VIDEO CONTENT-BASED QoE PREDICTION FOR HEVC ENCODED VIDEOS DELIVERED OVER IP NETWORKS

L. Anegekuh

Ph.D. October 2014

To my late parents and my boys: Lucson and Lemuel Anegekuh

VIDEO CONTENT-BASED QoE PREDICTION FOR HEVC ENCODED VIDEOS DELIVERED OVER IP NETWORKS

by

Louis Anegekuh

A thesis submitted to Plymouth University
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Computing and Mathematics
Faculty of Science and Environment

October 2014

VIDEO CONTENT-BASED QoE PREDICTION FOR HEVC ENCODED VIDEOS DELIVERED OVER IP NETWORKS

Louis Anegekuh

Abstract

The recently released High Efficiency Video Coding (HEVC) standard, which halves the transmission bandwidth requirement of encoded video for almost the same quality when compared to H.264/AVC, and the availability of increased network bandwidth (e.g. from 2 Mbps for 3G networks to almost 100 Mbps for 4G/LTE) have led to the proliferation of video streaming services. Based on these major innovations, the prevalence and diversity of video application are set to increase over the coming years. However, the popularity and success of current and future video applications will depend on the perceived quality of experience (QoE) of end users. How to measure or predict the QoE of delivered services becomes an important and inevitable task for both service and network providers.

Video quality can be measured either subjectively or objectively. Subjective quality measurement is the most reliable method of determining the quality of multimedia applications because of its direct link to users' experience. However, this approach is time consuming and expensive and hence the need for an objective method that can produce results that are comparable with those of subjective testing.

In general, video quality is impacted by impairments caused by the encoder and the transmission network. However, videos encoded and transmitted over an error-prone network have different quality measurements even under the same encoder setting and network quality

of service (NQoS). This indicates that, in addition to encoder settings and network impairment, there may be other key parameters that impact video quality.

In this project, it is hypothesised that video content type is one of the key parameters that may impact the quality of streamed videos. Based on this assertion, parameters related to video content type are extracted and used to develop a single metric that quantifies the content type of different video sequences. The proposed content type metric is then used together with encoding parameter settings and NQoS to develop content-based video quality models that estimate the quality of different video sequences delivered over IP-based network.

This project led to the following main contributions:

(1) A new metric for quantifying video content type based on the spatiotemporal features extracted from the encoded bitstream.

(2) The development of novel subjective test approach for video streaming services.

(3) New content-based video quality prediction models for predicting the QoE of video sequences delivered over IP-based networks. The models have been evaluated using subjective and objective methods.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award.

Publications:

The original work presented in this thesis has been published in refereed Journals and International Conferences.

REFEREED JOURNAL

1. L. Anegekuh, L. Sun and E. Ifeachor "Encoding and Video Content Based HEVC Video Quality Prediction", Journal of Multimedia Tools and Applications, Springer, Vol. 67, No. 3 Dec. 2013.

REFEREED INTERNATIONAL CONFERENCES

1. L. Anegekuh, L. Sun and E. Ifeachor, "A Screening Methodology for Crowdsourcing Video QoE Evaluation" IEEE Globecom conference, Austin, TX USA, 8 – 12 December 2014.
2. L. Anegekuh, L. Sun and E. Ifeachor, "Encoded Bitstream based Video Content Type Definition for HEVC Video Quality Prediction" IEEE ICC conference, Sydney, Australia, 10 – 18 June 2014.

Signed... Louis Anegekuh

Date... 30th October 2014

Word Count (45,719)

Acknowledgments

This thesis would not have been possible without the support and guidance of many people. First, I would like to thank my first supervisor and director of studies, Dr. Lingfen Sun for her professional guidance, encouragement and patience throughout this project, for the benefit of her wide knowledge and vision for this project and for the tremendous amount of time and efforts she spent to ensure the high quality of my papers and this thesis. I would also like to thank my second supervisor, Professor Emmanuel Ifeachor for his guidance, critical comments, time and effort spent with discussing my work and giving me the best advice ever – publish, get other people to pick the holes in your work, remain thorough! It was the advice given to me by him on my first day and continuous reminders ever since that motivated me to aim high and publish only on renowned conferences and journals as part of my PhD, his invaluable comments ensured high quality publications and this thesis, and for that I am forever grateful.

I would also like to thank the members of Signal Processing Multimedia Communications (SPMC) Research group: Dr. Emmanuel Jammeh, Dr. Is-Haka Mkwawa and my other colleagues for all the help, support and constructive discussions throughout this project. It has been a great pleasure to be with the SPMC group. I would also like to take the opportunity to thank the numerous reviewers who reviewed my papers. Without their constructive comments, I would not have been able to raise the standard of my publications.

I would also like to thank my family, brothers and sisters for their love and support and to my late father for being a source of inspiration to me. I would also like to thank my late mum who could not live to see this day. Her encouragement, support and the motivation made me work harder. Mum, I love you and will forever miss you!

Lastly, this thesis is dedicated to my beautiful kids and their mum for their endless love, support and for bearing with me!

Table of Contents

List of Tables

List of Figures

List of Abbreviations and Glossary

3G	Third Generation
4G	Fourth Generation
ACR	Absolute Category Rating
AMA	Average Motion Activity
ANOVA	Analysis of Variance
ANSI	American National Standards Institute
APA	Average Picture Complexity
AQoS	Application Quality of Service
AVC	Advanced Video Coding
CBR	Constant Bit Rate
CT	Content Type
DCR	Degradation Category Rating
DSIS	Double Stimulus Impairment Scale
ECDF	Empirical Cumulative Distribution Function
FR	Frame Rate
HD	High Definition
HDTV	High Definition Television
HEVC	High Efficiency Video Coding
IP	Internet Protocol
IPTV	Internet Protocol Television
ISO/IEC	International Organisation for Standardization/International Electronics Community
ITS	Institute for Telecommunication Science
ITU	International Telecommunication Union
ISO	International Organization for Standardization

IQR	Interquartile Range
LTE	Long-Term Evolution
MOS	Mean Opinion Score
MPEG	Moving Pictures Experts Group
MV	Motion Vector
NAL	Network Abstraction Layer
NS-3	Network Simulator Version 3
NTIA	National Telecommunications and Information Administration
NQoS	Network Quality of Service
PCA	Principal Component Analysis
PDP	Packet Data Protocol
PDU	Packet Data Unit
PQoS	Perceptual Quality of Service
PSNR	Peak Signal to Noise Ratio
QoE	Quality of Experience
QoS	Quality of Service
QP	Quantization Parameter
RMSE	Root Mean Squared Error
RTCP	Real Time Transport Control Protocol (IETF)
RTP	Real Time Transport Protocol (IETF)
RTT	Round Trip Time
SAD	Sum of Absolute Difference
SDP	Session Description Protocol
SI	Spatial Information
SSIM	Structural Similarity Index
SSCQE	Single Stimulus Continuous Quality Evaluation
ST	Spatio-Temporal

SVC	Scalable Video Coding
TB	Transport Block
TCP	Transport Control Protocol
TI	Temporal Information
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
VBR	Variable Bit Rates
VQM	Video Quality Metric
VoD	Video on Demand
VoIP	Voice over Internet Protocol
VQEG	Video Quality Expert Group

Chapter 1

1. Introduction

This chapter presents the motivations behind the project, the fundamental research questions, and the aims and objectives of the project. Furthermore, the chapter highlights the main contributions of this thesis.

The chapter is arranged as follows. Section 1.1 presents the motivations behind the project. The research questions are given in Section 1.2. Section 1.3 presents the project aims and objectives. The major contributions are summarized in Section 1.4. A brief overview and the organisation of the thesis are given in Section 1.5.

1.1 Motivations

Major advances in multimedia coding (e.g. compression efficiency) have enabled significant reduction in transmission bandwidth requirement [1]. For example, the recently released HEVC video codec [2] halves the transmission bandwidth requirement for the same quality when compared to H.264. Furthermore, H.264 also halves the transmission bandwidth requirement for the same video quality when compared to MPEG-2 video codec. Additionally, advances made in fixed and mobile network technologies have also resulted in an increase in available network bandwidths from 2Mbps for 3G networks and up to 100Mbps for 4G/Long-Term Evolution (LTE) networks. The reduction in multimedia transmission bandwidth requirements and the increased network capacities have increased the proliferation of multimedia applications and novel business opportunities.

However, the success of current and future video applications will depend on the end-users' perceived quality of experience (QoE). Thus, how to measure or predict the QoE of delivered services represents an important and inevitable task for both service and network providers.

Different approaches have been proposed to measure or predict the QoE of delivered video applications. These approaches employ either objective or subjective methods. Objective methods measure the quality of a video application by using different quality assessment models (i.e. media, parametric packet, bitstream layer models, etc.) [3] while subjective methods ask participants in a video application testing exercise to grade the quality of a video application on a five-point Mean Opinion Score (MOS) scale which may range from "bad" (1) to "excellent" (5). Subjective test measurements are typically conducted in a laboratory (controlled) environment where different opinions about the multimedia application (e.g. video) under test are collected from a test panel that is supervised by a moderator. Because of the controlled nature of this test, laboratory based subjective testing can be accurate and reflect the users' perceived quality of the application under test.

Although laboratory based subjective methods tend to offer a better indication of users' perceived quality, this approach can be time consuming and expensive because a large sample of participants is needed to obtain results that are statistically meaningful. Furthermore, it is also not possible to use subjective testing in real-time to estimate video quality.

Considering the costs and time demands posed by subjective tests, recently, researchers have proposed cheaper and less time consuming approaches such as crowdsourcing subjective testing method [4] [5]. Crowdsourcing is a process of assigning tasks that would have traditionally been undertaken by contractors to anonymous uncontrolled internet crowd [6]. Crowdsourcing in essence is an extension of outsourcing principles where tasks are assigned to anonymous internet crowd rather than to a specified group of employees. Because of the uncontrolled nature of crowdsourcing approach, results obtained from this test may be inaccurate if not properly screened.

Although subjective (i.e. laboratory or crowdsourcing based test) evaluation of quality may be the most reliable method of determining the QoE of users of multimedia applications, it could be time consuming and hence the need for objective methods that can produce results that are comparable with those of subjective testing. Objective measurements can be performed in an intrusive or non-intrusive way. Intrusive measurements require access to the original sequence in order to compare the original and the impaired sequence while non-intrusive measurements predict video quality from network and application related parameters and does not require access to the source video. Examples of intrusive quality measurements metric include Peak-Signal-to-Noise-Ratio (PSNR), Structural Similarity (SSIM) and Video Quality Metric (VQM) [7]. In real-time quality prediction, non-intrusive measurements are preferred to intrusive quality measurement approaches because of their abilities to measure quality without the need for the original video sequence.

Current non-intrusive video quality assessment approaches are mainly based on coder parameter settings and network quality. These approaches are limited because videos compressed and transmitted over an error-prone network have different quality measurements even under the same encoder setting and network quality. Parameters such as the video content type may have an impact on quality measurement and there is a need for video content type to be quantified and used as variable in quality measurement.

Existing non-intrusive video quality measurements that are based on encoder related QoS parameters are presented in [8] [9] [10] [11], the ones that consider network impairments are presented in [12] [13] [14] [15] and those that consider video content types only are given in [16] [17] [18].

Although there is an on-going effort to quantify and use video content type as a variable in modelling video quality, current efforts are focused mainly on classifying videos into a finite and in some cases a small set of video content groups. This may be limited because different videos have different spatiotemporal characteristics and in theory, the number of video content types is infinite. Quantifying video content types into a finite and discrete set of content types is limited. In fact, the grouping process itself may be inappropriate and sometimes inaccurate given that a video sequence may contain scenes that fall into several different groups.

This project specifically seeks to develop novel subjective test approaches (e.g. crowdsourcing testing) and to address the measurement of HEVC High Definition (HD) videos delivered over IP based network. The video quality measurement takes into account parameters related to encoder setting, network quality and video content type. The work is important because it provides the basis for the content type of a video sequence to be quantified and use as an additional variable to develop video quality measurement metrics that enables effective monitoring and provisioning of videos with acceptable quality

These models have potential applications in several areas, e.g.

- Prediction of initial encoded video quality by content and network providers.
- Prediction of end-to-end video quality in an objective and non-intrusive manner in real-time and for any video content type.

It should be noted that, HEVC Test Model under Consideration (TMuC) was released in 2010 [19], this model allows research to be carried out on initial testing on the codec. The first phase of this project was spent testing the model. This led to the identification of the parameters that were used to quantify video content type and subsequently used to develop quality prediction models that are based on HEVC encoded videos delivered over IP based networks.

1.2 Research Questions

This thesis seeks to address the following research questions:

Q 1) What is the impact of parameters associated with the video codec, IP network impairments, and video content type on video quality?

This led to a significant research to investigate the impact of parameters related to video codec settings (e.g. QP settings), network impairments (e.g. packet loss), and content type on video quality. A fundamental investigation of the impact of these parameters on perceived video quality is undertaken using both objective and subjective test results. The codec used in this thesis is the newly released HEVC codec. This work will be discussed in chapters 3, 4 and 6.

Q 2) What is the impact of video content type (CT) on quality? And how should the content type of a video sequence be objectively quantified?

Research has shown that video sequences encoded and streamed over an error-prone network experienced different levels of quality degradation and susceptibility to network impairments. This led to a fundamental research to investigate the impact of content type on video quality and to further develop metrics based on video spatiotemporal features to quantify the content types of different videos objectively. Considering the significance of video content type in video quality measurements, it is vital to find a method to quantify the content type of all types of video sequences. To do this, different approaches of quantifying video content type were investigated. Firstly, the pixel differences between successive frames of a video sequence were computed using the sum of absolute difference (SAD). Secondly, the spatiotemporal (ST) features of video sequences from an encoded bitstream were extracted to develop a metric that quantifies video content type. The content type metric developed was used together with encoding and network parameters to develop a non-intrusive prediction models.

This work will be discussed in chapter 4.

Q 3) How should subjective test data be screened to identify invalid test data?

Subjective video quality evaluations are typically conducted in a controlled laboratory environment where different opinions about a multimedia application under test are collected from test panels that are supervised by the test moderator. This lab based testing tends to be expensive and time consuming. Because of the expensive nature of this test, other testing methods such as crowdsourcing have been proposed, this approach subcontract subjecting testing tasks to anonymous internet users rather than contractors. Because of the anonymity and uncontrolled nature of the approach, data collected from crowdsourcing tend to be very unreliable. To identify and screen data from unreliable evaluators in subjective video testing environments (i.e. laboratory and internet users), testing platforms that capture evaluators' own data were developed. Based on the evaluators' data, an algorithm for screening unreliable was developed.

This work will be discussed in chapter 5.

Q 4) How should the perceived quality of videos delivered over IP based network be predicted non-intrusively?

This question was addressed by first establishing the relationships between QP, Content type (CT), PLR and video quality.

Once the relationships between video quality and QP and CT and PLR were established, this led to the following research question:

> *How should the QP, CT and PLR parameters be combined to predict the quality of different video sequences non-intrusively?*

This enabled the development of regression-based models to predict video quality. Work presented in chapter 6 describes the development of regression-based models for video quality prediction. The data sets used for model development were generated using both objective and subjective test methods. The data generated by objective methods were based on recommendations made by the Joint Collaborative Team on Video Coding (JCT-VC) which published a list that recommends the conditions under which HEVC should be tested [20] and ad-hoc subgroup (AHG 14) of the JCT-VC loss simulator [21]. Subjective data on the other hand, was based on laboratory and crowdsourcing testing [4] [5].

To evaluate the performance of the proposed video quality models, subjective data not used in model derivation were used. The quality of each sequence encoded with a given encoder setting and transmitted over a given network condition was computed. Using model computed quality values and the subjective data; a comparison analysis between the data sets was performed to determine the closeness between measured and predicted values.

This work is discussed in chapters 6 and 7.

1.3 Project aims and objectives

The main aims of this project are (1) to investigate and evaluate the impact of impairments caused by the encoder parameter settings, network quality of service (NQoS) and content type on video applications delivered over IP networks, (2) to objectively quantify the content type of different video sequences, (3) to investigate and develop novel subjective test approach for video streaming services and (4), to develop and evaluate novel reference-free models that are able to predict the quality of video applications delivered over IP based networks.

Specific objectives of the research are to:

- Investigate the different parameters (i.e. encoding, network and video content type parameters) that impact the delivery of HEVC encoded video sequences over IP based networks and to further identify the parameters that can be used for video quality prediction modelling.
- Develop a metric that quantifies the content type of different video sequences. The developed metric is used together with encoding and network parameters to develop non-intrusive models that are able to predict the quality of different video.
- Develop a screening algorithm that uses crowdsourcing subjective evaluator's own data to determine the validity of evaluator's scores. The valid data are used for model derivation and evaluation.

- Develop and evaluate novel non-intrusive video quality prediction models that predict the quality of different video sequences from a combination of content type, encoder settings and the quality of the transmission network.

1.4 Contribution of Thesis

The contributions of the thesis are as follows:

1. Video content type was found to be a significant parameter which determines how video sequences are impacted by both encoding and networking impairments. Based on this, a new metric was developed to quantify the content type of different video sequences. This new metric is based on the extraction of spatial and temporal features from the encoded bitstream. The developed content type metric is used together with encoding and network parameters to develop non-intrusive video quality prediction models.

(The associated publications are [22].

2. Crowdsourcing and laboratory based subjective testing platforms were designed. These testing platforms allowed rapid assessment of video quality by evaluators in a laboratory or crowdsourcing environment (i.e. internet crowd who have not geographical restrictions). Because crowdsourcing subjective testing approaches are anonymous and unsupervised, a screening algorithm was developed to identify and weed out unreliable evaluators. The subjective results database has been made publicly available at [23] to the research community as currently there is a shortage of video quality assessment database available that combines distortions caused by the encoder and IP network for different types of video content, especially for the newly released HEVC codec.

(The associated publication is [24])

3. New models to predict video quality non-intrusively are proposed. The models are based on a combination of parameters associated with the encoder, the IP network and

video content type. Two models have been developed; the first model is based on regression and uses objective metric i.e. PSNR. This model estimates the initial encoding quality of different video sequences. The parameters used to derive the initial encoding quality include the encoder QP settings and video content type. This prediction model has an accuracy of 95% when predicted PSNR values are compared with those of full reference PSNR. The second model is based on a combination of regression and impact of packet loss on encoded videos (derived using exponential function). This model is derived using subjective data obtained from laboratory and crowdsourcing testing environment. The model estimates the end-to-end quality (i.e. MOS) of different video sequences by taking into account encoder QP settings, network packet loss rate (PLR) and the video content type. The degraded videos under test were generated using HEVC encoder and based on the settings recommendations made by the Joint Collaborative Team on Video Coding (JCT-VC). Furthermore, packet loss was introduced into the encoded bitstreams using ad-hoc subgroup (AHG 14) of the JCT-VC loss simulator [21]. The proposed end-to-end quality prediction model has an accuracy of 93% when predicted MOS values are compared with those of subjective evaluation.

(The associated publications are [25])

1.5 Outline of Thesis

Figure 1.1 shows the outline of the thesis and is described as follows:

Chapter 2 provides an overview of the newly released HEVC video coding standard and also reviews literatures related to the work presented in this project, the techniques used in compressing raw video sequences, the streaming of HEVC encoded videos over IP network and the different methods used in evaluating video quality. Section 2.2 presents the literature review and an overview of HEVC encoding standard. Section 2.3 discusses the techniques used in compressing video sequences with emphasis on HEVC encoding standards. Section 2.4 discusses the streaming of encoded videos over IP networks. Section 2.5 discusses the different QoS parameters that may impact video quality. Section 2.6 provides an overview of different video quality assessment techniques. Section 2.7 summarizes the chapter.

Chapter 3 discusses the impact of QoS parameters on video quality. Section 3.2 presents the different experimental setups used to study the impacts of compression and network impairments on video quality. Section 3.3 compares H.264/AVC and HEVC encoding standards. Section 3.4 investigates the impact of encoding parameter settings such as QP on video quality. Section 3.5 investigates the impact of network impairments such as packet loss on video quality. Section 3.6 summarizes the chapter.

Chapter 4 presents the impact of video content type on video quality and objective methods that can be used to quantify the content type of different video sequences. The video sequences used to study the impact of content type on quality are presented in Section 4.2. An overview of video motion estimation is discussed in Section 4.3. Development of content type metric is presented in Section 4.4. Section 4.5 presents the evaluation of the proposed metric and a comparison with existing pixel-wise metrics. Section 4.6 summarizes the chapter.

Chapter 5 presents an algorithm for screening unreliable crowdsourcing subjective test evaluators, Data set generation, platform that was used for subjective test, based on crowdsourcing and a screening algorithm is presented in Section 5.2. Section 5.3 describes the steps taken to evaluate the performance of the proposed algorithm for screening crowdsourcing subjective test results and limitations of the proposed algorithm. Section 5.4 summarizes the chapter.

Chapter 6 presents the development of reference free video quality models for predicting the quality of HEVC encoded videos delivered over IP based networks. Section 6.2 describes the prediction and performance evaluation of initial video quality prediction model. The subjective encoded, end-to-end video quality prediction is presented in Section 6.3. Section 6.4 summarizes the chapter.

Chapter 7 presents the evaluation of the proposed quality prediction models. Section 7.2 presents work related to video quality evaluation. Section 7.2 presents the evaluation of the proposed end-to-end video quality model with subjective data. Section 7.3 presents a standalone video quality evaluation tool that enables the estimation of video quality over a time period. Section 7.4 summarizes this chapter.

Chapter 8 reviews the achievements of the project, concludes the thesis and suggests future work.

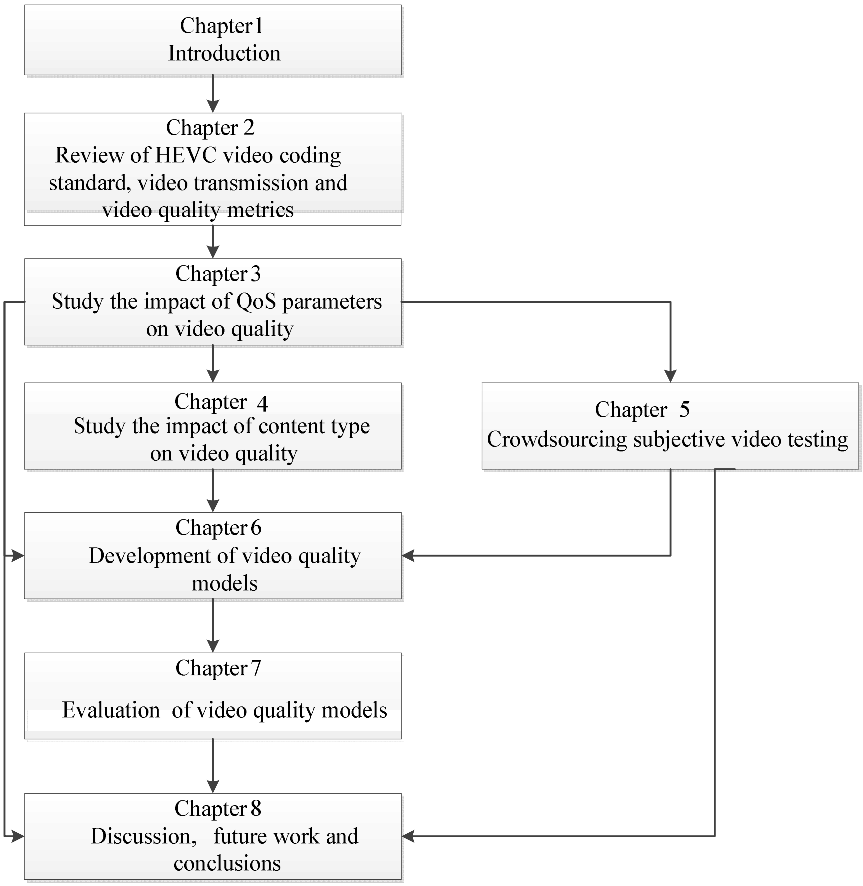

Chapter1
Introduction
Chapter 2
Review of HEVC video coding standard, video transmission and video quality metrics
Chapter3
Study the impact of QoS parameters on video quality
Chapter 4
Study the impact of content type on video quality
Chapter 5
Crowdsourcing subjective video testing
Chapter6
Development of video quality models
Chapter7
Evaluation of video quality models
Chapter8
Discussion, future work and conclusions

Figure 1.1 Outline of thesis

Chapter 2

2. Review of HEVC Video Coding Standard, Encoded Video Transmission and Video Quality Metrics

2.1 Introduction

Major innovations in multimedia devices, compression efficiency and in fixed/mobile network technologies [1] have led to a proliferation of video streaming services. For example, new mobile communication standards such as 4G/LTE offer transmission bandwidth of up to 100 Mbps. Furthermore, the newly released High Efficiency Video Coding (HEVC) standard [2] halves the transmission bandwidth requirement of encoded video for the same perceptual video quality when compared to H.264. Based on these innovations, the prevalence and diversity of video applications are set to increase e.g. video streaming services, interactive gaming, remote teaching etc.

Although advances have been made in compression efficiency and communication technologies, the delivery of these video applications to end-users still remain a challenge because video applications have strict QoS requirements which must be met by content providers in order to satisfy the end-users.

The aim of this chapter is to provide an overview of the newly released HEVC video coding standard and to review the techniques used in compressing raw video sequences, the streaming of HEVC encoded videos over IP network and the different methods used in evaluating video quality. This review is important because it lays the foundation for the work presented in later chapters. Section 2.2 presents the literature review and an overview of HEVC encoding standard. Section 2.3 discusses the techniques used in compressing video sequences with emphasis on HEVC encoding standards. Section 2.4 discusses the streaming of encoded videos over IP networks. Section 2.5 discusses the different QoS parameters that may impact video quality. Section 2.6 provides an overview of different video quality assessment techniques. Section 2.7 summarizes the chapter.

2.2 Literature review and overview of High Efficiency Video Coding - HEVC

This section presents a comprehensive literature review on the work presented in this thesis. Additionally, the section also presents an overview of the codec (HEVC) used throughout the project.

2.2.1 Impact of encoding and transmission impairments on video quality

The delivery of videos with acceptable quality to end users' devices depends initially on encoding parameter such as the QP. The impact that QP has on the encoded video quality is content dependent. When videos of different content types are encoded with the same QP setting, the resulting bitrate and quality are different.

