag{3.30}$$

The TCF curves for different sequences are shown in Figure 3.9. The subplots correspond to different frame dropping ratios (from 2 to 4), respectively. It can be seen that the TCF is also a function of bpp. But as mentioned above, TCF actually aims to model the impact of jerkiness by frame dropping, so it should not have a very strong relationship with bpp. The reason is that the TA values of the full-resolution sequences are used instead of those of the sequences after adaptation. This can be explained with Figure 3.10.

In Figure 3.10, the x-axis is the pixel bit-rate and the y-axis is the SNR video quality (which is only subjected to quantization artifact without considering the impact of frame dropping). The solid blue curve is the RD-curve for the video sequence before adaptation (V_o) and the red dotted curve is the RD-curve for the video sequence after adaptation (V_a). The impact of frame dropping on the spatial details is very limited and can be neglected, so the spatial activity remains the same before and after the adaptation. But as discussed earlier, the larger temporal distance of the successive frames after frame dropping results in a higher temporal activity. In this sense V_a is “harder” to be compressed than V_o, so at the same pixel bit-rate, a coarser quantization is needed for V_a and its SNR video quality is lower. This is why in Figure 3.10, the red curve is always below the blue curve. Before the adaptation, the video is encoded at a pixel bit-rate of bpp_0, corresponding to point P1 in Figure 3.10. After the frame dropping, the available pixel bit-rate becomes bpp calculated by Eq.(3.29). Since



(a) Drop Ratio 4

(b) Drop Ratio 3

(c) Drop Ratio 2

Figure 3.9: TCF vs. bpp without correction

the TA and SA values of the full-resolution video V_o are used for the estimation of SNR video quality, the predicted value still corresponds to the point on the blue curve (P2). From P1 to P2, the *bpp* value increases because more bits can be used to encode the pixels in the non-skipped frames. However, the actual operation point should be on the red curve at the same pixel bit-rate *bpp* (P3 in Figure 3.10). If the blue curve is still used for the quality estimation, a pixel bit-rate which is a little lower than *bpp* should be used (P4 in Figure 3.10). This means if TA and SA values of the full-resolution video V_o are used in Eq.(3.24) to estimate the SNR video quality in case of frame dropping, instead of using Eq.(3.29), *bpp* should be calculated as:

$$bpp = bpp_0 \cdot (\frac{FR_{max}}{FR})^{P_T} \tag{3.31}$$

where P_T is a content-dependent parameter within the range [0,1] and P_T increases as TA decreases. As an extreme case, when $TA \rightarrow 0$ (which means a static scene), frame dropping does not affect the spatial-temporal complexity anymore. In this case, the red and blue curves in Figure 3.10) should coincide with each other which means $P_T \rightarrow 1$. Therefore, in this work,

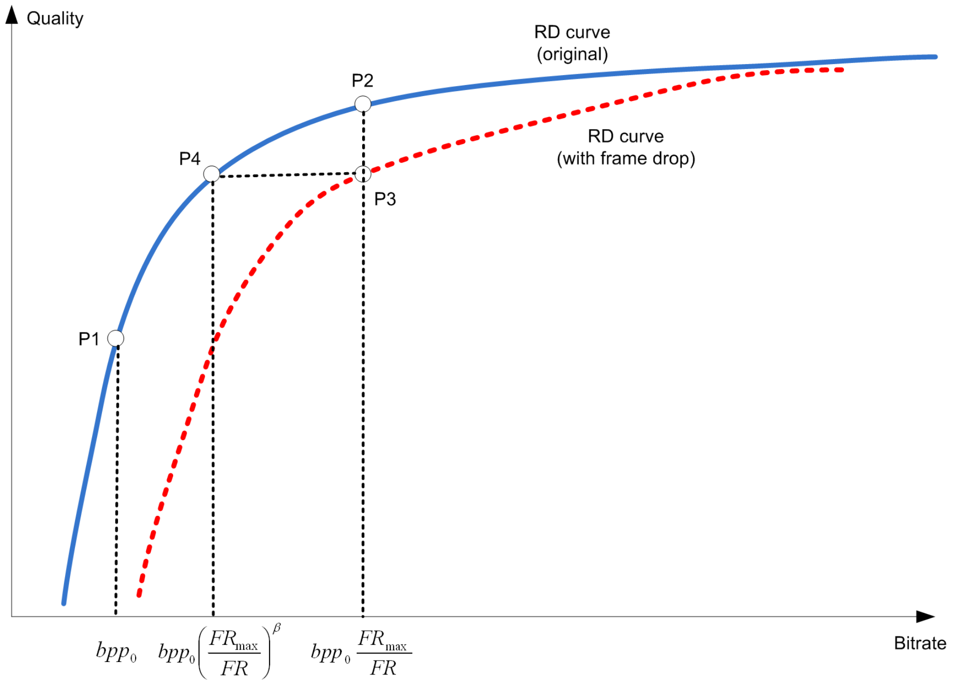

Quality
RD curve
(original)
P2
RD curve
(with frame drop)
P4
P3
P1
Bitrate
bpp_0
$bpp_0\left(\frac{FR_{\max}}{FR}\right)^{\beta}$
$bpp_0\frac{FR_{\max}}{FR}$

Figure 3.10: Illustration of the bpp correction process

an exponential function of TA is used to estimate P_T:

$$P_T = e^{-a_T \cdot TA} \tag{3.32}$$

where a_T is a model parameter which needs to be trained with subjective data.

From Figure 3.9, several further conclusions can be drawn:

1. The decrease of the DMOS value is smaller for low-motion video content such as HA/OB (suggested by a higher TCF value) and larger for high-motion videos such as SC/KO/FB. This indicates that the impact of frame dropping is content-dependent and it has a stronger negative impact for videos with higher temporal activity.
2. At the same reduction ratio, the impact of frame dropping is different for 60fps and 30fps sequences. This can be observed in Figure 3.9b. Actually, PJ is a sequence with much higher motion than PA, but the TCF values of PJ are higher than those of PA. This suggests that frame dropping might have a more negative impact on low frame rate videos (e.g. PA with 30fps) than high frame rate ones (e.g. PJ of 60fps). This is due to the fact that, although the reduction ratio is 3 in both cases (60fps to 20fps for PJ and 30fps to 10fps for PA), a frame rate of 10fps already causes some uncomfortable viewing experience while a frame rate of 20fps is still acceptable for most viewers.

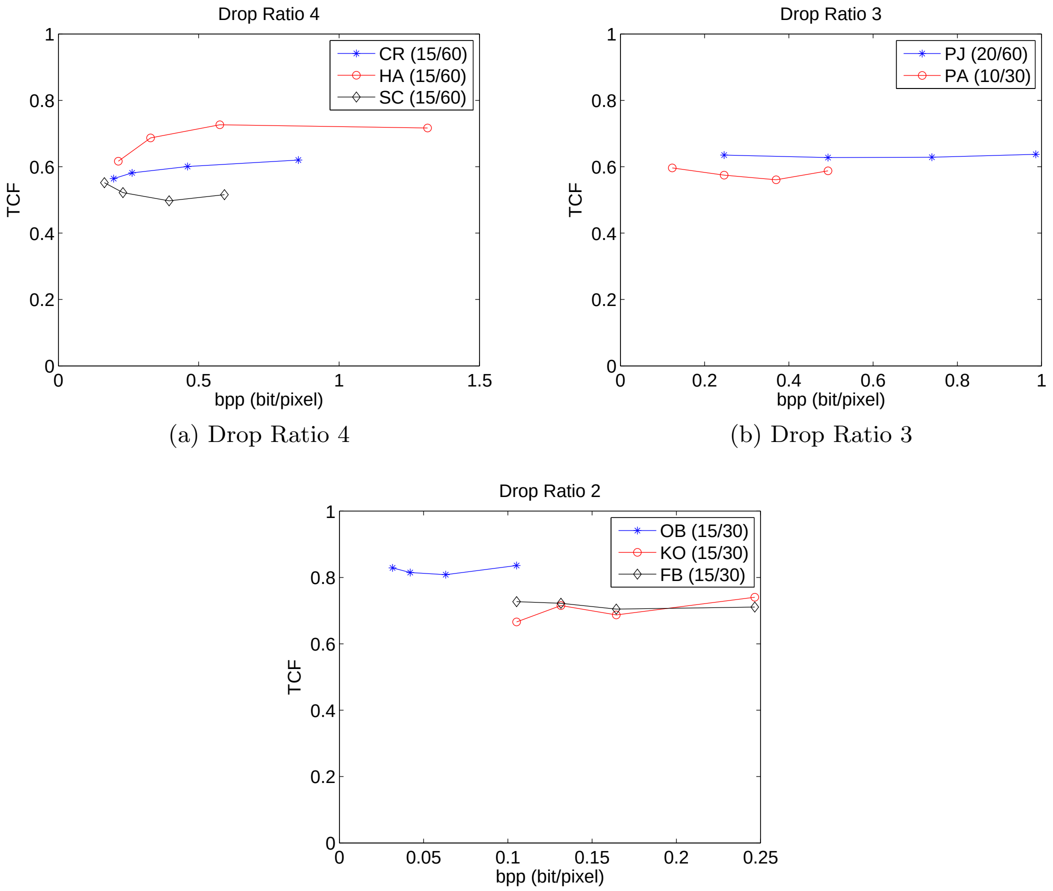
According to the two considerations above, our proposed model for TCF is given as:

$$TCF = \frac{FR}{FR_{max}} \cdot \frac{1 + b_T \cdot FR_{max}/TA}{1 + b_T \cdot FR/TA} \tag{3.33}$$

where b_T is a model parameter which needs to be trained. It can be seen that when $FR = FR_{max}$ or $TA \to 0$, TCF reaches 1, which is in accordance with the fact that at full frame rate or for a static scene, the overall quality should be the same as the SNR quality.

The proposed temporal quality metric in Eq.(3.28) is trained again by using Eq.(3.24) and Eq.(3.31) to estimate $SNRVQ$ and using Eq.(3.33) for the TCF. When the temporal quality metric is fitted to subjective ratings from all 8 SRCs, the obtained model parameters are $a_T = 0.0518$ and $b_T = 0.7889$.

Now, the TCF values can be calculated again with Eq.(3.30) and the results are shown in Figure 3.11. It can be seen that the curves are much more flat and this means that the TCF can now be modelled independently of *bpp* which justifies the functional form of the TCF model in Eq.(3.33).



(a) Drop Ratio 4

(b) Drop Ratio 3

(c) Drop Ratio 2

Figure 3.11: TCF vs. bpp with correction

3.4.2.2 Performance Analysis of the Temporal Quality Metric

In this section, the performance of the proposed temporal video quality metric (referred to as MDVQM_TVQ) is evaluated in the presence of frame dropping. The metric is compared with two other metrics that also consider the impact of frame rate reduction:

- VQMTQ_TVQ: the metric in [OMLW11] (see Eqs.(3.1)-(3.3))
- STVQM_TVQ: the metric in [PS11] (see Eqs.(3.6)-(3.8))

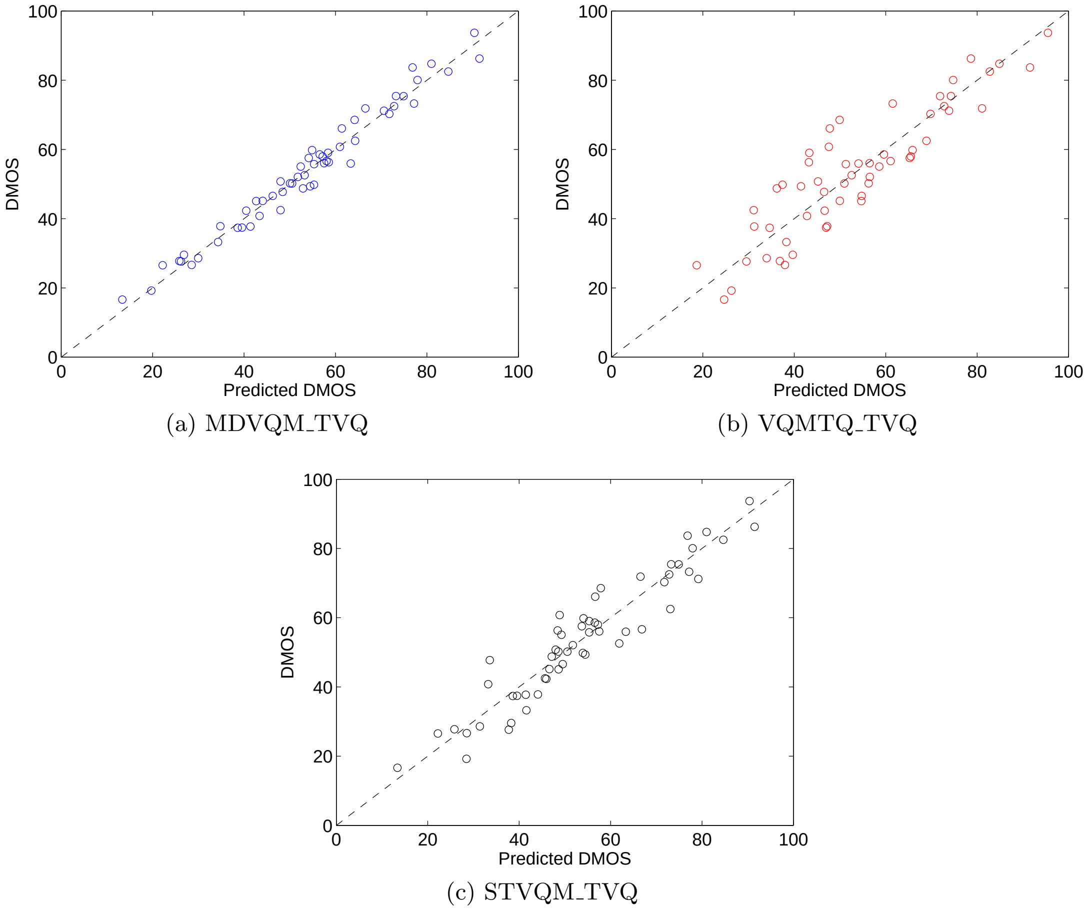
For the modelling of the TCF, both MDVQM_TVQ and STVQM_TVQ use the feature TA extracted from the video sequence and two model parameters need to be trained. In VQMTQ_TVQ, two features from the video (Motion Direction Activity (MDA) and Displaced Frame Difference (DFD)) are used and three parameters need to be trained from the subjective data. For MDVQM_TVQ, the TA value is calculated from the processed videos, while for STVQM_TVQ and VQMTQ_TVQ, the required feature values are derived from the reference videos. For all the metrics, the model parameters are trained based on the subjective ratings of the sequences in the SNR and TR group obtained in our subjective tests.

To give an intuitive view of the estimation accuracy, Figure 3.12 illustrates the linear correlation between the predicted quality values and the actual DMOS values. From the figures, it can be seen that the predictions from all the three quality metrics have a high linear correlation with the subjective ratings and the proposed MDVQM_TVQ provides the best performance. These observations are confirmed and quantified by the statistical performance metrics given in Table 3.9-3.11. MDVQM_TVQ outperforms the other two comparison metrics with a higher PC value and a smaller RMSE value. The results from the significance test show that this observed performance difference is statistically significant.

Table 3.9: Pearson correlation values of the temporal quality metrics

Metric	*PC*	*LB PC*	*UB PC*	*Sig. Level*
VQMTQ_TVQ	0.9011	0.8362	0.941	1
STVQM_TVQ	0.9468	0.9105	0.9686	0.999
MDVQM_TVQ	0.9855	0.9752	0.9915	-

Similar to the evaluation of the SNR quality metric, cross validation is performed to evaluate the metric for unknown data sets. Again, the sequences SC/KB/PA are used as verification sequences for the cross validation. Table 3.12 shows the results of the cross validation. The results confirm that the proposed temporal quality model also provides the best prediction performance for unknown data.



(a) MDVQM_TVQ

(b) VQMTQ_TVQ

(c) STVQM_TVQ

Figure 3.12: Actual DMOS vs. predicted DMOS from the temporal quality metrics

Table 3.10: RMSE values of the temporal quality metrics

Metric	*RMSE*	*LB RMSE*	*UB RMSE*	*Sig. Level*
VQMTQ_TVQ	7.8957	6.666	9.686	1
STVQM_TVQ	5.8041	4.9002	7.1202	1
MDVQM_TVQ	3.0627	2.5857	3.7572	-

Table 3.11: Outlier ratios of the temporal quality metrics

Metric	*OR*	*CI*	*Sig. Level*
VQMTQ_TVQ	0.6071	0.1279	0.9993
STVQM_TVQ	0.5	0.131	0.996
MDVQM_TVQ	0.2321	0.1106	-

Table 3.12: Cross validation results for the temporal quality metrics

Test	V.Seq.	VQMTQ_TVQ		STVQM_TVQ		MDVQM_TVQ	
		PC	RMSE	PC	RMSE	PC	RMSE
Test1	Soccer	0.9963	7.6071	0.9431	7.0040	0.9789	5.6327
Test2	Kobe	0.8860	6.3960	0.9428	5.4634	0.9897	1.9002
Test3	Peda	0.9860	3.2990	0.9675	5.4790	0.9969	2.2346

3.4.3 Spatial Quality Model

The third option to adapt a video stream is to reduce the spatial resolution (spatial down-sampling). Reducing the frame size allows for a finer quantization because the amount of information to be compressed is reduced, thus it can alleviate several artifacts such as blocking and ringing, etc. But on the other hand, spatial down-sampling is an irreversible process and will introduce blurring into the video if it is up-sampled to the original spatial resolution. In this section, the impact of spatial resolution on the perceived video quality is studied.

3.4.3.1 Design of the Spatial Quality Model

In [OXMW11], it was observed that the overall video quality with spatial down-sampling can be decomposed as:

$$SVQM = SNRVQ \cdot SCF \tag{3.34}$$

where $SNRVQ$ is the quality for a video which is subjected only to quantization effects. And SCF is a Spatial Correction Factor, which captures the impact of spatial down-sampling. Roughly, it can be considered that $SNRVQ$ represents the SNR quality when the video is down-sampled, encoded and displayed at the reduced spatial resolution without resampling back to the original size. So there is no blurring effect introduced. Then SCF simulates the negative impact of the blurring effect introduced when the video is up-sampled and displayed at the original size.