In the existing literature, a large body of research work has investigated the impact of encoder parameter settings on video quality for coding standards such as MPEG-2, MPEG-4 and HEVC. For example, work presented in [8] investigated the impact of quantizer scale factor (referred to as MQUANT) on MPEG-2 encoded videos, the authors concluded that video quality and MQUANT show a linear relationship. In [9], the authors predicted the quality of different MPEG-4 encoded videos by studying the combined impact of application and network related parameters and in [10], the work proposed a no-reference PSNR estimation method based on Laplacian mixture probability density function (PDF) for HEVC encoded video. The proposed method of quality estimation took into account different distribution characteristics of transform coefficient values in various Coding Unit (CU) depth levels of the quadtree. A full reference model called MOVIE index is presented in [26], the model design is based on spatiotemporal features of a video sequence.

Besides the encoding impact on video quality, video contents delivered to end users are also impacted by the quality of the network. This network quality is generally referred to as network QoS (NQoS). In this thesis, the NQoS parameter taken into account is packet loss rate (PLR). Increased PLR may lead to error propagation and degradation in video quality.

The susceptibility of video sequences to PLR is content dependent as videos transmitted over a lossy network shows different levels of impairments in terms of quality [27].

The impact of network impairments on video quality has been addressed by a large number of researchers. For example, work presented in [28], used three different approaches that targeted the trade-offs between estimation accuracy and computational complexity to monitor and estimate video quality based on packet loss using Mean Squared Error (MSE). Similarly, work presented in [12] proposed a macroblock level degradation caused by motion compensation and spatio-temporal error concealment reference free video model based on MSE metric. Authors in [29] further extended this approach by including pixel information into the estimation of video quality. The above approaches used MSE to evaluate the impact of network layer distortions on video quality. Work presented in [27] extracted bitstream features from MPEG-2 videos to determine the level of perceivable degradation in quality caused by packet loss. The authors classified and predicted the visibility of packet loss by using tree classifier - Classification and Regression Trees (CART) and Generalized Linear Model (GLM) respectively. Authors in [7] reviewed the evolution of video quality metrics and presented a state of the art metric that combines video bitstream with network losses. In [30], the authors designed a framework for evaluating HEVC streamed video over a bandwidth limited network, the authors used Peak Signal-to-Noise Ratio (PSNR) quality metric to evaluate the impact of bandwidth reduction on different classes of videos.

Besides using packet-header information (parametric packet-layer models) to estimates video quality, Human Visual Systems (HVS) have been used to estimate the degradation caused by network impairments on video quality. For example, the work reported in [31] incorporated visual attention and saliency information into the GLM method to improve the rates of impairment recognition. Authors in [32] proposed a saliency awareness model to estimate the annoyance of packet loss on video quality. The approach used visual attention information to supplement video quality metric. HVS-based systems used an elaborate mechanism to capture human perception of distortion and spatial differences.

Although, most of the work presented in the literature has either studied the impact of encoder parameter settings on video quality or the impact of network impairments on video or how both parameters impact video quality. The work presented in this thesis is different in that, it focuses on identifying how the spatiotemporal characteristics of a video impact the way through which encoder parameter settings and network impairments impact video quality. The impact of encoding and network impairments on video quality is discussed in chapter 3.

2.2.2 The impact of video content type on video quality

In general, the impairments due to encoding and network transmission are content dependent considering that, videos that are encoded with the same encoder parameter settings and transmitted over the same network have different quality measurements. Recently, video quality researchers have identified other parameters, especially those that are associated with video content types to design video quality models. For example, work presented in [16] and [17] extracted both spatial and temporal motion features of video sequences to determine how video content types impact video quality. The extracted motion features were subsequently used to classify videos into groups and to further develop a content-based video quality metric. Authors in [18] used spatiotemporal features of HEVC video sequences to group and compare the impact of packet loss on H264 and HEVC encoding standards. Although there is an ongoing effort to quantify and use video content type as a variable in modelling video quality, current work is mainly focused on classifying videos into different groups. This approach is limited because different videos have different spatiotemporal characteristics and the limited groups cannot cover all video content types. Furthermore, the grouping process itself may be misleading and inaccurate as a single video may contain scenes that fall into several groups at the same time.

Furthermore, in [33], the authors concluded that video content type has a significant impact on video quality and is the second most significant QoS parameter after encoder type and settings. Authors in [34] developed a bitstream layer novel model to estimate the visibility and the impact of packet loss on H.264/AVC, HD and SD video quality assessment by extracting MV features information from H.264/AVC encoded bitstream to account for spatiotemporal characteristics of video content and classification of packet loss events using support vector regression (SVR) [35]. Although the work presented by these authors used extracted MV, this is fundamentally different from the work presented in this project in that, the extracted MV were only used to identify macroblocks (MBs) that were impacted by loss without any consideration for the overall motion activities and complexities of video sequences.

Beside bitstream layer objective models, a large body of research work has also used human vision system (HVS) to design objective video quality assessment metric. This approach uses the human eye and brain pathway to define video content type. For example, work presented in [36] proposed an HVS-based non-intrusive quality metric based on macroblocks error detection weighted by temporal and spatial saliency maps computed at the decoder side of video delivery process. This HVS-based approach used salient areas in a video sequence to identify different content types. HVS-based systems used an elaborate mechanism to capture human perception of distortion and spatial differences. However, these systems have limitations in modelling the temporal aspects of human vision and distortions in video [37]. Additionally, HVS-based systems only model temporal changes that occur in the early stages of processing the visual cortex [38].

The approach used in this project is different from HVS-based content type metric in that, the TI and SI parameters used in content type metric is based on features/parameters extracted from an encoded bitstream. Additionally, the metric has less computational overhead considering that the variables needed to derive the metric are already calculated in the encoding process.

Although some of the work presented in existing literatures have either used spatiotemporal information to determine video content type or in some cases classified videos into groups in terms of motion. These approaches are limited and fundamentally different from the work presented in this project where a single metric that quantifies the content types of different video sequences is developed and used as an additional variable to develop a non-intrusive video quality model. The impact of video content type on video quality is discussed in chapter 4.

2.2.3 Crowdsourcing video testing

Crowdsourcing is a process of assigning tasks that would have traditionally been undertaken by contractors to an anonymous internet crowd [6]. Crowdsourcing in essence is the further development of outsourcing principles where tasks are assigned to anonymous internet crowd rather than to a specified group of employees. These tasks are small in nature (microtasks), can be accomplished within a few minutes or hours and do not require long term employment. The employer (or test coordinator) has no influence on who can get contracted to do the job. However, employees (i.e. crowd) can be restricted by countries, language and interests. For paid jobs, the employer deposits the payment to the crowdsourcing platform that in turns pays the employees whose job has been approved by the employer. In principle, the crowdsourcing platform mediates between the employers and crowd. Examples of crowdsourcing platforms include Amazon Mechanical Turk[1] (MTurk), InnoCentive[2] and Microworkers[3].

In this work, Microworkers crowdsourcing platform is used because it supports international workers. Additionally, this platform enables the grouping of employees (i.e. employees are grouped by country, ratings, interests and expertise). Because the work presented in this thesis involved video quality assessment, members with the highest success rates in Microworkers YouTube group were chosen for subjective evaluation.

Crowdsourcing results are inherently marred by poor quality. Different design strategies have been proposed to improve on the quality of crowdsourcing results. For example, authors in [39] demonstrated that the quality of results for image labelling tasks could be improved through the design of the task. To prove this, authors varied the number of images per task and campaign and the payments per task in order to screen and achieve the best results. This is limited because, a proper design of task can improve results; however, the method used by the authors cannot be used to screen out unreliable workers. Additionally, authors in [4] proposed a crowdsourcing framework based on paired comparison where a participant is asked to compare two stimuli when evaluating the quality of a multimedia application. In [5], the authors proposed a YouTube QoE based modelling that takes into account stalling as key factors that influence quality, the authors further developed a subjective testing methodology for testing online applications based on crowdsourcing. To detect unreliable results, the authors asked evaluators series of questions related to the watched video and participants with wrong answers to the questions were rejected. The authors further monitored browsers events to detect evaluators who failed to watch video sequences before scoring. The approaches used by the

[1] http://www.mturk.com
[2] http://www.innocentive.com

[3] http://Microworkers.com

authors above are fundamentally different from the approach used in this project in that, they do not fully address the challenges listed in the previous section. For example, none of the strategies deployed by the authors are able to detect participants with multiple accounts and syndication. The approach in this thesis, on the other hand, gathers IP addresses, device, date, location, time and voting pattern information to determine the genuineness of an evaluator through a screening algorithm. Crowdsourcing video quality assessment and the development of crowdsourcing screening algorithm are discussed in chapter 5.

2.2.4 Video quality prediction

The exponential growth of IP based multimedia applications such as video streaming applications (e.g. IPTV) on different devices makes video quality predictions at the user level very desirable. Several studies on video quality prediction can be found in literature. Existing video quality prediction models consider video content features, or the effects of distortions caused by the encoder or network impairments. However, these models are either content blind or have identified video content by classifying videos into groups. Content-blind models are limited because research has shown that video sequences encoded and streamed over an error-prone network have different quality measurements even under the same encoding and network settings (as shown in chapter 3). Additionally, content-blind models can only predict video quality based on 'average' video content, which may not be accurate for specific videos, e.g. fast movement and slow movement. In the existing literature, researchers have also considered a video content type when developing video quality models by classifying videos into groups. This approach is limited because different videos have different spatiotemporal characteristics and the limited groups cannot cover all video content types. Furthermore, the grouping process itself may be misleading and inaccurate as a single video may contain scenes that fall into several groups at the same time.

Although the impact of the encoding and transmission processes on video quality have been investigated subjectively and objectively for MPEG-2, H.264 and HEVC encoding standards [40] [30], video content type which has been identified to have a significant impact on video quality has not been explicitly developed (without classifying videos into groups) and used as an additional parameter in video quality modelling especially for the newly released HEVC standard. For example, work presented in [16] and [17] extracted both spatial and temporal motion features of video sequences to identify how video content types impact video quality. The extracted motion features were subsequently used to classify videos into groups and to further develop a content-based video quality metric. Additionally, authors in [18] used spatiotemporal features of HEVC video sequences to group and compare the impact of packet loss on H264 and HEVC encoding standards.

Work in [26] presents full reference video quality prediction models based on video content features for H.264 video. Reduced reference metrics presented in [41] used raw video features to predict video quality. In [42], a video quality prediction model that combines video content type (defined by grouping videos) application and network level parameters for UMTS network based on PSNR to MOS conversion is proposed. This is limited because the grouping of videos may not adequately reflect the content type of a video.

Whereas, the authors in [43] presented an approach to estimate video quality based on content adaptive parameters and content classification. Content classification is based on motion characteristics determined by MV information and pixel wise sum of absolute differences (SAD). The reported approach classifies videos into four groups. A full reference model called MOVIE index is presented in [26], the model is based on spatiotemporal features of a video sequence. Authors in [34] developed a bitstream layer novel model to estimate the visibility and the impact of packet loss on H.264/AVC HD and SD video quality. The authors extracted MV from H.264/AVC encoded bitstream to account for the spatiotemporal characteristics of different videos and to classify packet loss events using support vector regression [35]. In [44], the author proposed a multipass system for predicting the SSIM of MPEG-2 compressed video. In [90], the authors present a non-reference metric based on spatiotemporal features to estimate blurring of images and video. Authors in [43] present a video quality prediction model for H.264 videos. The model is based on sender bitrate, frame rate and content types. In [45], a theoretical framework is presented for MPEG4 video quality prediction. The framework takes into account the sender bitrate. In [12], the authors proposed NORM (NO-Reference video quality Monitoring), this algorithm assesses the quality degradation of H.264/AVC video affected by channel errors. The proposed metric works at the receiver side where both the original and the impaired video content are unavailable. In [46], a video quality metric based on quantization errors, frame rate and motion speed is proposed. A metric is presented in [47] to estimate the quality of H.264 encoded video sequences using a video decoder. The work used two parameters together (quantization parameter and contrast measure) within the H.264 decoder to give an estimation of subjective video quality. The prediction models presented in these works are from application layer parameters only (encoder based distortion and/raw content features). The derivation of video quality prediction models is discussed in Chapter 6.

2.2.5 Performance evaluation of video quality prediction models

The popularity and success of current and future video applications will depend on the delivered video quality or users' Quality of Experience (QoE). How to measure, monitor or predict the quality of delivered services or QoE becomes an important and inevitable task for both service and network providers. Different approaches have been proposed to measure and evaluate user's satisfaction of multimedia applications. The quality of a video sequence can be evaluated either subjectively or objectively. Recently, researchers have used different approaches to test, evaluate and validate different non-intrusive video quality prediction models. For example, authors in [48] validated the their QoE prediction model by comparing model predicted MOS with MOS scores published by authors in [49]. Additionally, the authors also validated the model by comparing model predicted DMOS with DMOS from subjective data in (live wireless database) [50]. Authors in [51] used the testing dataset to validate their proposed video quality assessment model. Work presented in [52] determined the accuracy of their proposed audiovisual quality estimation model by computing the linear correlation (R^2) and Root Mean Squared Error (RMSE) between model predicted MOS and actual MOS values. Besides validation of quality prediction models, researchers have also used different tools (or applications) to assess the quality of video applications and in so doing, test the performance of different video quality models. For example, work presented in [53] described a tool set for evaluating the quality of MPEG videos transmitted in a simulation network environment. The tool developed by these authors is an extension of the popular EvalVid framework [54]. Authors in [55] designed and implemented a novel open-source tool, named QoE Monitor. This tool consists of new modules for the NS-3 [56] simulator that can be used to perform objective and full-reference quality-of-experience (QoE) assessments. The evaluation of the derived video quality prediction models is presented in chapter 7.

2.2.6 Overview of High Efficiency Video Coding - HEVC

HEVC is the newest video coding standard developed by the working group of ISO/IEC MPEG and ITU-T VCEG (Video Coding Experts Group), jointly published as ISO/IEC 23008-2 and ITU-T Recommendation H.265 [57]. The main goal of the standard is to enable significant improvement in compression efficiency when compared to existing standards, i.e. reduction in bitrate in the range of 50% for almost the same perceptual quality. HEVC standard is based on the hybrid coding approach (i.e. spatial and temporal prediction, together with transform coding). Over the years, compression has improved (Figure 2.1) because increased computational capabilities.

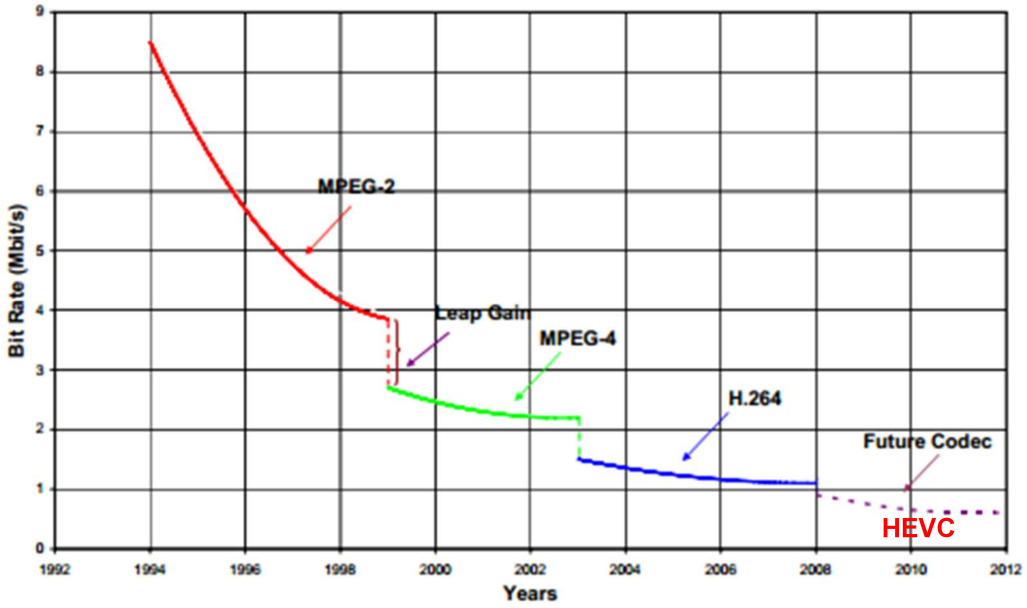


Figure 2.1.Evolution of video codec

The features of HEVC encoding standard are as follows:

Coding Tree Units

The core of the coding layer prior to HEVC encoding standard was based on macroblock. This unit contained one 16 x 16 block of luma sample and two 8 x 8 blocks of chroma sample in a 4:2:0 colour sample. In HEVC, the macroblock (used in previous standards) is analogous to coding tree unit (CTU). The size of the CTU is selected by the encoder and can be larger than a conventional macroblock. The CTU is made up of a luma coding tree block (CTB) and the corresponding chroma CTBs. The size of the luma CTB can be chosen as 64 x 64, 32 x 32 or 16 x 16 samples. HEVC supports the splitting of the CTBs into smaller blocks using quatree-like signalling and tree structure [58].

Coding Units

To increase the coding efficiency of HEVC, each CTU can be split into multiple coding units (CUs) of different sizes. An example of CTU splitting into CUs is shown in Figure 2.2. A CU is made of a square block of luma samples, the two associate blocks of chroma samples and the syntax elements. Each CU can be further partitioned into prediction units (PUs) and a tree of transform units (TUs).

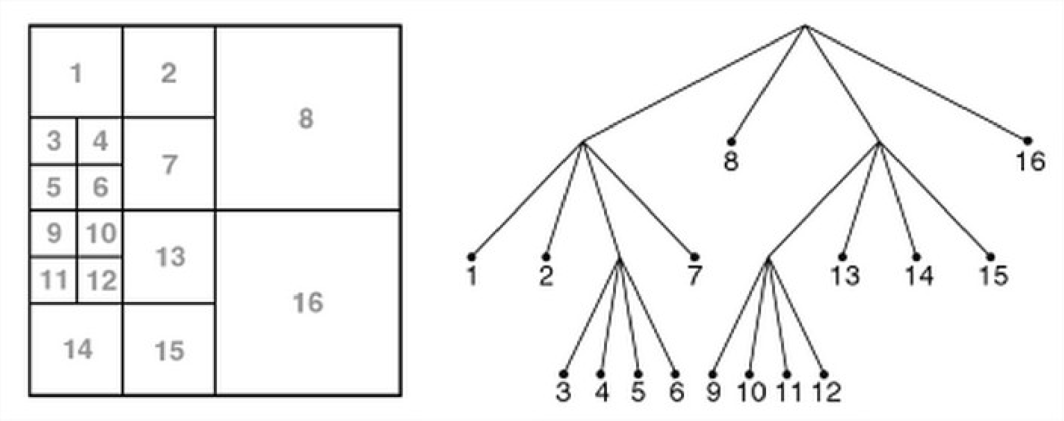


Figure 2.2 Example of the splitting of a 64 x 64 CTU into CUs of 8 x 8 to 32 x 32 luma samples, the numbers indicate the coding order of the CUs

Prediction Units

At the CU level, the decision is made whether to code a picture using intra or inter picture prediction. The prediction unit (PU) has its root at the CU level because a CU can be split into one, or more PUs depending on the prediction mode. The size of the PU can be as large as the root CU and as small as 4 x 4 in luma block size. The PU forms the basis for prediction.

Transform Units

Similar to PU, the transform unit (TU) has its root at the CU level. Each TU is responsible for the transformation of the residuals from PU. In essence, the TU is the basic unit for the transformation and quantization process. The size of the PU defines the size and the shape of the TU.

Motion Vectors

In HEVC, motion vectors (MV) are used to determine the offset by which reference blocks to the current block are located. Typically, MVs are used for inter prediction and are determined per PU and each PU references one or two blocks. Furthermore, HEVC introduces a mechanism known as merge mode [58] where a list of candidates (i.e. previously coded neighbouring PUs) is created for PU being encoded. These candidates are either temporally or spatially close to the current PU. The encoder will typically signal which candidate from the merge list will be used and the motion information from the current PU is copied from the candidate. It should be noted that, HEVC also support advanced motion vector (AMVP) prediction technique which also build a list of candidate MVs.

Picture and Slice

A picture is an image captured in a time t and can be divided into one or several slices as shown in Figure 2.3. Slices are independently coded, and the three types of slices include I, P and B.

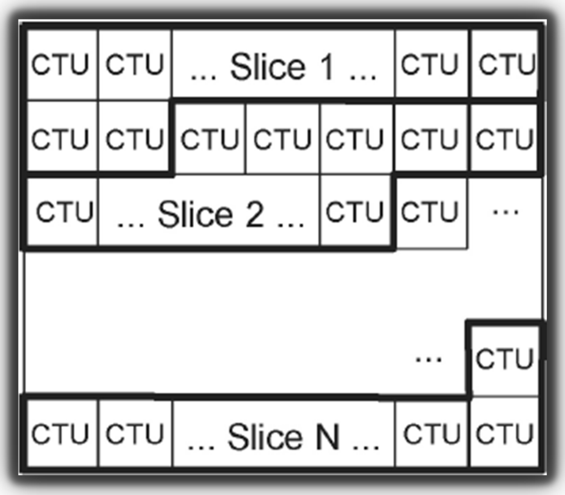

CTU CTU ... Slice 1 ... CTU CTU
CTU CTU CTU CTU CTU CTU CTU
CTU ... Slice 2 ... CTU CTU ...
... CTU
CTU CTU ... Slice N ... CTU CTU

Figure 2.3. HEVC slices

I Slice: This slice uses only intrapicture prediction to code all the CUs of the slice.

P Slice: This slice uses intrapicture and interpicture prediction, however, only one MV per PU is allowed for interprediction.

B Slice: CUs in the B-slice can be coded with inter prediction with biprediction.

Group of Pictures

A group of pictures (GOP) is a series of successive pictures within a coded video stream, it specifies the order in which the video frames are arranged, this is important because it enables capabilities such as random-access during coding. Typically, a GOP will start with an I-frame, followed by P-frame and then the B-frame. The I and P frames are used for referencing. The B-frame on the other hand, uses either the I or the P frame as reference. An example of a GOP structure is shown in Figure 2.4.

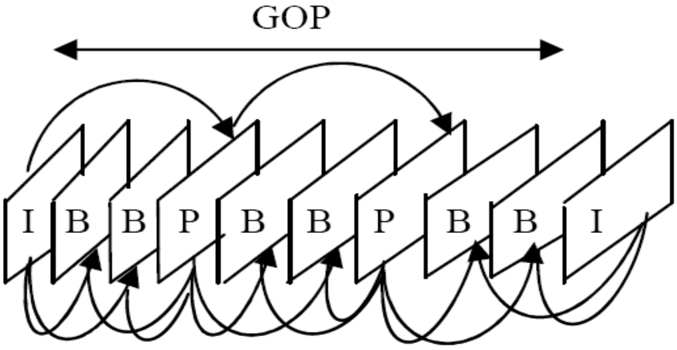


Figure 2.4. GOP structure

2.3 Video coding techniques

Uncompressed video requires a huge amount of transmission and storage space. For example, to store or transmit an HD video of 720 x 1280 resolution, 30 frames per second and 60 minutes full colour will require around 298,60GB of storage space and approximately 82.94MB/s bandwidth to transmit the video. Considering that end users are limited by the available bandwidth and storage space, it is therefore imperative to compress videos to cost effectively transmit or store them.

A video sequence can be compressed because of the spatial and temporal redundancies that are inherent in a video, i.e. high correlation between neighbouring pixels in a single frame and the correlation between adjacent frames. A video compression algorithm operates by removing the spatial and temporal redundancies that exist in a video. However, the process of compressing a video is inherently lossy as the reconstructed signal from the compressed bitstream is often not identical to the original signal. For IP video streaming, the compression standards in use today include: H.263 [59] standardized by ITU, MPEG-4 part 2 [60] standardized by International Organization for Standardization (ISO) Motion Picture Expert Group (MPEG), H.264 [61] (also known as Advanced Video Coding (AVC) and MPEG-4 part 10), standardized by the Joint Video Team (JVT) of experts from both ISO/IEC (International Electrotechnical Commission) and ITU, and the emerging, the newest HEVC [57]. These coding standards have similar principles of compression.

This thesis focused on the newly released HEVC video codec, which uses a new coding structure called coding tree units (CTU) structures. The CTU in HEVC replaces macroblocks that are used in H.264/AVC standard. Unlike H.264/AVC, that divides frames into macroblock of 16x16, 16x8, 8x16, 8x8, 8x4, 4x8 and 4x4, HEVC deploy Coding Units (CU) of 64x64, 32x32, 16x16 and 8x8 pixels. By using larger CU in images with similar characteristics, HEVC can easily achieve high efficient compression through intra-prediction and transforms.

2.3.1 Video and colour sampling

The first step in video compression is sampling in spatial, temporal and colour domain. Sampling in the spatial domain refers to the number of pixels in each of the frames according to frame resolution, whereas, in the temporal domain, sampling refers to the number of frames per second, depending on the frame rate (expressed as frames per second (fps) which is the unique consecutive frames produced within a time unit). In the colour domain, sampling refers to colour space (RGB) and the number of bits used to denote the colour of a single pixel often referred to as colour depth.

A video frame contained pixels, the intensity of pixels in a picture are scalar values, i.e. a unit of data. The captured RGB (Red, Green and Blue) picture is represented by three M x N colour component matrices that consist of q-bit (q is usually 8) and the YUV signals which are created from the original RGB. Each pixel is made up of one luma (Y) component and two chroma components (U, V). The brightness of the pixel is dependent on the luma component while the other two chroma components relate to the colour of the pixel. Considering that the human visual system is more sensitive to brightness (luma) than colour (chroma), the chroma components are under sampled [62] to minimize the storing and bandwidth requirements of a video. The subsampling scheme is usually expressed as a three part ratio, for example, 4:1:1 or 4:2:0; this subsampling shows that for four luma pixels there is only one blue and one red chroma pixels. Therefore, by subsampling the YUV colour components, the data rate can be reduced.

The original bitrate (R_{raw}) of an uncompressed video can be calculated by taking into account the colour components (RGB), the frame rate (fr) and the frame resolution (M x N) and the bit depth (q) as shown in Table 2.1.

Table 2.1 RGB based bitrate calculation

YUV formats	RGB	Frame rate (fr)	Frame Resolution (MxN)	Bit depth q	Bitrate (Mb/s)
Uncompressed (4.4.4)	3	30	720 x 1280	8	663.6
4:2:0	1.5	30	720 x 1280	8	331.8

Although there is a 50% reduction in the size of the bitrate when the YUV format is changed from 4:4:4 to 4:2:0 (as shown in Table 2.1), there is still need to further compress the video as it is not feasible to transmit a video of this size over today's IP networks. The following sections will describe how further compression can be achieved in order to reduce the data rate.