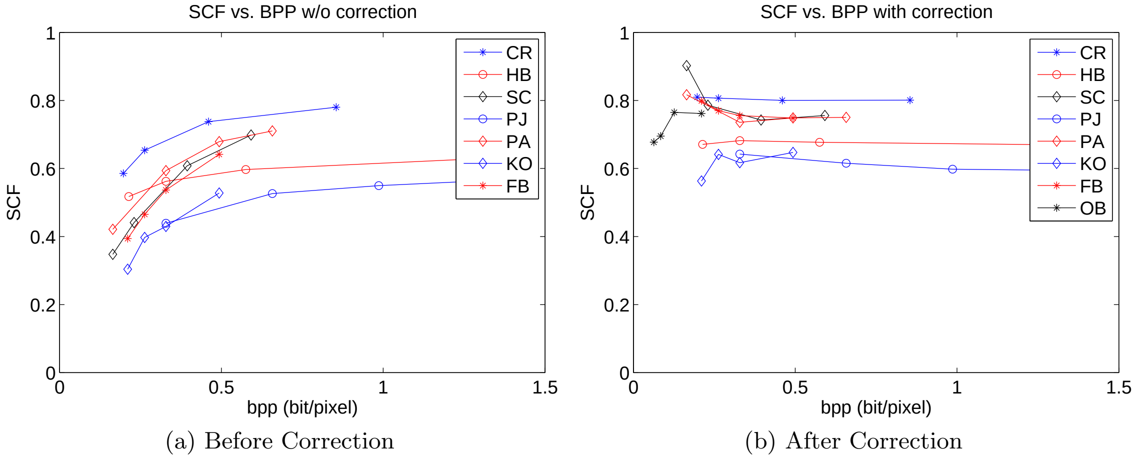
Similar to the case of frame rate reduction, when calculating SNRVQ using the model in Eq.(3.24), the SA values of the full-resolution (4CIF) sequences are used (because before the decision for video adaptation is made, spatial down-sampling has not been done yet and the SA value of the CIF sequence is unknown). The pixel bit-rate after transcoding with spatial down-sampling can be calculated by:

$$bpp = bpp_0/SF \tag{3.35}$$

where bpp_0 is the pixel bit-rate when the video is encoded at the target bit-rate with the original resolution. SF denotes the spatial Scaling Factor which is the ratio between the reduced and the original spatial resolution. In our case, $SF = CIF/4CIF = 0.25$.

Inserting Eq.(3.35) into Eq.(3.24), the estimated $SNRVQ$ in Eq.(3.34) can be obtained and SCF is derived as:

$$SCF = DMOS/SNRVQ \tag{3.36}$$



(a) Before Correction

(b) After Correction

Figure 3.13: SCF curves with/without correction

Figure 3.13a shows the obtained SCF values by this way and it can be seen that the SCF values depend heavily on the pixel bit-rate. Similar to the design of the temporal quality metric, it is desired that the SCF is modelled independently of the bit-rate, so a correction to the bpp value is introduced for the estimation of $SNRVQ$:

$$P_S = e^{-a_S \cdot SA} \qquad bpp = bpp_0 \cdot (SF)^{P_S} \tag{3.37}$$

where a_S is a model parameter which needs to be trained using subjective test results.

From Figure 3.13a, it can be seen that the SCF value is content dependent. For contents with higher spatial details, the SCF value is lower, indicating that the quality of this kind of video is affected more seriously by spatial down-sampling. Based on these observations, the proposed SCF model is given as:

$$SCF = (SF)^{b_S \cdot SA} \tag{3.38}$$

where b_s is a model parameter which needs to be trained. It can be seen that when $SF = 1$ or $SA \to 0$, SCF reaches 1, which indicates that without spatial down-sampling or for sequences with very few spatial details, the overall quality should be the same as the SNR quality.

Using the subjective ratings from all 8 SRCs, the two model parameters a_S and b_S are trained by least-square non-linear fitting. The obtained values are $a_S = 0.0222$ and $b_S = 0.0035$.

Figure 3.13b shows the obtained SCF values with the correction given in Eq.(3.37). It can be seen that after the correction, the curves are quite flat at middle or high bit-rate (when *bpp* is greater than 0.3 bits/pixel), indicating that the SCF is independent of the bit-rate. However, at the low bit-rate end, the curves become a little irregular. This can be attributed to the difficulty in rating the videos when the quality is very low. For example, when there are very obvious artifacts in the video, it is hard to decide whether to give it a rating of 20 or 30. But these ratings do have great impact on the obtained SCF values. It is assumed that the curve will become flat and regular when the number of test subjects is larger.

3.4.3.2 Performance Analysis of the Spatial Quality Model

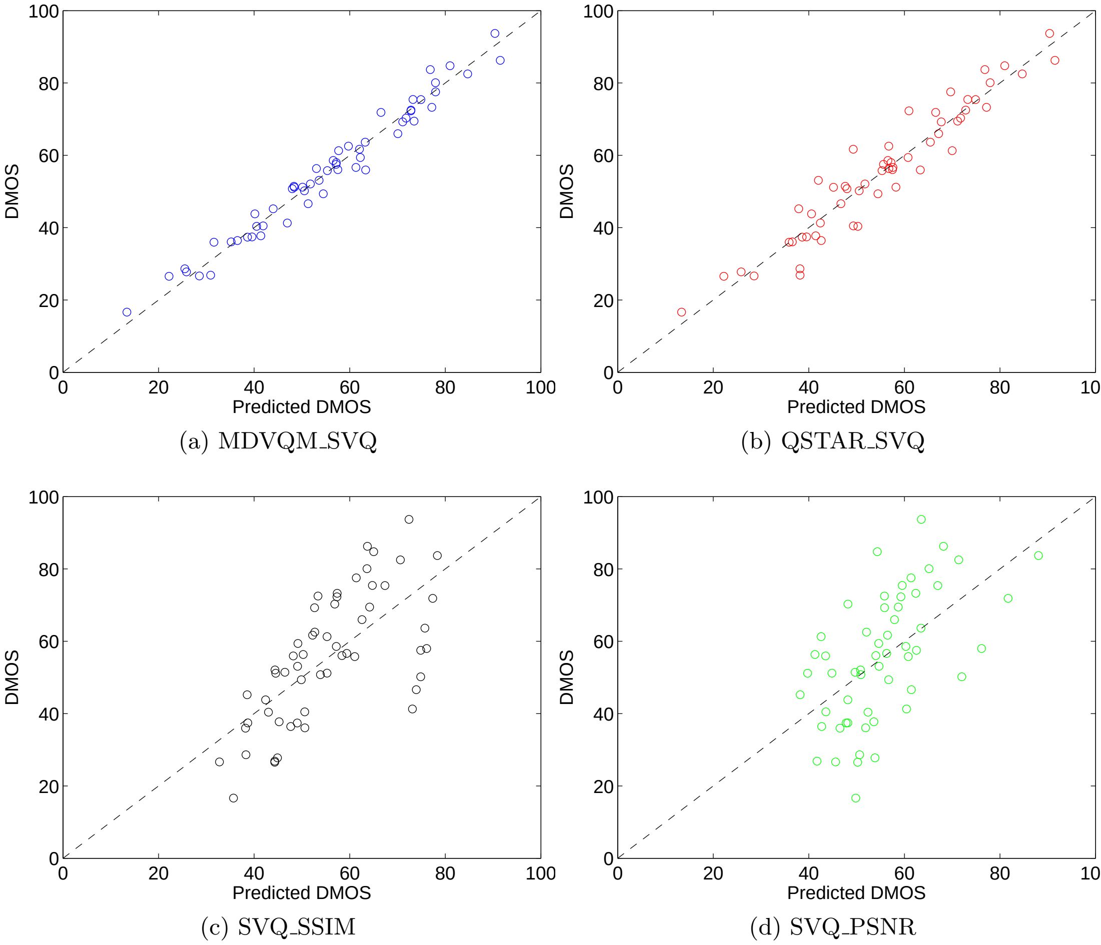
In this section, the performance of the proposed spatial quality metric (referred to as MDVQM_SVQ) is evaluated and compared with that of three other quality metrics: PSNR, SSIM and the spatial quality metric proposed in [OXMW11] (see Eq. (3.4), referred to as QSTAR_SVQ).

Figure 3.14 shows the linear relationship between the actual DMOS and the predicted quality values from the models in comparison. Comparing Figure 3.14c with Figure 3.8c, it can be found that the SSIM index becomes less accurate in case spatial down-sampling is performed. This indicates that, although down-sampling only introduces spatial artifacts to the video, SSIM alone is uncapable of capturing this impact on video quality caused by spatial down-sampling. In comparison, the two metrics which explicitly model the impact of spatial resolution, i.e., MDVQM_SVQ and QSTAR_SVQ do provide much better quality prediction than SSIM and PSNR. The statistical performance metrics and results of significance tests are summarized in Table 3.13-3.15. All the performance metrics indicate that the proposed spatial quality metric provides the best prediction among the metrics in comparison. The difference of performance is statistically significant, as suggested by the significance test results.

Also, our SCF model requires only one video feature (SA value) and two parameters that need to be trained from the subjective ratings, while the model QSTAR_SVQ requires four parameters. The better results of the proposed model comes from the fact that the impact of frame down-sampling is dependent on the characteristic of the video which is considered in MDVQM_SVQ, whereas no video feature is considered in QSTAR_SVQ.

Table 3.13: Pearson correlation values of the spatial quality metrics

Metric	*PC*	*LB PC*	*UB PC*	*Sig. Level*
PSNR	0.5749	0.3675	0.7278	1
SSIM	0.6765	0.5031	0.7976	1
QSTAR_SVQ	0.9298	0.8827	0.9584	0.9926
MDVQM_SVQ	0.9846	0.9737	0.991	-



(a) MDVQM_SVQ

(b) QSTAR_SVQ

(c) SVQ_SSIM

(d) SVQ_PSNR

Figure 3.14: Actual DMOS vs. predicted DMOS for the spatial quality metrics

Table 3.14: RMSE values of the spatial quality metrics

Metric	*RMSE*	*LB RMSE*	*UB RMSE*	*Sig. Level*
PSNR	14.4586	12.2068	17.7371	1
SSIM	13.0131	10.9864	15.9638	1
QSTAR_SVQ	6.5194	5.5041	7.9977	0.9999
MDVQM_SVQ	3.1252	2.6385	3.8339	-

Table 3.15: Outlier ratios of the spatial quality metrics

Metric	*OR*	*CI*	*Sig. Level*
PSNR	0.7857	0.1075	1
SSIM	0.7679	0.1106	1
QSTAR_SVQ	0.5179	0.1309	0.9687
MDVQM_SVQ	0.25	0.1134	-

Again, the performance of the metrics on unknown data sets is evaluated through cross validation. The sequences SC/KB/PA are used as the verification sequence in the three validations separately. Table 3.16 summarizes the results of the cross validation:

Table 3.16: Cross validation result for the spatial quality metrics

Test	Veri.Seq.	QSTAR_SVQ		MDVQM_SVQ	
		PC	RMSE	PC	RMSE
Test1	Soccer	0.9647	5.6103	0.9897	3.0565
Test2	Kobe	0.9195	7.3253	0.9881	2.8517
Test3	Peda	0.9360	6.4076	0.9876	2.8659

3.4.4 Spatial-Temporal Quality Model

As mentioned in Section 3.3, subjective tests are conducted to study the impact on perceived video quality when both TR and SR are changed simultaneously (TEST II). In this section, the interaction between spatial and temporal impairments and the corresponding impact on overall video quality are examined.

As shown in the previous sections, the impact of quantization is separable from that of TR and SR changes. So the overall spatial-temporal video quality metric MDVQM is designed as:

$$P_T = e^{-a_T \cdot TA} \qquad P_S = e^{-a_S \cdot SA} \tag{3.39}$$

$$bpp = bpp_0 \cdot (SF)^{P_S} \cdot (\frac{FR_{max}}{FR})^{P_T} \tag{3.40}$$

$$MDVQM = SNRVQ(bpp, TA, SA) \cdot STCF(TA, SA) \tag{3.41}$$

where $STCF$ is a Spatial-Temporal Correction Factor which simulates the negative impact of jerkiness and blurring artifacts introduced by frame dropping and spatial down-sampling.

Earlier metrics in the literature assume that the impacts of frame dropping and spatial down-sampling are separable [OXMW11], so that $STCF$ is modelled as the product of TCF

and SCF. If this assumption is adopted, the $STCF$ can be derived from the TCF and SCF models proposed in the previous sections as:

$$STCF = TCF \cdot SCF \tag{3.42}$$

where, TCF and SCF can be calculated according to Eq.(3.33) and Eq.(3.38), respectively.

To verify the accuracy of the above model, the subjective ratings from TEST II are used as the validation data set. For $SNRVQ$, TCF and SCF, the obtained model parameters given in Section 3.4.1 to 3.4.3 are used. Figure 3.15(a) shows the linear relationship between the actual subjective ratings and the predicted DMOS values using the $STCF$ model in Eq.(3.42) (referred to as PROD). The performance metrics of model PROD are given in the first lines of Tables 3.17-3.19. From the results, it can be seen that the prediction accuracy is not satisfactory. The RMSE value is high and the predicted values are often far below the actual value as shown in Figure 3.15(a).

Based on this observation, instead of directly using the product of TCF and SCF as in Eq.(3.42), three other models for the STCF are examined:

$$STCF = \sqrt{TCF \cdot SCF} \tag{3.43}$$

$$STCF = \max\left(TCF, SCF\right) \tag{3.44}$$

$$STCF = \min\left(TCF, SCF\right) \tag{3.45}$$

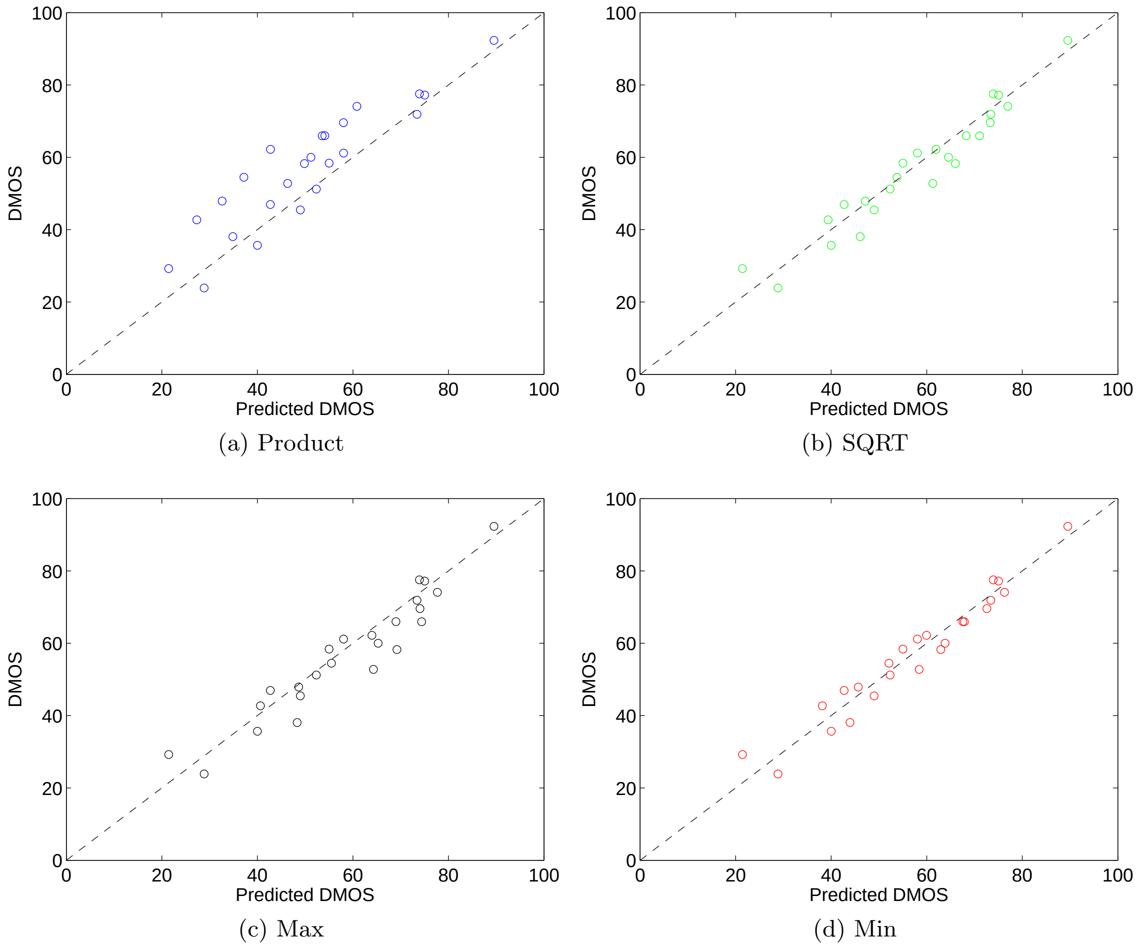
and the overall video quality is then predicted by Eq.(3.41).

Again, the subjective ratings from TEST II are used to validate these candidate models. The metric predictions and the actual DMOS values are shown in Figs.3.15(b)-(d). The performance metrics of different models are given in Tables 3.17-3.19.

From the results, it can be seen that the minimum function in Eq.(3.45) achieves a better overall performance than the other three comparison $STCF$ models with a higher PC value and lower RMSE and OR values. The results from the significance tests (also shown in Tables 3.17-3.19) indicate that this performance difference between MIN and PROD is statistically significant, but the statistical significance of the difference among SQRT, MAX and MIN is below the typical 95% significance level.

Further, the performance of the metrics is examined on each individual verification sequence as shown in Table 3.20. It can be seen that, although the minimum function does not always perform the best (e.g., the performance of SQRT and MAX is better for the sequence FOOTBALL), the performance of it is much more stable than that of the other models. Based on the above observation, the minimum function given in Eq.(3.45) is selected to calculate $STCF$ in our overall spatial-temporal video quality metric MDVQM (see Eq.(3.41)).

This indicates that when both TR and SR are reduced, the perceived video quality is mostly affected by the prevailing (more significant) distortion, either temporal or spatial.



(a) Product

(b) SQRT

(c) Max

(d) Min

Figure 3.15: Linear relationship between the actual DMOS and the predicted DMOS for different STCF models

Table 3.17: Pearson correlation values of the spatial-temporal quality metrics

Metric	*PC*	*LB PC*	*UB PC*	*Sig. Level*
PROD	0.9093	0.7989	0.9604	0.9455
SQRT	0.9647	0.9189	0.9848	0.3068
MAX	0.9525	0.8918	0.9795	0.6207
MIN	0.9723	0.9360	0.9881	-

Table 3.18: RMSE values of the spatial-temporal quality metrics

Metric	*RMSE*	*LB RMSE*	*UB RMSE*	*Sig. Level*
PROD	9.4365	7.3683	13.1276	1.0000
SQRT	4.4236	3.4541	6.1540	0.7656
MAX	5.3801	4.2009	7.4845	0.9493
MIN	3.7962	2.9641	5.2810	-

Table 3.19: Outlier ratios of the spatial-temporal quality metrics

Metric	*OR*	*CI*	*Sig. Level*
PROD	0.5417	0.1993	0.9144
SQRT	0.2917	0.1818	0.0000
MAX	0.3750	0.1937	0.4567
MIN	0.2917	0.1818	-

Table 3.20: Performance of video quality metrics using different STCF models (Eqs.(3.42-3.45)) when both TR and SR are changed

Veri.Seq.	PROD		SQRT		MAX		MIN	
	PC	RMSE	PC	RMSE	PC	RMSE	PC	RMSE
PEDA	0.9758	6.8067	0.9307	6.3021	0.9021	8.0971	0.9543	4.8327
FOOT	0.7446	12.2296	0.9841	2.7401	0.9902	2.6633	0.9743	3.2484
RUSH	0.9091	8.4407	0.9713	3.3885	0.9646	3.7655	0.977	3.0538

3.5 Summary

In this chapter, a no-reference objective video quality metric MDVQM is presented, which considers the impact of both spatial and temporal quality impairments on the overall perceived video quality. The metric is based on the pixel bit-rate, frame rate, frame resolution as well as spatial and temporal video features (SA and TA values) that can be easily computed from the video sequences. Different from previous works, the situation in which frame rate and frame resolution change at the same time is also investigated. Verification with the data collected from our subjective tests shows that MDVQM provides accurate predictions for the perceptual video quality. The performance is significantly better than that of the comparison metrics.