2.3.2 HEVC video compression mechanism

Newer coding standards such as HEVC support a hybrid of spatial and temporal prediction and transform coding. An encoding algorithm that produces an HEVC compliant bitstream would typically proceed with the following 8 steps (4 each for both encoder and decoder) to encode and decode a video sequence [58]. These steps are shown in Figure 2.5 and further discussed in details afterward.

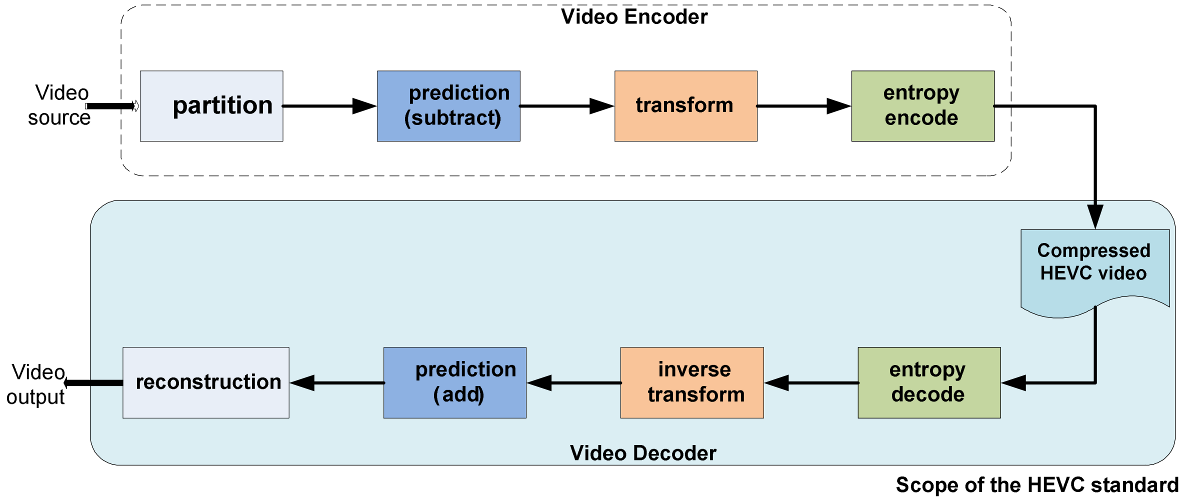


Figure 2.5. Structure of HEVC encoder and decoder

The following can be deduced from Figure 2.5:

- A video coding standard [57] is needed to ensure the interoperability of the encoder and decoder. This standard specifies the compressed format and the method of decoding the compressed video. However, this standard does not specify how a video should be encoded.
- Partitioning is the first process in HEVC encoding. This process splits each picture into block shaped regions called slices. The slices are further broken up into square units known as coding tree unit (CTU) which can be up to 64x64 pixels in size. A video codec will typically process one CTU at a time. The codec store and process a CTU as three components; Y (luma or brightness), Cr (chroma red) and Cb (chroma blue). The Cr and

Cb components are stored using halve the resolution of the Y component because the human eye is more sensitive to brightness than to colour. It should be noted that a CTU can be further partitioned into square regions known as Coding Units (CUs) using a quadtree structure [58] as shown in Figure 2.6. The CUs are the basic units for both intra and inter prediction in HEVC.

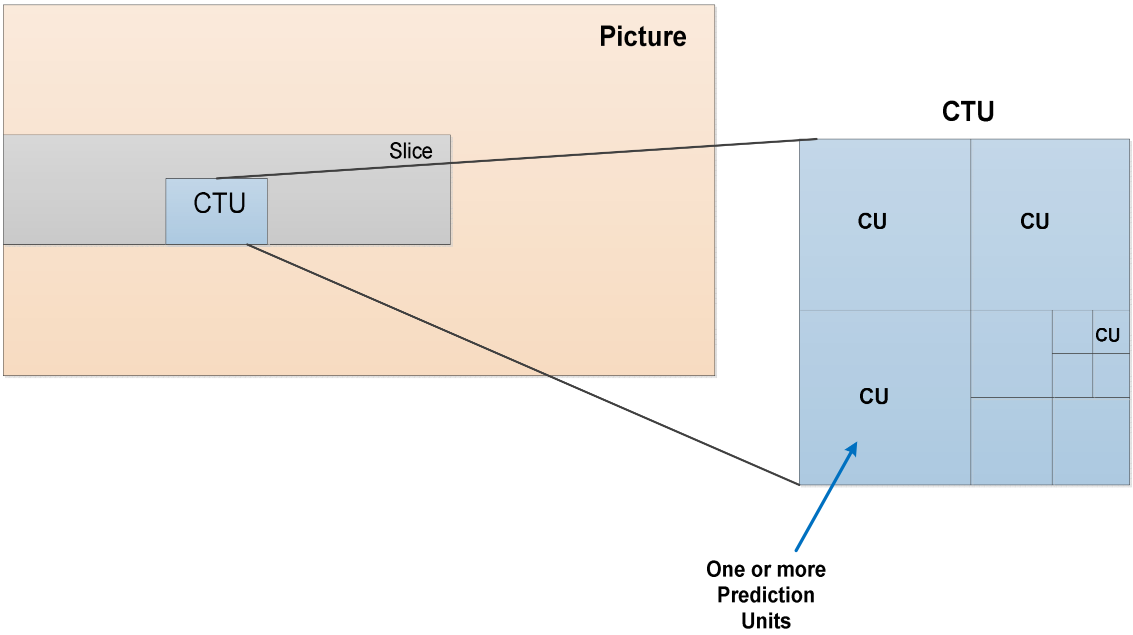


Figure 2.6. Picture, slice, CTU and CUs

- The prediction process begins when the CU is further partitioned into Prediction Units (PUs). HEVC encoder uses both inter and intra predictions. Inter prediction makes full use of motion compensated prediction to predict PUs from other pictures in the stream i.e. prediction from image data in one or two reference pictures (found before or after the current display). Intra prediction on the other hand, predicts PUs from neighbouring data in the same picture. The idea is to generate prediction from previous frame which is already available to the encoder and decoder. This prediction is achieved by searching the previous video frame one block at a time for pixels that matched the current frame. When a good match is found, the pixel is shifted to create the prediction. Furthermore, the predicted frame is subtracted from the original frame to form a residual frame.

- The remaining data (i.e. residual) after intra and interpicture prediction i.e. the differences between the original block and its prediction is transformed using a Discrete Cosine Transform (DCT). The transformed coefficients are scaled, quantized, entropy encoded and transmitted with the prediction information.
- At the entropy encoding stage, the coded bitstream is further encoded using Context Adaptive Binary Arithmetic Coding (CABAC). At this phase of the encoding process, the transformed data is organised and compressed into separate components. These components include quantized (quantization is controlled by the parameter (QP) which range from 0 to 51), transform coefficients, prediction modes, partition information, motion vectors and other header data. This is the final stage of encoding. The video can now be stored or transmitted as a bitstream.
- To decode the encoded video, the process starts by entropy decode extracting the encoded element of the video sequence. This is followed by rescaling and inverting the transform stage of the encoding process. Partitioned units of the original picture are restored, components are added to the output of the inverse transforms (i.e. predicted and the predictions). The final step is to display the video which has been reconstructed by the decoder.

2.3.3 Video compression artefacts

In coding schemes that rely on motion compensation and on a block-based Discrete Cosine Transform (DCT) with the subsequent quantization of the coefficients, degradation in quality is typically caused by the quantization of the transform coefficients which is controlled by the QP. The QP controls the amount of spatial detail that is retained and the encoded bitrate [63].

Increasing the QP value of an encoder leads to detail aggregation and drop in encoded video bitrate and quality.

Even though, other factors such as motion prediction or decoding buffer size impact the visual quality of encoded video, these factors do not directly introduce visual degradation in quality. The following artefacts (though not exhaustive) are associated with video compression:

Blurring: This is exhibited by the reduction of edge sharpness and a loss of spatial details. This is commonly due to coarse quantization, which suppresses high-frequency coefficients. An example of blurriness is shown in Figure 2.7.

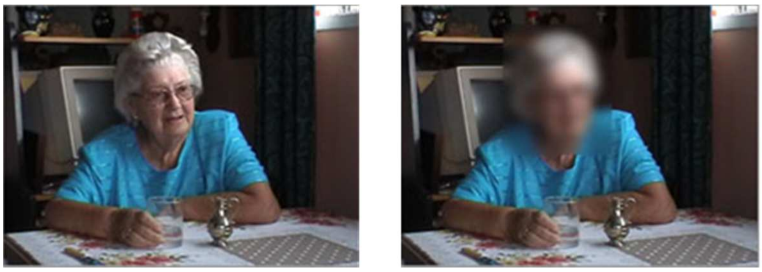


Figure 2.7. Blurring effect

Blocking artefact: This refers to patterns of blocks in a compressed video. This is due to discontinuities at the boundaries of adjacent blocks in block-based coding schemes where individual blocks are independently quantized. Newer coding standards such as HEVC employ a deblocking filter followed by Sample adaptive offset (SAO) filter to reduce deblocking artefacts. An example of blocking is shown in Figure 2.8.

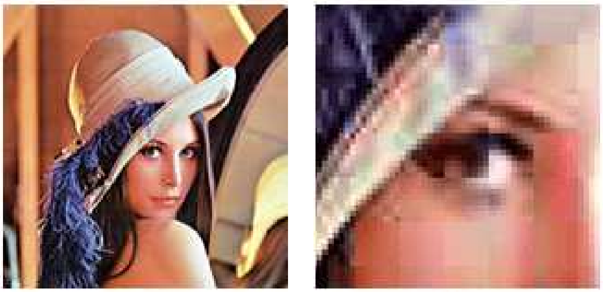


Figure 2.8. Blocking artefact

Jerkiness: This refers to object motion disorder caused by insufficient motion compensation or low temporal resolution. Typically, this happens as a result of poor performance of poor motion estimation.

Colour bleeding: This is the smearing of colours between different strong areas of the chrominance. This happens because of the suppression of high-frequency coefficients chrominance components. Colour bleeding will typically extend over an entire CTU or macroblock.

2.4. Video streaming over IP-based network

Streaming refers to the continuous transmission of data such as audio and video applications from a server to a client over an IP based network. Because of increased availability of internet bandwidth and compression efficiency, audio and video streaming technologies have increased over the years. Streaming applications such as video-on-demand and IPTV offer real-time services.

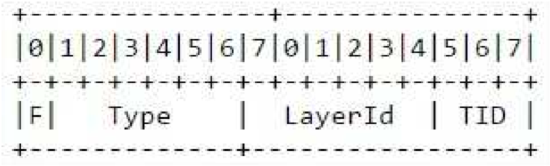
The streaming of video applications over IP packet networks uses the Real-Time Protocol (RTP) together with Real-Time Control Protocol (RTCP) and User Datagram Protocol (UDP) [40]. It should be noted that, the RTP, RTCP and UDP operate at the application, session and transport layer respectively. RTP provides generic transport capabilities, RTCP deals with feedback, synchronization and user interface. The UDP on the other hand, is responsible for the delivery of the application to the client. Furthermore, UDP also provides error detection through checksum, however, it does not provide Automatic Repeat reQuest (ARQ) mechanism to perform retransmissions.

Although not exhaustive, a streaming service will includes, a set of streaming protocols, media codecs, session description and transport protocols. To stream, an end user obtains a Universal Resource Identifier (URI) that is suitable for his/her terminal. The URI comes from the World Wide Web (WWW) browser; it specifies the streaming server and its address. After the session has been established, session parameters such as the session name, date and time, addresses and ports of terminals, data format, bandwidth requirements etc. are exchange between the end points using Session Description Protocol (SDP). The SDP file is usually delivered with a link inside the Hyper-Text Mark-up Language (HTML) page that users download.

Stream services require at least a content server and a streaming client. However, additional components such as caching, portal, profile and proxy servers can be included to improve the quality of service. For example, a portal server will generally provide search and browsing facilities for convenient access to content, while the profile server is used to store end users terminal capabilities and preferences.

2.4.1 Video payload

A hybrid compression algorithm designed architecture comprises of video coding layer (VCL) and the Network Abstraction Layer (NAL). The VCL represents the video content while the NAL provides the ability to map VCL data onto various network transport layers (i.e. RTP/IP, ISO, etc.). Additionally, the NAL also provide the framework for packet loss resilience. In essence, the slice output of VCL is encapsulated by the Network Abstraction Layer (NAL) of the video encoder to form a Network Abstraction Layer Units (NALU) [40]. NALU are classified into different types, for example, NALU containing data from the VCL are classified as VCL NALU, while NALU containing other associated data, such as sequence and picture parameters, filler data, access unit delimiter, display parameters, picture timing, Supplemental Enhancement Information (SEI) are classified as non-VCL NALU [58]. The structure of a NALU is presented in Figure 2.9.



```
+---------------+---------------+
|0|1|2|3|4|5|6|7|0|1|2|3|4|5|6|7|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|F|   Type    |  LayerId  | TID |
+-------------+-----------------+
```

Figure 2.9. Structure of HEVC NAL unit header

Data to be streamed (i.e. video stream) is encapsulated in NALUs in a way suitable for specific networks. The first field in the structure of the NALU is known as the F (forbidden_zero_bit), in HEVC, this is required to be zero at all time while a value of 1 means a syntax violation. The inclusion of the first bit is to enable the transmission of HEVC encoded videos over MPEG-2 transport systems. The 6 bit type field defines the various NALU types. The layer ID field also known as nuh_layer_id, is a 6 bit field (required to be zero in HEVC). This field is intended for future use. The 3 bits TID field specifies the temporal identifier of NALU plus 1. NALU types of less than 32 are considered to be VCL NALU, these NALU types contained data related to the output of VCL. NALU types of more than 32 are the non-VCL NALUs carrying associated data, such as Supplemental Enhancement Information (SEI), picture parameter set (PPS), access unit delimiter etc.

Different transport protocols and file formats can easily be encapsulated by NALUs. The NALU can be transmitted over a single or multiple RTP sessions. Different packetization modes are supported by the RTP payload. For example, the non-interleaved mode, transmits NAL units in a NALU decoding order while the interleaved mode allows the transmission of NALUs in an out of NALU decoding order [64]. It should be noted that, in non-interleaved mode, the same RTP packet can encapsulate several NALUs of the same picture while in interleaved mode; the same RTP packet can encapsulate NALUs belonging to different pictures. Both interleaved and non-interleaved modes allow for fragmentation of a single NALU into several RTP packets.

2.4.2 Architecture and protocol of IP-based network

This sub section looks at the general constraints that apply to the transmission of the HEVC encoded videos over IP-based network. Current IP-based networks can be divided into two main categories; the unmanaged (e.g. the internet) and managed IP networks [40]. The work presented in this thesis considered both networks and is intended for IP-based video streaming applications such as IPTV. Figure 2.10 shows the conceptual transmission of video applications over IP-based network. It consists of three parts - the sender, the IP-based core network and the receiver.

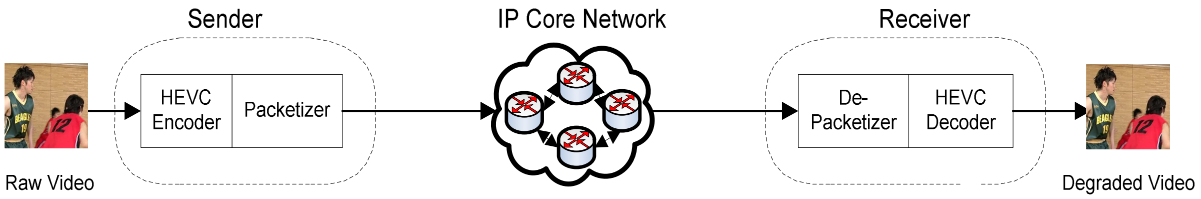


Figure 2.10. Video transmission over IP-based network

At the sender, the raw video content is first digitized and encoded by the encoder to create a bitstream which is further packetized to form the payload part of a packet (e.g. RTP packet). The packet for transmission is formed by adding headers (e.g. IP/UDP) to RTP packets. These packets are then transmitted over IP-based networks which may be impacted by different network impairments (e.g. packet loss, delay, jitter, etc.). On the receiver side, video frames are extracted by stripping off the packet headers from the payload using the depacketizer. These packets are then decoded to recover the received video.

The next two sections give an overview of IP network protocols which are used in video streaming application.

2.4.3 IP network protocol environment

The protocol hierarchy for conversational and streaming applications include;

- **Physical and data link layer**

Considering that IP networks operate over different physical and data link layer protocols and are largely designed to abstract from those fundamental protocols. These two layers are considered in this thesis.

- **Network layer**

In general, IP networks use the internet protocol (IP) [65]. A lot of research work has been carried out on IP design considerations, the actual design and properties of the internet protocol. For the purpose of this thesis, it suffices to say that IP packets are carried independently from the sender to the receiver through a series of routers. The IP is also responsible for splitting and reassembly of service data units (SDUs) that are larger than the maximum transmission unit (MTU) (more discussion on MTU later). The time to send and receive packet varies from packet to packet and the routers in between the sender and receiver can freely discard packets (depending on buffer size and processing time). When packets are discarded or dropped by the router, the receiver observes this as packet loss. Therefore, services offered by IP are known as best effort service. The size of the IP header is 20 byte and it is protected by a checksum to maintain its integrity. It should be noted that, no protection is performed on the payload itself.

- **Transport Layer**

IP networks commonly use two transport layer protocols which include User Datagram Protocol (UDP) and Transmission Control Protocol (TCP). These two protocols are responsible for error control of the payload and application addressing through the port number.

UDP offers a simple and unreliable transport service. UDP header comprises of a checksum which enables the detection and removal of bit errors. Like IP, UDP offers the same best effort services where packets may be duplicated, lost and re-ordered enroute from source to destination.

TCP, on the other hand, offers a byte-based, guaranteed transport service which is based on the concept of re-transmission and timeout mechanisms for error control. Because of this characteristic (delay in the process of re-transmission), TCP is not appropriate for real-time communication.

- **Application Layer Transport**

Considering that, the work presented in this project involves video streaming, the application layer protocol taken into consideration is the Real Time Transport protocol (RTP) [66].

RTP is used for most audio and video applications to transmit data. RTP will typically run over existing transport protocols such as UDP. The function of RTP is to provide applications that take place in real-time with end-to-end delivery services such as payload type identification and delivery monitoring. The timing information of the data as received by the receiver from the sender can be reconstructed from the information provided by RTP. RTP messages contain a message sequence number which allows applications to detect packet loss, packet reordering or packet duplication.

- **Maximum Transmission Unit**

This is the largest size of a packet that can be transmitted from the source to the destination over an IP network without being broken or recombined at the transport and network layer. In general, it is advisable to have coded slice sizes that are not closer or bigger than the size of the MTU because it minimizes the loss probability of the slice. Furthermore, MTU also enhanced the relationship between the payload/header overhead. It is very difficult to identify the end-to-end MTU size of a transmission route between two IP end nodes because the sizes may change dynamically during connection. Based on this, most research accepts an MTU size of 1500 bytes for wired IP links.

In Section 2.3 and 2.4, it was reviewed that compression and transmission of videos over error-prone networks result in visual quality degradation. This degradation in quality may have an impact on how the video is consumed by end users. The upcoming chapters will focus on the different approaches that can be used to determine the impact of visual quality degradation on end users and to further develop quality metrics that can be used to estimate the degradation of quality as a result of compression and transmission of encoded videos.

2.5 QoS Parameters that may impact Video Quality

Several technical factors influence video quality. These factors can be characterized as QoS parameters. In this thesis, the QoS parameters that are considered can be split into QoS affected by the network and QoS affected by the encoder. The impact of the QoS parameters on video quality as perceived by the users is referred to as the Quality of Experience (QoE). QoE is difficult to measure as it goes beyond the boundary of measurable QoS parameters to how users feel about a particular video service or application. This thesis focuses only on the QoS parameters associated with the encoder (or application layer) and the network level parameters (or network layer). Network factors that impact video quality includes network packet loss, network link bandwidth and other factors such as network jitter and delay. When grouped together, these factors can be defined as Network QoS (NQoS). The network factor considered in this thesis is network packet loss. The other factors such as network bandwidth, jitter and delay are not taking into consideration because delay is more significant in voice applications. Voice applications are sensitive to delay and only tolerate a one-way delay of 150ms [67]. Delay exceeding this time frame makes the user feel that the communication is lost. However, with video, delay is not significant, especially due to bigger cache memory now available. Additionally, with the advent of network technologies such as LTE, which has the capability to offer up to 100Mbps, delay and probably network bandwidth, will be even less significant. Hence the reason this project only focused on packet loss rate (PLR) as the network impairment parameter.

The codec related parameter considered in this thesis is the Quantization Parameter (QP), the QP controls the amount of spatial detail that is retained and the encoded bitrate [63] and it is the main initial quality determinant.

The impact that QP has on the encoded video quality is content dependent considering that when videos of different content types are encoded with the same QP setting, the resulting bitrate and quality are different.

The quality of a video delivered over a network is also impacted by impairments such as packet loss rate (PLR). The susceptibility of video sequences to PLR is content dependent. Different encoded videos of different content types transmitted over an error-prone network will have different levels of quality degradation [27].

The content type of a video sequence is determined by its temporal (e.g. movements) and spatial (e.g. brightness, edges, blurriness etc.) information.

The characteristics of encoding parameter settings, network related impairments and video content type and their impact on video quality will be thoroughly discussed in chapter 3 and chapter 4.

2.6 Video quality assessment

Video quality can be assessed using either subjective or objective methods. Subjective quality is the users' perception of quality (ITU-T P.910) [68]. Mean Opinion Score (MOS) is the most widely used metric for subjective testing. The most reliable method of measuring video quality is through subjective test approach. On the other hand, objective measurement can be performed in an intrusive or non-intrusive manner.

2.6.1 Subjective video quality assessment

Video Quality Experts Group (VQEG) and the International Telecommunication Union (ITU) have both defined the subjective methods as a testing method whereby a number of evaluators (viewers) are selected to watch video clips under test in a controlled environment. These evaluators are asked to grade the quality of the video clips on a five-point Mean Opinion Score (MOS) scale which may range from "bad" (1) to "excellent" (5). Subjective testing can be time consuming and expensive because a large sample of participants is needed to obtain results that are statistically meaningful. Considering the costs and time demands by this testing method, recently, uncontrolled testing environments such as crowdsourcing has emerged as a cheaper and quicker alternative to traditional laboratory based quality evaluation for video streaming services (crowdsourcing will be fully discussed in chapter 5).

Subjective test methods are described in ITU-R T.500-13 (2012) [69] and ITU-T Rec. P.910 (1999) [68]. These methods guide subjective test coordinators on the type of viewing conditions, the benchmark for evaluators and the selection of test materials, the procedure for assessment and methods to statistically analyse testing results. ITU-R Rec. BT.500-13 described subjective methods that are specialized for television applications, while ITU-T Rec. P.910 is proposed for multimedia applications.

The most popular and broadly used subjective methods are:

- **Double Stimulus Impairment Scale (DSIS)**

  The evaluators in this test method are shown the impaired sequence after the unimpaired (also known as the reference clip). The reference clip is shown before the degraded pair. Evaluators grade the impaired sequences according to a scale of impairment as, “imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying”. This scale is also known as the 5 point scale where 1 indicates worst quality (“very annoying”) and 5 the best (“imperceptible”). This testing method is typically used in a situation where there is a minimum impairment difference between the impaired and the unimpaired sequence e.g. testing the impact of codec settings on video quality.

- **Single Stimulus Methods** – Evaluators in this test method are shown multiple separate scenes. This method can be run as a single stimulus where the test scenes are not repeated or single stimulus where there is a repetition of the test scenes. This approach of quality assessment uses three different scoring methods, among which include [69]:

    - Adjectival: This is the same as described in DSIS, however, half scales are allowed in adjectival testing method.
    - Numerical: This uses an 11-grade numerical scale.
    - Non-categorical: This uses a continuous scale with no numbers. For example, a large range of numbers can be used i.e. 0 – 100.
- **Stimulus Comparison Method**: This method is typically used in a situation where two matched monitors are available. Based on the comparison of the scenes, the differences between the scenes are scored using either of the following ways:
    - Adjectival: This corresponds to a 7-grade scale which is labelled from +3 to -3: The scale is interpreted as, “much better, better, slightly better, the same, slightly worse, worse, and much worse”.
    - Non-categorical: This is very similar to adjectival; however, this approach uses a continuous scale which has no numbers.
- **Single Stimulus Continuous Quality Evaluation (SSCQE):** The evaluator in this approach of quality evaluation watches a video for about 20 – 30 minutes with no reference to the source video. The sequences under test are graded using a slider continuously perceived quality. The grading scale ranges from ‘bad’ to ‘excellent’. This matches a numerical scale that ranged from 0 to 100.
- **Double Stimulus Continuous Quality Scale (DSCQS):** With DSCQS, the evaluators watch many pairs of short videos which could be 10 seconds long. The sequences under test will typically be made up of test and reference sequences. Each pair appears twice, with random order of both the test and the reference. The evaluators have no knowledge of the existence of reference sequences and are asked to grade both the reference and the test sequences separately on a continuous quality scale. This scale ranges from ‘bad’ to ‘excellent’. This is equivalent to a numerical scale from 0 to 100.

This testing method is described in detail in the ITU-R Rec. T.500-13 document and is mainly intended for television signals.

Other evaluation methods such as Absolute Category Rating (ACR) and Degradation Category Rating (DCR) for multimedia services are described in ITU-T Rec. P.910 [68] and are based on the slight modifications and adaptations of the aforementioned evaluation methods.

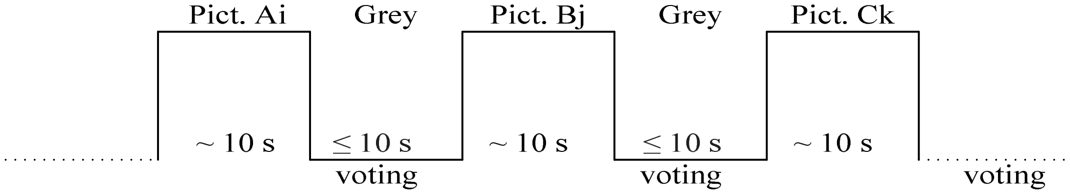
- **Absolute Category Rating (ACR) method**

The evaluators in this test method watch a video sequence without watching the original sequence. After watching the sequence, the evaluators are asked to give their quality rating in terms of opinion score. The quality rating from evaluators is based on an opinion scale as shown in Table 2.2. To obtain the Mean Opinion Score (MOS), the average of the opinion scores from the evaluators is calculated.

Table 2.2 Opinion scale for ACR test

Category	Video Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The time pattern used to present video sequences to evaluators is shown in Figure 2.11. The voting time is equal to or less than 10s.