Chapter 4

Improved ρ-domain Rate Control for H.264/AVC Video

In many video applications, compressed video streams are delivered under a certain rate restriction. Therefore rate control (RC) plays a very important role in order to meet the rate requirement as well as maintain a good picture quality.

H.264/AVC is a widely deployed international video coding standard. By utilizing many coding options such as variable block size, intra prediction, quarter-pel motion compensation, multiple reference frames, etc., the coding efficiency is significantly improved. Compared with previous video coding standards (MPEG4 or H.263), a bit rate reduction of 50% can be achieved [WSBL03].

ρ-domain rate control [HM01, HM02a] has been shown to be simple and effective for DCT-based hybrid video codecs. When it is applied to H.264/AVC, improvements need to be made because of the large amount of header information and the QP-dependent Rate Distortion Optimization (RDO). In this chapter, new rate models to estimate the size of header information in H.264/AVC coded video streams are proposed. A two-stage rate control algorithm is presented which combines the proposed header rate model and the ρ-domain source model. In comparison with previous header rate models and rate control algorithms, the proposed approach improves the PSNR of the decoded video, meets the target bit rates more accurately and results in smaller quality fluctuation inside one frame.

The remainder of this chapter is organized as follows. A review of related work is given in Section 4.1. Section 4.2 presents the proposed rate control algorithm based on the ρ-domain model. Experimental results are presented and discussed in Section 4.3. Section 4.4 gives a summary of this chapter.

4.1 Related Work

Although rate control is not a normative part of any video coding standards, it is an essential part of video codecs which are used in practical applications. The purpose of rate control is to maximize the video quality under certain resource constraints (such as file size, transmission rate or delay, etc.). According to whether or not the instantaneous bit-rate is allowed to vary significantly, rate control can be classified into Variable Bit-Rate (VBR) algorithms and Constant Bit-Rate (CBR) algorithms. VBR algorithms have the flexibility to allocate more resources to more complex scenes within a sequence, therefore a better video quality can be achieved. However, there are many scenarios where variation of the content complexity is not known or strict constraints are put on the instantaneous bit-rate. In these situations, CBR algorithms are used to ensure the constraints are met. VBR algorithms are often used in storage applications where the video is consumed locally and CBR algorithms are widely used in video streaming or video communication (telephony/conference) applications. In this work, considering the real-time video adaptation scenario, the focus is put on CBR situations.

Nowadays, H.264/AVC is the dominant video coding standard and many rate control algorithms have been proposed for it. Most of these algorithms are based on a certain functional relationship between the bit-rate and the encoding parameters (mainly the quantization parameters). The encoding process utilizes this function to adjust the encoding parameters in order to meet the target bit-rate. For example, in [MGWL03, MLW03], Li et al. propose a rate control algorithm employing a quadratic model, which has been adopted by the Joint Video Team for its reference implementation of H.264/AVC (JM codec) [Tea]:

$$\hat{D}_i = a \cdot D_{i-1} + b \tag{4.1}$$

$$R = \frac{c \cdot \hat{D}}{QS} + \frac{d \cdot \hat{D}}{QS^2} + h \tag{4.2}$$

where $\hat{D}_i$ is the predicted Mean Absolute Difference (MAD) of frame i. As shown in Eq.(4.1), $\hat{D}_i$ is estimated using a linear function of the actual MAD value of frame $i-1$ (D_{i-1}). QS denotes the quantization step-size and h is the size of header information. a,b,c and d are model parameters which need to be updated with the statistics of the encoded frames. Another widely used open source implementation of H.264/AVC is X264 [X26]. In [MV07], the rate control algorithm for X264 is introduced. Before encoding a frame, motion estimation is performed on a half-resolution version of the frame and the Sum of Absolute hadamard Transformed Difference (SATD) of the residual signal is calculated as a measure of frame complexity. The initial QP value is then determined by this SATD value empirically. During the encoding process, the QP values are updated for each macroblock (MB) according to the difference between the target frame size and the actual number of bits that have been generated. The above algorithms do not utilize the frame statistics of the current frame, so they often suffer from relatively large errors in terms of rate control accuracy.

In [HM01, HM02a], He et al. observed that for DCT-based video coding (H.263/MPEG4), the coding bit-rate has a linear relationship with the percentage of coefficients which are quantized to zero:

$$R(\rho) = \theta \cdot (1 - \rho) \tag{4.3}$$

where R denotes the coding bit-rate, ρ is the percentage of zero coefficients after quantization, and θ is a content-depend constant. To use this relationship in rate control, a one-to-one mapping between ρ and the quantization parameter is needed. This mapping can be derived from the specific quantization scheme used in the video codecs. Taking the H.263 video coding [ITU05b] as an example, the quantized coefficients are calculated as:

$$L = \begin{cases} Round(\frac{COF}{8}) & : \quad \text{if } COF \text{ is a DC coefficient in an intra-MB} \\ UTSQ(2q, 2q; COF) & : \quad \text{if } COF \text{ is a AC coefficient in an intra-MB} \\ UTSQ(2q, 2.5q; COF) & : \quad \text{if } COF \text{ is a coefficient in an inter-MB} \end{cases} \tag{4.4}$$

where COF denotes the unquantized transform coefficients and q is the quantization parameter. UTSQ denotes the Uniform Threshold Scalar Quantization:

$$UTSQ[q, \delta; c] = \begin{cases} 0, & \text{if } |c| \leq \delta \\ \left\lceil \frac{c - \delta}{q} \right\rceil, & \text{if } c > +\delta \\ \left\lfloor \frac{c + \delta}{q} \right\rfloor & \text{if } c < -\delta \end{cases} \tag{4.5}$$

where δ is the dead-zone threshold. Then the relationship between ρ and q can be derived as:

$$\rho(q) = \frac{1}{L} \sum_{|c|<2q} H_I(c) + \frac{1}{L} \sum_{|c|<2.5q} H_P(c) \tag{4.6}$$

where $H_I(\cdot)$ and $H_P(\cdot)$ are the histograms of the unquantized DCT coefficients for intra-coded and inter-coded MBs respectively. L is the number of coefficients in the current video frame.

The ρ-domain rate control algorithm proposed in [HM02a] utilizes Eq.(4.3) and Eq.(4.6) as a rate model for the rate control of H.263 and MPEG-4 video codecs. Compared with other algorithms, the ρ-domain rate model is very simple and can provide more accurate control of the coding bit-rate. However, when it is used for H.264/AVC video coding, several issues need to be resolved first.

The first issue is the inter-dependency between RDO and rate control. In H.264/AVC, up to 7 block sizes are supported for motion estimation. Small blocks improve the accuracy of motion estimation and reduce the energy of the residual signal, but leads to more motion information (reference frame ID, motion vectors (MVs)). Therefore, a trade-off needs to be found. This is typically done by a rate-distortion optimized way:

$$C = D + \lambda \cdot R_{mot} \tag{4.7}$$

where C means the cost of encoding the MB, D denotes the distortion (normally calculated as the Sum of Absolute Difference (SAD)), R_{mot} is the estimated size of the encoded motion information and λ is a Lagrange multiplier which depends on the choice of QP values [SW98]. Hence, the QP value is required by the rate-distortion optimized motion estimation. But on the other hand, ρ-domain rate control determines the QP value based on the statistics of the transformed coefficients, which requires to perform motion estimation before the selection of QP. This contradiction leads to the so-called "chicken-egg-dilemma".

The second issue comes from the increased amount of header information in H.264/AVC. By utilizing various coding options in H.264/AVC, the energy of the intra/inter prediction errors is significantly reduced. But at the same time, more bits are spent to signal these coding options (such as block size for an inter MB and intra prediction mode for an intra MB). At high bit-rates, the impact is not very serious since the texture information dominates the bitstream. But at low bit-rates, the header information occupies a large portion of the total bit-rate, which causes the accuracy of ρ-domain rate control to be reduced.

In [HW08], He and Wu propose to use the average QP value of the previous frame for the estimation of λ in Eq.(4.7) to break the inter-dependency between rate control and RDO, so that ρ-domain rate control can be used for H.264/AVC. The authors assume that the size of header information is also proportional to ρ, so that the ρ-domain rate control originally proposed for H.263 can also be applied to H.264/AVC. However, as will be discussed in Section 4.2, this assumption is not always true, especially for low bit-rate cases. In [KSK07], Kwon et al. propose a method to estimate the size of motion information in H.264/AVC. The method is combined with the quadratic rate model in [MGWL03, MLW03] and the experimental results show that it performs better than the rate control method in JM8.1 [Tea] which uses the same source rate model.

In the following, a method to estimate the size of header information in H.264/AVC is proposed. The proposed method is used together with the ρ-domain rate model to improve the accuracy of rate control for H.264/AVC codecs. A two-stage encoding structure is also employed to decouple rate control and RDO.

4.2 Proposed Rate Control Algorithm

In [HW08], the ρ-domain rate control algorithm is adapted for the H.264/AVC encoder. The authors claim that for H.264/AVC video coding, the total bit-rate consumed by a frame follows a similar linear relation with ρ. However, as has been mentioned, H.264/AVC introduces several advanced prediction schemes which can reduce the prediction error but the size of overhead information is also increased. Typically this header information overhead changes from frame to frame and is not addressed in the ρ-domain rate model. Figure 4.1 shows the relationship between the percentage of non-zero coefficients (NNZs) and the size of a frame. To run the experiment, the X264 encoder [X26] is used to encode the sequence FOREMAN and MOTHER&DAUGHTER (CIF@25fps) with CAVLC (Context Adaptive Variable Length Coding). The sequences are encoded using different QPs from 25 to 45. The results for high bit-rates and low bit-rates are shown separately. For high bit-rates, the used QP values range from 25 to 33. For low bit-rates, the QP values are from 34 to 45.

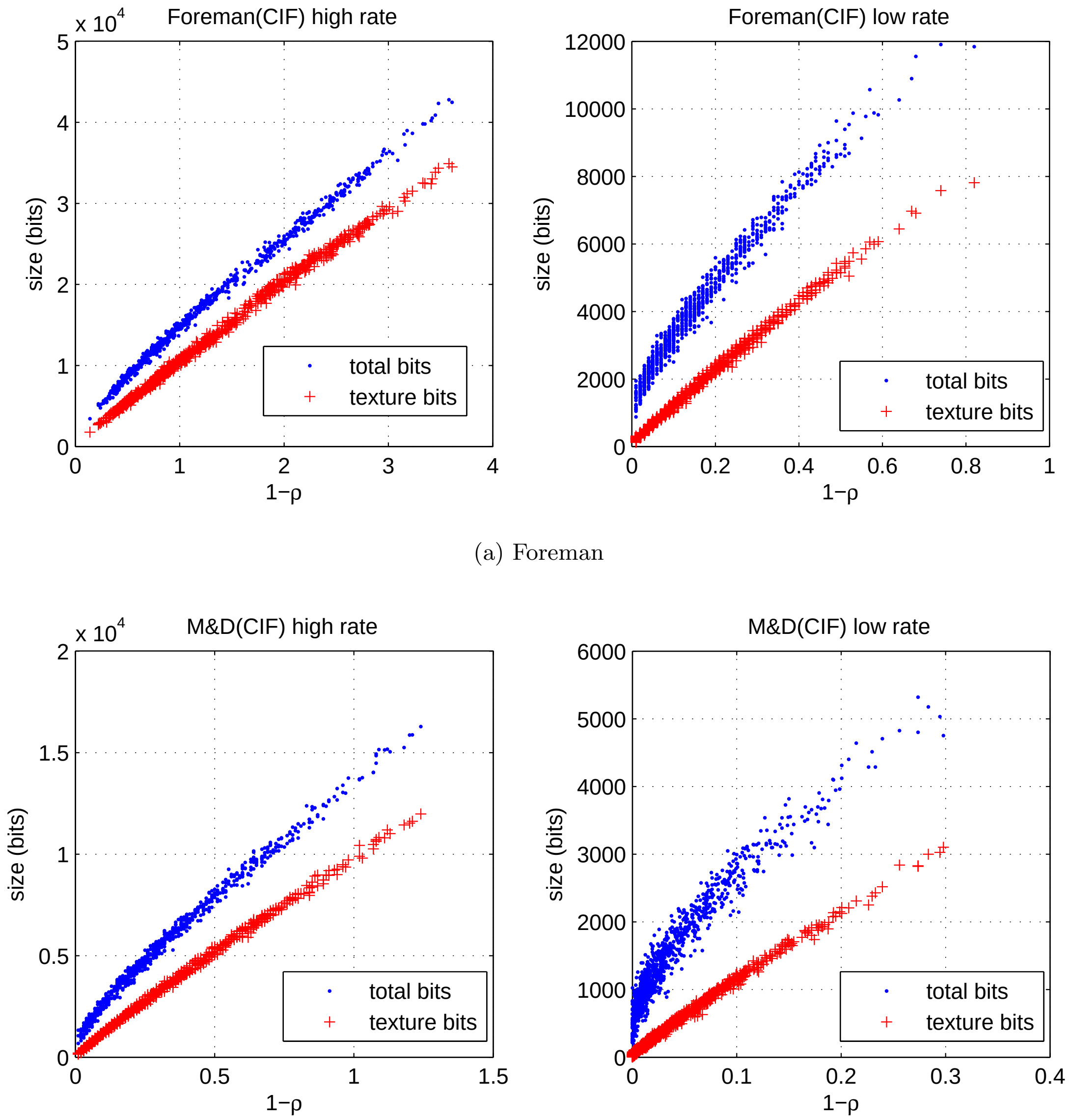
The X-axis shows (1-ρ), which is the percentage of non-zero coefficients. The Y-axis shows the number of consumed bits. The red crosses show the size of a frame and the blue dots show the size of the texture information (residual information). It can be seen that although the size of the texture information is strictly proportional to (1-ρ), the total size of a frame does not follow such a rule, especially at low bit-rates. The difference between the red and blue points is simply the size of header information in each frame. To make ρ-domain rate control more accurate for H.264/AVC, a precise estimation of the size of header information is very important. The number of header bits changes significantly for different frames and it is hard to derive a closed-form mathematic model to relate the number of header bits with the parameter ρ. In the following, selected observations from the experiments are presented and the size of the header information is estimated in an adaptive manner.

4.2.1 Header Information in H.264/AVC

Header information in H.264/AVC includes the NAL (Network Abstraction Layer) header, the sequence header (PPS and SPS), the slice header and the MB header. The NAL header is rather small (1 byte per NAL unit). PPS and SPS are not sent very often. A single slice header is also very small but when a frame is divided and encoded into multiple slices, it might also occupy a certain percentage of the total frame size. Compared with the MB header, the size of a slice header is stable and easy to estimate. Normally, a slice contains either a constant number of MBs or a constant number of bits. The size of the slice header can be estimated as:

$$R_{SH} = \begin{cases} \frac{N_{MB}}{N_{MBpS}} \cdot b_{SH} & \text{(fixed MB number)} \\ \\ \frac{R_T}{N_{BpS}} \cdot b_{SH} & \text{(fixed slice size)} \end{cases} \tag{4.8}$$

where R_{SH} is the estimated size of slice headers of the current frame, N_{MB} is the total number of MBs in the frame, N_{MBpS} is the number of MBs per slice, R_T is the total number of bits



(a) Foreman

(b) Mother & Daughter

Figure 4.1: Relationship between the percentage of non-zero coefficients (1-ρ) and the number of generated bits (total bits and texture bits) for test sequences. (a) Foreman. (b) Mother&Daughter(M&D)

allocated to the frame, N_{BpS} is the number of bits allocated to each slice, and b_{SH} is the average slice header size in the previous frames.

4.2.2 Rate Model for Inter MB Headers

The MB header is the most important header information. It contains the encoding parameters for inter MBs (such as the MVs, the reference frame IDs, etc.) and for intra MBs (such as intra prediction type, etc.). Since the header of different MB types contains different information, the numbers of header bits for inter and intra MBs need to be estimated separately.

In [KSK07], a linear rate model for the size of the header information in inter MBs is proposed. The authors claim that there is a strong relationship between the header size of inter MBs and the number of non-zero horizontal/vertical MVs. The authors also consider that the size of the coded block pattern (CBP bits) has a strong relationship with the number of non-zero coefficients. Specifically, the size of the header information in inter MBs is modelled as:

$$R_{hdr,p} = \gamma \cdot (N_{nzMVe} + \omega \cdot N_{MV}) \tag{4.9}$$

where ω is fixed to 0.3 for single frame motion estimation, N_{nzMVe} is the number of non-zero motion vector elements, N_{MV} is the total number of MVs, and γ is a parameter to estimate. The experiments in [KSK07] show that this model works well for many sequences. But our experiments show that this model does not always predict the number of header bits very accurately.

In our experiments, the video sequences (250 frames long) are encoded at different bitrates using the original rate control algorithm of the X264 encoder. Figure 4.2 shows the result for the CIF sequence FOREMAN encoded at 512kbps and Mother&Daughter (M&D) encoded at 384kbps. The x-axis gives the value of $(N_{nzMVe} + \omega \cdot N_{MV})$ and the y-axis is the size of the information. The blue points correspond to the total header size and the red crosses correspond to the size of motion information. It can be seen that for both the total header size and the motion information size, the predictions provided by the model in Eq.(4.9) do not correlate well with the actual value.

This estimation error results from the fact that the H.264/AVC encoder performs CAVLC not directly on MVs but on the differential MVs (MVDs), which are the differences between the actual MV and a predicted MV. So the size of the motion information should have a stronger linear relationship with the statistics of the MVDs. Let N_{nzMVDe} and N_{zMVDe} denote the number of non-zero and zero MVD elements respectively (e.g. if two MVDs (-1,0) and (0,0) are considered, then $N_{nzMVDe} = 1$, $N_{zMVDe} = 3$), a new rate model is proposed for motion information which is similar to Eq.(4.9) but based on the statistics of the MVDs:

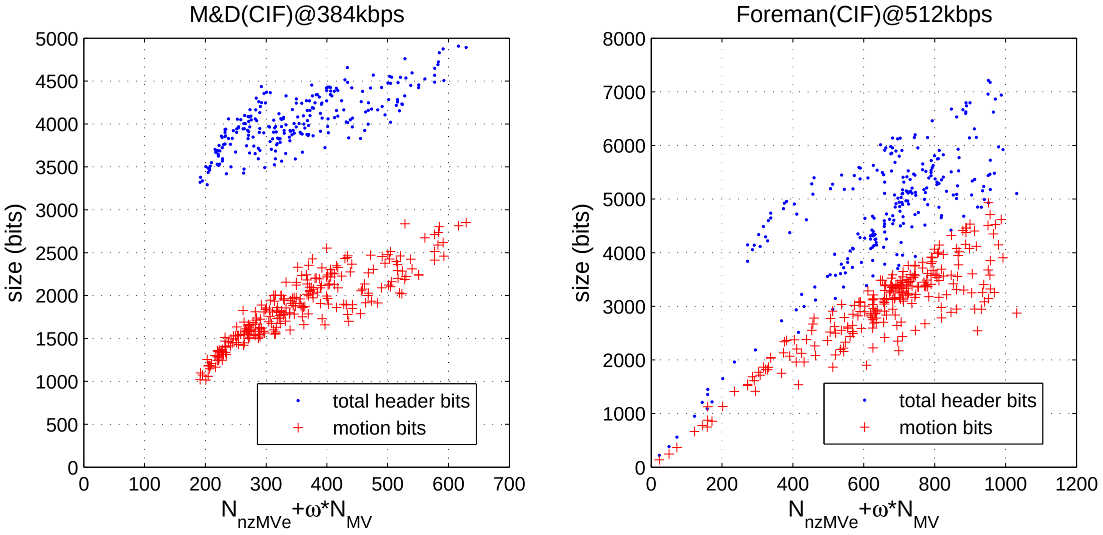


Figure 4.2: Performance of the MV-based rate model in Eq.(4.9)

$$R_{mot,p} = \gamma_{mot} \cdot (N_{nzMVDe} + \omega_{mot} \cdot N_{zMVDe}) \tag{4.10}$$

where ω_{mot} is a fixed weighting factor and γ_{mot} is a parameter to be estimated. It is observed from our experiments that ω_{mot}=0.2 works well for all the sequences when only one reference frame is used.