Ai	Sequence A under test condition i
Bi	Sequence B under test condition j
Ck	Sequence C under test condition k

Figure 2.11 Stimulus presentation in the ACR method

- **Degradation Category Rating (DCR) Method**

Unlike ACR testing method that struggles to pick up slight difference quality (e.g. between 3 and 4), DCR testing method is effective in a situation where a sequence under test has minimum quality different. The degradation in quality is rated by the evaluators by comparing the degraded video sequence to the original (reference). The rating scales or the degradation levels are shown in Table 2.3.

Table 2.3 Opinion scale for DCR test

Category	Video Quality
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

The time pattern used to present video sequences to evaluators is shown in Figure 2.12. The voting time is equal to or less than 10s.

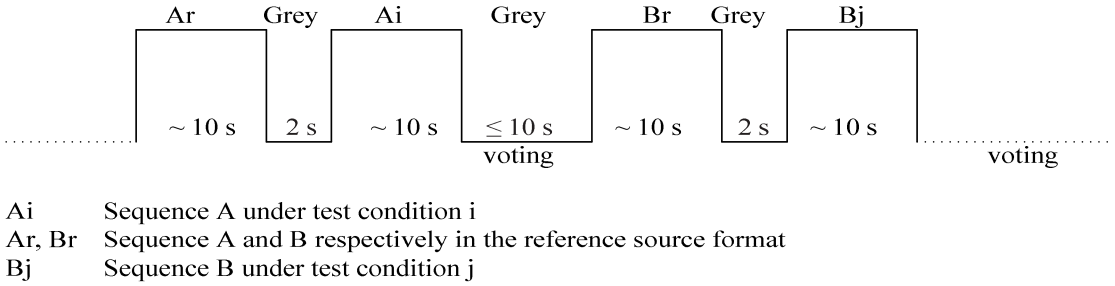

Ar
Grey
Ai
Grey
Br
Grey
Bj
~ 10 s
2 s
~ 10 s
≤ 10 s
voting
~ 10 s
2 s
~ 10 s
voting
Ai Sequence A under test condition i
Ar, Br Sequence A and B respectively in the reference source format
Bj Sequence B under test condition j

Figure 2.12 Stimulus presentation in the DCR method

Pair Comparison Method

In pair comparison (PC) quality evaluation method, a repeated comparison between the sequences under test is performed. The sequences are combined using all possible combinations and presented in all possible orders. The evaluators provide a preference between each pair rather than grading the sequences with continuous or discrete scores.

The time pattern used to present video sequences to evaluators is shown in

Figure 2.13. The voting time is equal to or less than 10s.

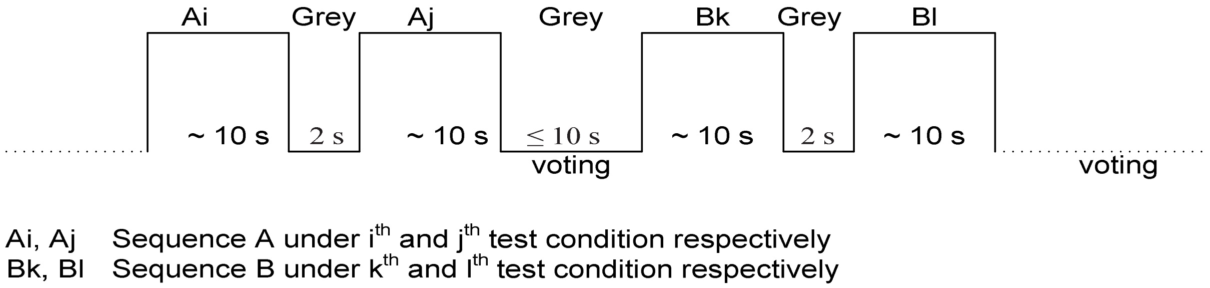

Ai
Grey
Aj
Grey
Bk
Grey
Bl
~ 10 s
2 s
~ 10 s
≤ 10 s
voting
~ 10 s
2 s
~ 10 s
voting
Ai, Aj Sequence A under i^th and j^th test condition respectively
Bk, Bl Sequence B under k^th and l^th test condition respectively

Figure 2.13 Stimulus presentation in the PC method

2.6.2 Subjective tests procedures

To collect data or test, video applications subjectively, factors such as scene characteristics of the application under test, replication, and presentation have to be taken into consideration and incorporated into the design of the experiment. A general description of these factors is presented below.

Scene characteristics

The selection of sequences for subjective testing should be a representation of the data that need to be collected. To ensure subjective test evaluators are not bored during the testing exercises, it is important to select videos with different scenes. Additionally, it is also important to have the same test sequences for all the evaluations.

Replications

ITU-T P.910 recommends the replication of the video sequences; these sequences should be repeated at least twice. The same test sequence can also be repeatedly shown to the evaluators three or four times. This is important because it makes it possible to determine the reliability of subjects and the results they produced.

Order of presentation

The order in which the sequences under test are presented to the evaluators should be randomized and different for the evaluators taking the same test. Although randomized, when the results obtained from the evaluators are being analysed, the presentation order need to be taken into consideration. This is because if an evaluator viewed a ‘bad’ (degraded) sequence, then viewed a ‘fair’ sequence, they may rate is as ‘good’. The most commonly used randomization technique is Latin Squares [70] which is the technique used throughout this project to randomize the video sequences that were presented to subjective viewers.

Evaluators

ITU-T recommends 4-40 evaluators, In general, at least 15 evaluators should participate in the experiment and the evaluators should not be expert in video quality evaluation.

Viewing conditions

The viewing conditions for subjective testing should be uniform for all evaluators, e.g. display equipment, seating position, etc. However, the advent of alternative subjective testing methods such as crowdsourcing makes it difficult to have uniform viewing conditions.

Instructions to evaluators

The evaluators should be briefed about the intended application of the test to be undertaken. These instructions must be in writing form to explain fully what is required from them.

Training session

To familiarise the evaluators to the test, a training session should be included before the start of the testing exercise.

Evaluation

The evaluation of video quality depends on the test method and the type of evaluation scale used as explained in the test methods description section. Different grading scales are possible, for example, five-graded, seven-graded for comparison or even with more points or continuous scales can be used. However, the scale must be clearly explained to evaluators during the training phase of the test. In addition to the grading scale, it is also important to decide if the assessment of video quality concentrates on objects in the video or the whole video sequence.

Subjective video quality assessment can be made up of two phases which include, the initial and the test phase. The initial phase gets the evaluators to familiarise themselves with the test (i.e. instructions and training phase), while the test phase is where the evaluators watch the video sequences and give their opinion scores. Typically, a subjective test session should last for less than half an hour. To avoid brain fatigue or tiredness, the sessions must be split up to make provision for breaks. The general structure of subjective test assessment is illustrated in Figure 2.14.

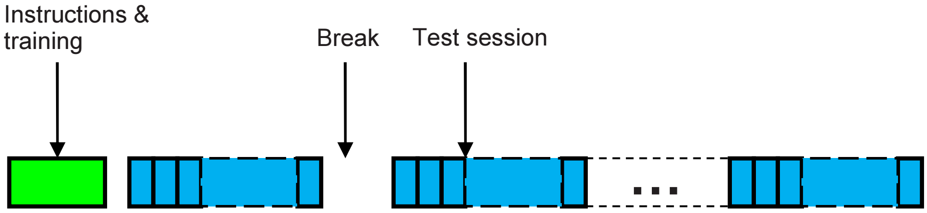


Figure 2.14 The general structure of Subjective tests

2.6.3 Objective quality assessment

The subjective tests are expensive and time consuming because a large sample of evaluators is needed to obtain results that are statistically meaningful. These challenges have limited the implementation of subjective test assessment methods, especially for research purposes. Additionally, subjective test cannot be used in real-time video quality evaluation. Objective testing methods on the hand are quick and easy to set up, thus making them highly desirable for video quality evaluation. Study groups such as VQEG SG9 [71] is dedicated to finding effective objective methods that can be used to obtain results that are comparable to those of subjective video quality evaluation.

Objective video quality measurements are divided into three main areas, this includes, full-reference, reduced-reference and no-reference. Both full-reference and reduced reference approaches of quality measurement are intrusive and are described in section 2.6.4 while the no-reference approach is a non-intrusive measurement and is described in section 2.6.5.

2.6.4 Intrusive video quality assessment methods

Intrusive approach of video quality measurement can be defined as full reference and reduced reference.

Full Reference Method

The full reference (FR) methods require access to the source or original video.

This approach of quality measurement works by comparing the original video to the received degraded video. The comparison between the original and the source video provide an objective value which is used as an indicator of quality. They are impractical for real-time monitoring and estimation where access to the original video is not possible. However, in a lab environment, FR can be used to evaluate reference free prediction models.

Examples of full reference video quality evaluation metrics include, PSNR, SSIM and VQM. In this thesis PSNR has been used extensively for comparison and evaluation of the video quality prediction models.

PSNR

Peak-Signal-to-Noise-Ration (PSNR) is a popular objective video quality metric. PSNR is a full reference quality metric where an objective value for each video frame is obtained by comparing the original frame to the reconstructed video frame. PSNR is defined by the Mean Squared Error (MSE) of two images as shown by Eq. 2.1.

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{2.1}$$

The PSNR value approaches infinity when MSE is close to zero. Therefore a higher PSNR value will indicate a higher video quality, while a low PSNR value will indicate high numerical differences between frames and hence low video quality as defined in Eq. 2.2.

$$\begin{aligned} PSNR &= 10\log_{10}\left(\frac{MAX_1^2}{MSE}\right) \\ &= 20\log_{10}\left(\frac{MAX_1}{\sqrt{MSE}}\right) \end{aligned} \tag{2.2}$$

MAX

MAX_I is the maximum possible pixel value of the image.

SSIM

Developed by Wang et al [72], the Structural Similarity Index Measurement (SSIM) is a full reference quality metric that provides a quality index measure of the similarity between two images. The SSIM metric is calculated using Eq. 2.3. The measure between two images *x* and *y* of common size *N*×*N* is given in Eq. 2.3.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{2.3}$$

This formula is applied only on luma to evaluate the image quality with a maximum value of 1 indicating excellent quality.

Structural dissimilarity (DSSIM) is a distance metric derived from SSIM and is given in Eq. 2.4.

$$DSSIM(x,y) = \frac{1}{1 - SSIM(x,y)} \tag{2.4}$$

VQM

Video quality metric (VQM) is an objective video quality measurement metric developed by the Institute for Telecommunication Science (ITS), the research, engineering branch of the National Telecommunications and Information Administration (NTIA). VQM is adopted as ANSI and ITU standards [73]. VQM measures quality by taking into account the original and the processed videos. The measurement uses the following steps [74]:

- Calibration: - In preparation for features extraction, this step calibrates the sampled video. In so doing, the spatial, temporal, contrast and brightness offset of the video are estimated and corrected with respect to the original sequence.
- Video quality features extraction: - using a mathematical function, quality features, i.e. spatial, temporal and chrominance properties are extracted from the video sequence.
- Video quality parameters calculation: - The quality parameters that describe the quality changes in the sequences are calculated by comparing features extracted from the distorted sequence with those from the reference sequence.
- VQM computation: - The VQM rating is calculated by using a linear combination of the parameters computed in earlier steps.

A block diagram of NTIA VQM general model is shown in Figure 2.15.

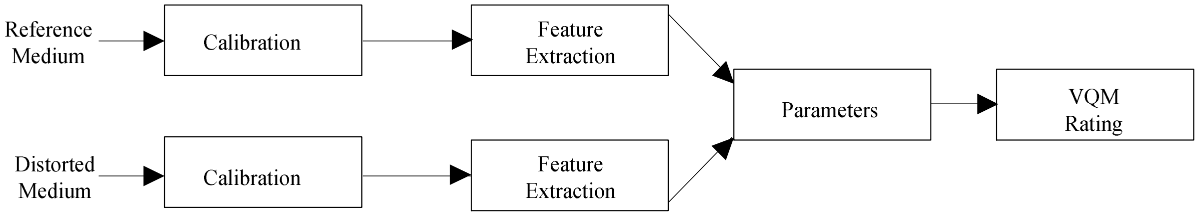


Figure 2.15 Block diagram of NTIA general model

Reduced Reference Methods

Unlike the full reference quality measurement method, reduced reference method uses only some features extracted from the original video sequence. Therefore, if a reduced-reference method is to be used in an IPTV system, the user requirement should specify the side channel, through which the feature data are transmitted.

2.6.5 Non-intrusive video quality assessment

In contrast to both full and reduced reference quality measurement methods, reference free video quality measurement methods do not require access to the original video sequence. The video quality is estimated using information extracted from either the degraded bitstream or from the decoded video clip. In order to monitor the perceptual video quality at the receiver side, it is difficult to use the full-reference methods as they require access to the original video which may not always be available. This makes reference-free methods an attractive option for quality estimation, especially on end users' side as they do not have access to the source video sequence.

In general, objective quality metrics are classified into five main categories by ITU standardization activities [3], these include:

1) Media-layer models: This category of models uses the video signal to compute QoE without requiring any information about the system under test. This type of objective measurement is suitable for codec comparison and optimization scenarios
2) Parametric packet-layer models: These models use packet-header information for QoE prediction without having access to the media signals. This is a lightweight solution for QoE prediction considering that the models do not have to process the media signals.
3) Parametric planning models: QoE prediction using this type of models is based on quality planning parameters for networks and terminal devices. As a result, prior knowledge about the system under test is required.
4) Bitstream-layer models: The type of models uses information from the encoded bitstream and packet layer to predict QoE.
5) Hybrid models: These models are based on a combination of two or more models mentioned above.

The work presented in this thesis takes into account bitstream and packet layers as well as a combination of both layers in designing video quality prediction models.

P.NBAMS

Video quality assessment model for multimedia streaming (e.g. IPTV) scenario has been finalised in ITU-T P.NBAMS also known as Non-Intrusive Bitstream model for the Assessment of Performance of Multimedia Streaming (P.NBAMS) [75]. P.NBAMS is standardised as ITU-T P.1202.2 [76]. It is a reference free bitstream layer objective quality assessment model that estimates the quality of a video sequence by using parameters from the encoded bitstream as input.

P.NBAMS operates in two modes, which include P.NBAMS mode 1 and 2.

P.NBAMS mode 1

Figure 2.16 shows a block diagram of P.NBAMS mode 1. This mode is also known as parsing mode where compressed video bitstream information is extracted by demultiplexing IP video streaming data. The extracted information is analysed and used for quality estimation without the bitstream being fully decoded [75].

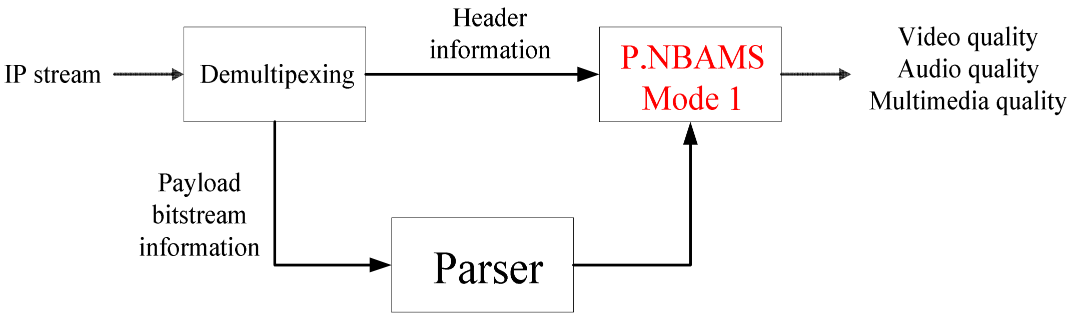


Figure 2.16 P.NBAMS mode 1

P.NBAMS mode 2

Figure 2.17 shows the block diagram of P.NBAMS mode 2. This mode is also known as decoding mode where in addition to the parsed bitstream information, the quality is estimated by partially or fully decoding the encoded bitstream. This mode uses both the parsed bitstream and decoding information (i.e. pixel information) to estimate video quality.

In general, P.NBAMS mode 2 improves the accuracy of quality estimation, but at the expense of higher computational efforts.

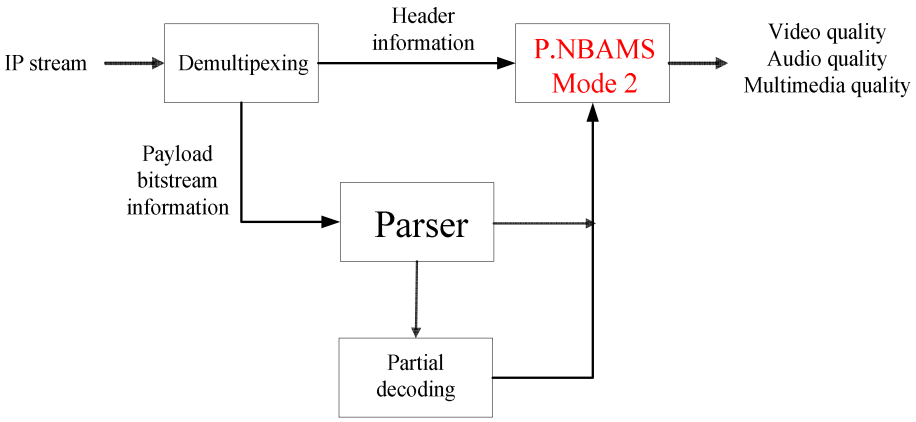


Figure 2.17 P.NBAMS mode 2

Regression-based method

The regression based approach of quality estimation uses a number of parameters in the regression analysis and a model is fitted according to the correlation coefficient that measures the goodness of the fit and the root mean squared error (RMSE). The block diagram of the regression based video quality prediction model developed in this project is shown in Figure 2.18.

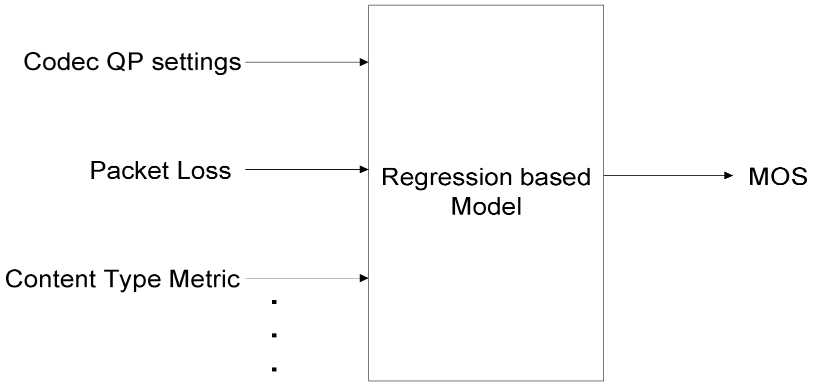


Figure 2.18 Conceptual diagram of regression-based model for quality prediction

2.7 Summary

The purpose of this chapter has been to review the different coding standards, the current literature related to the work presented in this thesis, the different application and network layer parameters that may have impact on video quality and to present the most up to date subjective and objective video quality measurement methods. The description of the subjective video quality measurement (e.g. MOS), and objective video quality measurement, including both intrusive – full reference (e.g. PSNR and SSIM) and reduced reference and non-intrusive video quality measurement (e.g. regression-based) have been presented.

This chapter is important because it has set the background for video quality assessment over IP-based networks and will direct the PhD studies that will be presented from chapters 3 to 7.

Chapter 3

3. The Impact of Encoding and Network Impairments on Video Quality

3.1 Introduction

The streaming of different video contents over IP-based networks is expected to dominate the Internet traffic over the coming years [77]. However, the delivery of these services is very challenging because of the strict QoS requirements of video applications. These requirements may include the level of video compression and the available internet bandwidth. In general, increased video compression may lead to degradation in video quality while limited bandwidth in IP network may also lead to impairments such as network packet loss which may as a consequence impact the quality of video applications.

Determining the impact of these parameters on video quality as perceived by the end user is paramount to content providers as the success of current and future video applications will depend on the delivered video quality.

The aim of this chapter is to study the application and network related parameters that impact the delivery of video applications with acceptable quality, i.e. the minimum quality which a customer could accept the service to end users. A study of these QoS parameters is important because it provides a measure on how the quality of videos delivered over an IP network is impacted and it also underpins the work that will be presented in upcoming chapters.

The chapter is organized as follows. Section 3.2 presents the different experimental setups used to study the impacts of compression and network impairments on video quality. Section 3.3 compares H.264/AVC and HEVC encoding standards. Section 3.4 investigates the impact of encoding parameter settings such as QP on video quality. Section 3.5 investigates the impact of network impairments such as packet loss on video quality. Section 3.6 summarizes the chapter.

3.2 Experimental setup

Figure 3.1 shows the block diagram of the system that was used to provide a realistic investigation of the impact of compression and transmission on video quality.

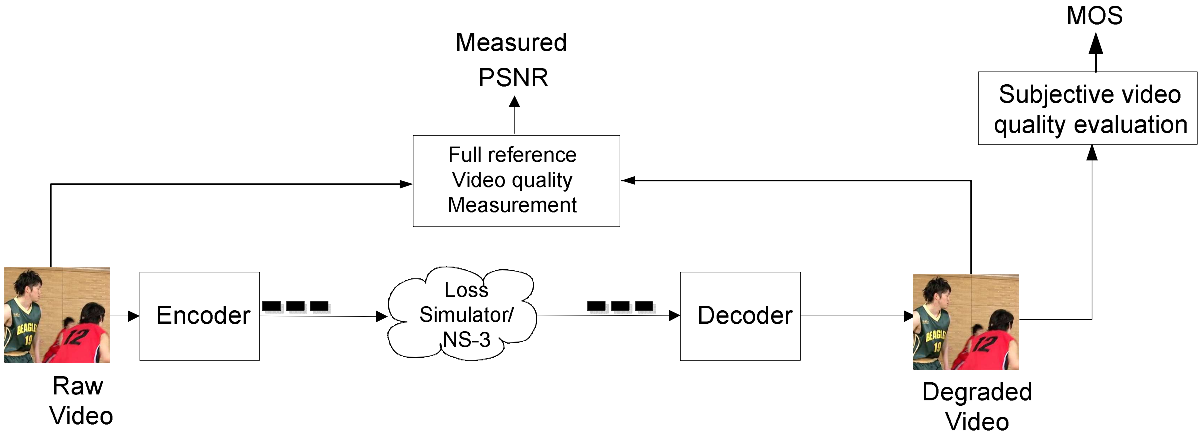

MOS
Measured PSNR
Subjective video quality evaluation
Full reference Video quality Measurement
Loss Simulator/ NS-3
Encoder
Decoder
Raw Video
Degraded Video

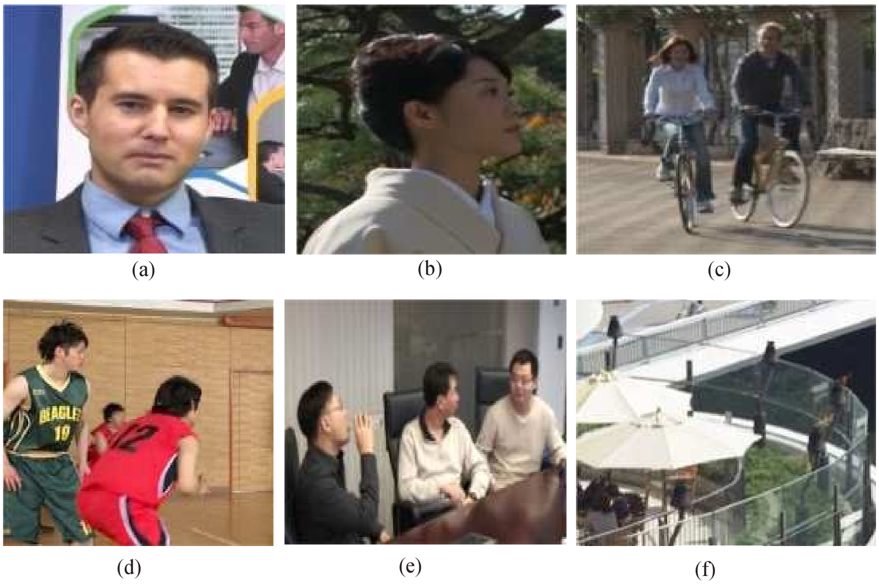
Figure 3.1. Block diagram of a system for the impact of encoder setting and network impairment on video quality

3.2.1 Encoding process

The encoding process involves the compression of a raw video sequence with HEVC version 5.0 and 10.0 codecs using different QP values. The QP values are chosen as recommended by the Joint Collaborative Team on Video Coding (JCT-VC) [20]. Several video sequences which were chosen from the set recommended HEVC test sequences were used. A snapshot of the sequences that were used for subjective test is shown in Figure 3.2. The video sequences are selected based on their spatial and temporal information to enable different classes of video sequences to be used.

Given the need for low complexity and low delay applications for mobile services, Low Complexity (LC) and Low Delay (LD) profile of HM 5.0 and 10.0 was used. A spatial resolution of 1280 x 720 and temporal resolution of 30 frames per second were used for encoding.

HEVC Test model 5.0 and 10.0 used in this work do not have any packet loss concealment implemented and the decoder crashes when decoding bitstreams that have been impaired by packet loss, especially when high packet loss rates have been used. This problem was solved by firstly trapping the decoder errors and secondly, a simple error concealment method was implemented by replacing lost frames with previously received frames.


(a) (b) (c)

(d) (e) (f)

Figure 3.2. Snapshots of video sequences (a) Johnny, (b) Kimono1 (c) ParkScene (d) BasketballDrive (e) Vidyo1 and (f) BQterrace

3.2.2 Network simulation

In the absence of an open source real-time HEVC encoder that can be used to stream and evaluate the impact of network impairments on HEVC encoded videos, in this thesis, the IP-based network was simulated using ad-hoc subgroup (AHG 14) of the JCT-VC loss simulator [21] and NS-3 network simulator [55]. AHG 14 simulator uses random packet loss and takes into account loss scenarios using ITU-T G.10505/TIA 9216 [78] [79].

A packet loss within the four ITU recommended loss patterns (i.e. 3%, 5%, 10% and 20%) [80] was used. Packet loss was introduced into the HEVC encoded video bitstream using the loss simulator [21]. Encoded videos which have losses introduced were decoded to obtain video sequences that may be degraded.

Degraded and original video sequences were used for Single Stimulus Continuous Quality Evaluation (SSCQE) subjective testing.

- **NS-3 video streaming framework**

To test the different NQoS parameters that may impact video quality, an NS-3 streaming platform was built. The simulation framework is shown in Figure 3.3. This framework allows the encoded bitstream to be packetized, streamed, reconstructed, decode reconstructed video bitstream and compute quality. It should be noted that the NS-3 streaming framework used in this project is an extension of the evalvid streaming framework [54]. The framework was extended by adding the video reconstruction component for NS-3.