In Figure 4.3, the relationship between $(N_{nzMVDe} + \omega_{mot} \cdot N_{zMVDe})$ and the total header size (blue points) as well as the size of the motion information (red crosses) are presented. From the results, it can be seen that the size of the motion information can be very well predicted using the proposed model. The estimation errors of the motion information size for different sequences are presented in Table 4.1 using the R^2 value [DF99]. The R^2 value between the actual values $\vec{y}$ and the model predictions $\vec{x}$ is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{4.11}$$

where $\overline{y}$ is the mean value of the actual data in $\vec{y}$ and n is the size of the data set. The R^2 value is used to measure the deviation between the predictions of a model and the actual data values. It takes the value from 0 to 1 and the better the prediction, the closer the R^2 value to 1. From the results in Table 4.1, it can be seen that the proposed rate model for the motion information in Eq.(4.10) provides better predictions than the model in Eq.(4.9).

Another observation from the result is that, although size of the motion information can be predicted quite accurately using the proposed model, the linear relationship between

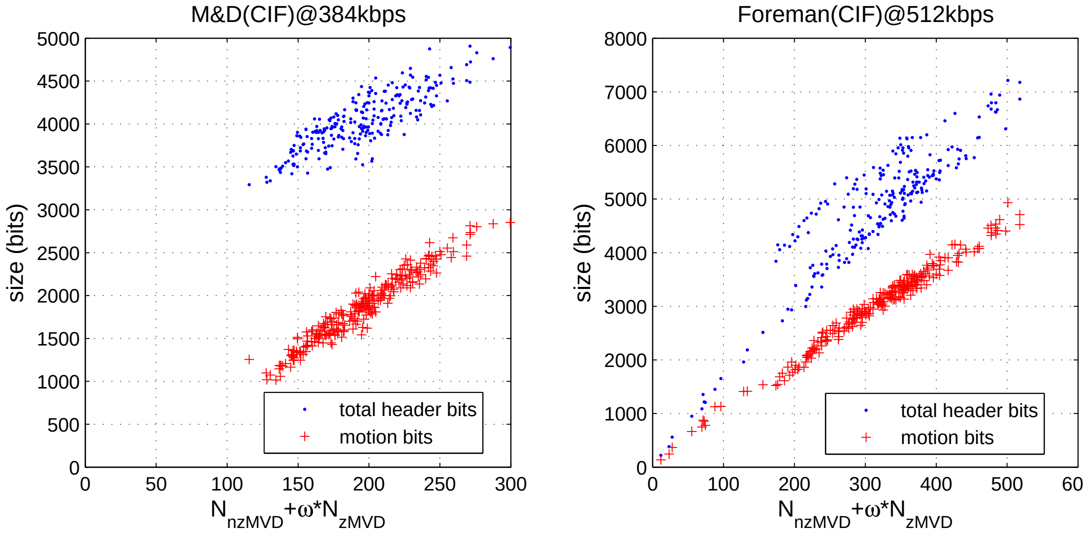


Figure 4.3: Performance of MVD-based rate model in equation (4.10)

Table 4.1: Performance comparison of the two rate models for motion bits in Eq.(4.9) and Eq.(4.10)

Seq.	Bit-rate (kbps)	R^2 Value	
		Model in Eq.(4.9)	Model in Eq.(4.10)
M&D (CIF)	384	0.5742	0.9219
Foreman (CIF)	512	0.7373	0.9767
Football (CIF)	640	0.7119	0.9730
Carphone (CIF)	384	0.7800	0.9772

$(N_{nzMVDe} + \omega_{mot} \cdot N_{zMVDe})$ and total header size is much weaker. This estimation error results from other header information in inter MBs.

As specified in H.264/AVC, an inter MB contains the following header information:

- Motion Information (MVDs and Refs)
- MB Type information (16x16, 16x8, 8x16, 8x8.)
- Coded Block Pattern (CBP)
- QP value for the MB

To give an example, the average percentage of different types of header information in inter MBs for the sequence Foreman (CIFx30fps, encoded at 512kbps) is shown in Figure 4.4. From Figure 4.4, it can be seen that the size of the QP information is very limited compared

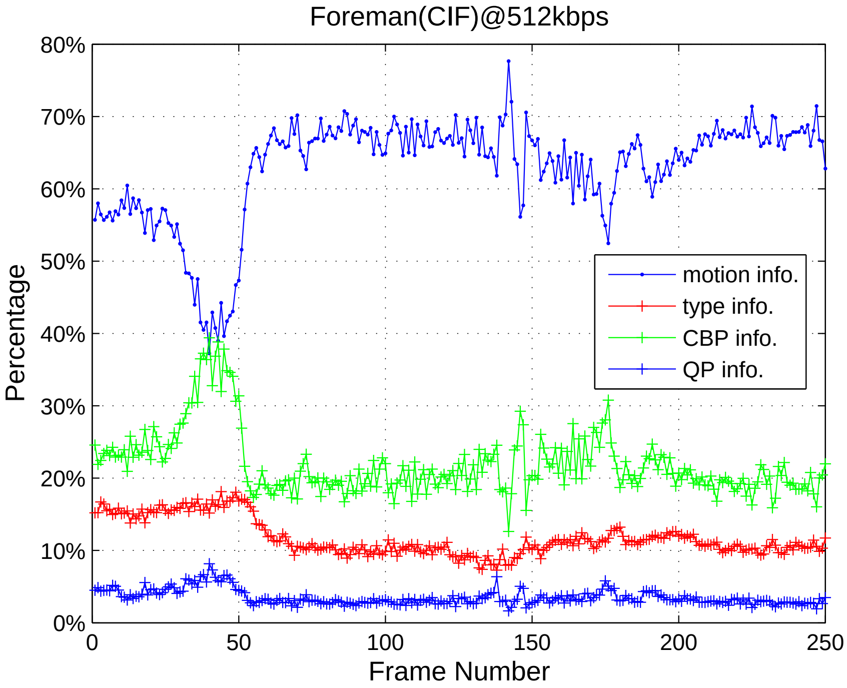


Figure 4.4: distribution of bits in the MB header

to other header information. So the number of bits consumed by QP information is simply estimated using the average size of QP information per MB in the previous frames (b_{QP}):

$$R_{qp,p} = N_p \cdot b_{QP} \tag{4.12}$$

where, N_p is the number of inter MBs in the current frame.

The percentage of MB type information is a little higher but since the number of MB types for an inter MB is limited (4 for MB Type and 4 for sub-MB Type) and code sizes for each MB type or sub-MB type are fixed, once the number of MBs of each type or subtype is known, the size for this header information can be estimated by:

$$\begin{aligned} R_{type,p} &= \sum_{i\in A} N_{p,i} \cdot b_{type,i} + \sum_{j\in B} N_{p8,j} \cdot b_{subtype,j} \\ A &= \{P16\times16, P16\times8, P8\times16, P8\times8\} \\ B &= \{D8\times8, D8\times4, D4\times8, D4\times4\} \end{aligned} \tag{4.13}$$

where $N_{p,i}$ is the number of inter MBs of type i (as given in set A), $N_{p8,j}$ is the number of 8x8 blocks of sub-block type j (as given in set B), $b_{type,i}$ and $b_{subtype,j}$ are the number of bits used for encoding the corresponding MB type and sub-MB type information for inter MBs and inter-8x8 blocks, respectively. $b_{type,i}$ and $b_{subtype,j}$ are fixed values specified in the standard. For example, one bit is used for encoding the $P16\times16$ mode and three bits are used for encoding the $P16\times8$ mode and the $P8\times16$ mode.

The percentage of the CBP information in different frames fluctuates heavily and for some frames it occupies a significant portion of the header size. As has been mentioned, in [KSK07], the authors consider that the size of the CBP information has a strong relationship with the size of the texture (residual) information. Based on ρ-domain theory, the size of texture information has a strong linear relationship with the percentage of non-zero coefficients (1-ρ). So this assumption can be verified by observing the relationship between (1-ρ) and the size of the CBP information. Figure 4.5 shows the experimental results for the CIF sequence FOREMAN encoded using different QP values. The X-axis shows the percentage of non-zero coefficients in a frame (1-ρ) and the Y-axis shows the size of the CBP information. It can be observed that the linear relationship is not as strong as expected. To predict the number of CBP bits based on (1-ρ) can introduce large estimation error.

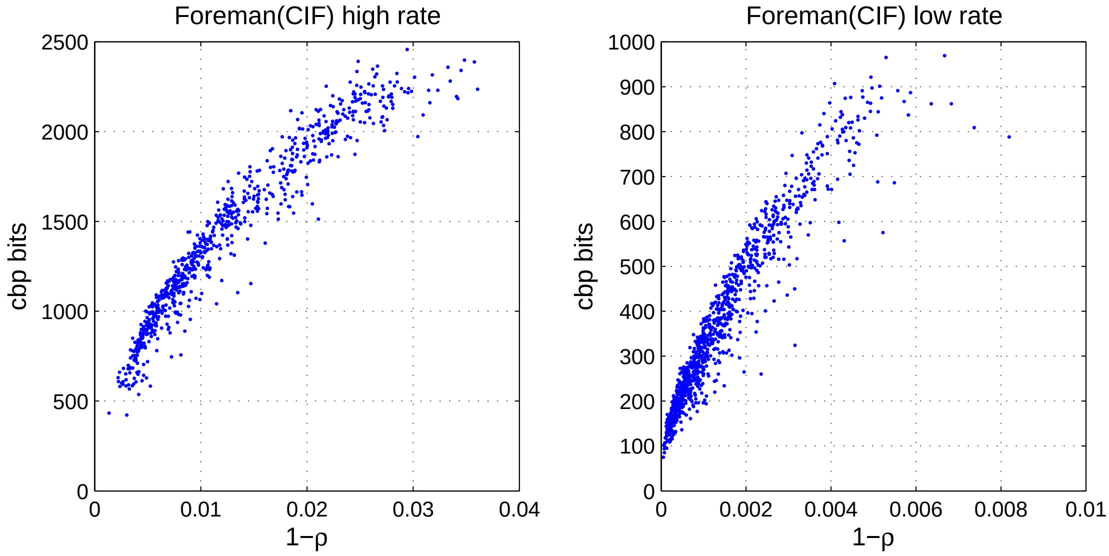


Figure 4.5: Relationship between the percentage of non-zero coefficients (1-ρ) and the size of the CBP information for the sequence Foreman (left: high bit-rate range, right: low bit-rate range)

A possible explanation for this is that the number of CBP bits depends not only on the number of non-zero coefficients but also on the distribution of these coefficients. An extreme example is that, assuming that there are six non-zero coefficients in a MB, the number of CBP bits would be different when these coefficients are evenly distributed within all the six 8x8 blocks (four luma blocks and two chroma blocks) in a MB from that when these coefficients all belong to one 8x8 block. Hence, to accurately estimate the size of the CBP information, the distribution of non-zero coefficients should also be an important factor.

Let's define the zero MBs as the MBs in which all the quantized coefficients are zeros. Then inspired from the linear rate model for the motion information in [KSK07], the number of zero and non-zero MBs can be used as an indicator of the distribution of non-zero coefficients. Let N_{nzMB} and N_{zMB} be the number of non-zero and zero MBs respectively, a linear rate model for the CBP information is proposed as:

$$R_{cbp,p} = \gamma_{cbp} \cdot (N_{nzMB} + \omega_{cbp} \cdot N_{zMB}) \tag{4.14}$$

where ω_{cbp} and γ_{cbp} have a function similar to that in equation (4.10).

Figure 4.6 shows the relationship between $(N_{nzMB} + \omega_{cbp} \cdot N_{zMB})$ (X-axis) and the size of the CBP information (Y-axis). ω_{cbp} is set empirically to 0.1 based on the experiments.

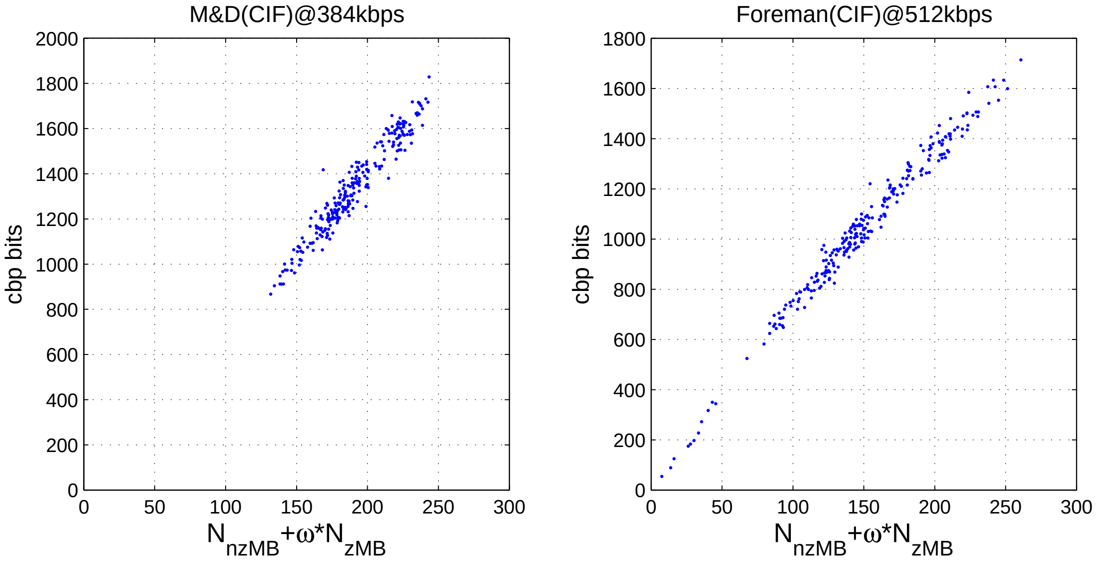


Figure 4.6: Experimental results for Eq.(4.14)

From Figure 4.6, it can be seen that there is a very strong linear relationship as suggested in Eq.(4.14). The estimation errors for different test sequences measured by the R^2 values are shown in Table 4.2. For all the sequences, the R^2 values are very close to 1, suggesting that our linear rate model for the CBP information works well.

Table 4.2: Performance of the proposed rate model for CBP bits

Seq.	Bit-rate (kbps)	R^2 Value of Eq.(4.14)
M&D (CIF)	384	0.9347
Foreman (CIF)	512	0.9725
Football (CIF)	640	0.9749
Carphone (CIF)	384	0.9740

The header bits for an inter MB can be modelled as:

$$R_{hdr,p} = R_{mot,p} + R_{qp,p} + R_{type,p} + R_{cbp,p} \tag{4.15}$$

where $R_{mot,p}$, $R_{qp,p}$, $R_{type,p}$ and $R_{cbp,p}$ are estimated using Eqs.(4.10)(4.12-4.14), respectively.

4.2.3 Rate Model for Intra MB Headers

As there are very few intra MBs in a frame, especially at low bit-rates, and the size of the header information of intra MBs is limited compared to the size of the texture information in the same MB, the size of header information of intra MBs in a frame is estimated by:

$$R_{hdr,I} = N_{i16\times16}\cdot b_{i16\times16} + N_{i4\times4}\cdot b_{i4\times4} \tag{4.16}$$

where $N_{i16\times16}$ and $N_{i4\times4}$ are the number of intra-16x16 MBs and intra-4x4 MBs, respectively. $b_{i16\times16}$ and $b_{i4\times4}$ are the average size of the header information of intra-16x16 and intra-4x4 MBs in the previous frame.

4.2.4 Two-Stage Rate Control Algorithm

In this section, a two-stage rate control algorithm is proposed. In the first stage, motion estimation and mode decision are performed to collect necessary statistics of the MBs in the current frame. In the second stage, the proposed rate models for header information in Sections 4.2.2 and 4.2.3 are combined with ρ-domain rate control theory to accurately control the size of the current frame.

1. Frame Level Bit allocation
   Sophisticated frame level bit allocation algorithms can be used here. But since the main purpose is to verify the accuracy of the proposed rate models for header information, a simple frame level allocation method is used. The target size for a frame is determined by:
   $$R_T = \frac{r}{F}$$
   where r is the target bit-rate of the video stream in the unit of bits/s, and F is the frame rate in the unit of fps (frames/s).
2. Stage One: Analysis Stage
   In this stage, motion estimation and mode decision are conducted for all the MBs in the current frame using the average QP value of the previous frame in the RDO process. Then, the prediction residuals are transformed into the DCT domain for ρ-domain analysis. After the analysis, the model parameters N_{nzMVD} and N_{zMVD} are counted.
3. Stage Two: Actual Encoding Stage

a) Before encoding each MB, the size of the header information except the CBP information for inter MBs is estimated for the remaining MBs as discussed in Section 4.2.2 and 4.2.3. The reason why the CBP information is excluded here is that the number of zero MBs depends on the selected QP, so it is estimated later when candidate QPs are examined according to the ρ-domain rate control method.

b) The estimated header size is subtracted from the remaining available bit budget for the current frame R_T to determine the available bits R_{avail}. Then all the possible QP values are examined. The size of the texture information R_{tex} is estimated using the original ρ-domain rate control model and the size of the CBP information R_{cbp} is estimated by Eq.(4.14). The smallest QP which results in $(R_{cbp} + R_{tex}) \leq R_{avail}$ for the current MB is selected.

c) After encoding each MB, the bit budget R_T is updated by substracting the actual number of bits used for encoding the current MB. Also, the parameters in the header rate model $(\gamma_{cbp}, \gamma_{mot}, b_{QP}, b_{i4}, b_{i16})$ and ρ-domain rate model (θ) are also updated accordingly.

d) The above procedure is repeated for all MBs in the current frame.

4. After encoding the current frame, the model parameters for the current frame are saved to be used for the first MB in the next frame. At the beginning of the first frame, default values ($\gamma_{cbp} = 4$, $\gamma_{mot} = 10.3$, $\theta = 5.4$) are used to initialize the model.