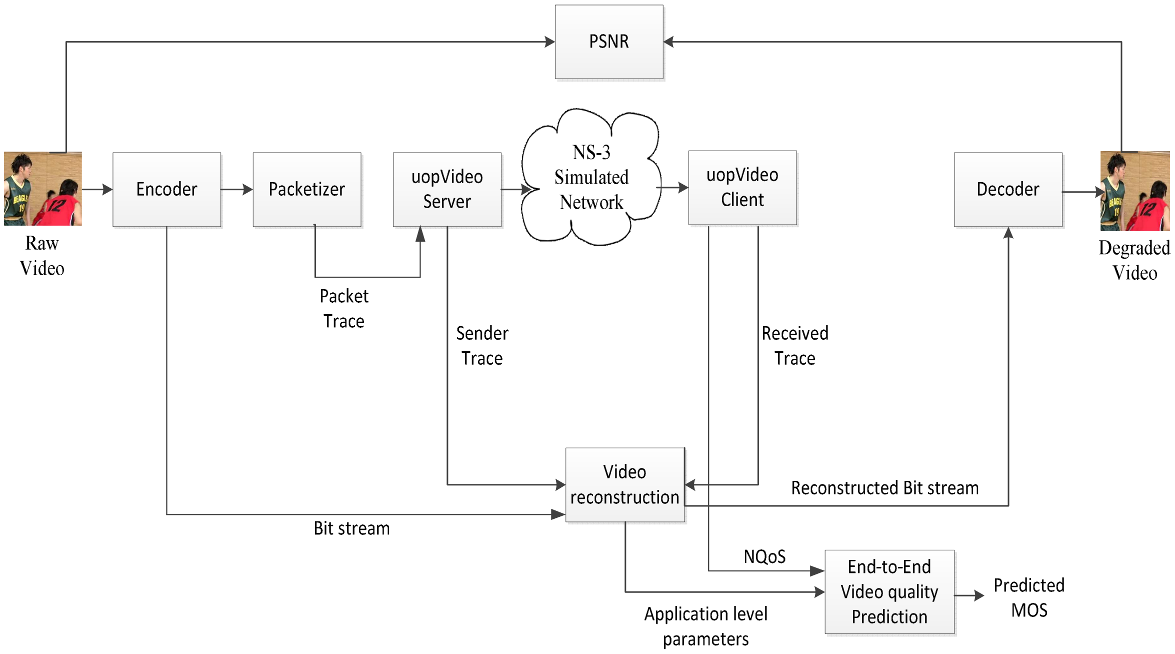


Figure 3.3 NS-3 based framework for video quality evaluation

- **Main components of the framework**

Packetizer: The encoded bitstreams are packetized into RTP packets using the packetizer component which also generates a packet trace file. Tracefiles contained information about packetid, type, size, number of packets (NALU) and the timestamps. The format of the generated trace is shown in Table 3.1 (an extract of Vidyo1 video sequence encoded at QP 17).

Table 3.1 Format of generated trace file

Frame ID	Frame type	Packet Size	Number of UDP packets	Send Time (sec)
0	I	102774	71	0.000
1	B	9775	7	0.033
2	B	13009	9	0.067
3	B	9933	7	0.100
…	…	…	….	….

uopVideo server component: The packets are transmitted/simulated through a point-to-point NS-3 network by the server component. The main task of this component is to create RTP segments, transmit these segments through UDP packets over a simulated point-to-point network. For each transmitted UDP packet, the timestamp, packetid, packet type and the packet payload size are recorded in a sender tracefile. This sender tracefile is used at a later stage for reconstruction.

uopVideo client component: This component receives packets and remove header information. Additionally, the client component creates a received tracefile that contains packetid, protocol type, packet size, and timestamp. This tracefile is used by the reconstruction component to generate a decodable bitstream. Specifically, the client component receives packets, check sequence numbers to determine which packet did not arrive (i.e. packet loss during transmission), and the time of arrival (i.e. timestamps) to determine the one way delay (owd), and jitter. In essence, the Network Quality of Services (NQoS) is obtained from this component.

Video reconstruction component: Once the transmission is over, this component is used to reconstruct the received video. To achieve this, the sender, receiver tracefile and the original encoded bitstream are used. The reconstruction of a potentially degraded video file is processed by copying packets from the sender encoded video (original bitstream) and omitting those packets that have been indicated as lost by the client component.

Decoder: The degraded reconstructed video is then decoded. A simple error concealment was used where lost frames are replaced by the last successful decodable frames [81].

Quality evaluation: The quality of the degraded video sequence can be estimated intrusively (performing a comparison between the original sequence and the degraded sequence using PSNR). Additionally, the quality can also be estimated non-intrusively (using video quality prediction models).

- **NS-3 simulation setup**

In this section, an NS-3 client/server architecture that enables the simulation of video traces transmission over a point-to-point (p2p) network is used. A dumbbell topology where several flows follow the same path and compete for the same bandwidth in a single link was used in this work. The experimental setup is shown in Figure 3.4. The architecture consists of two bottleneck nodes. In addition to video streaming, the architecture also includes TCP background traffic. This traffic is based on Pareto On/Off traffic generator which is an application embedded in NS-3. Pareto traffic is generated according to Pareto an on/off distribution [82] where a fixed rate of packets is sent during ON periods and no packets are sent during OFF periods. The bandwidth between the client/server nodes and the bottleneck nodes was fixed at 100Mbps while the bandwidth between the routers varied between 1, 5 and 10Mbps. The router buffer type is based on a first in first out (FIFO) queue with a buffer size of 100 packets (i.e. a typical buffer size for a fibre line delay [83]). The Maximum Transmission

Unit (MTU) is set to 1450 bytes. This will equate to approximately 1490 bytes when the headers of IP/UDP/RTP header (i.e. 20 + 8 + 12 = 40 bytes) are added.

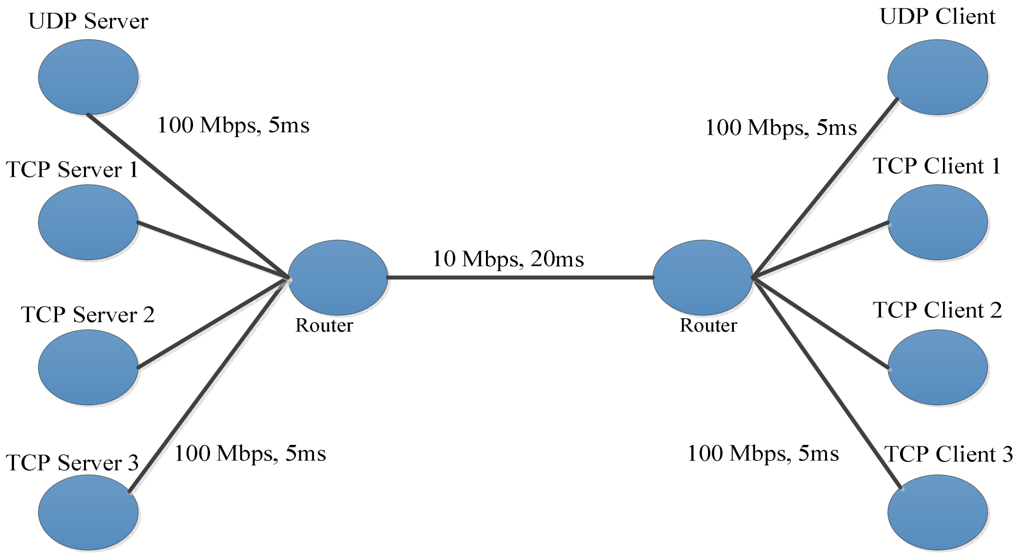


Figure 3.4 NS-3 video quality evaluation topology

It should be noted that, the use of NS-3 may be limited because no real video packets are transmitted (that can enable the implementation of a concealment scheme when packet occurs) as only the sender, receiver tracefile and the original encoded bitstream are used to reconstruct the video which is decoded and the quality calculated.

3.3 Performance comparison between H.264 and HEVC

In this thesis, HEVC coding standard is used because it is the newest codec and it is anticipated to be the direct replacement of H.264/AVC, which is the codec widely used in applications such as broadcast of HD TV signals, camcorders, Blu-ray Discs, mobile video application streaming. H.264/AVC provides a huge bitrate savings when compared to MPEG-2. However, with increasing diversity of video services and the emergence of videos beyond HD formats, e.g. High and Ultra-High Definition (HD and UHD) (i.e. 3840 x 2160 (4K) and 7680 x 4320 (8K) resolutions), HEVC encoding standard that increases the coding efficiency when compared to H.264/AVC and have the designed capabilities to encode videos with higher resolutions of up to 4K and 8K is deemed to be a direct replacement of H.264/AVC.

To determine the performance differences between HEVC and H.264/AVC encoding standards, a comparison between the two standards in terms of PSNR, bitrate and encoding run time is performed.

The performance comparison is based on results obtained from the encoding of the sequences in Figure 3.2. These sequences were encoded using similar encoding configurations from both H.264 and HEVC standards. The comparison starts by looking at the terminological differences between the two standards as shown in Table 3.2.

Table 3.2 Terminological difference between H.264 and HEVC

H.264 terminology	HEVC terminology	Definition
Frame	Frame	Complete video frame
Macroblock (MB)	Coding Tree Unit (CTU)	A basic coding unit with square region
Block	Coding Unit (CU)	A divided MB or CTU
MB partition	Prediction Block (PB) or Prediction Unit (PU)	A rectangular area predicted using inter or intra prediction
Block (transform)	Transform Block (TB) or Transform Unit (TU)	Block of samples to be transformed
Slice	Slice	A continuous sequence of MBs or CTUs
-	Tile	A rectangular shape set of CTUs that can be decoded in parallel

To obtain the results needed to compare both codecs in terms of PSNR, bitrate and encoding run time, the configuration parameters are carefully selected such that similarity was ensured in order to avoid being biased towards one of the codecs. The parameter settings for both codecs are presented in Table 3.3.

Table 3.3 Parameter settings for JM and HM reference software encoder

Parameters	H.264	HEVC
Encoder Version	JM 18.6	HM 9.2
Profile	Main	Low Complexity and Low Delay
R/D Optimization	Enabled	Enabled
YUV format	4:2:0	4:2:0
GOP	4	4
Search Range	32	64
Intra Period	1 Sec	1 Sec
Coding Unit Size / depth	-	64/4
Transform Unit Size (Min/Max)	-	4/32
Sample adaptive offset (SAO)	-	Enabled
Frame rate	30fps	30fps
Resolutions	1280x720	1280x720
Quantization Parameter (QP)	17,22,27,32,37,42,47	17,22,27,32,37,42,47
Number of frames	240	240

The test sequences and Quantization Parameter (QP) settings used for comparison are based on Joint Collaborative Team on Video Coding (JCT-VC) recommendations. All videos were encoded on a High Performance Computing (HPC) that has Dual Intel Xeon Quad Core E5620 with 2.4 GHz Processor and 12GB RAM.

3.3.1 Results and discussion on HEVC and H.264 performance comparison

The performance of H.264 and HEVC encoding standards was compared in terms of video quality (PSNR), the bitrate and the encoding run time for each encoded sequence.

The bitrate of each encoded sequence gives an indication of the bitrate differences between the two encoding standards. The PSNR values indicate the quality that can be achieved with the given bitrate while the encoding run time is indicative of the complexity of the encoding standard. The results are shown in Figure 3.5 and Figure 3.6. The results in Figure 3.5 show that video sequences encoded with HEVC required almost half the bitrate of those encoded with H.264 to achieve almost the same PSNR values. However, it took almost as twice the encoding runtime to encode the same videos at a given QP when HEVC encoded run time is compared with H.264 run time (Figure 3.6.). Results also show that the bitrate of all video sequences decreases with increased QP for the two encoding standards.

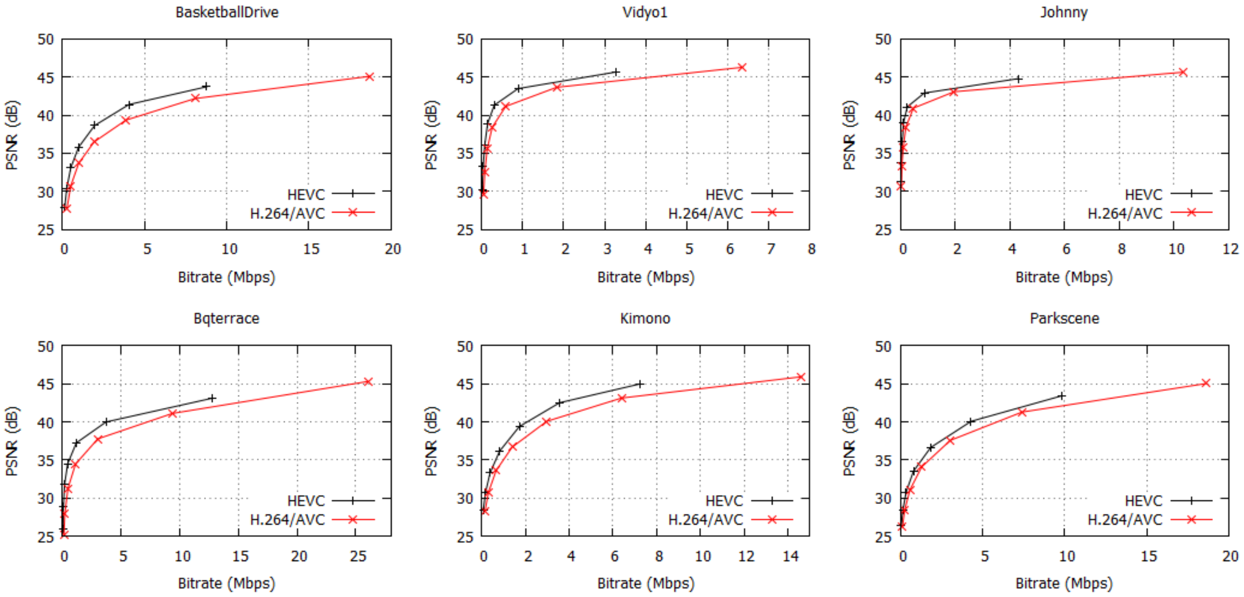


Figure 3.5. Bitrate versus PSNR

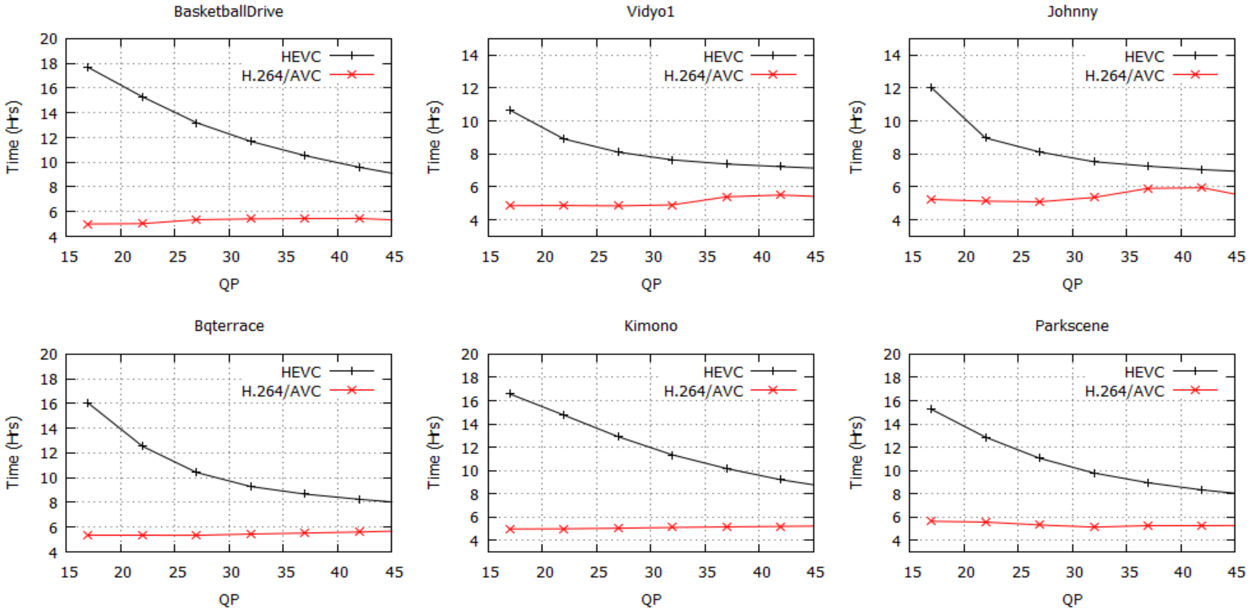


Figure 3.6 QP versus encoding run time

3.4 Impact of video compression on video quality

This section studies the impact of compression on video quality as perceived by end users. The study takes into account the encoder parameter setting such as quantization parameter (QP). The QP controls the amount of spatial detail that is retained during the encoding process and it also determines the encoded bitrate of a video sequence and its quality.

3.4.1 Objective impact assessment of HEVC compression video quality

The initial video quality depends on encoding settings. As expected, lower QP results in higher bitrate (BR) which may lead to increased video quality as shown in Figure 3.7. Results also show that as the QP increases the BR reduces which may as a consequence leads to a reduction in video quality. In addition to reduction in video bitrate, results also show that for the same QP setting, sequences with high temporal complexity and high spatial complexity have higher bitrate when compared to those with low temporal complexity and low spatial complexity. For example, at QP 17 Johnny has a bitrate of around 3.3Mb/s, Vidyo1 4.1 Mb/s Kimono1 6.8Mb/s and BQterrace 13Mb/s.

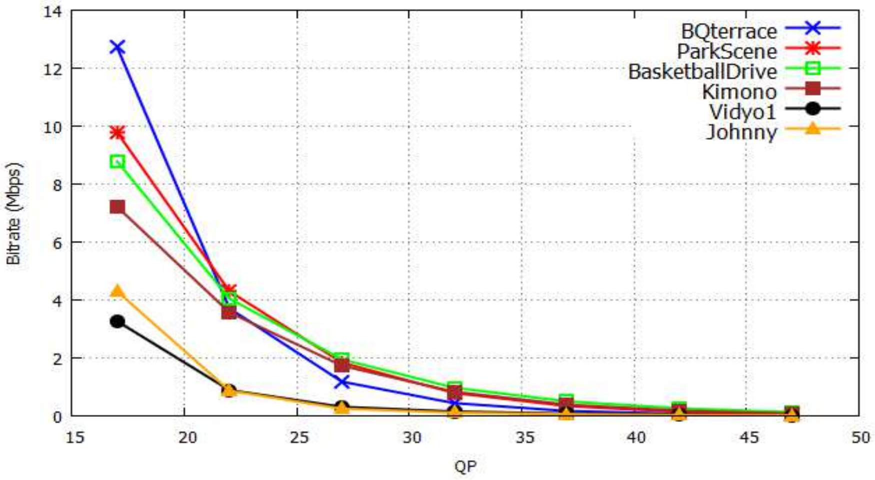


Figure 3.7. Impact of QP on video Bitrate

Figure 3.8 shows the impact of encoder QP settings on video quality. The results show that, for all encoded videos, the quality drops with increased compression. However, the drop in quality is steeper for sequences with high temporal complexity and high spatial complexity when compared to those with low temporal complexity and low spatial complexity. What is also evident from the result presented is that when encoding video, one would need to be mindful of the level of compression because it determines the initial video quality i.e. too much compression might cause higher degradation in initial video quality. For example, to achieve a video quality greater than 37 dB, it is clear from the results (Figure 3.8) that encoding the videos with a QP value of more than 37 will not achieve a video quality of 37dB. Furthermore, all sequences require a different compression level for the same quality because of the differences in spatiotemporal characteristics. For example, Johnny and Vidyo1 video sequences (head and shoulder videos) with lower motion/complexity require lower compression for almost the same quality when compared to ParkScene, BaskballDrive BQterrace and Kimono1 video sequences (fast moving videos) with higher motion activities/complexities. Based on the results presented in Figure 3.7 and Figure 3.8, it can be concluded that the initial video quality depends on encoder setting parameters such as the QP and impact. The impact that QP has on the encoded video quality is content dependent, considering that when videos of different content types are encoded with the same QP setting, the resulting bitrate and quality are different.

It should be noted that, fixed QP settings used to test the impact of encoder parameter on video quality might be limiting considering that in a real world scenario, the QP values vary in order to meet the required bitrate and subsequently video quality. However, the range of QPs used (as recommended by JCT-VC) to encode each sequence is representative of the varied QPs that can be used to meet the required video bitrate and quality requirements of a video.

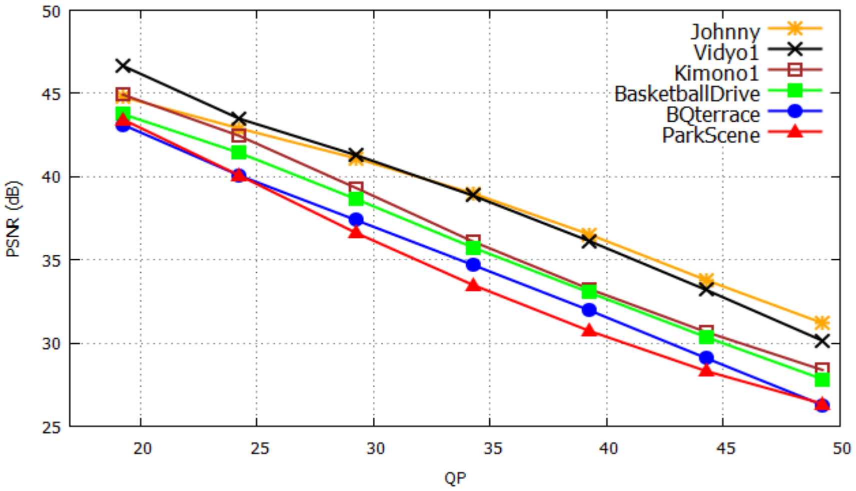


Figure 3.8. Impact of QP on video quality

3.4.2 Subjective impact assessment of HEVC compression on video quality

Although objective video quality assessment metrics such as PSNR are less computationally intensive, they may be limited because they do not factor in the human visual perception. On the other hand, subjective video quality assessment techniques are able to indicate the human visual perception of quality.

In this section, the Double Stimulus Impairment Scale (DSIS) subjective quality rating approach [84] was used to evaluate the impact of compression on video quality. DSIS testing method is commonly used in a situation where the quality differences between the unimpaired and the impaired multimedia application under test is minimal. The subjective test plan followed the ITU recommendations for subjective video testing [69] [68]. A total of 42 impaired videos of 9 seconds each were generated from six reference video for subjective testing.

- **Subjective test design**

Evaluators in this test method are shown the impaired video after the unimpaired video sequence. To restrict the test time of the subjective evaluation to range between 10 and 15 minutes as recommended by ITU, the sequences were randomly split into three datasets. The unimpaired and impaired videos were uploaded to three identical subjective test websites [85] with each website containing two columns, one for unimpaired sequences and the other for the impaired sequences. The impaired video sequences column also had unimpaired videos (used as hidden reference videos) needed to validate test scores. It should be noted that DSIS ratings for reference sequences were only used for referencing and not for analysis.

The voting period was non-restrictive, as participants had the option to watch a video clip more than once before committing to submit their final vote. To track how long participants took to complete the test, all testing websites had hidden timers. And to minimize memory effect, all sequences were randomized using Latin square randomization technique [70].

- **Test participants**

The subjective test was conducted at the Plymouth University computer lab and it took about 2 ½ to complete the test. The lab had 20-inch Philips - 201E1SB LCD monitors with a native display of 1600 x 900 pixels and the highest colour selection. The lab also had a white colour background. All sequences under test were displayed in the original sizes (1280 x 720). Participants gave their ratings to video sequences through computer keyboard or mouse.

A total of 63 students from the School of Computing and Mathematics (SoCM) from Plymouth University took part in the test. This included 43 undergraduates and 20 graduate students and a mix of males and females with majority male. Although no attempt was made to measure the computer skills and familiarity with video testing of evaluators, participants were all computer science students who had a high degree of computer literacy.

- **Test procedure**

As shown in Figure 3.9, the assessor used the first five minutes of a test session to explain to evaluators the type of assessment, the grading scale, the sequences and timing and the voting. This was immediately followed by the distribution of printout papers (random order of distribution) that contained the test web links. This procedure was repeated for all test sessions which had between 10 and 17 evaluators.

To gauge the visible artefacts caused by HEVC encoder settings on video, DSIS ratings were mapped onto a MOS scale from 1 to 5 where 1="very annoying", 2="annoying", 3="slightly annoying", 4="perceptible but not annoying" and 5="imperceptible" as recommended by ITU P.910 [68]. On average, it took between 10 and 15 minutes to complete a test (5 minutes to read test instructions and answer simple questions, 9 seconds to watch original video clip, 2 seconds to switch and watch the degraded video and 10 seconds non-restrictive voting time). After watching both non-degraded and the degraded video sequences, participants were asked; "On a scale of 1 to 5, grade the difference between the two videos in terms quality", where 1 indicates worst quality ("very annoying") and 5 the best ("imperceptible"). To simplify numerical analysis and plotting of graphs, individual DSIS ratings were averaged to obtain the Mean Opinion Score (MOS).