4.3 Experimental Results

The proposed rate control algorithm is implemented in X264, which is an open source implementation of H.264/AVC. The encoder is configured to conform to the baseline profile. CAVLC is used for entropy coding. Extensive simulations are conducted using various standard test sequences. For each sequence, 250 frames are encoded. The first frame is encoded as an I frame and the following frames are encoded all as P frames. For fair comparison, the QP value for the I frame is determined in the same manner as in the original X264 rate control (CBR mode). As mentioned above, a simple frame level bit allocation which depends on the target bit-rate and the frame rate is employed in order to examine the accuracy of the header size prediction and rate control.

The proposed algorithm is compared with three other rate control algorithms:

- **X264**: The original CBR mode rate control algorithm in X264
- **ORIG**: The original ρ-domain rate control algorithm without estimation of header information size

- **MVHE**: ρ-domain based rate control algorithm, the rate model in [KSK07] is used to estimate the size of header information

Several performance metrics are used for the evaluation: video quality in PSNR, accuracy of the rate control, QP fluctuation within one frame.

4.3.1 Video Quality in PSNR

Table 4.3: Performance comparison of the rate control algorithms

Seq.	Target Bit-rate (kbps)	X264		ORIG		MVHE		Proposed	
		BR (kbps)	PSNR (dB)	BR (kbps)	PSNR (dB)	BR (kbps)	PSNR (dB)	BR (kbps)	PSNR (dB)
M&D (QCIF)	48	47.42	36.51	47.09	36.45 (−0.06)	47.95	36.58 (+0.07)	47.92	36.59 (+0.08)
	96	95.43	39.89	95.04	39.94 (+0.05)	95.62	40.06 (+0.18)	95.60	40.07 (+0.19)
	128	127.48	41.33	126.84	41.36 (+0.03)	127.42	41.51 (+0.18)	127.48	41.53 (+0.20)
Foreman (QCIF)	96	94.95	33.43	90.80	33.42 (−0.02)	94.54	33.59 (+0.16)	95.01	33.72 (+0.29)
	128	126.50	34.81	125.88	34.74 (−0.07)	126.20	34.94 (+0.13)	126.40	35.03 (+0.22)
	192	189.81	36.65	188.74	36.64 (−0.01)	189.16	36.76 (+0.11)	189.58	36.86 (+0.21)
M&D (CIF)	128	127.38	37.23	127.38	37.39 (+0.15)	128.08	37.30 (+0.07)	127.90	37.42 (+0.18)
	192	191.17	39.18	191.08	39.30 (+0.13)	191.48	39.35 (+0.17)	191.56	39.40 (+0.22)
	256	254.87	40.43	254.67	40.63 (+0.20)	255.04	40.65 (+0.22)	255.23	40.70 (+0.27)

Table 4.3 shows a subset of the test results in terms of PSNR. The bit-rate of the encoded streams and the PSNR gain of the three ρ-domain rate control algorithms over the X264 rate control are also presented. It can be seen that for the original ρ-domain rate control, the PSNR gain is not always positive. The reason is that the X264 rate control uses a buffer to smooth the bit-rate changes for different frames, so although the overall average bit-rate of the sequence is very close to the target value, the fluctuation of frame sizes within the sequence is very large. This will be discussed soon below. For MVHE and the proposed algorithm, a positive PSNR gain can always be achieved. And the gain of the proposed algorithm is larger than that of MVHE.

4.3.2 Accuracy of the Rate Control

According to Table 4.3, all the algorithms can control the average bit rate quite accurately with a control error smaller than 2%. But this is only one side of the story. As has been mentioned, the size of every single frame should be controlled accurately. Figure 4.7 shows the size of the first 100 P-frames for the QCIF sequence FOREMAN encoded at 192kbps, 30fps (i.e. each frame should be encoded with 400 bytes).

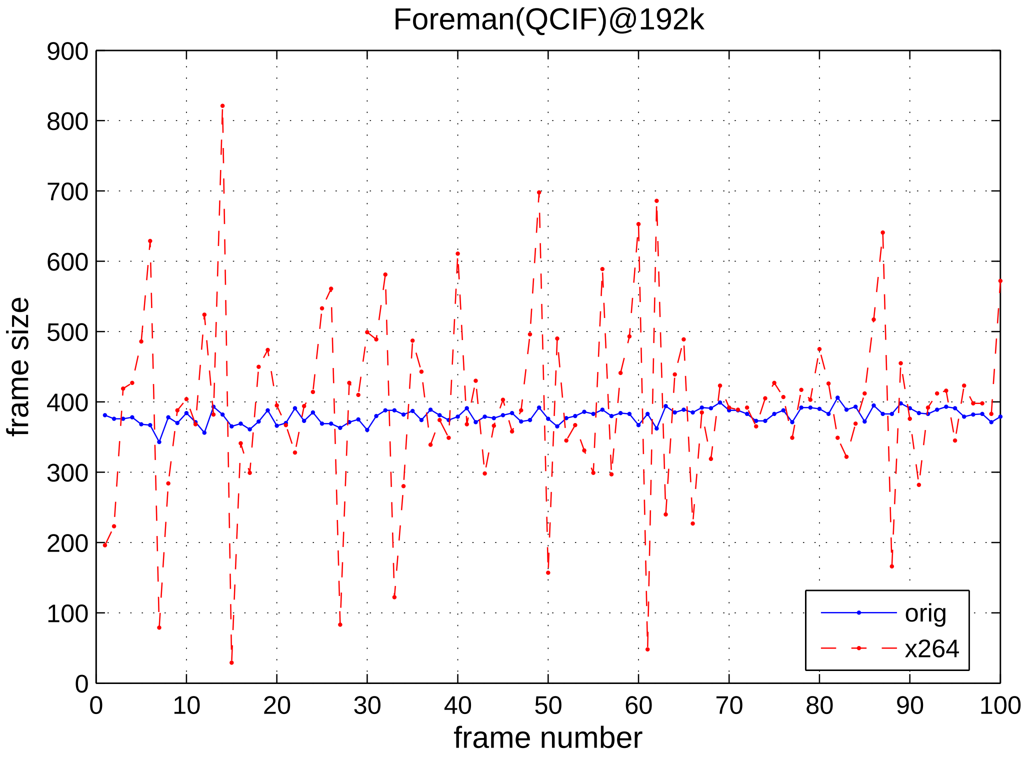
Figure 4.7a shows the comparison between the original ρ-domain rate control and the X264 rate control. It can be seen that the frame size fluctuation of X264 is much larger than the original ρ-domain rate control, which demonstrates the advantage of ρ-domain rate control. Further, Figure 4.7b shows the size of the encoded frames for the three ρ-domain algorithms. It can be seen that the size fluctuation of MVHE and the proposed algorithm is smaller than that of the original ρ-domain rate control. This improvement is due to the accurate estimation of the header size. Table 4.4 presents the average deviation of the actual frame size from the target frame size. It can be seen that for all the test sequences, the proposed algorithm gives the smallest deviation, which means it can control the frame size most accurately.

Table 4.4: Average deviation of the actual frame size from the target frame size

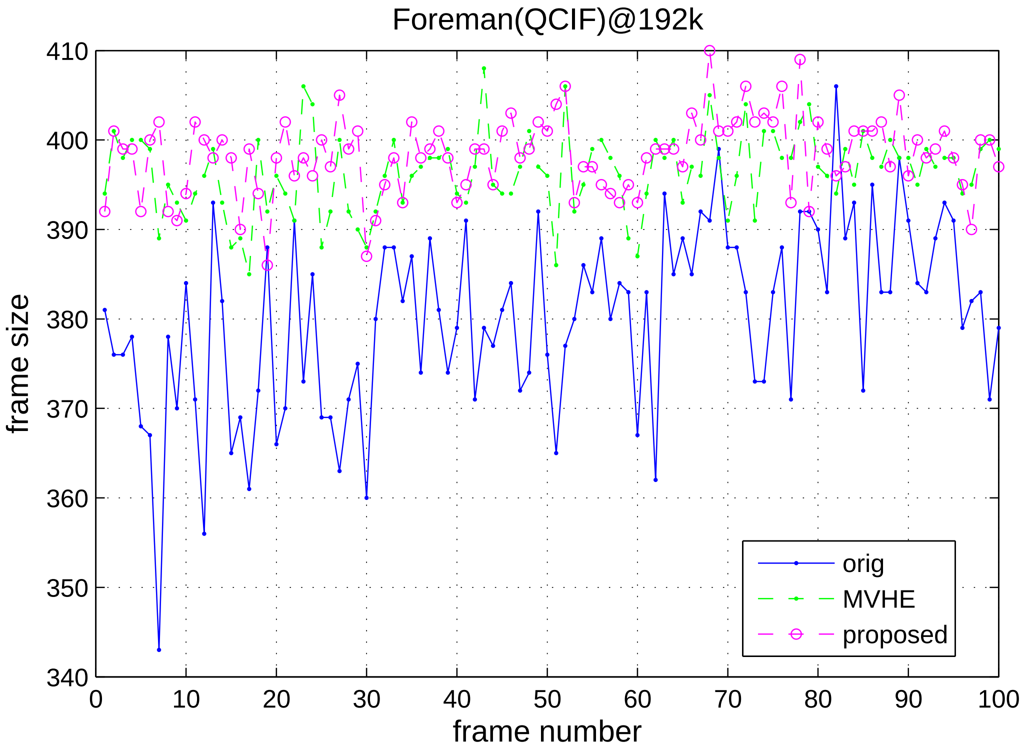
Seq.	Target Bit-rate	Target Frame Size	ORIG		MVHE		Proposed	
	(kbps)	(Byte)	Dev. (Byte)	Percent (%)	Dev. (Byte)	Percent (%)	Dev. (Byte)	Percent (%)
Foreman (QCIF)	96	400	26.94	6.74	9.75	2.44	7.96	1.99
M&D (CIF)	192	800	10.82	1.35	10.32	1.29	5.83	0.73
Football (CIF)	384	1600	11.44	0.72	5.70	0.36	5.54	0.35
Foreman (CIF)	192	800	24.74	3.09	7.25	0.91	5.73	0.72

4.3.3 QP Fluctuation among MBs within a Frame

All the MB level rate control algorithms allow the QP to be adjusted for each MB to meet the target frame size accurately. On the other hand, if the QP changes too frequently, more bits need to be spent to signal the QP changes between successive MBs in the bitstream. Also, higher QP variation causes more significant quality fluctuations within a frame, which might be annoying sometimes. Figure 4.8 shows an example of QP variation within one frame for the three ρ-domain rate control algorithms. It can be seen that the proposed algorithm (green



(a) Comparison between ORIG and X264



(b) Comparison between ORIG, MVHE and the proposed method

Figure 4.7: Comparison of the frame size fluctuation of different rate control methods

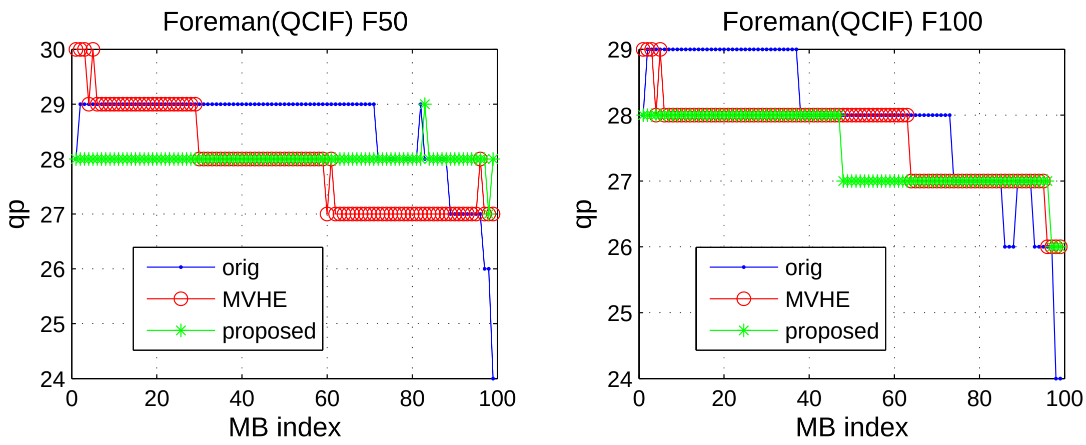


Figure 4.8: Comparison of QP fluctuation within one frame

line with star points) results in much smaller QP variation within the frame compared with the original ρ-domain rate control (blue line with diamond points) and MVHE (red line with round points). Table 4.5 shows the average variance of the QP values within a frame. The average maximum difference between the QP values within a frame (maximum difference is the difference between the largest QP and smallest QP within a frame) is also given. The results show that for most of the cases, the proposed algorithm results in a smaller variance and maximum difference, which indicates again a smaller fluctuation of QP values within a frame.

The smaller variation allows us to spend more bits on the residual signal and improve the picture quality. It also proves that the proposed algorithm predicts the total size of header and residual information more accurately than the other two algorithms, so that an appropriate QP is selected from the beginning and does not need to be changed for the last MBs dramatically to meet the target frame size.

In summary, compared with the other rate control algorithms, our proposed algorithm gives the best video quality, the smallest frame size control error and the smallest QP variation within a frame. This also proves the effectiveness of the proposed header estimation method.

4.4 Summary

In this chapter, an efficient rate control algorithm for H.264/AVC with accurate header information estimation is proposed. The approach uses a two-stage encoder structure to resolve the inter-dependency between RDO and ρ-domain rate control. The header information is estimated using an improved rate model which considers different components in a MB header (type information, motion information, CBP information, etc.). Experimental results show the proposed algorithm can achieve better rate control accuracy and video quality compared

Table 4.5: Average variance and maximum difference of QP values within a frame

Seq.	Target Bit-rate (kbps)	ORIG		MVHE		Proposed	
		Var.	Max. Dif.	Var.	Max. Dif.	Var.	Max. Dif.
M&D (QCIF)	48	1.09	4.31	0.82	3.32	0.45	2.50
	96	1.27	4.37	0.44	2.29	0.31	1.95
	128	1.15	4.15	0.32	2.07	0.33	1.89
Foreman (QCIF)	96	0.52	3.35	0.69	3.37	0.16	1.33
	128	0.48	3.52	0.51	2.83	0.19	1.31
	192	0.48	3.73	0.38	2.51	0.17	1.49
M&D (CIF)	128	1.94	4.56	0.38	2.51	0.37	2.17
	192	1.71	4.01	0.28	2.18	0.33	2.01
	256	1.13	3.68	0.28	2.13	0.31	2.03
Football (CIF)	640	0.93	3.89	0.51	2.64	0.50	2.56
	800	0.83	3.52	0.48	2.66	0.40	2.43
	1000	0.66	3.28	0.39	2.53	0.32	1.97

with other rate control algorithms. The proposed rate control algorithm is used in the QoE-driven MDA scheme presented in Chapter 5 to achieve accurate rate adaptation.

Chapter 5

QoE-driven Multi-Dimensional Video Adaptation

In Chapter 3 and Chapter 4, the two most important components, i.e. perceptual quality estimation and rate control, of the video adaptation system shown in Figure 1.2 are studied respectively. Based on these studies, a QoE-driven MDA scheme is proposed in this chapter to determine the optimal combination of different adaptation operations and optimize the perceptual video quality. The QoE is estimated based on the objective quality metric MDVQM presented in Chapter 3, which has been shown to provide a good estimation of perceptual quality for videos in the presence of both spatial and temporal impairments. The presented QoE-driven video adaptation scheme automatically examines the impact of different adaptation strategies and makes the best decision for video adaptation. The ρ-domain rate control algorithm presented in Chapter 4 is also integrated into the system for accurate rate adaptation.

5.1 System Overview

As mentioned in Chapter 3, three major parameters can be adjusted to perform video adaptation: the quantization parameter, the temporal resolution of the video (frame rate) and the spatial resolution of the video (frame size). They affect the SNR, temporal and spatial quality of the video, respectively. The contribution of the three quality measures to the overall perceived video quality depends heavily on the characteristics of the video content. The aim of a QoE-driven video adaptation scheme is to find a compromise among different quality measures to maximize the perceived video quality (QoE) for constrained system resources. Therefore, in the proposed video adaptation scheme, three operating modes are considered:

- SNR Mode: The video adaptation algorithm does not change the spatial and temporal resolution of the video. Rate adaptation is performed by adjusting only the QP values.
- Temporal Mode (T-Mode): The video adaptation algorithm reduces the frame rate to maintain a good SNR quality of the encoded frames. For T-Mode, there are different sub-modes corresponding to different ratios of frame rate reduction (e.g., to reduce the frame rate to 1/2, 1/3 or 1/4 of the original frame rate in our demo implementation). The sub-mode providing the best quality is selected for T-Mode.
- Spatial Mode (S-Mode): The video adaptation algorithm reduces the spatial resolution of the video. For simplicity, the down-sampling factor is fixed to 2 both horizontally and vertically in our demo implementation, so that the frame size is 1/4 of the original size.

Figure 5.1 shows the workflow of the proposed adaptation scheme. The incoming video stream is decoded and a scene change detector is applied on the decoded frames. In the presence of a scene change, the first frame after the scene change is encoded as an intra frame and the current adaptation mode will be kept unchanged (left path in Figure 5.1). Otherwise, the normal mode decision process is conducted as follows (right path in Figure 5.1).

The decoded frame is first used to calculate the spatial and temporal complexity measures (TA/SA) required by the video quality metric (see Eq.(3.10)(3.9)). In order to make globally optimal decisions, the TA and SA values are averaged over a window of 5 frames. The mean TA/SA values are used to estimate the resulting video quality as proposed in Chapter 3. Note that if there is a scene change, the mean TA/SA values are reset. This is because different scenes have quite different spatial and temporal characteristics and by resetting the TA/SA values, a more accurate estimation for the current scene can be achieved.

Then, the elapsed time since the last adaptation mode change is compared with a threshold "waiting time" (referred to as T_{wait}). T_{wait} is used to control the minimum duration the algorithm needs to wait before it is allowed to change the adaptation mode again. The purpose for introducing T_{wait} is to avoid too frequent mode changes. The considerations are two-folds: Firstly, jumping between different adaptation modes may cause jitter effects of the frame quality and affect the user experience. Secondly, the change of spatial resolution requires to reset the encoder status and the first frame after this kind of adaptation mode change must be encoded as an intra frame. If this happens too frequently, the number of intra frames will increase and the coding efficiency of the adapted video stream will be affected. In the implementation, a waiting period of 1 second is used.

If the elapsed time is longer than T_{wait}, quality estimation and mode decision is conducted to select the best adaptation mode. Otherwise, the current adaptation mode is kept unchanged.

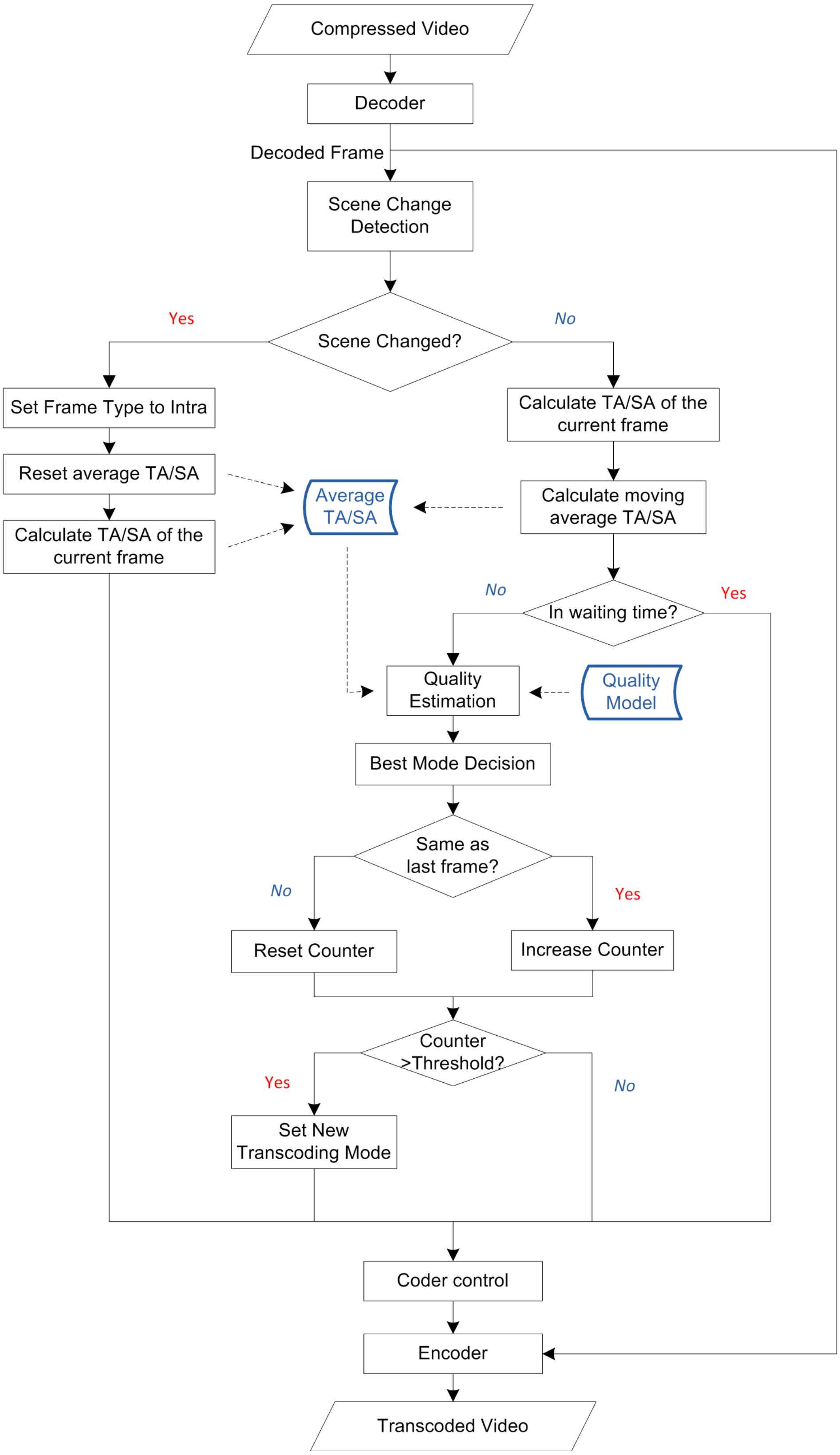


Figure 5.1: Workflow of the video adaptation algorithm

For the quality estimation, the mean TA/SA values are passed to the proposed quality metric in Section 3.4. The estimated quality values for the sub-modes of the temporal mode are first calculated and compared with each other to determine the optimal frame rate for the temporal mode (Section 3.4.2). The estimated quality value for the temporal mode (referred to as TVQ_{best}) is then compared with the estimated video quality for the spatial mode (SVQ, Section 3.4.3) and for the SNR mode (SNRVQ, Section 3.4.1) to determine the optimal adaptation mode. The adaptation mode is changed only if 5 successive frames indicate the same optimal adaptation mode. The consideration here is also to avoid frequent mode changes as mentioned earlier.

After the whole mode decision process mentioned above, the adaptation mode for the current frame is determined and the encoder re-encodes the current frame according to the decision.

5.2 Performance Evaluation of the Video Adaptation Scheme

To evaluate the performance of the QoE-driven video adaptation algorithm, a prototype system with a cascaded transcoding architecture is implemented based on the open source X264 video codec.

Since the spatial adaptation mode changes the resolution of the frames, image scaling needs to be implemented to down- and up-sample the frames. In our implementation, the image scaling algorithm proposed in [Rie] is adopted. For image down-sampling, it simply calculates the average of four adjacent source pixels to get the target pixel value. For image up-sampling, it uses a directional interpolation method, which performs the interpolation according to the direction of the gradient of each pixel so that the interpolation is done "along the edge" instead of "across the edge". This can avoid the typical strong blurring artifacts introduced by bi-linear interpolation. Also, the rate control algorithm in the original X264 codecs is replaced with the ρ-domain rate control algorithm introduced in Chapter 4 to achieve a better accuracy for rate adaptation.

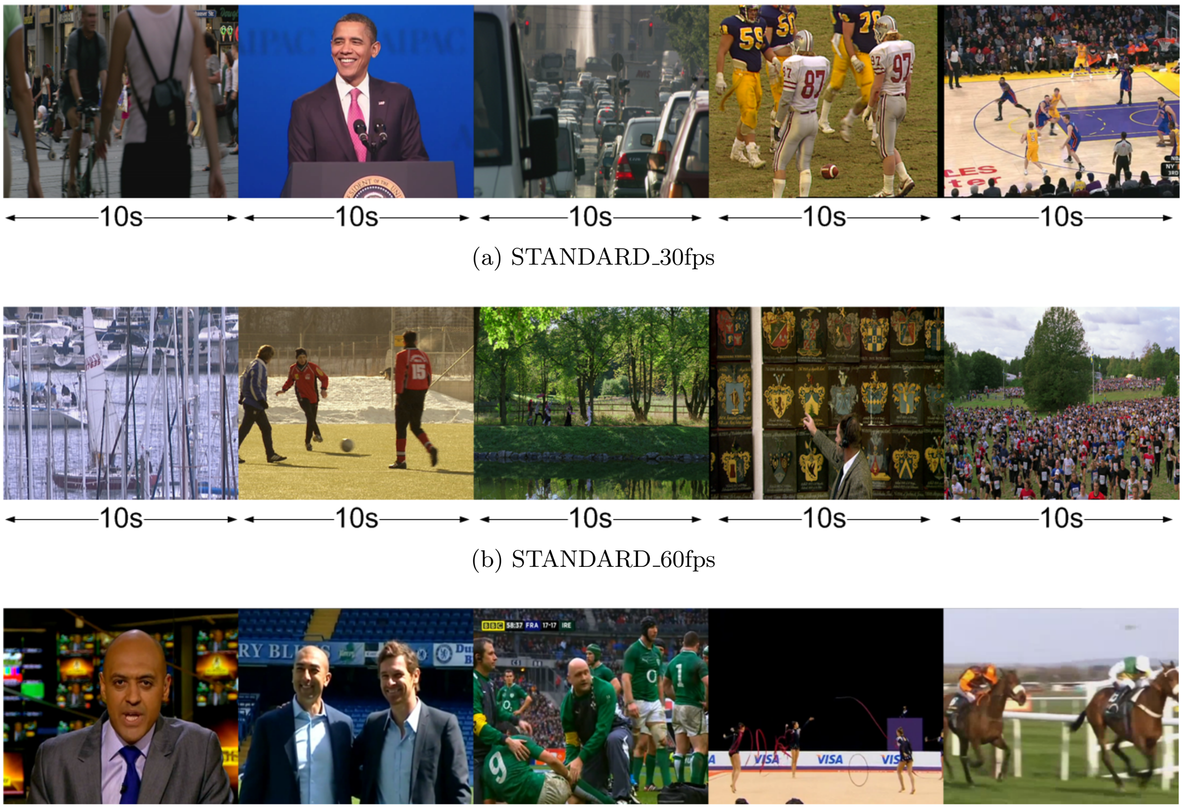
The developed transcoder prototype is used to adapt videos containing different types of content. In the following, three test videos are selected to test the performance of the demo system.

- The first one is a combination of 5 standard test sequences (PEDESTRIAN, OBAMA, RUSH HOUR, FOOTBALL, KOBE) with 30fps frame rate (referred to as STANDARD_30fps). Each sub-scene contains 300 frames (corresponds to a duration of 10 seconds), so the total length is 50 second.

- The second one is a combination of 5 standard test sequences (HARBOUR, SOCCER, PARKJOY, SHIELDS, CROWD RUN) with 60fps frame rate (referred to as STANDARD_60fps). Each sub-scene contains 500 frames.

- The third test sequence is a video clip of sport news from BBC downloaded from Youtube (referred to as BBCNEWS_30fps). The frame rate of this video is 30fps. The sub-scenes in this video are of different duration.

Example scenes from the test videos are shown in Figure 5.2. The scenes are arranged according to their order in the test videos.



(a) STANDARD_30fps

(b) STANDARD_60fps

(c) BBCNEWS_30fps

Figure 5.2: Example frames of test sequences

These videos are first encoded at a relatively high bit-rate so that the output video has a very good quality (which can avoid the impact of the quality of the source videos on the performance evaluation of the adaptation scheme). The developed transcoder is used to adapt the videos to a relatively low bitrate. The encoding bit-rates of the input and output video streams are summarized in Table 5.1. Adapted videos employing four different adaptation strategies are generated for comparison:

- Proposed adaptive mode decision scheme, which adaptively selects the best video adaptation mode.

- SNR-only scheme, which only uses SNR-mode for adaptation.
- TR-only scheme, which only uses T-mode for adaptation.
- SR-only scheme, which only uses S-mode for adaptation.

Table 5.1: Encoding bit-rates of the input and output video streams for performance evaluation

test video	input bitrate (kbps)	output bitrate (kbps)			
		BR1	BR2	BR3	BR4
STANDARD_30fps	4000	1500	1000	768	512
STANDARD_60fps	8000	1920	1440	1000	768
BBCNEWS_30fps	6000	1500	1000	768	512

To compare the performance of different adaptation schemes, the quality improvement is measured quantitatively as:

$$IR = \frac{\overline{VQ}_{QoE}}{\overline{VQ}_{na}} - 1 \tag{5.1}$$

where $\overline{VQ}_{QoE}$ is the resulting mean video quality of the proposed QoE-driven adaptation scheme and $\overline{VQ}_{na}$ is the mean video quality when one of the three non-adaptive adaptation schemes is used. The video quality is estimated using the video quality metric MDVQM proposed in Chapter 3.

Tables 5.2-5.4 show the quality improvement of the proposed adaptation scheme against the non-adaptive strategies when adapting the original video stream to different lower bit-rates. The improvement is measured in two ways. The first one is the improvement of the mean quality value over the whole video. This is referred to as "overall" quality improvement. The second one is the improvement of the mean video quality over the periods in which different decisions are made by the comparison schemes. For example, when the proposed adaptive scheme is compared with SNR-only scheme, only the parts in the video where the algorithm uses an adaptation mode other than SNR mode is considered. In Table 5.2-5.4, this is referred to as "optimized" part.

From the results in Table 5.2-5.4, it can be seen that compared with the conventional SNR-only scheme (which changes only QP for video adaptation), the proposed QoE-driven adaptive scheme can achieve an overall quality improvement of up to 10%. Also, the "overall" improvement of video quality decreases when the target adaptation bit-rate gets higher. This is easy to understand, because when the temporal/spatial complexity of the video content is low or the target bit rate is relatively high, then there is no need to do any special adaptation operations other than changing the QP. In this case, the overall quality improvement against

Table 5.2: Video quality improvement of the QoE-driven adaptation scheme against non-adaptive strategies (for video STANDARD_30fps transcoded from 4Mbps)

Bitrate	QoE vs SNR		QoE vs SR		QoE vs TR	
	Optimized	Overall	Optimized	Overall	Optimized	Overall
1500kbps	14.78%	0.69%	24.57%	22.97%	23.64%	23.64%
1Mbps	21.74%	3.07%	21.10%	16.98%	18.59%	18.59%
768kbps	23.67%	5.37%	19.74%	13.73%	15.07%	15.07%
512kbps	22.75%	8.68%	20.91%	11.04%	10.66%	10.59%

Table 5.3: Video quality improvement of the QoE-driven adaptation scheme against non-adaptive strategies (for video BBCNEWS_30fps transcoded from 6Mbps)

Bitrate	QoE vs SNR		QoE vs SR		QoE vs TR	
	Optimized	Overall	Optimized	Overall	Optimized	Overall
1500kbps	40.92%	4.24%	27.42%	22.71%	16.30%	16.30%
1Mbps	63.88%	7.80%	27.39%	21.22%	15.14%	15.14%
768kbps	58.01%	8.04%	28.44%	21.44%	12.67%	12.67%
512kbps	46.57%	6.34%	29.58%	22.87%	8.39%	8.36%

Table 5.4: Video quality improvement of the QoE-driven adaptation scheme against non-adaptive strategies (for video STANDARD_60fps transcoded from 8Mbps)

Bitrate	QoE vs SNR		QoE vs SR		QoE vs TR	
	Optimized	Overall	Optimized	Overall	Optimized	Overall
1920kbps	54.49%	5.19%	39.32%	32.68%	11.59%	11.42%
1440kbps	31.77%	5.80%	36.02%	30.49%	7.86%	6.86%
1Mbps	9.79%	8.13%	36.89%	32.08%	15.51%	3.53%
768kbps	11.81%	10.71%	39.05%	35.53%	14.06%	1.87%

the SNR-only scheme is low. Therefore, the level of "overall" quality improvement depends heavily on the characteristics of the video content and target bit rate. When the proposed QoE-driven adaptive scheme is compared with the other two non-adaptive (TR-only and SR-only) schemes, significant quality improvements can also be observed. This indicates the importance of choosing a proper adaptation method.

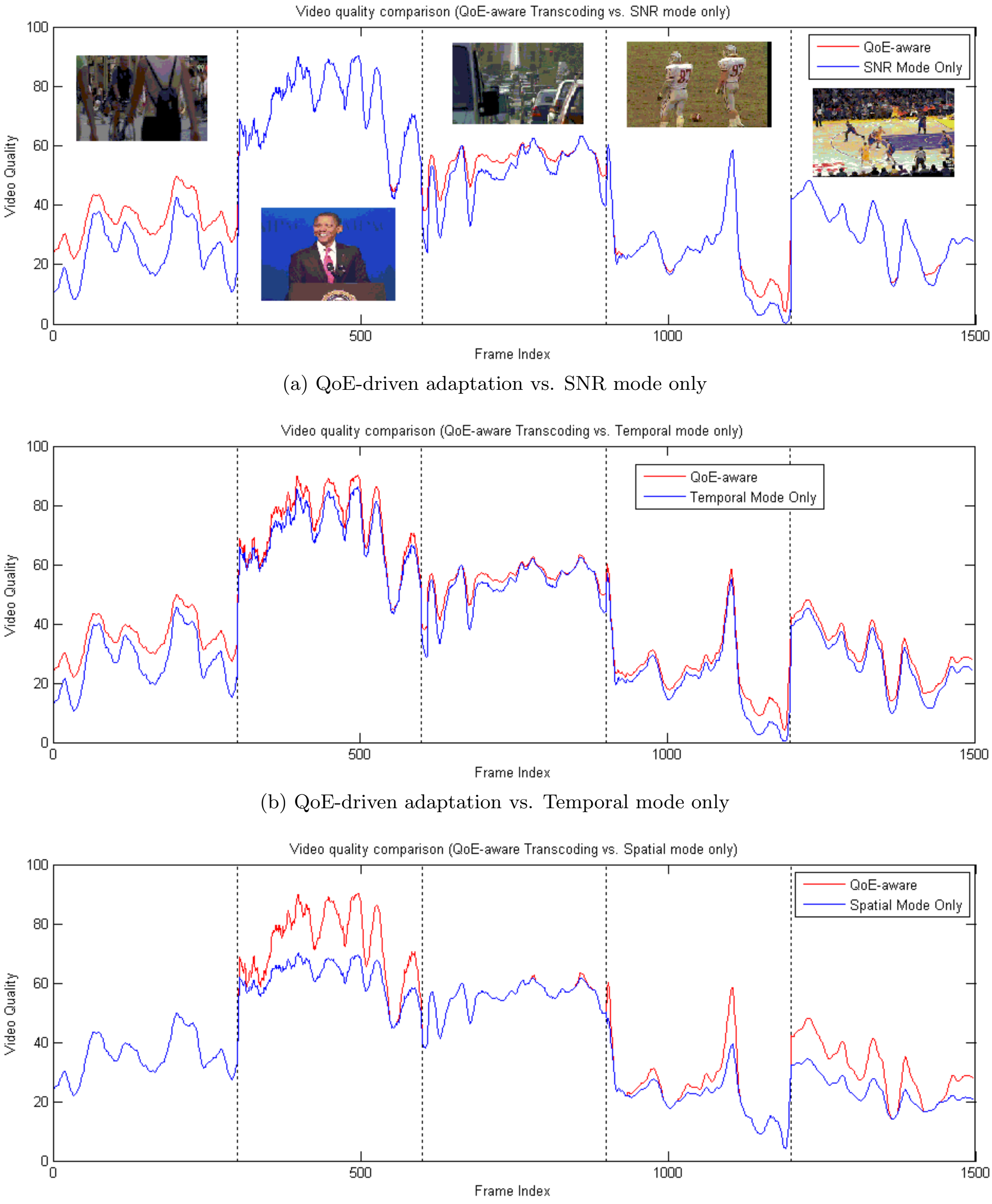
To make the comparison more clear, Figure 5.3 shows the change of video quality over time when different adaptation strategies are used to transcode the video STANDARD_30fps from 4Mbps to 512kbps. The x-axis is the frame index and the y-axis is the video quality calculated by the metric MDVQM. The proposed QoE-driven adaptation scheme is compared with three other non-adaptive strategies, i.e., SNR mode only(Figure 5.3a), temporal mode only (Figure 5.3b) and spatial mode only (Figure 5.3c). The red curves show the quality change when the QoE-driven adaptation scheme is used and the blue curves show the case when the other strategies are used. Sometimes the curves overlap because the QoE-driven adaptation scheme has chosen the same adaptation mode as the one chosen by the comparison scheme. It can be seen that the red curves are always higher than or overlap with the blue curves, which proves that the proposed QoE-driven adaptation scheme can adaptively make the best adaptation decision to optimize the video quality. Figure 5.4 and Figure 5.5 show the change of video quality over time for the other two test videos (i.e. STANDARD_60fps and BBCNEWS_60fps). In Figure 5.3 and Figure 5.4, the period of different sub-scenes contained in the test videos is marked by showing example pictures of the sub-scenes to make the figures more intuitive.

From the results, it can be seen that for 30fps sequences, the temporal mode is not used very often. Most of the time, the transcoder operates in SNR- or spatial-mode. The reason for this is that 30fps is already a threshold below which the human visual system tends to recognize the jerkiness caused by frame dropping. So reducing the frame rate from 30fps to 15fps or lower is not a preferred way for the video adaptation. This behavior of the transcoder accords with the result from the subjective tests.