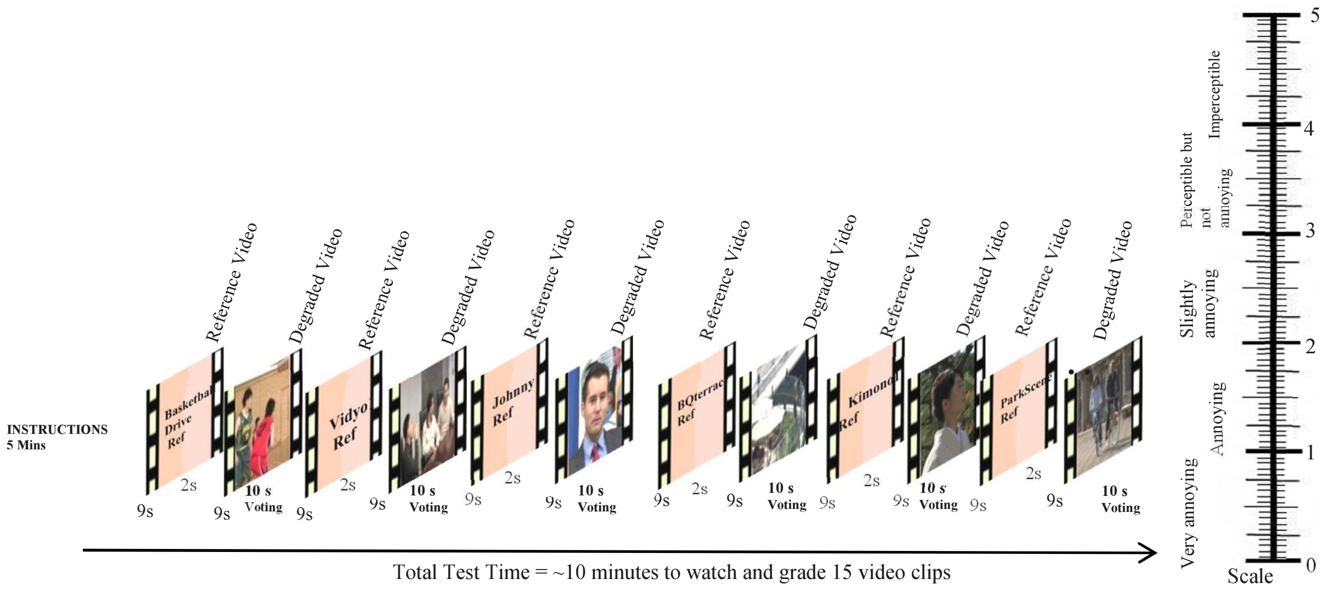


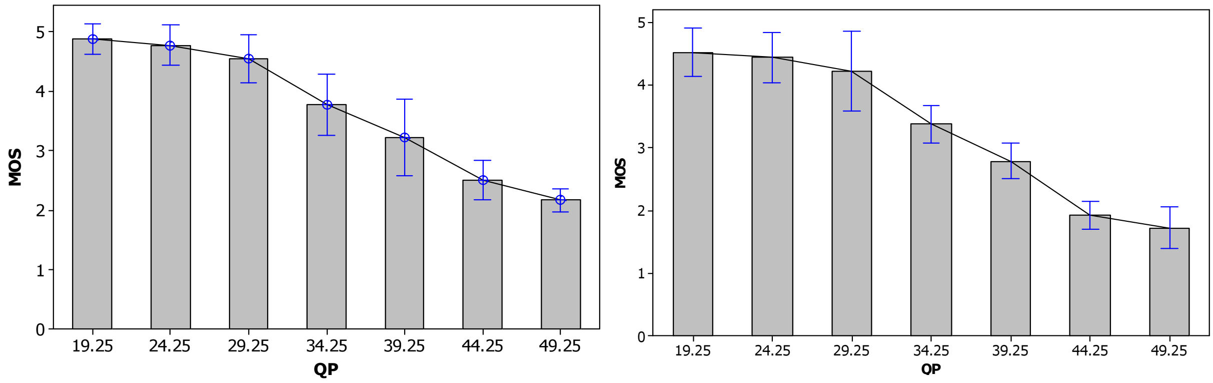
Figure 3.9 DSIS testing with a discrete impairment scale

- **Outlier detection**

To ensure validity of subjective results, hidden reference videos and a timer were incorporated into the websites. Evaluators whose DSIS ratings for the reference video were lower than the degraded video were automatically rejected. Additionally, evaluators who failed to enter their first name were asked to go back and do so before their scores could be accepted. To identify evaluators who did not watch all the videos before scoring, the testing websites were incorporated with a hidden timer. The timer calculated the entire test time by subtracting the start from the finish time. Test scores of participants whose test time was less than 10 minutes were automatically rejected. In total, 9 out of the 63 evaluators were rejected because of poor scoring or lack of credible ratings.

3.4.3 Results and analysis on the impact of compression on video quality

To test the quality of the subjective data, the distribution of the 95% confidence intervals (CI) for all DSIS ratings from valid evaluators is computed. The CI for the mean for Vidyo1 and ParkScene is shown in Figure 3.10. The average size of confidence intervals is 0.130 on a MOS scale of 1 – 5 for all video sequences. This indicates a good agreement between evaluators.



(a)

(b)

Figure 3.10 95% confidence interval for the MOS for (a) Vidyo1 and (b) BQterrace video sequences

The impact of encoder settings on video quality as perceived by users is presented in Figure 3.11 and Figure 3.12.

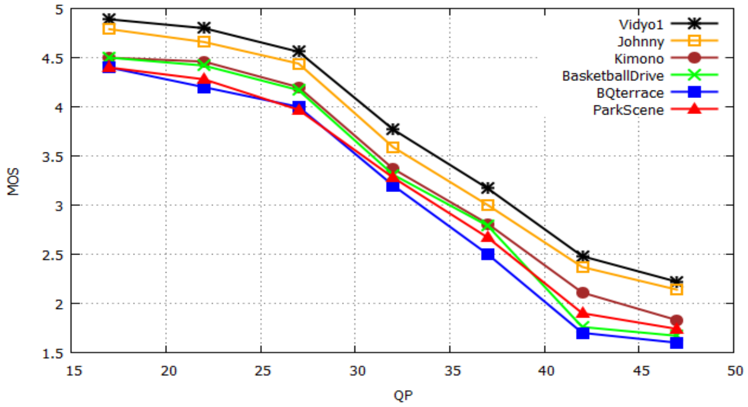


Figure 3.11. Impact of QP on MOS

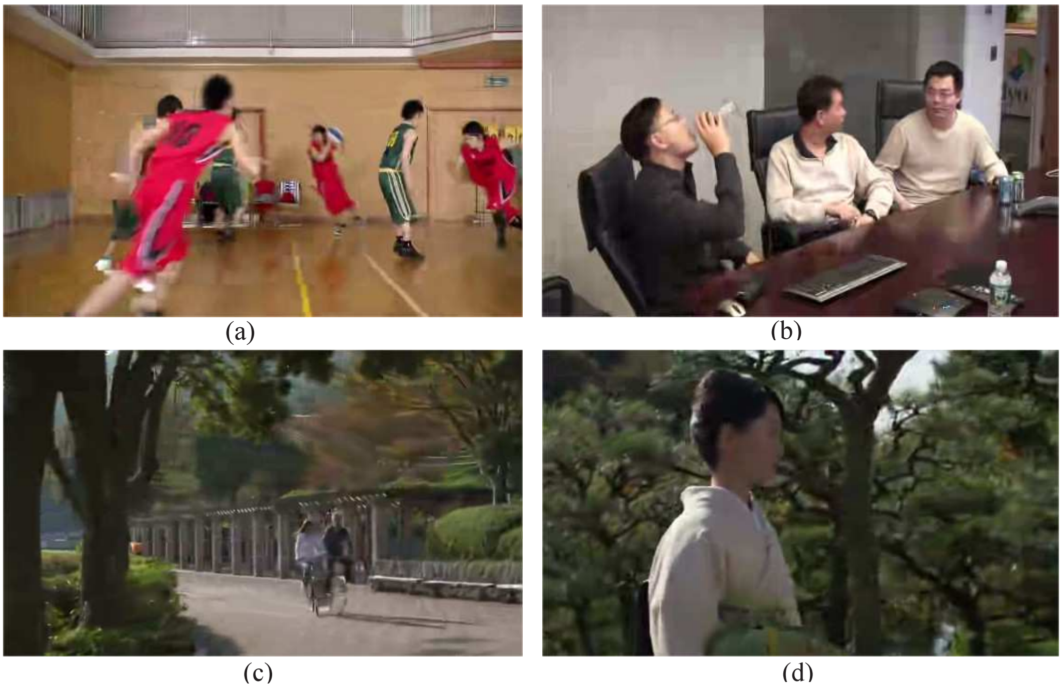


Figure 3.12. Perceptual quality comparison of (a) BasketballDrive, (b) Vidyo1, (c) ParkScene and (d) Kimono1 video sequences encoded at QP 47

The results show that increased compression (QP) leads to high visible quality degradation for all sequences. Furthermore, results indicate that because of differences in spatiotemporal characteristics, video sequences are impacted differently under the same encoder setting. For example, Johnny and Vidyo1 had a MOS value of 4.9 and 4.8 respectively, when compared to BasketballDrive, BQterrace, and Kimono1 and ParkScene MOS score of 4.2, 4.6, 4.7 and 4.7 respectively under the same QP (QP of 17). However, all sequences show a high level of acceptability as the video QP decreases. The threshold for acceptable MOS for encoded video was determined by calculating the mean MOS scores for six video sequences (Mean of 3.7 MOS).

It is also evident from the results that perceived video quality depends on encoder settings and content type. Therefore, one has to be mindful when selecting and encoding video as the initial quality of encoded video increases with decreased QP.

It should be noted that, although, the impact of devices on video quality was not investigated in this project, the used of Philips LCD monitors with screen size of 1600 x 900 for subjective testing might be limiting and have impact on video quality grading by subjective test evaluators.

3.5. Impact of NQoS on video quality

Using the framework for video streaming discussed above, the impact of network level parameters on video quality is evaluated. Specifically, the impact of QP, bandwidth and the number of TCP connections (i.e. congestion) on video quality in terms of how these parameters impact packet loss and subsequently video quality is evaluated.

- **Impact of network bandwidth on packet loss**

This section evaluates how different videos with different content complexities and level of motions are impacted by bandwidth reduction in terms of packet loss. To do this, several video sequences of different spatiotemporal complexity were encoded. They were encoded using exactly the same temporal resolution and exactly the same spatial resolutions and exactly the same QP. To ensure NQoS parameters such as congestion had no interference on the evaluation, the architecture for this experiment was only made up of UDP server and client and the bottleneck routers.

The results presented in Figure 3.13 show that for same QP setting, PLR decreases as the bandwidth increases. However, the decreased in PLR varies with spatiotemporal characteristics as sequences with high temporal complexity and high spatial complexity have a higher packet loss when compared to those with low temporal complexity and low spatial complexity. For example, at a bandwidth of 5mbps, the packet loss rate for Johnny and Vidyo1 is 0, Kimono1 38%, BasketballDrive 42%, ParkScene 52% and BQTerrace 62%.

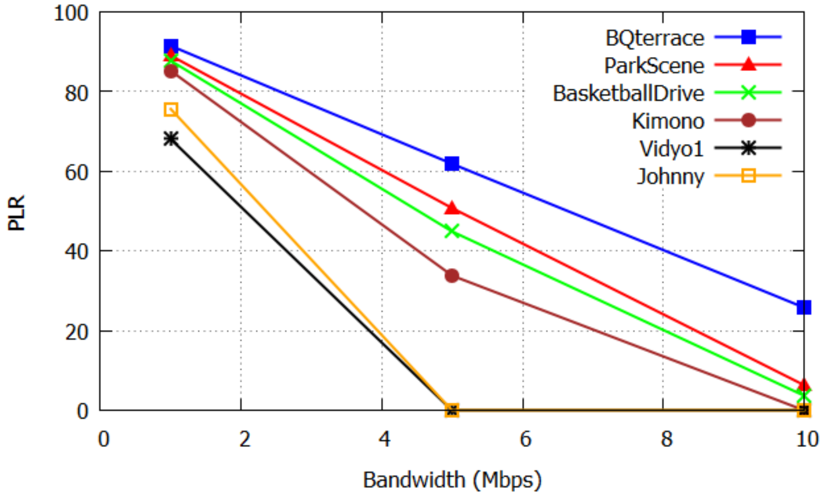


Figure 3.13 Impact of bandwidth on packet loss

- **Impact of network congestion on packet loss**

Modern networks are heterogeneous and made up of different applications. This section tests the impact of congestion on network packet loss rate. Although TCP is self-adjusting to network conditions i.e. reduces its sending rate when packet loss is detected, the results presented in Figure 3.14 show that PLR increases as the number of TCP connections increases from 2 to 6. This is because TCP only reacts to packet loss by adjusting its transmission rate, however, the time it takes to adjust to changing network condition may determine how many packets get loss. Results also show that the increase in PLR varied with spatiotemporal characteristics of the sequences. Sequences with high temporal complexity and high spatial complexity have higher a packet loss rate when compared to those with low temporal complexity and low spatial complexity. Furthermore, the results also show that PLR increases as the QP decreases.

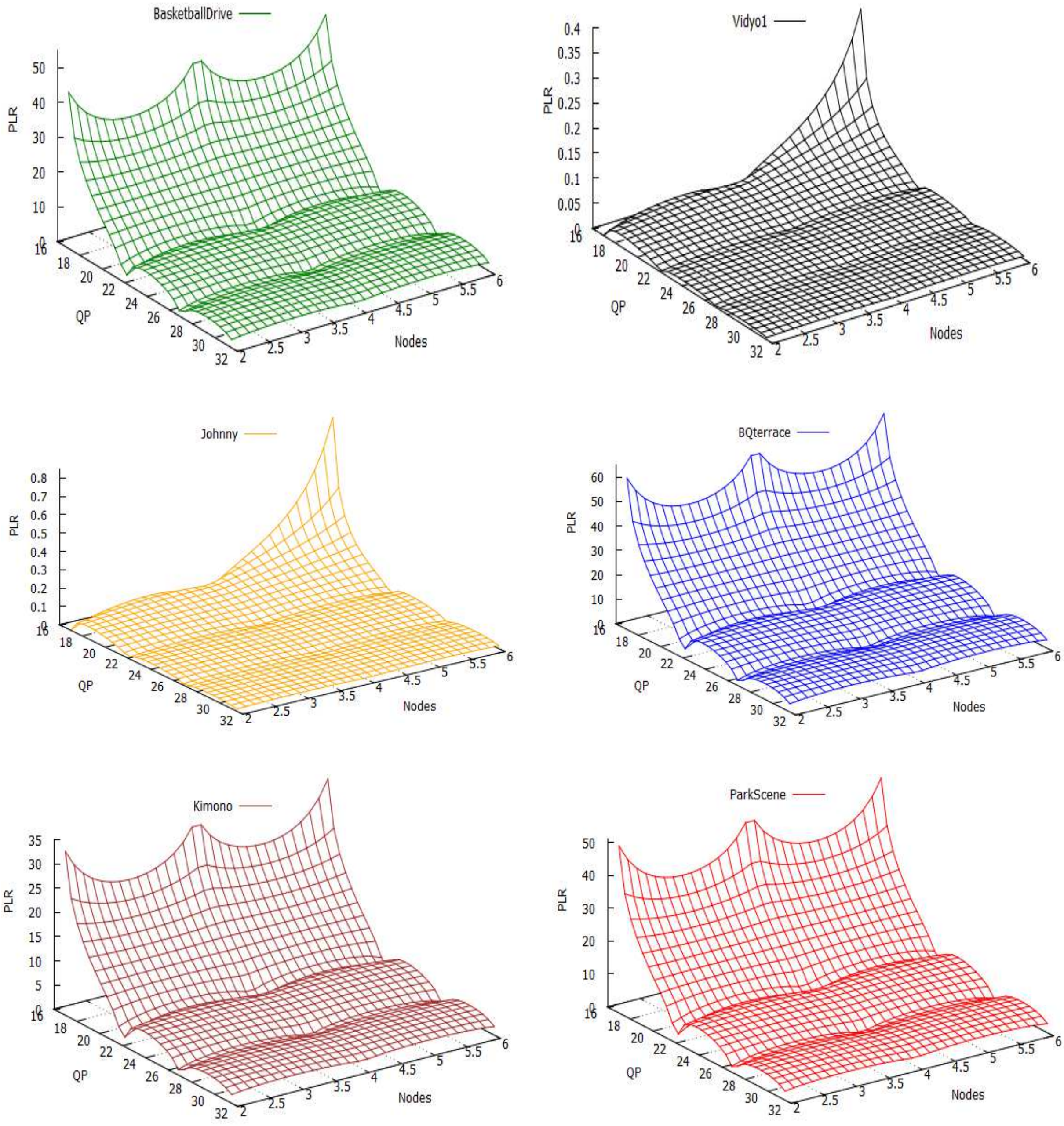
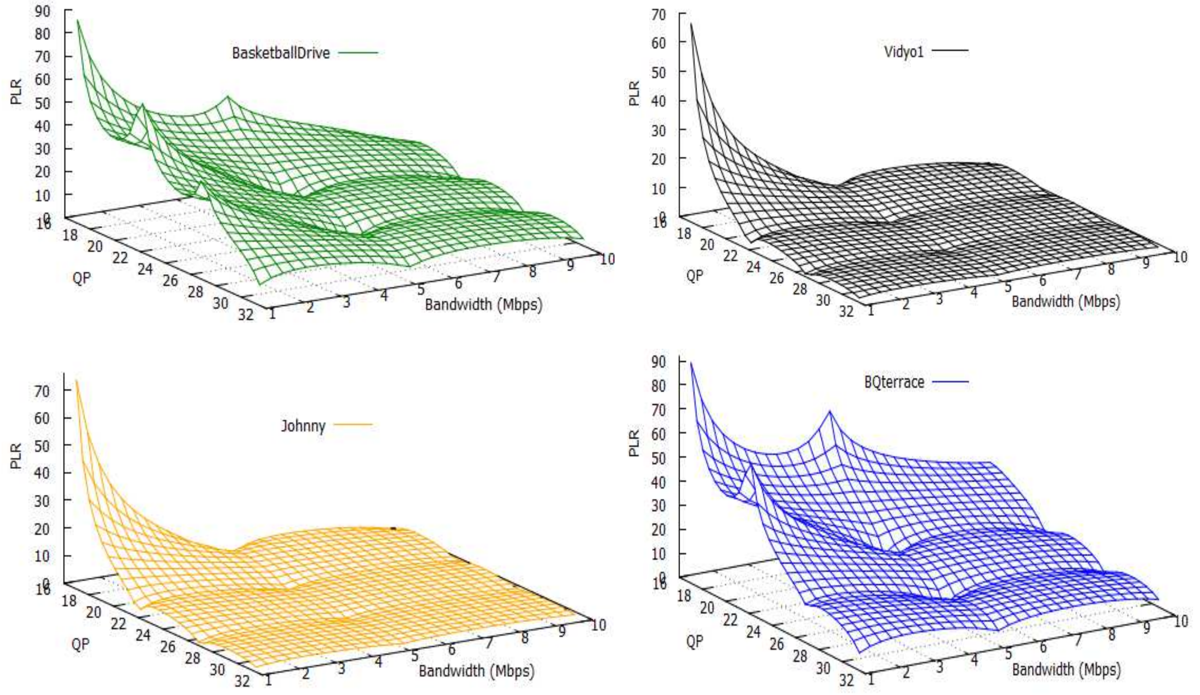


Figure 3.14. Impact of QP and congestion on PLR

- **Impact of encoder QP settings on packet loss**

To evaluate the impact of QP on packet loss, several video sequences of different spatiotemporal complexity were encoded. They were encoded using exactly the same temporal resolution and exactly the same spatial resolutions and different QP settings. Each bitstream, was packetized and streamed (or simulate) from the UDP server to the UDP client over the different bandwidth using NS-3 simulation framework. The number of TCP connections was fixed.

The results are presented in Figure 3.15. The results show that packet loss increases as QP values decreases, especially for videos with high complexity and motion. This can be attributed to two factors. Firstly, the available bandwidth is fixed for the experiment and secondly, video bitrate which is determined by the QP and complexity varies with QP. For example, lower QP values lead to higher bitrate which as consequence require more bandwidth for transmission and in so doing, PLR increases (Figure 3.15) as the bitrate increases considering that the bandwidth is fixed. The results also show that high complexity and high motion videos show more loss as the QP decreases than videos with lower complexity and lower motion.



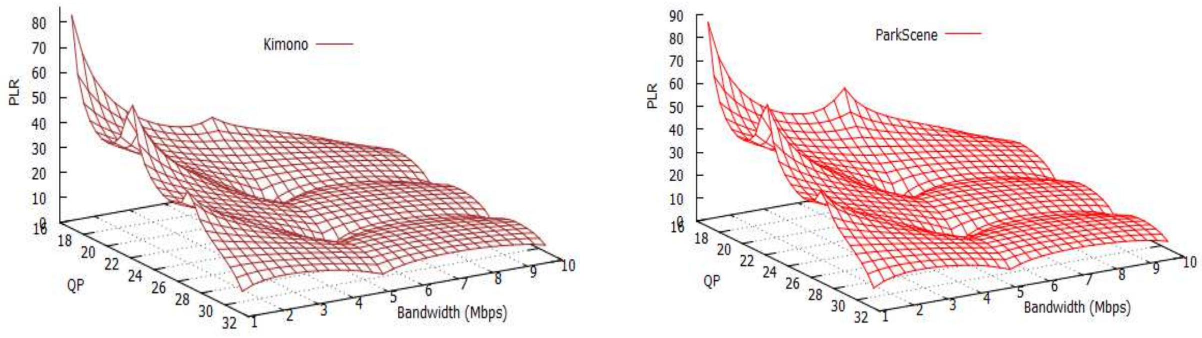


Figure 3.15. Impact of QP and bandwidth on PLR

3.5.1 Objective impact assessment of packet loss on video quality

In this section, videos encoded with the same QP setting and impaired using ad-hoc subgroup (AHG 14) of the JCT-VC loss simulator are used for quality assessment. The quality of each impaired sequence was measured in terms of full reference PSNR measurement. The results are shown in Figure 3.16. It should be noted that, the AHG loss simulator was used to ensure that all the sequences under test were impaired with the same PLR. Additionally, the packet loss rate range used in this thesis assumes the presence of error correction algorithms such as Forward Error Correction (FEC) in the transmission network.

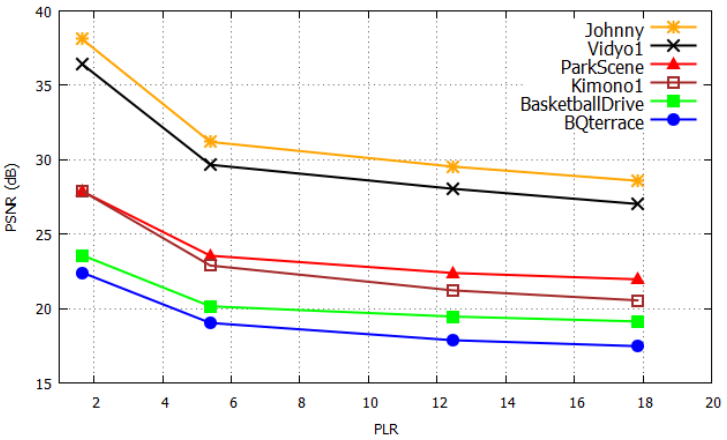


Figure 3.16. Impact of PLR on video quality

The results show that for the same QP setting, the PSNR for each sequence decreases with increased PLR. Results also show that for the same PLR, the PSNR estimation for each sequence is different. For example, with PLR of 5.4, Johnny sequence has a PSNR of 32dB, Vidyo1 30dB, Kimono1 24dB, BasketballDrive 20dB and BQterrace 18dB. This impairment difference may be attributed to the spatiotemporal complexity of a sequence which is different for each sequence.

3.5.2 Subjective impact assessment of packet loss on video quality

Considering that objective quality metrics such as PSNR may be limited to adequately estimate the visual impact of packet loss on video quality. In this section, a further subjective evaluation is carried out to determine the impact of packet loss on videos delivered to end users in an error prone network.

- **Subjective test**

To evaluate the impact of packet loss on video quality, the ITU recommended SSCQE testing method [69] was used. This testing method shows the video being evaluated to evaluators and mimics the perceptual response of typical video application consumers who have no access to the original video sequence.

- **Testing platform**

To evaluate the impact of packet loss on video quality, web based testing platforms were built and used to upload the sequences under test for subjective evaluators to watch and give their opinion scores. The test was restricted to 10 minutes as recommended by ITU and the test sequences were randomly split (using Latin square randomization technique [70]) into fourteen sets of fifteen videos each. Randomization reduces human memory effect [69]. Each set of videos was uploaded to a subjective testing platform (website) [23]. In total, fourteen identical websites were built for testing. Each video was uploaded to its own webpage and rated independently on a discrete five level scale from "bad" (1) to "excellent" (5). To rate a video sequence after watching, participants were asked; "On a scale of 1 to 5, what is the quality of the degraded video?" where 1 indicates the worst quality ("bad quality") and 5 indicated the best ("Excellent"). The individual ratings were mapped onto a Mean Opinion Score (MOS) scale from 1 to 5 where 1="very bad", 2="poor", 3="fair", 4="good" and 5="excellent" as recommended by ITU-R BT.500-13 [69]. To simplify numerical analysis and plotting of graphs, individual SSCQE ratings were averaged to obtain the MOS.

The voting box (i.e. text box) for each website was restricted to a range between 1 and 5; evaluators whose scores were beyond this range were gently reminded with the following message "please grade the video with values between 1 and 5".

The voting period on the testing websites was non-restrictive as evaluators could watch a video sequence more than once before scoring. Evaluators confirmed their scores to a video by clicking on the "Next" button present on each testing web page.

- **Test participants and procedure**

The subjective test was conducted over a month period (with six sessions of approximately 16 participants each) at Plymouth University computer network laboratory.

A total of 97 students from the School of Computing and Mathematics (SoCM) at took part in the test. This includes 72 undergraduates and 25 graduate students and a mix of males and females with majority male. Although participants' familiarity with video testing was not tested, the students who took part in the test had a high degree of computer literacy.

The assessor began the test by using the first 5 minutes to explain and demonstrate to evaluators how the test should be conducted. This was closely followed by distributing the printout white papers that contained the urls (web links) of each testing website to participants. To ensure each evaluator only visited and viewed a website that contained test sequences that were presented in a different order, the presentation order of url to evaluator was also randomized.

- **Outlier detection**

The MOS scores obtained from the subjective testing were scanned for unreliability and inconsistency by calculating the time an evaluator spent on a web page containing a video sequence. Following ITU recommendation, the video sequences under test were approximately

9 seconds long and because ITU recommends 10 seconds voting period and 2 seconds to switch over to another video sequence, a reliable evaluator was expected to spend in a range of 18 - 21 seconds on a web page before proceeding to the next video. A participant whose time was below this range on any of the web pages containing the videos was automatically rejected. Because of the screening methodology used, 15 out of the 97 evaluators subjective testing were rejected.

The frequency distribution of valid MOS scores is shown in Figure 3.17. The distribution shows that MOS is biased towards low MOS values. This is because participants gave low ratings to high motion and complexity videos which are more susceptible to encoding and network impairments than low motion and complexity videos. It should be noted that two (Vidyo1 and Johnny video sequences) out of the six video sequences under test had relatively stable movement (i.e. head and shoulder).