On the other hand, for 60fps sequences, since the original frame rate is relatively high, the impact of frame rate reduction from 60fps to 30fps is not very significant. So the temporal-mode is used more often than for the 30fps sequences. This observation can also be confirmed by the results in Table 5.2-5.4: For the two 30fps test videos, the performance of the TR-only scheme is much worse than the SNR-only scheme (which is proved by a much higher overall quality loss against the proposed scheme of the TR-only scheme than the SNR-only scheme). However, according to the results for the test video STANDARD_60fps (Table 5.4), the overall performance of the TR-only scheme is very close to that of the SNR-only scheme, sometimes even better especially at low bit-rates. This indicates for videos with a relatively high frame rate (e.g. 50/60fps), the temporal-mode (reducing the frame rate) could be a preferred choice for adaptation.

Figure 5.6a shows sample frames from BBCNEWS_30fps adapted using the proposed adaptive QoE-driven scheme (right) and the conventional SNR-only scheme (left). For this frame, the adaptive scheme chooses to encode the frame in SR-mode (which means the frame is down-sampled and then encoded at the lower resolution) while the SNR-only scheme choose to adjust simply the quantization parameter. The achieved quality improvement can be seen clearly. In Figure 5.6b and Figure 5.6c comparisons of adapted frames for test videos STANDARD_30fps and STANDARD_30fps are also presented. Similar quality improvements can also be observed from the pictures.

The above results have shown the effectiveness of the proposed QoE-driven adaptation

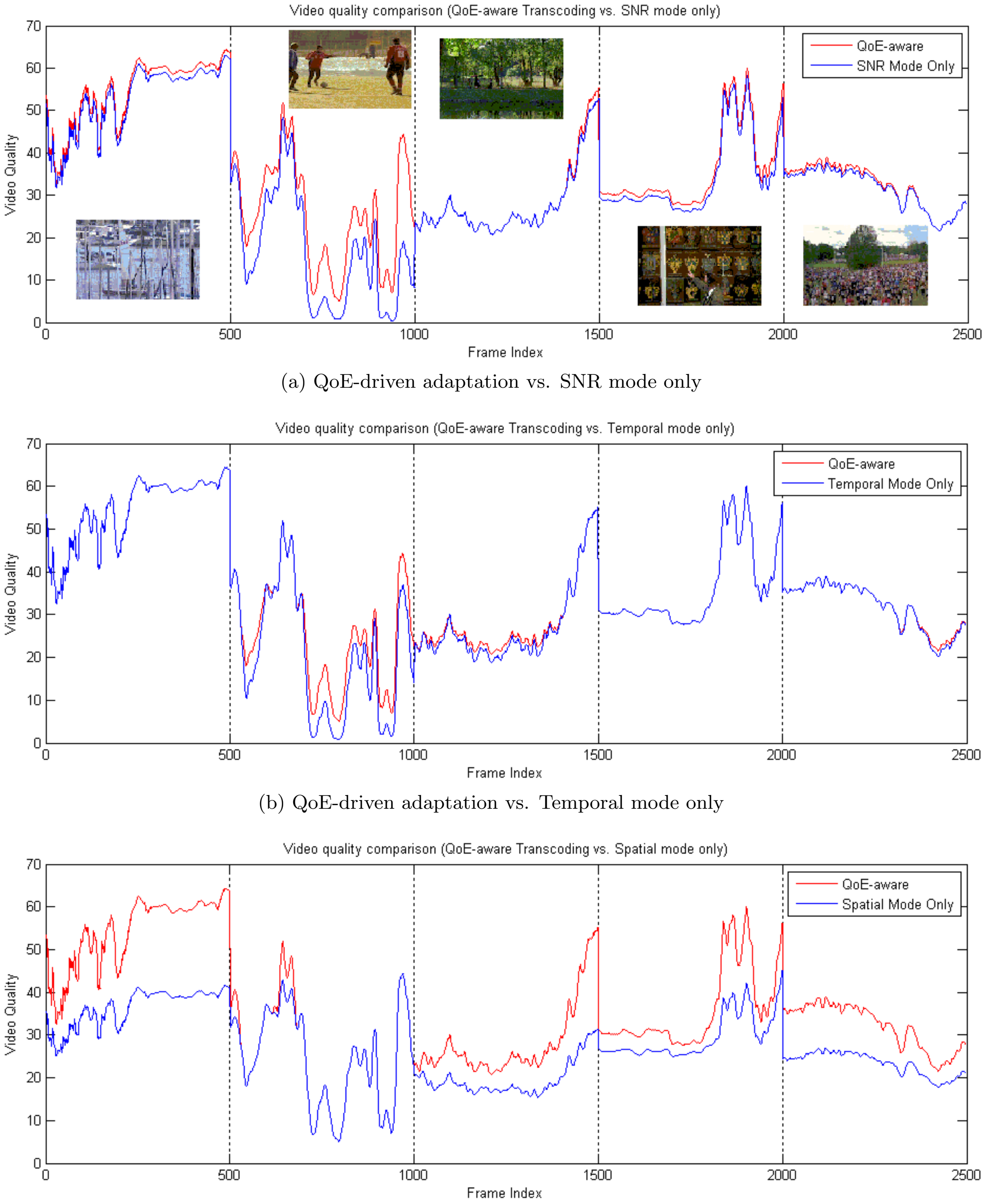


(a) QoE-driven adaptation vs. SNR mode only

(b) QoE-driven adaptation vs. Temporal mode only

(c) QoE-driven adaptation vs. Spatial mode only

Figure 5.3: Video quality comparison between the QoE-driven adaptation scheme and three non-adaptive strategies (SNR-only, Spatial-only, Temporal-only) for the video STANDARD_30fps (transcoded from 4Mbps to 512kbps)

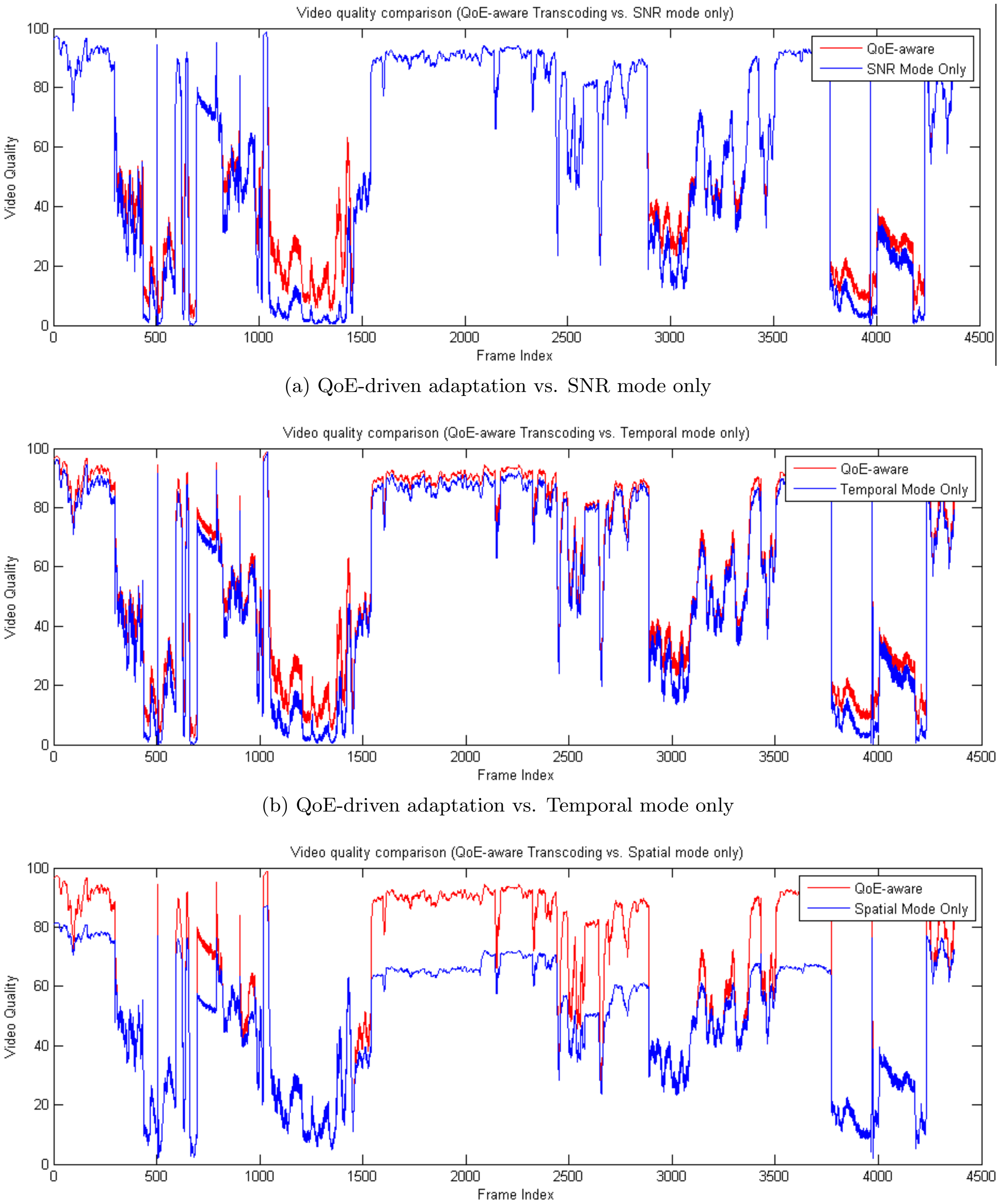


(a) QoE-driven adaptation vs. SNR mode only

(b) QoE-driven adaptation vs. Temporal mode only

(c) QoE-driven adaptation vs. Spatial mode only

Figure 5.4: Video quality comparison between the QoE-driven adaptation scheme and three non-adaptive strategies (SNR-only, Spatial-only, Temporal-only) for the video STANDARD_60fps (transcoded from 8Mbps to 1Mbps)



(a) QoE-driven adaptation vs. SNR mode only

(b) QoE-driven adaptation vs. Temporal mode only

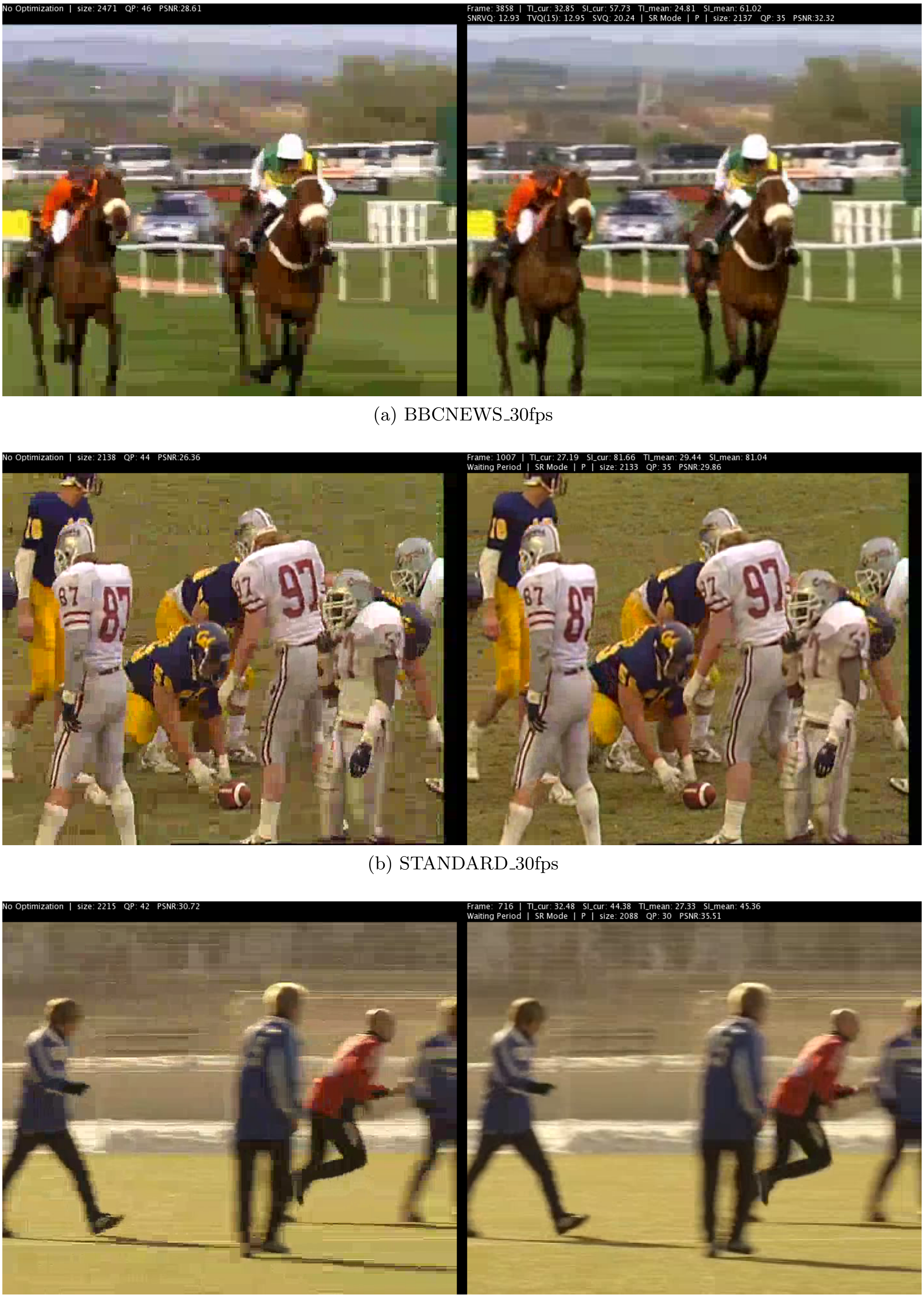
(c) QoE-driven adaptation vs. Spatial mode only

Figure 5.5: Video quality comparison between the QoE-driven adaptation scheme and three non-adaptive strategies (SNR-only, Spatial-only, Temporal-only) for the video BBC-NEWS_30fps (transcoded from 6Mbps to 512kbps)

scheme.

5.3 Summary

In this chapter, a QoE-driven video adaptation scheme is proposed to adaptively make the optimal adaptation decision according to the characteristics of the video content and the channel resources. Three adaptation modes, i.e. SNR Mode, Temporal Mode and Spatial Mode, are considered in the proposed scheme. The metric MDVQM proposed in Chapter 3 is used to estimate perceived video quality and select the suitable adaptation operations. The frequency of mode changes is restricted to avoid the spatial and temporal flickering effect. The performance of the proposed scheme is evaluated using video sequences with different characteristics. The results show that the proposed adaptive MDA scheme outperforms all the other non-adaptive approaches. Compared with the traditional adaptation scheme where only QP is adjusted, the improvement of the overall perceived video quality is up to 10%.


No Optimization | size: 2471 QP: 46 PSNR:28.61
Frame: 3858 | TI_cur: 32.85 SI_cur: 57.73 TI_mean: 24.81 SI_mean: 61.02
SNRVQ: 12.93 TVQ(15): 12.95 SVQ: 20.24 | SR Mode | P | size: 2137 QP: 35 PSNR:32.32
No Optimization | size: 2138 QP: 44 PSNR:26.36
Frame: 1007 | TI_cur: 27.19 SI_cur: 81.66 TI_mean: 29.44 SI_mean: 81.04
Waiting Period | SR Mode | P | size: 2133 QP: 35 PSNR:29.86
No Optimization | size: 2215 QP: 42 PSNR:30.72
Frame: 716 | TI_cur: 32.48 SI_cur: 44.38 TI_mean: 27.33 SI_mean: 45.36
Waiting Period | SR Mode | P | size: 2088 QP: 30 PSNR:35.51

(a) BBCNEWS_30fps

(b) STANDARD_30fps

(c) STANDARD_60fps

Figure 5.6: Sample frames of the test videos adapted using SNR-only mode(left) and QoE-driven adaptation scheme(right)

Chapter 6

Conclusion and Future Work

6.1 Conclusions

In this thesis, various aspects of QoE-driven multi-dimensional video adaptation are investigated. The main objective of this work is to find the optimal combination of multi-dimensional adaptation operations and optimize the perceived video quality (QoE) for the end-users. Specifically, two important issues are investigated: the objective estimation of QoE and rate control for accurate rate adaptation.

To understand how the perceived video quality is affected when the spatial and temporal resolution of the video content are changed separately or even jointly, extensive subjective experiments are conducted using diverse video contents. Based on the subjective data from the experiments, a no-reference video quality metric (MDVQM) for multi-dimensional video adaptation is proposed to estimate the perceived video quality under different combinations of spatial, temporal and SNR resolution. The overall video quality is modelled as the product of separate items, with each of the items trying to catch the impact of quantization, frame dropping and spatial down-sampling, respectively. The results of performance evaluation indicate that the analytical quality metric can provide accurate quality estimation in the presence of different spatial/temporal impairments.

A ρ-domain rate control algorithm with header size estimation for H.264/AVC video is also proposed for the task of accurate rate adaptation. A two-stage encoder structure is used to resolve the inter-dependency between RDO and ρ-domain rate control for H.264/AVC encoding. The size of header information is estimated by an improved rate model which considers various components in a MB header. Experimental results show that the proposed algorithm achieves better rate control accuracy and video quality when compared with the original ρ-domain rate control algorithm.

Finally, combining the proposed video quality metric and rate control algorithm, a QoE-driven approach is proposed for multi-dimensional video adaptation to optimize the perceived video quality. The resulting QoE under different adaptation modes is estimated by the proposed video quality metric and the optimal combination of adaptation operations is determined by considering the predicted QoE and computational complexity. Performance evaluations have shown that the proposed QoE-driven MDA scheme can provide significant QoE improvements when compared with conventional video adaptation schemes.

6.2 Future Work

In this section, the potential extensions and applications of the work presented in this thesis are discussed in three directions.

6.2.1 Video Quality Metric Design

The proposed metric uses the pixel-bit-rate as a video feature for the estimation of the video quality. This feature is a good indicator of the video quality and can be read directly from the compressed bitstream which reduces the complexity of the metric. One issue of using the pixel-bit-rate is that it is encoder dependent and therefore, the parameters in the metric may need to be re-trained for each encoder type (such as MPEG-4 ASP and VC-1). It will be interesting to look into more general video features to improve the generality of the metric.

In the proposed metric, the spatial quality modelling which simulates the impact of spatial down-sampling is only verified by a down-sampling ratio fixed to 2 (from 4CIF to CIF). It is worth extending the discussion to more flexible spatial down-sampling ratios.

The temporal and spatial activity measures used in the proposed metric are calculated from the decoded video frames which need a complete decoding process. From a complexity point of view, it is desirable to estimate the measures without decoding or with only partial decoding of the compressed stream. This could be done, for example, by using information from the residual signal or the statistics of transformed coefficients in intra-coded frames.

6.2.2 Rate Control Algorithm

For simplicity, a simple frame level bit allocation scheme is used, which allocates the bit budget evenly to the inter-coded frames. The video quality could be further improved by employing a more advanced frame level bit allocation algorithm based on the characteristics of the video content. This is of particular interest to applications where constant video quality is more desired than a strict constant bit-rate constraint.

6.2.3 QoE-driven MDA scheme

The proposed MDA scheme allows us to adapt the video content actively in the spatial and temporal domain to meet the target bit-rate. However, too frequent changes of the SNR, spatial or temporal quality may cause serious flicker effect which degrades the perceived video quality. Researches have shown that the perceptual effect of flicker largely depends on the frequency and amplitude of the quality variations [ZKSS03][NEE$^+$11]. Therefore, the time, the frequency and the amplitude of the adaptation operations should be carefully considered. In the proposed MDA scheme in Chapter 5, this issue is only addressed empirically by setting a fixed interval for successive adaptation operations. It is worth investigating how the corresponding flicker effects affect the perceptual video quality.

In this thesis, the proposed video quality metric is only applied to the MDA problem for a single video stream. Actually, the metric can also be applied to a multi-stream transmission scenario, where multiple concurrent video streams are transmitted to a user or even multiple users. In such scenarios, the metric can support the joint adaptation of a set of video streams in consideration of their respective characteristics to achieve fairness among the streams and users while maintaining a good perceived video quality.

Bibliography

Publications by the author

[SZPD12] E. Steinbach, F. Zhang, Y. Peng, and W. K. Deng. Adaptive wireless video transmission systems and methods. *U.S. Patent 20120079329*, March 2012. [cited at p. 6]

[ZS11] F. Zhang and E. Steinbach. Improved p-domain rate control with accurate header size estimation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 813–816, Prague, Czech Republic, May 2011. [cited at p. 6]

General publications

[BHTHB06] M. D. Brotherton, Q. Huynh-Thu, D. S. Hands, and K. Brunnström. Subjective multimedia quality assessment. *IEICE Transactions*, 89-A(11):2920–2932, 2006. [cited at p. 11]

[BM10] B. Belmudez and S. Möller. Extension of the g.1070 video quality function for the mpeg-2 video codec. In *Proc. International Workshop on Quality of Multimedia Experience*, pages 7–10, Trondheim, Norway, June 2010. [cited at p. 22]

[CAL96] S. Cheung, M. Ammar, and X. Li. On the use of destination set grouping to improve fairness in multicast video distribution. In *Proc. IEEE International Conference on Computer Communications*, pages 553–560, San Francisco, CA, USA, March 1996. [cited at p. 2]

[Cha13] D. M. Chandler. Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, pages 1–53, January 2013. [cited at p. 22]

[Che] Cheetah Technologies. Cheetah v-factor source monitor. available: `http://www.cheetahtech.com`. [cited at p. 23]

[CMB02] I. Cox, M. L. Miller, and J. A. Bloom. *Digital Watermarking.* Morgan Kaufmann, New York, 2002. [cited at p. 21]

[CSRK11] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Transactions on Broadcasting*, 57(2):165–182, June 2011. [cited at p. 16]

[CT07] J. Y. C. Chen and J. E. Thropp. Review of low frame rate effects on human performance. *IEEE Transactions on Systems, Man, and Cybernetics*, 37(6):1063–1076, November 2007. [cited at p. 26]

[Dal93] S. Daly. The visible differences predictor: An algorithm for the assessment of image fidelity. In A. B. Watson, editor, *Digital Images and Human Vision*, pages 179–206. MIT Press, 1993. [cited at p. 15]

[DF99] J.L. Devore and N.R. Farnum. *Applied Statistics for Engineers and Scientists.* Duxbury, New York, USA, 1999. [cited at p. 68]

[DK82] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach.* Prentice-Hall, London, GB, 1982. [cited at p. 44]

[EF95] A. M. Eskicioglu and P. S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43:2959–2965, December 1995. [cited at p. 15]

[FCM05] M. C. Q. Farias, M. Carli, and S. K. Mitra. Objective video quality metric based on data hiding. *IEEE Transactions on Consumer Electronics*, 51(3):983–992, August 2005. [cited at p. 21]

[FM05] M. Farias and S. Mitra. No-reference video quality metric based on artifact measurements. In *Proc. IEEE International Conference on Image Processing*, volume 3, pages III: 141–144, Genoa, Italy, September 2005. [cited at p. 22]

[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. 28]

[Gei93] S. Geisser. *Predictive Inference.* Chapman and Hall, New York, USA, 1993. [cited at p. 44]

[Gir93] B. Girod. What's wrong with mean-squared error. In A. B. Watson, editor, *Digital Images and Human Vision*, pages 207–220. MIT Press, 1993. [cited at p. 15]

[GSR10] M.N. Garcia, R. Schleicher, and A. Raake. Towards a content-based parametric video quality model for iptv. In *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, USA, January 2010. [cited at p. 22]

[HM01] Z. He 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. 61, 63]

[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. 61, 63]

[HR10] S. Hemami and A. Reibman. No-reference image and video quality estimation: Applications and human-motivated design. *Journal of Image Communication*, 25(7):469–481, August 2010. [cited at p. 22]

[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, Honolulu, HI, USA, August 2005. [cited at p. 10]

[HTG08] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *IET Electronics Letters*, 44(13):800–801, June 2008. [cited at p. 15]

[HW08] Z. He and D. O. Wu. Linear rate control and optimum statistical multiplexing for h.264 video broadcast. *IEEE Transactions on Multimedia*, 10(7):1237–1249, November 2008. [cited at p. 64]

[HYTT07] T. Hayashi, K. Yamagishi, T. Tominaga, and A. Takahashi. Multimedia quality integration function for videophone services. In *Proc. IEEE Global Telecommunications Conference*, pages 2735–2739, Washington, DC, USA, November 2007. [cited at p. 22]

[ITU98] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications, 1998. [cited at p. 8, 9, 13, 14]

[ITU99] ITU-R Recommendation BT.500-11. Methodology for the subjective assessment of the quality of television pictures, 1999. [cited at p. iii, 8, 9, 10, 13, 14]

[ITU00] ITU-T Recommendation J.143. User requirements for objective perceptual video quality measurements in digital cable television, 2000. [cited at p. iii, 16, 17, 18]

[ITU04a] ITU-R Recommendation BT.1683. Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference, 2004. [cited at p. 19]

[ITU04b] ITU-T Recommendation J.144. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, 2004. [cited at p. 19, 21]

[ITU05a] ITU-R Document 6Q/131-E. Technical report: Comparison of dscqs and acr, October 2005. [cited at p. 9]

[ITU05b] ITU-R Recommendation H.263. Video coding for low bit rate communication, 2005. [cited at p. 63]

[ITU07] ITU-T Recommendation BT.1788. Methodology for the subjective assessment of video quality in multimedia applications, 2007. [cited at p. iii, 8, 10, 12, 13, 14, 31, 33, 34, 35]

[ITU08a] ITU-T Recommendation J.246. Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference, 2008. [cited at p. 19, 20]

[ITU08b] ITU-T Recommendation J.247. Objective perceptual multimedia video quality measurement in the presence of a full reference, 2008. [cited at p. 19, 20, 38]

[ITU10] ITU-T Recommendation J.249. Perceptual video quality measurement techniques for digital cable television in the presence of a reduced reference, 2010. [cited at p. 19, 20, 21]

[ITU11a] ITU-T Recommendation J.341. Objective perceptual multimedia video quality measurement of hdtv for digital cable television in the presence of a full reference, 2011. [cited at p. 19, 20, 38]

[ITU11b] ITU-T Recommendation J.342. Objective multimedia video quality measurement of hdtv for digital cable television in the presence of a reduced reference signal, 2011. [cited at p. 19, 20]

[ITU12] ITU-T Recommendation G.1070. Opinion model for video-telephony applications, 2012. [cited at p. 21]

[KHD12] C. Keimel, J. Habigt, and K. Diepold. Hybrid no-reference video quality metric based on multiway plsr. In *Proc. European Signal Processing Conference*, pages 1244–1248, Bucharest Romania, August 2012. [cited at p. 23]

[KJSR08] C. S. Kim, S. H. Jin, D. J. Seo, and Y. M. Ro. Measuring video quality on full scalability of h.264/avc scalable video coding. *IEEE Transactions on Communications*, E91-B(5):1269–78, May 2008. [cited at p. 28]

[KKHD11] C. Keimel, M. Klimpke, J. Habigt, and K. Diepold. No-reference video quality metric for hdtv based on h.264/avc bitstream features. In *Proc. IEEE International Conference on Image Processing*, pages 3382–3385, Brussels, Belgium, September 2011. [cited at p. 22, 23]

[KSI09] A. Khan, L. Sun, and E. C. Ifeachor. Content-based video quality prediction for mpeg4 video streaming over wireless networks. *Journal of Multimedia*, 4(4):228–239, August 2009. [cited at p. 22]

[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. 64, 67, 71, 75]

[KSW05] F. Kozamernik, P. Sunna, and E. Wyckens. Subjective quality of internet video codecs - phase 2 evaluations using samviq, January 2005. [cited at p. 13]

[Lab] Multimedia Communications Lab. Media networking. `http://www.surrey.ac.uk/cvssp/activity/labs/i-lab/media_networking.htm`. [cited at p. iii, 2]

[LLS$^+$05] Z. Lu, W. Lin, B. C. Seng, S. Kato, S. Yao, E. Ong, and X. K. Yang. Measuring the negative impact of frame dropping on perceptual visual quality. In *Proc. SPIE Human Vision and Electronic Imaging*, volume 5666, pages 554–562, San Jose, CA, USA, January 2005. [cited at p. 27]

[LSR$^+$10] J. S. Lee, F. D. Simone, N. Ramzan, Z. Zhao, E. Kurutepe, T. Sikora, J. Ostermann, E. Izquierdo, and T. Ebrahimi. Subjective evaluation of scalable video coding for content distribution. In *Proc. ACM International Conference on Multimedia*, pages 65–72, Firenze, Italy, October 2010. [cited at p. 27]

[Lub97] J. Lubin. A human vision system model for objective picture quality measurements. In *Proc. International Broadcasting Convention*, pages 498–503, Amsterdam, Netherlands, September 1997. [cited at p. 15]

[MB11] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, December 2011. [cited at p. 23]

[MGWL03] S. Ma, W. Gao, F. Wu, and Y. Lu. Rate control for avc video coding scheme with hrd consideration. In *Proc. IEEE International Conference on Image Processing*, pages 793–796, Barcelona, Spain, September 2003. [cited at p. 62, 64]

[MJ96] S. McCanne and V. Jacobson. Receiver-driven layered multicast. In *Proc. ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 117–130, Stanford, CA, USA, August 1996. [cited at p. 2]

[MLW03] S. Ma, Z. Li, and F. Wu. Proposed draft of adaptive rate control. Document JVT-H017, May 2003. [cited at p. 62, 64]

[MSL99] R. Mohan, R.J. Smith, and C-Sheng Li. Adapting multimedia internet content for universal access. *IEEE Transactions on Multimedia*, 1(1):104–114, March 1999. [cited at p. 1]

[MV07] L. Merritt and R. Vanam. Improved rate control and motion estimation for h.264 encoder. In *Proc. IEEE International Conference on Image Processing*, pages 309–312, San Antonio, TX, USA, September 2007. [cited at p. 62]

[NCA06] A. Ninassi, P. Le Callet, and F. Autrusseau. Pseudo no reference image quality metric using perceptual data hiding. In *Proc. SPIE Human Vision and Electronic Imaging*, volume 6057, pages 146–157, San Jose, CA, USA, January 2006. [cited at p. 21]

[NEE$^+$11] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen. Flicker effects in adaptive video streaming to handheld devices. In *Proc. ACM International Conference on Multimedia*, pages 463–472, Scottsdale, AZ, USA, November 2011. [cited at p. 97]

[OMLW11] Y. Ou, Z. Ma, T. Liu, 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. 28, 45, 50]

[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*, Scottsdale, AZ, USA, January 2009. [cited at p. 28]

[OXMW11] Y. Ou, Y. Xue, Z. Ma, and Y. Wang. A perceptual video quality model for mobile platform considering impact of spatial, temporal, and amplitude resolutions. In *Proc. IEEE Image, Video, and Multidimensional Signal Processing Workshop*, pages 117–122, Ithaca, NY, USA, June 2011. [cited at p. 28, 38, 52, 54, 56]

[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*, Kyoto, Japan, September 2008. [cited at p. 11]

[PS11] Y. Peng and E. Steinbach. A novel full-reference video quality metric and its application to wireless video transmission. In *Proc. IEEE International Conference on Image Processing*, pages 2517–2520, Brussels, Belgium, September 2011. [cited at p. 29, 38, 45, 50]

[PW02] M. H. Pinson and S. Wolf. Video quality measurement techniques. Technical Report TR-02-392, National Telecommunications and Information Administration (NTIA), June 2002. [cited at p. 38]

[PW08] M. H. Pinson and S. Wolf. Techniques for evaluating objective video quality models using overlapping subjective data sets. Technical Report TR-09-457, NITA, November 2008. [cited at p. 37, 43]

[QG08] H.-T. Quan and M. Ghanbari. Temporal aspect of perceived quality of mobile video broadcasting. *IEEE Transactions on Broadcasting*, 54(3):641–651, September 2008. [cited at p. 27]

[RCNR07] M. Ries, C. Crespi, O. Nemethova, and M. Rupp. Content based video quality estimation for h.264/avc video streaming. In *Proc. Wireless Communications and Networking Conference*, pages 2668–2673, Hong Kong, China, March 2007. [cited at p. 22]

[RGSM$^+$08] A. Raake, M.N. Garcia, J. Berger S. Moller, F. Kling, P. List, J. Johann, and C. Heidemann. T-v-model: Parameter-based prediction of iptv quality. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1149–1152, Las Vegas, NV, USA, April 2008. [cited at p. 22]

[Rie] T. Riemersma. Quick image scaling by 2. `http://www.compuphase.com/graphic/scale2.htm`. [cited at p. 84]

[SB10] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, February 2010. [cited at p. 20]

[Sto74] M. Stone. Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36(2):111–147, 1974. [cited at p. 44]

[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. 64]

[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):269–80, September 2010. [cited at p. 28]

[Tea] Joint Video Team. JVT Reference Software Encoder. available: `http://www.bs.hhi.de/seuhring/tml/download`. [cited at p. 62, 64]

[THB08] A. Takahashi, D. Hands, and V. Barriac. Standardization activities in the itu for a qoe assessment of iptv. *IEEE Communications Magazine*, pages 78–84, March 2008. [cited at p. 21]

[UE07] H.-J. Zepernick U. Engelke. Perceptual-based quality metrics for image and video services: A survey. In *Proc. 3rd EuroNGI Conference on Next Generation Internet Networks*, pages 190–197, Trondheim, Norway, May 2007. [cited at p. 16]

[VQE00] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, April 2000. [cited at p. 11, 21, 38, 43]

[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. 11, 21, 38, 43]

[VQE07] VQEG. Multimedia group test plan draft version 1.19, 2007. [cited at p. 9, 13]

[VQE08] VQEG. Final report of vqeg's multimedia phase i validation test, September 2008. [cited at p. 11, 19, 41]

[VQE09] VQEG. Validation of reduced-reference and no-reference objective models for standard definition television, phase i, June 2009. [cited at p. 11, 21]

[Wan05] Y. Wang. Resource constrained video coding/adaptation, 2005. [cited at p. 3]

[WB02] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, March 2002. [cited at p. 15]

[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. 18, 19, 41]

[WCL04] Y. Wang, S.-F. Chang, and A. Loui. Subjective preference of spatio-temporal rate in video adaptation using multi-dimensional scalable coding. In *Proc. IEEE International Conference on Multimedia and Expo*, volume 3, pages 1719–1722, Taipei, Taiwan, China, June 2004. [cited at p. 26]

[WHM01] A. B. Watson, J. Hu, and J. F. McGowan. Digital video quality metric based on human vision. *Journal of Electronic Imaging*, 10(1):20–29, January 2001. [cited at p. 15]

[Win98] S. Winkler. A perceptual distortion metric for digital color images. In *Proc. IEEE International Conference on Image Processing*, pages 399–403, Chicago, IL, USA, October 1998. [cited at p. 15]

[Win99] S. Winkler. A perceptual distortion metric for digital color video. In *Proc. SPIE Human Vision and Electronic Imaging*, volume 3644, pages 175–184, San Jose, CA, USA, January 1999. [cited at p. 15]

[WJP+93] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf. An objective video quality assessment system based on human perception. In *Proc. SPIE Conference on Human Vision, Visual Processing, and Digital Display*, volume 1913, pages 15–26, San Jose, CA, USA, February 1993. [cited at p. 20]

[WL07] Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *Journal of the Optical Society of America A (Optics, Image Science, Vision)*, 24(12):B61–B69, 2007. [cited at p. 19, 21]

[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. 19, 41]

[WM08] S. Winkler and P. Mohandas. The evolution of video quality measurement: From psnr to hybrid metrics. *IEEE Transactions on Broadcasting*, 54(3):660–668, October 2008. [cited at p. 15, 16, 23]

[WP99] S. Wolf and M. H. Pinson. Spatial-temporal distortion metric for in-service quality monitoring of any digital video system. In *Proc. SPIE Conference on Multimedia Systems and Applications*, volume 3845, pages 266–277, Boston, MA, USA, November 1999. [cited at p. 21]

[WSB03a] 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, pages 1041–1078. CRC Press, Boca Raton, FL, September 2003. [cited at p. iii, 15, 16]

[WSB03b] Z. Wang, E. Simoncelli, and A. Bovik. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, volume 2, pages 1398–1402, Pacific Grove, CA, November 2003. [cited at p. 19]

[WSBL03] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576, July 2003. [cited at p. 61]

[WSV+03] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield. Toward optimal rate control: a study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate. In *Proc. SPIE Visual Communications and Image Processing*, volume 5150, pages 198–209, Lugano, Switzerland, June 2003. [cited at p. 27]

[X26] X264. available: `http://developers.videolan.org/x264.html`. [cited at p. 62, 65]

[Xip] Xiph.org. `http://media.xiph.org/video/derf/`. [cited at p. 30]

[XOMW10] Y. Xue, Y. Ou, Z. Ma, and Y. Wang. Perceptual video quality assessment on a mobile platform considering both spatial resolution and quantization artifacts. In *Proc. Packet Video Workshop*, pages 201–208, Hong Kong, China, December 2010. [cited at p. 28]

[YH06] K. Yamagishi and T. Hayashi. Opinion model for estimating video quality of videophone services. In *Proc. IEEE Global Telecommunications Conference*, pages 1–5, San Francisco, CA, USA, November 2006. [cited at p. 22]

[YMM00] T. Yamashita, M.Kameda, and M.Miyahara. An objective picture quality scale for video images (pqs video) - definition of distortion factors. In *Proc. SPIE Visual Communications and Image Processing*, volume 4067, pages 801–809, Perth, Australia, May 2000. [cited at p. 20]

[Youa] Youtube Video. L.a lakers highlights vs new york knicks (12/29/11). available: `http://www.youtube.com/watch?v=cfyykDEt8fs`. [cited at p. 30]

[Youb] Youtube Video. President obama at 2012 aipac policy conference. available: `http://www.youtube.com/watch?v=A0rFbP6KvxY&hd=1`. [cited at p. 30]

[ZCL$^+$08] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh. Cross-dimensional perceptual quality assessment for low bit-rate videos. *IEEE Transactions on Multimedia*, 10(7):1316–1324, November 2008. [cited at p. 27]

[ZKSS03] M. Zink, O. Kunzel, J. Schmitt, and R. Steinmetz. Subjective impression of variations in layer encoded videos. In *Proc. International Workshop on Quality of Service*, pages 137–154, Monterey, CA, USA, June 2003. [cited at p. 97]