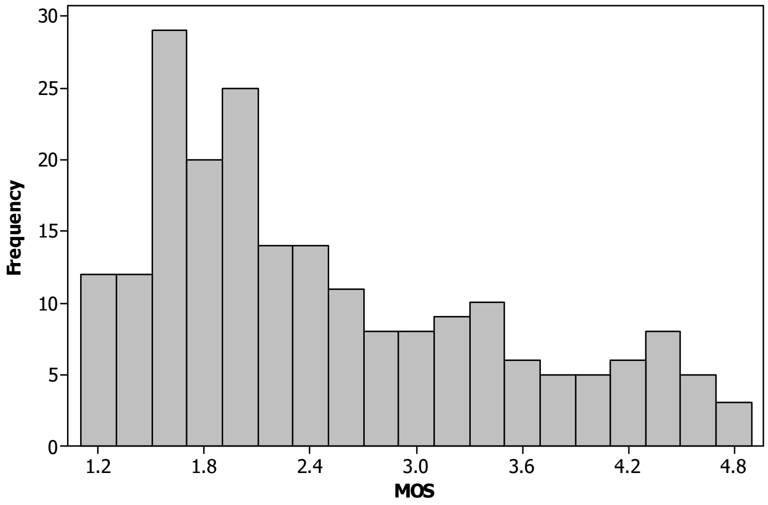


Figure 3.17 Frequency distribution of subjective MOS

3.5.3 Results and analysis on the impact of PLR on video quality

The results obtained from subjective testing were analysed using Principal Component Analysis (PCA) to determine the combined impact of encoder QP settings and PLR on video quality.

- **Principal component analysis**

Although the two parameters (considered in this project) that impact the quality of a video sequence delivered to end-users operate at the application and network layer, i.e. QP and PLR respectively. The combined impact assessment on how these parameters affect the quality of a video is important because it establishes how these parameters can be used in modelling video quality. To determine the combine impact of QP and PLR on video quality, a well-known multivariate statistical method, Principal Component Analysis (PCA) [86] is used. PCA statistical method involves calculating eigenvalues and their corresponding eigenvectors of correlation or the covariance matrix. Correlation matrix is used where the data set under consideration have different variables, while covariance matrix is used where data has the same set of variables. In this section, the correlation matrix was used to determine the PC loadings because the variables under consideration have different variances. In PCA, the loadings are correlation coefficients between original variables and the PC scores, the variable with the highest loading in the PC1 has the highest impact on the principal component and the sum of squares of all loadings for a PC is 1 [87]. Results of PCA plot and loadings for the QP, PLR and MOS of the six videos under test are shown in Figure 3.18 and Table 3.4.

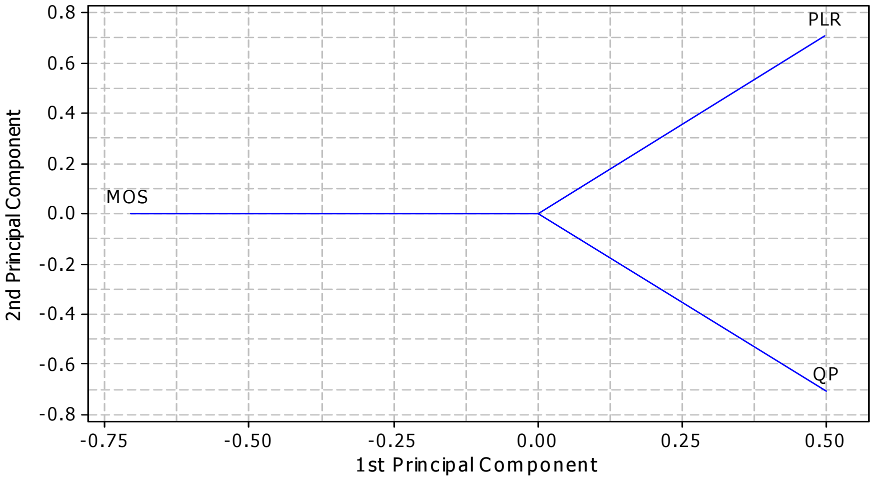


Figure 3.18 PCA loading plot

Table 3.4 PCA table of loadings

Variable	PC1	Squared loadings	PC2	Squared loadings
QP	0.501	0.251	-0.71	0.497
PLR	0.499	0.249	0.71	0.503
MOS	-0.707	0.500	0.00	0.000
Sum		1		1

It should be noted that, in PCA, the horizontal axis represents the first principal component (PC1) while PC2 is represented by the vertical axis. PC1 explains the maximum variance in the data from the origin, while PC2 is orthogonal to PC1 and explain the unexplained part of the data. From the results presented in Figure 3.18 and Table 3.4, it can be observed that both QP and PLR have positive loadings in PC1 while MOS have negative loadings in PC1. This indicates that QP and PLR are negatively correlated with MOS (i.e. increasing QP and PLR leads to a decrease in MOS). In terms of which of the variable has a stronger impact on quality, results in Table 3.4, indicates that QP has the highest PC1 value, this is followed by PLR which has the least value in PC1. However, results in Figure 3.18 and Table 3.4 show that the two variables impact the principal component by not having a 0,0 cordination on the plot. This is important because it quantifies the variables that impact MOS and show how they can be used in modelling quality.

Figure 3.19 and Figure 3.20 shows the impact of PLR on video quality. The results show that for the same QP setting, the MOS for each sequence decreases with increased PLR. Results also show that for the same PLR, the MOS estimation for each sequence is different. For example, with PLR of 5.4, Johnny sequence has a MOS of around 2.6, Vidyo1 2.4, Kimono1 2.4, BasketballDrive 2.3 and BQterrace 2.3. The results indicate that the impairment due to packet loss on video quality is different for each sequence. This impairment difference may be attributed to the spatiotemporal complexity of a sequence.

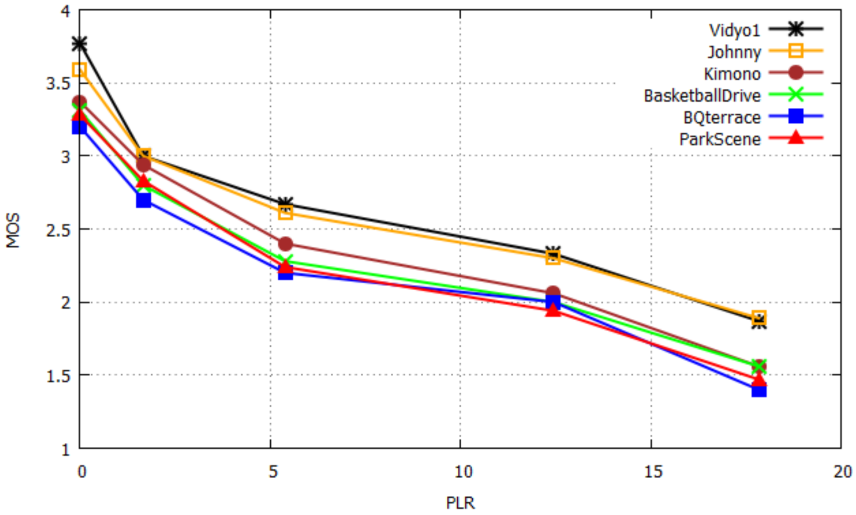


Figure 3.19 Impact of PLR on MOS

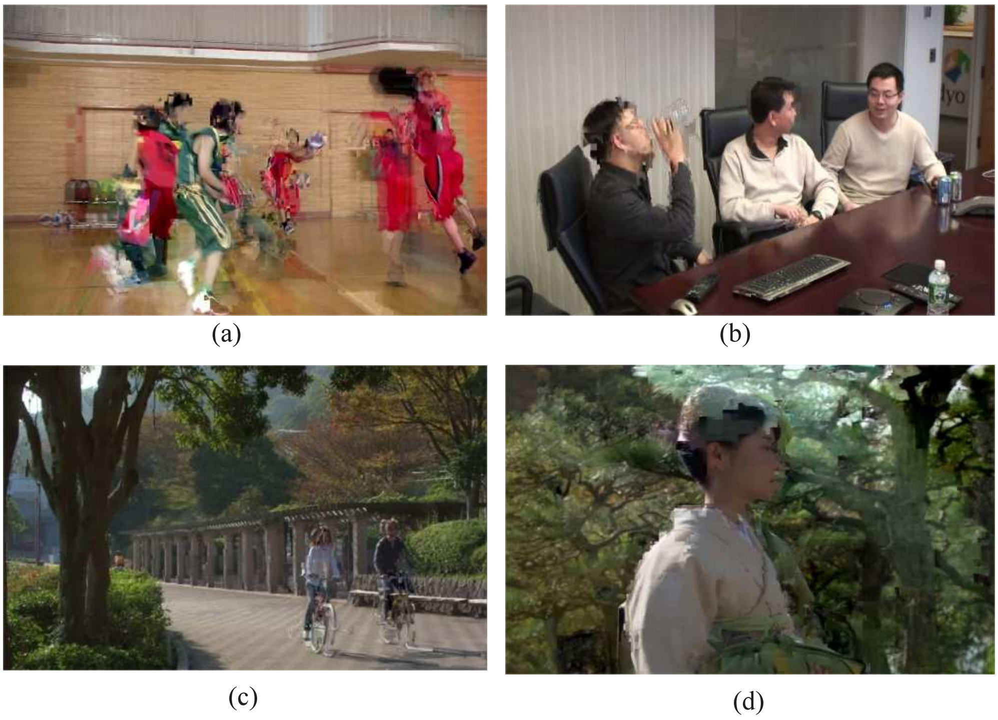


Figure 3.20. Perceptual quality comparison for (a) BasketballDrive, (b) Vidyo1, (c) ParkScene and (d) Kimono1 video sequences encoded at QP 17 and PLR of 17.84%

3.6. Summary

In this chapter, an investigation has been carried out to determine the impact of encoding QP settings and network packet loss rate on video. The investigation was carried out by using datasets generated by objective and subjective methods. The results show that at the application layer, the encoder QP setting has a great impact on initial video quality, whereas at the network layer, results show that PLR significantly impact video quality. Based on the principal component analysis, results show that the encoding QP setting has more impact on video quality than PLR. The most crucial findings from the results was that, although, the quality of all sequences was impacted by the encoder QP setting and network packet loss, the impact of these two processes on video quality is content dependent as videos encoded with the same QP and transmitted/induced/simulated with the same PLR show different quality measurements. Considering that one of the objectives of this thesis is to develop video quality models that are able to estimate the quality of different video applications delivered to end users' devices. It is therefore important to investigate the impact of content type on video quality and if possible, quantify video content type and use as a variable in the modelling process. The conclusions arrived at in this chapter will enable further investigation on the impact of video content type on quality and to determine how the parameters that impact video quality can be used to develop non-intrusive video quality prediction models (presented in chapters 6). Given that the impact of video quality is content dependent, the next chapter, presents an investigation on the impact of content type on video quality.

Chapter 4

4. The Impact of Content Type on Video Quality

4.1 Introduction

In chapter 3 an investigation on the impact of QoS parameters on video quality was presented. Results show that the quality of an encoded video sequence, for a given codec setting and for a given network quality of service (NQoS), varies with the spatiotemporal characteristics of the content of the video sequence. The level of video quality degradation during encoding and transmission is high for higher motion content when compared to video sequences of lower spatiotemporal complexity. Therefore, in order to accurately model video quality requires that the video content type be quantified and used as a variable in the modelling process. The aim of this chapter is to investigate the impact of video content type on quality. To do this, a metric for quantifying the spatiotemporal characteristics of a video sequence is derived. This metric is used to investigate how video content type impacts the quality of different video sequences. Additionally, the metric is subsequently used as a variable alongside other parameters (i.e. encoding and transmission parameters) to design non-intrusive video quality prediction models. (See chapters 6 and 7).

A video sequence can be decomposed into spatial and temporal complexities. It can, for example, have a high temporal complexity and a high spatial complexity or low temporal complexity and low spatial complexity. The way the encoded video sequence is impacted by encoding and network impairments may vary according to the spatiotemporal characteristic. Results in chapter 3 show the impact of encoder parameter settings (e.g. QP settings) and network impairment (e.g. PLR) on video quality for a number of video sequences that have different spatiotemporal characteristics. All the video sequences were encoded using the same codec with exactly the same QP, frame rate and frame resolution settings and impaired with the same PLR. The results show that all the video sequences have different video quality measurements or estimations for the same codec settings and for the same PLR. This variation in quality can be attributed to video content type which is determined by temporal (e.g. movements) and spatial (e.g. brightness, edges, blurriness etc.) information.

The two approaches commonly used approaches to quantify or identify video content types are the extraction of motion features from the encoded bitstream and computation of pixel-wise differences between successive video frames. These two approaches are fundamentally different in terms of computational efforts and implementation. Pixel-wise differences between consecutive frames involve calculating the differences between the current and the previous frames in terms of pixels. This approach uses matching metrics such as Sum of Absolute Difference (SAD), Sum of Square Difference (SSD) and Mean Absolute Deviation (MAD) to find the differences between adjacent frames. On the other hand, an encoded bitstream carries motion information such as motion vectors (MV) and other parameters (e.g. frame type, number of bits of coded frames, QP etc.) which can be extracted and used to quantify video content type in terms of motion activities and the complexity of a video sequence. The extraction of motion information from the encoded bitstream is considered to be simple with less computational overheads because the parameters needed for video content type identification are computed during the encoding process [88].

In this chapter, a single metric that quantifies video content type is developed. This metric takes into account the spatiotemporal information of different videos extracted from the encoded bitstream. Based on the content type metric, a content-based video quality prediction model that takes into consideration encoder parameter settings (e.g. QP) and impairments caused by NQoS (e.g. PLR) is developed (see chapter 6).

The rest of the chapter is organized as follows: The video sequences used to study the impact of content type on quality are presented in Section 4.2. An overview of video motion estimation is discussed in Section 4.3. Development of content type metric is presented in Section 4.4. Section 4.5 presents the evaluation of the proposed metric and a comparison with existing pixel-wise metrics. Section 4.6 summarizes the chapter.

4.2 Selection of video sequences

The video sequences chosen to study the impact of content type on video quality and to further develop a metric that quantifies video content type ranged from very little movement to fast moving sports clips. The choice of video sequences is based on JCT-VC recommended sequences for HEVC testing [20]. Additionally, the video sequences were also selected to reflect the varying spatiotemporal characteristics of the content typically offered by content providers e.g. news and fast moving sports content.

The description of the different content types is given in Table 4.1.

Table 4.1 Video sequences and their description

Sequence	Characteristics
BasketballDrive	Basketball playing scene
Vidyo1	Three men making a conference call
Johnny	Head & shoulders newscaster
BQterrace	People sitting in a terrace with background bridge, river and moving cars
Kimono1	A lady walking with fast moving forest background
ParkScene	A man and a lady cycling in a park with fast moving background
RaceHorses	Horses walking with riders on it

4.3 Overview of video motion estimation

Motion estimation (ME) can be defined as the process of determining the movement of blocks between adjacent frames. This process consumes a greater proportion of computational resources during video coding. ME is employed by the encoder to reduce temporal redundancy [89]. The basic premise of estimating motion in a video sequence is that successive video frames are identical to each other except when changes are induced by movement of objects within frames [90]. In a situation where there is no movement, it is fairly easy for any motion estimation metric (e.g. template matching, block matching and optical flow) to efficiently predict the current frame to be a duplicate of the previous one. However, when there is movement, the motion estimation metric has to adequately represent the differences or the changes between the video frames. This is realised by performing a comprehensive 2-D (2-dimension) spatial search on each luminance macroblock to determine the best match. During video encoding process, the search of the 2-D area can be based on full and exhaustive or limited pixel search range. Full search range yields the best matching results, but at the expense of computational overheads. The search of the 2-D area also involves the assignment of motion vectors (MV) to the macroblock to indicate how far the block has to move horizontally and vertically before making a match. In essence, a predicted block that has moved forward and backward may contain two MVs. With this in mind, it is therefore safe to say that MVs carry vital motion information which is used by the decoder to reconstruct the original video sequences. In this thesis, the MVs are used to determine the temporal component of the video content type metric. This component is then used together with the spatial information to determine the individual video content type. In the context of this thesis, ME is important because it gives an overview on how the features (e.g. MVs) needed to quantify the content type of a video are formed or derived.

4.4 Video content type metric

In order to quantify the content type of a video sequence, spatiotemporal features were extracted from the video bitstream. The block diagram for quantifying video content type is presented in Figure 4.1. For simplicity, the HEVC decoder was modified and used for feature extractions.

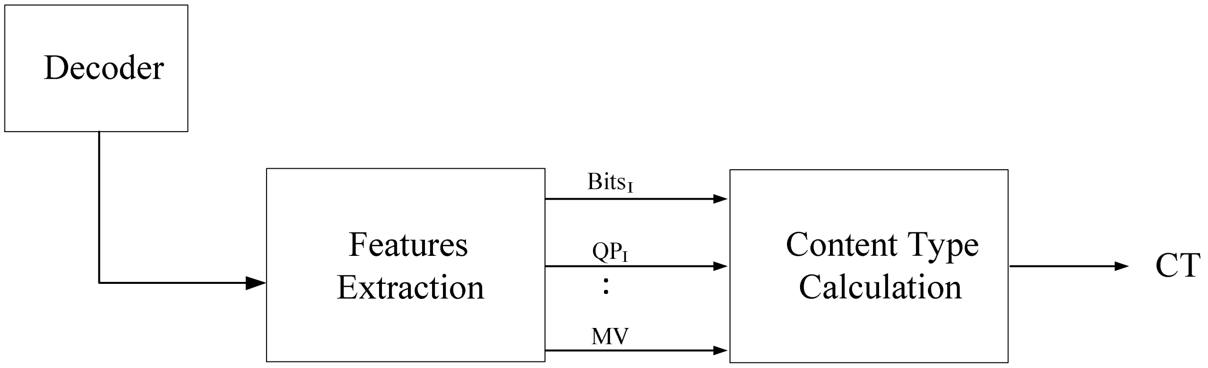


Figure 4.1 Block diagram of video content type computation

The content type is decomposed into two metrics. These metrics gauge the spatial and temporal components of a video sequence. The spatial component indicates the brightness, edges and blurriness of a video while the temporal component relates to the amount of movement or motion activities of a video sequence. For two different videos, the temporal complexity might be different in terms of movement of objects and the spatial complexity might be different in terms of brightness, the edges and the blurriness.

4.4.1 Spatial complexity and video quality

Picture Complexity (PC) metric is used to quantify the spatial information (SI) of a video sequence. This metric is derived from Intra (I) encoded video frame of a video sequence. The number of bits of coded I-frame ($Bits_I$) depends to a large extent on the spatial complexity of the video sequence. For a given QP, the number of bits of the coded I frame will be a function of the spatial complexity of the sequence. Based on this assertion, we model the spatial complexity of video sequences in terms of the QP and the number of bits of coded I frame (i.e. QP_I and $Bits_I$). The spatial complexity of video sequences is therefore modelled in terms of the QP of I frame and the number of bits of an I-frame. This is modelled as,

$$PC = f(QP_I, Bits_I) \tag{4.1}$$

In order to study the variation in the number of bits of coded I-frame to be a function of the video sequence complexity, the number of I-frame bits for different video sequences which were encoded with different QP settings was evaluated. For this evaluation, the temporal and spatial resolutions were kept exactly the same. The results of the evaluation show that the number of bits of each coded I frame decreases with increased QP (i.e. indirectly proportional relationship). Furthermore, the decrement in the number of bits of the coded I frame is different for each QP setting. Based on this evaluation, the spatial complexity of a video sequence is modelled in terms of picture complexity (PC) metric. Formalised as,

$$PC = \left(\frac{Bits_I}{QP_I} \right)^{\alpha} \tag{4.2}$$

where α is the modelling parameter and represents the percentage of bits reduction (PBR). To determine the PBR, several video sequences of different spatiotemporal complexity were encoded with different QP settings that ranged from 1 to 50. For each QP setting, the number of bits of the coded I frame was extracted and their corresponding QP recorded. The reduction in the number of bits in terms of PBR is calculated as,

$$PBR = \frac{(Bits_{x_{QP_n}} - Bits_{x_{QP_m}})}{Bits_{x_{QP_n}}} \times 100 \tag{4.3}$$

where $Bits_{x_{QP_n}}$ and $Bits_{x_{QP_m}}$ are the number of bits (I-frame) of previous and current QP settings respectively. The PBRs for the sequences were averaged to determine the overall percentage of reduction in the number of bits. Results of PBR computation are shown in Table 4.2.

Table 4.2 percentages of bits reduction

Video sequence	PBR (%)
BasketballDrive	0.11
Vidyo1	0.11
Johnny	0.11
BQterrace	0.10
Kimono1	0.11
ParkScene	0.12
Weighted Average of PBR	0.11

Based on the weighted PBR average, α was determined to be 0.11.

Considering that an encoded video sequence can have L Intra coded frames, and that each of the I-frame may have different QP settings and the number of bits of coded I may also be different, an Average Picture Complexity (APC) metric is computed to quantify the spatial complexity as,

$$APC = \frac{1}{L}\sum_{i=1}^{L}\left(\frac{Bits_{I_i}}{QP_{I_i}}\right)^{\alpha} \tag{4.4}$$

To determine the impact of QP on the number of coded bits ($Bits_I$) and APC, several video sequences of different spatiotemporal complexity were encoded. They were encoded using exactly the same temporal resolution and exactly the same spatial resolutions. The videos were encoded with different QP settings. For each QP setting, the APC was computed and the QP value and the number of bits of coded I-frame recorded. The results are presented in Table 4.3. The results show that for the same QP setting, the sequences with the high complexity have the highest $Bits_I$ (this is because after intra-prediction and discrete cosine transform, the I-frame of sequences with more complexity will have more non-zero high frequency coefficients and will require more bits to code such frame) and APC while the video with lower complexity has the lowest $Bits_I$ and APC.

Table 4.3 Impact of QP on Bits$_I$ and APC

	Johnny		**BasketballDrive**		**BQTerrace**	
QP	Bits$_I$	APC	Bits$_I$	APC	Bits$_I$	APC
1	4199598	5.35	4574843	5.40	5681766	5.53
12	1479188	3.63	1872364	3.73	3011226	3.93
37	63479	2.27	79960	2.33	277576	2.67
47	19944	1.95	24128	1.99	62768	2.21

The relationship between APC and video quality for a given encoder setting was studied. To do this, several video sequences that have different spatiotemporal characteristics were encoded. For each encoder setting (QP), the APC and the PSNR were calculated. The results are presented in Table 4.4.

Table 4.4 Relationship between QP, APC and PSNR

	Johnny		**BasketballDrive**		**BQTerrace**	
QP	APC	PSNR	APC	PSNR	APC	PSNR
1	5.35	68.29	5.40	68.10	5.53	67.28
12	3.63	47.82	3.73	47.71	3.93	46.96
37	2.27	36.52	2.33	33.05	2.67	31.73
47	1.95	31.20	1.99	27.86	2.21	26.00

Results in Table 4.4 show that for the same QP setting, sequences with high spatial complexity have the highest APC than the video with lower complexity. Although high complexity videos have higher APC, results also show that for the same QP setting, the video with the lower complexity has a higher PSNR calculation when compared to the video with high complexity. Additionally, results also show that, the PSNR for the three sequences decreases with increased QP, however, the decrement in PSNR is higher for sequences with higher complexity than the video with low complexity.

The impact of frame resolution on APC was also investigated. To do this, several video sequences with different QP settings were encoded. For each QP and frame resolution settings, the APC was computed. The results are presented in Table 4.5 and Table 4.6.

Table 4.5 Impact of frame resolution on APC BasketballDrive

BasketballDrive				
QP	Frame Resolution	APC	Frame Resolution	APC
22	1280 x 720	2.95	1920 x 1080	3.27
27	1280 x 720	2.72	1920 x 1080	2.91
32	1280 x 720	2.51	1920 x 1080	2.66
37	1280 x 720	2.33	1920 x 1080	2.46

Table 4.6 Impact of frame resolution on APC RaceHorses

RaceHorses				
QP	Frame Resolution	APC	Frame Resolution	APC
22	416 x 240	2.69	832 x 480	3.08
27	416 x 240	2.51	832 x 480	2.85
32	416 x 240	2.32	832 x 480	2.64
37	416 x 240	2.14	832 x 480	2.42

The results in Table 4.5 and Table 4.6 show that for the same QP setting, the APC is different for each spatial resolution. Considering that APC values for the different QP settings are influenced by the spatial resolution, this influence of frame resolution on SI (APC) is removed by normalising the SI metric (APC). The normalised APC (nAPC) is shown in Eq. 4.5.

$$nAPC = \frac{1}{L}\sum_{i=1}^{L}\left(\frac{bits_{I_i}}{Max_{bits}} \times \frac{1}{QP_{I_i}}\right)^{\alpha} \qquad \text{nAPC} \in [0,1] \qquad (4.5)$$

where Max_{bits} denotes the maximum possible number of bits in a video frame (i.e. $Width \times Height \times depth \times colour\ sampling$).

To determine if normalisation removed the dependency of APC on spatial resolution, the nAPC of each sequence was computed, the results are presented in Table 4.7 and Table 4.8.

Table 4.7 Impact of frame resolution on nAPC BasketballDrive

BasketballDrive				
QP	Frame Resolution	nAPC	Frame Resolution	nAPC
22	1280 x 720	0.50	1920 x 1080	0.50
27	1280 x 720	0.46	1920 x 1080	0.45
32	1280 x 720	0.42	1920 x 1080	0.41
37	1280 x 720	0.39	1920 x 1080	0.38

Table 4.8 Impact of frame resolution on nAPC RaceHorses

RaceHorses				
QP	Frame Resolution	nAPC	Frame Resolution	nAPC
22	416 x 240	0.58	832 x 480	0.57
27	416 x 240	0.54	832 x 480	0.53
32	416 x 240	0.50	832 x 480	0.49
37	416 x 240	0.46	832 x 480	0.45

The results show that after normalization, there is no significant difference between APC values from the different spatial resolution indicating that the dependency of APC on spatial resolution has been removed.

4.4.2 Temporal complexity and video quality

To develop a metric that quantifies the temporal information (TI) in terms of motion or movement in a video, the Motion Vector (MV) information inherent in the encoded video bitstreams was extracted. MVs are generated in the process of motion estimation during video compression and transmitted as side information in the encoded bitstream [58]. The generated MVs are used by the decoder to reconstruct the original video sequences and therefore reflect the motion activity of a video sequence as estimated by the encoder [91].

In this thesis, the ratio of non-zero to total number of MV (i.e. MV whose x or y coordinates are not equal to zero) is used to quantify the amount of motion in terms of motion activity (MA) because the ratio of non-zero MVs gives a good correlation (computed in terms of R^2 between the two variables) with video bitrate for all the sequences used. The direct correlation results between bitrate and ratio of non-zero MVs is shown in Table 4.9.

Table 4.9 Direct correlation between video bitrate and MV metrics

Cor. Coef.	**Johnny**	**BasketballDrive**	**BQTerrace**
R^2	0.89	0.97	0.88

The MA metric is used to quantify the TI of a video sequence; this metric is derived from the ratio of non-zero motion vectors (MVnz) to the total number of counted MVs (MVc) per picture. The ratio of non-zero MVs depends to a large extent on the amount of motion as estimated by the encoder. Based on this assertion, the temporal complexity of video sequences in terms of non-zero MVs and total number of MVs is modelled. This is modelled as,

$$MA = f(MV_c, MV_{nz}) \tag{4.6}$$

In order to study the amount of motion in a video to be a function of the total number of motion vectors and the non-zero motion, different videos with exactly the same QP settings, temporal and spatial resolutions are encoded. The motion vectors from each encoded bitstream are extracted and the total number of motion vectors and the non-zero motion vectors evaluated. Based on this evaluation, it was determined that, the total number and the non-zero motion vectors vary with the temporal complexity of a sequence and in essence, the temporal complexity of a video sequence is therefore modelled, quantified as the intensity of motion activity (MA) metric as shown in Eq. 4.7,

$$MA = \frac{MV_{nz}}{MV_c} \tag{4.7}$$

Considering that an encoded video sequence can have M bi-directional (B) and N predictive (P) coded frames, and that each of this frame might have a different number of motion vectors, an Average Motion Activity (AMA) is computed to quantify the temporal complexity as,

$$AMA = \left(\frac{1}{M} \sum_{j=1}^{M} \frac{MV_{nz_j}}{MV_{c_j}} + \frac{1}{N} \sum_{k=1}^{N} \frac{MV_{nz_k}}{MV_{c_k}} \right) \tag{4.8}$$

where M and N represent the number of B and P frames of a video sequence respectively. In essence, the average MA of B and P frames are added together to get the AMA as shown in Eq. 4.8.

To study the impact of QP on AMA, several video sequences of different temporal complexity are encoded. They were encoded using different QP settings, same temporal resolution and the same spatial resolutions. For each video sequence, the AMA is computed. The results are shown in Table 4.10. The results show that for the same QP setting, sequences with high temporal complexity have the highest AMA (because of increased motion activity in terms lesser correlation between successive frames) while the video with low temporal complexity has the lowest AMA. Results also show that the AMA of all the sequences decreases with increased QP.

Table 4.10 Impact of QP on AMA

	Johnny	**BasketballDrive**	**BQTerrace**
QP	AMA	AMA	AMA
1	0.62	0.83	0.87
12	0.61	0.74	0.82
37	0.51	0.57	0.64
47	0.36	0.54	0.47

The relationship between the AMA and video quality was studied. To do this, several video sequences with different spatiotemporal characteristics were encoded. The frame rate and frame resolutions were kept constant. For each QP setting, the AMA and the video quality were computed. Results in Table 4.11 show that for the same QP, videos with high temporal complexity videos have the lowest PSNR value when compared to the video with lower temporal complexity. Additionally, results also show that the PSNR for the three sequences decreases with increased QP, however, the decrement in PSNR is higher for sequences with high temporal complexity than the video with low complexity.

Table 4.11 Relationship between QP and AMA and PSNR

	Johnny		**BasketballDrive**		**BQTerrace**	
QP	AMA	PSNR	AMA	PSNR	AMA	PSNR
1	0.62	68.29	0.83	68.10	0.87	67.28
12	0.61	47.82	0.74	47.71	0.82	46.96
37	0.51	36.52	0.57	33.05	0.64	31.73
47	0.36	31.20	0.54	27.86	0.47	26.00

4.4.3 Video content type metric derivation

The content Type (CT) metric is used to quantify the spatial and temporal information of a video sequence. This metric is derived from the normalised average picture complexity (nAPC) and average motion activity (AMA). Considering that video quality depends to a large extends on the content type, the spatiotemporal complexity of video sequences is modelled in terms of nAPC and AMA. This is modelled as,

$$CT = f(nAPC, AMA) \tag{4.9}$$

The nAPC and AMA of several video sequences were evaluated to establish the relationship between APC and MA. The results are shown in Table 4.12.

The results show that the nAPC and AMA of the sequences decreases with increased QP. Results also show that sequences with high AMA have the highest nAPC thus indicating that nAPC increases with AMA and vice versa.

Table 4.12 Comparison between nAPC and AMA

	Johnny		BasketballDrive		BQTerrace	
QP	nAPC	AMA	nAPC	AMA	nAPC	AMA
1	0.90	0.62	0.91	0.83	0.93	0.87
12	0.61	0.61	0.63	0.74	0.66	0.82
37	0.38	0.51	0.39	0.57	0.45	0.64
47	0.33	0.36	0.33	0.54	0.47	0.47

Based on these results, it was established that nAPC and AMA have a directly proportional relationship. The spatiotemporal complexity of a video sequence is therefore modelled, quantified as a content type (CT) metric as shown in Eq. 4.10.

$$CT = AMA \times nAPC \tag{4.10}$$

It should be noted that, AMA is multiplied by nAPC because the results in section 4.4.1 and 4.4.2 show that the picture complexity (in terms of APC) increases with increased motion activity (in terms of AMA) of a video.

The final CT equation is presented as follows when nAPC and AMA are combined.

$$CT = \left(\frac{1}{M} \sum_{j=1}^{M} \frac{MV_{nz_j}}{MV_{c_j}} + \frac{1}{N} \sum_{k=1}^{N} \frac{MV_{nz_k}}{MV_{c_k}} \right) \times \frac{1}{L} \sum_{i=1}^{L} \left(\frac{bits_{I_i}}{Max_{bits}} \times \frac{1}{QP_{I_i}} \right)^{\alpha} \tag{4.11}$$

where L represents the number of I-frame in the spatial domain while M and N denotes the number of B and P frames respectively in the temporal domain.

Considering that for a given QP setting, the quality of a video sequence is dependent on the content type, the combined impact of AMA and nAPC (in terms of CT) on video quality was determined. To do this, the PSNR and CT of different video sequences were computed. Results are shown in Table 4.13.

Table 4.13 Impact of QP on CT and PSNR

	Johnny		**BasketballDrive**		**BQTerrace**	
QP	CT	PSNR	CT	PSNR	CT	PSNR
1	0.56	68.29	0.76	68.10	0.81	67.28
12	0.37	47.82	0.46	47.71	0.54	46.96
37	0.19	36.52	0.22	33.05	0.29	31.73
47	0.12	31.20	0.18	27.86	0.17	26.00

Results in Table 4.13 show that for the same QP, sequences with high CT have the lowest PSNR while the video with a lower CT has the highest PSNR. Additionally, results also show that both CT and PSNR for the three sequences decreases with increased QP, however, the decrement in PSNR is higher for sequences with higher complexity than the video with low complexity.

- **Content type calculation over time**

To demonstrate how the proposed approach of quantifying video content type can be used over time to quantify the content type of a video sequence i.e. after a given number of frames. Several video sequences that have different spatiotemporal characteristics were encoded with the same QP, different Spatial and temporal resolutions. The CT of each sequence was computed after a given number of frames. To determine how many frames can be used, the CT of 5, 10, 15 and 20, 25 and 30 frames were computed. The results of the tested number of frames are shown in Figure 4.2. The results show that 15 frames iteration has a better frequency.

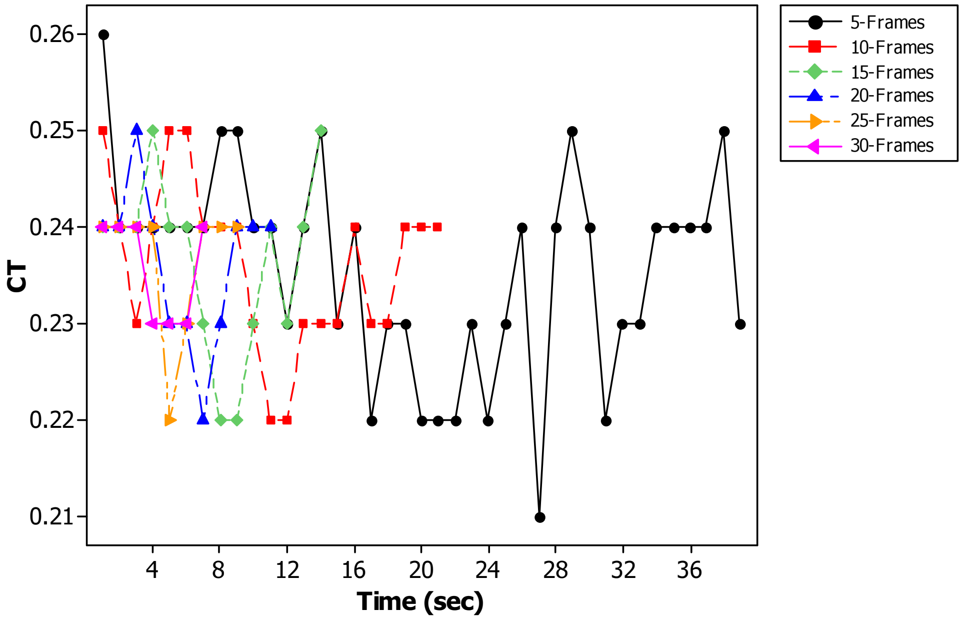


Figure 4.2 CT calculation for different frames for Johnny video sequence

Based on this assertion, 15 frames were chosen and used to demonstrate the implementation of CT computation over time. 15 frames represent a 0.5 second time period for a video encoded at 30fps temporal resolution and 0.3 seconds for videos encoded at 50 frames per second. The results of CT calculations over time for different video sequences are shown in Figure 4.3. The results show that the CT varies over time and videos with the highest spatiotemporal characteristics have the highest CT when compared to videos with the lowest CT.

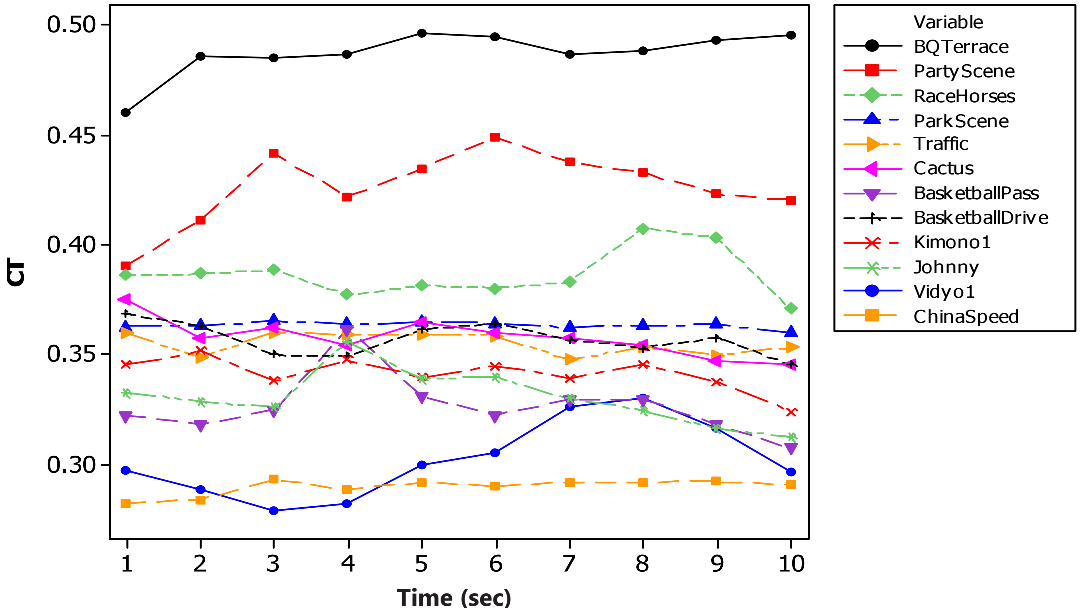


Figure 4.3 CT calculation over time

4.5 Comparison of the proposed CT-based metric with existing metrics

4.5.1 Comparison with other bitstream based content metrics

The proposed content type metric was compared with other metrics that quantifies the spatiotemporal complexity of videos in terms of composite complexity index (CPC), motion vector count (expressed as MVcount) and the absolute amount of motion defined by magnitude of motion (expressed as MVmag).

In this work, the derivation of the CPC is based on the content definition method used by [92] to quantify the complexity index of MPEG videos. This approach uses the composite complexity index (CC) in MPEG to calculate CPC for each frame type (e.g. I, P and B frames) in order to allocate a suitable amount of bitrate.

To determine the CPC, several video sequences of different spatiotemporal characteristics with different QP settings were encoded. The spatial and temporal resolutions were kept constant. For each QP setting the CPC was computed as,

$$CPC = \sum_{i=1}^{M} \left(B_i \times \frac{Q_i}{2} \right) \qquad (4.12)$$

where B denotes the number of bits of coded I, B and P frame, Q denotes the corresponding QP of each frame and M is the total number of I, B and P frames.

The MVcount of each sequence was determined by counting the number of motion vectors extracted from the encoded bitstream, while the MVmag was determined by calculating the sum of displacements of each motion vector extracted from the encoded bitstream. Based on the spatiotemporal metrics of CPC, MVcount (proposed by this project) and MVmag [27], two metrics were derived to gauge the content type of different sequences in terms of motion activity (MA). The first metric (MA_1) estimates video content type using a regression equation that combines CPC and MVcount. This is formalised as,

$$MA_1 = a + b \times (MV_{count}) + c \times (CPC) \qquad (4.13)$$

where a = 3.436, b = 0.115 and c= 0.136 are weighted coefficients that apply to all video sequences. The weights of the coefficients were determined by calculating the R^2 of correlation between the MVcount, CPC and video quality.

The second metric (MA_2) estimates video content type using a regression equation that combines CPC and MVcount and MVmag. This is formalised as,

$$MA_2 = \alpha + \beta \times MV_{count} + \delta \times CPC + \gamma \times MV_{mag} \qquad (4.14)$$

where $\alpha = 3.436$, $\beta = 0.115$, $\delta = 0.216$ and $\gamma = 0.0036$ are weighted coefficients that apply to all video sequences.

Comparison between CT and MA_1 and MA_2 based metric using PSNR as a benchmark

The quality of a video sequence is influenced by the spatiotemporal complexity of the sequence. Based on this assertion, a direct correlation is performed between quality (in terms of full reference PSNR) and the content type metrics (i.e. CT, MA_1 and MA_2) to determine the performances of the different metrics. To do this, the full reference PSNR, CT, MA_1 and MA_2 were computed for sequences that were encoded with different QP settings, same number of frames and the same spatial and temporal resolutions. A direct correlation in terms of R^2 is performed between the computed CT, MA_1, MA_2 and PSNR. The results are shown in Table 4.14. The results show that CT-based metric has a better correlation with video quality when compared to the MA_1 and MA_2-based metrics. These results are important because they established which of the content metric has the most influence on video quality and can be used for quality modelling.

Table 4.14 R^2 Correlation between content metric and PSNR

Cor. Coefficient	CT-based	MA_1-based	MA_2-based
R^2	0.63	0.41	0.39

4.5.2 Comparison between CT and pixel-wise content metric

To compare the proposed CT metric with existing pixel-wise content type definition, template matching (TM) algorithm in MATLAB and Sum of Absolute Difference (SAD) metric is used to compute the pixel differences between successful frames of several videos. The videos were encoded with different QP settings, same spatial and temporal resolutions. Template matching algorithm used is less processor intensive and is able to identify regions in an image with high motion activities, i.e. region of interest (ROI) [93]. The SAD metric on the other hand, was chosen because the work presented in existing literature also used the SAD metric to compute the pixel differences between video frames [43] [17].

The process of using TM and SAD to estimate the differences between pixels in a video frame involves comparing the absolute differences between each pixel in the template with the corresponding pixel within a sub image in a source image. The differences between the templates and the sub images are then summed up to create a similarity metric. For example, take a 2-D $a \times b$ (i.e. a by b pixels in size) template, T (x, y) that has to be matched within a source image S (x, y) with size $c \times d$ where $c > a$ and $d > b$. The SAD distance of each pixel with location (x, y) is calculated as,

$$SAD(x,y) = \sum_{K=0}^{(a-1)} \sum_{j=0}^{(b-1)} | S(x+k, y+j) - T(k,j) | \tag{4.15}$$

The smaller the SAD measurements between a template and a sub image, the higher the similarity between the template and the sub image at that particular location within the source image. Conversely, a larger SAD value will indicate less similarity between the sub image and the search template. Furthermore, a SAD estimation of zero will indicate that the local image and the template are identical with no motion activity between them.

To estimate the amount of motion activities in a video sequence in terms of pixel-wise differences between successive frames, the SAD of videos encoded with different QP settings were calculated. Results of SAD computation are shown in Table 4.15. The results show that SAD values of the sequences decrease with increased QP. Results also show that for the same QP setting, the sequences with the highest spatiotemporal activities have the highest SAD values when compared to the sequence with the lowest spatiotemporal activity. SAD based results are similar to CT-based results in that, the same videos show the similar trends, i.e. BQTerrace video has the highest CT value and Johnny has the lowest CT values.

Table 4.15 Impact of QP on SAD

QP	**Johnny**	**BasketballDrive**	**BQTerrace**
	SAD	SAD	SAD
22	1205363	8237614	9352530
27	1158456	8040411	9135383
32	1117858	7816071	8892938
37	1049957	7491549	8509505

The relationship between the SAD and video quality is further analysed. To do this, several video sequences with different spatiotemporal characteristics were encoded. The frame rate and frame resolutions were kept constant. For each QP setting, the SAD and the video quality were computed. The results are shown in Table 4.16.

Table 4.16 Relationship between QP and FS and APC

	Johnny		BasketballDrive		BQTerrace	
QP	SAD	PSNR	SAD	PSNR	SAD	PSNR
22	1205363	42.90	8237614	41.43	9352530	40.07
27	1158456	41.10	8040411	38.65	9135383	37.39
32	1117858	39.00	7816071	35.75	8892938	34.70
37	1049957	36.52	7491549	33.05	8509505	31.73

The results show that for the same QP setting, videos with highest SAD have the lowest PSNR values when compared to the video with lower SAD. Additionally, results also show that the PSNR for the three sequences decreases with increased QP, however, the decrement in PSNR is higher for sequences with high SAD than the video with low SAD.

To further understand the relationship between SAD and quality, more videos with different spatiotemporal characteristics were encoded. The frame rate and frame resolutions were kept constant. For each QP setting, the SAD and the video quality were computed. Figure 4.4 shows the results of SAD plotted against video quality.

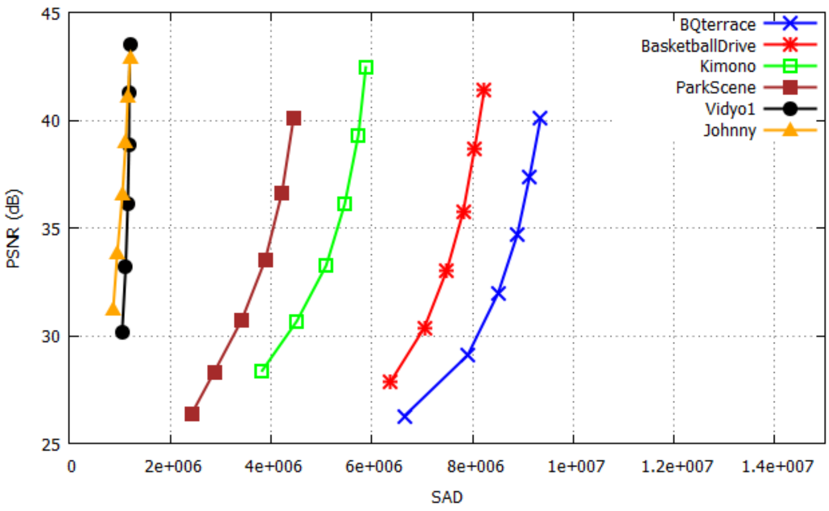


Figure 4.4 Impact of SAD on video quality

The results show that for almost the same SAD value, the video quality is different for each video sequence. However, the results show no clear relationship between SAD and PSNR for each individual video.

- **Comparison between CT-based and pixel-wise metric using PSNR as a benchmark**

A direct correlation between quality (in terms of full reference PSNR) and the two metrics (i.e. CT and SAD based metric) is performed to determine which of the metric has a better correlation with PSNR. To do this, the full reference PSNR, CT and the SAD were computed for several sequences that were encoded with different QP settings, same number of frames and the same spatial and temporal resolutions. A direct correlation in terms of R^2 is then calculated between the computed CT/SAD and PSNR. The results are shown in Table 4.17.

Table 4.17 R^2 Correlation between content metric and PSNR

Cor. Coefficient	CT-based	SAD-based
R^2	0.63	0.0013

The results show that CT-based metric has a better correlation with video quality when compared to the SAD-based metric.

Considering that the CT-based content type metric outperformed MA_1, MA_2 and SAD based metrics in terms of correlation with video quality. In this project, the CT-based metric is used to quantify the content type of different video sequences. This metric is subsequently used alongside the encoder parameter settings and network quality of service (NQoS) to develop quality models that estimate the quality of different video sequences.

4.6. Summary

In this chapter, an investigation has been undertaken to determine the impact of video content type on video quality. Based on the results obtained, it was concluded that although the quality of videos generally depends on the encoder settings and network impairments, the content type of a video actually determined how these two processes impact video quality. Thus necessitating the need to quantify the content type of a sequence and used as a variable when modelling video quality. Based on the aforementioned results, a new metric called content type (CT) is proposed, this metric is able to quantify the content types of different video sequences. The metric is based on temporal and spatial information of a sequence defined by the ratio of non-zero motion vectors to total number of motion vectors and the picture complexity of video sequences respectively. Based on the proposed metric, it was observed that videos with high CT have a high number of bits of coded I-frame than the video with lower CT. It was also observed that although videos with high CT have a higher number of bits, videos with low CT values have higher PSNR values than those with higher CT under the same encoding and network impairment.

The proposed metric was compared with existing motion amount (MA_1 and MA_2) and pixel-wise content type metrics that are based on a motion vector count (MVcount), composite picture complexity (CPC), magnitude of motion vector displacements and SAD. The results show that MA_1, MA_2 and SAD based metrics correlate very poorly with video quality (in terms of PSNR). Although SAD values do not correlate well with video quality, results further indicate that pixel-wise SAD metric can be used to classify videos in groups. Results also show a better correlation between video quality and the computed CT values.

Considering that the CT-based metric outperformed the MA and SAD based metrics, in chapter 6, the proposed metric will be used as an additional variable to develop new non-intrusive video quality models that can be used by content providers to estimate the quality of different videos delivered to end users' devices.

Chapter 5

5. Crowdsourcing Video Testing Screening Algorithm

5.1 Introduction

In recent years, research has shown that the measurement of end users' quality satisfaction for multimedia services continues to be a real challenge for content and network providers.

Different approaches have been proposed to measure and evaluate users' satisfaction of multimedia applications. These approaches employ either objective or subjective methods. Objective methods such as Peak Signal-to-Noise Ratio (PSNR), Mean Square Error (MSE) and Perceptual Evaluation of Video Quality (PEVQ) are less computationally intensive. However, these approaches are limited because their quality measurements do not factor in human visual perception directly. On the other hand, subjective methods ask participants in a video testing exercise to grade the quality of a multimedia application on a five-point Mean Opinion Score (MOS) scale which may range from "bad" (1) to "excellent" (5). Subjective test measurements are typically conducted in a laboratory (controlled) environment where different opinions about a multimedia application under test are collected from test panels that are supervised by the test moderator. Because of the controlled nature of this test, lab based subjective testing can be very accurate and reflect the end users' perceived quality of the application under test.

Although lab based testing tends to offer a better indication of end users' perceived quality, it can be time consuming and expensive because a large sample of participants is needed to obtain results that are statistically meaningful.

Given the costs and time demands posed by laboratory tests, recently, researchers have proposed cheaper and less time consuming approaches such as crowdsourcing subjective testing method that can produce results that are similar to those of lab based testing [4] [5].

The term “crowdsourcing” is a combination of “crowd” and “outsourcing” used to describe the steps taken to subcontract tasks to an anonymous group of people on the internet rather than contractors. This method of subjective testing has been enhanced by the dawn of ubiquitous Internet connectivity which has made it possible for the Internet crowd to participate in this form of activities (e.g. multimedia application testing) using their personal computers, laptops and tablet devices.

Although crowdsourcing is cheaper, less time consuming and makes it possible for researchers to reach a wide audience for subjective testing, the following challenges still remain unresolved:

The trustworthiness of evaluators: Because of the unsupervised nature of the approach, participants in crowdsourcing exercises may not pay attention to the task and in doing so, it would be difficult to trust results from crowdsourcing testing. This continues to be a major challenge to using this approach of quality evaluation. Furthermore, no mechanism has been devised to verify if the grading of multimedia applications (e.g. video clips) by crowdsourcing workers truly reflects their perceived quality.

Evaluators with multiple accounts for economic benefits: Given the anonymity and remunerations involved in this form of testing, it remains a challenge for test coordinators to determine the identities of participants in a crowdsourcing testing exercise especially the identity of those using multiple accounts to maximize their income.

Syndication: This is the act of forming groups or communities to compete and complete tasks on crowdsourcing platforms before other workers see or participate in the same exercise. How syndicate members perform the tasks at hand can hugely impact the quality of the results considering that not all interested members of the internet crowd are given the chance to share their opinions about the application under test.

Multimedia usage and application monitoring: The uncontrolled nature of the test makes it difficult to determine if participants are fully involved in quality observation before grading. For example, how to determine if an evaluator has completely watched a video clip before grading remains a huge challenge.

Although recent research in crowdsourcing has recognized some of the challenges listed above and have proposed different approaches to solving them. At present, no crowdsourcing testing platform has extracted information such as IP addresses, device and browser type, time spent on a website and scores to hidden reference sequences to design an algorithm that can be used to determine the validity and trustworthiness of crowdsourcing evaluators.

In this chapter, the challenges listed above are addressed by proposing a screening algorithm that uses an evaluator's own data to determine the trustworthiness of his/her scores, detect if multiple accounts have been used, determine if an evaluator is a member of a syndicate, and finally, determine the usage of the application under test.

The novelty in this approach is in capturing evaluator's own data and then utilising a screening algorithm based on this data to determine the reliability of crowdsourcing video QoE testing.

This chapter is organized is as follows. Data set generation, platform that was used for subjective test, based on crowdsourcing and a screening algorithm is presented in Section 5.2. Section 5.3 describes the steps taken to evaluate the performance of the proposed algorithm for screening crowdsourcing subjective test results and limitations of the proposed algorithm. Section 5.4 summarizes the chapter.

5.2 Video quality assessment

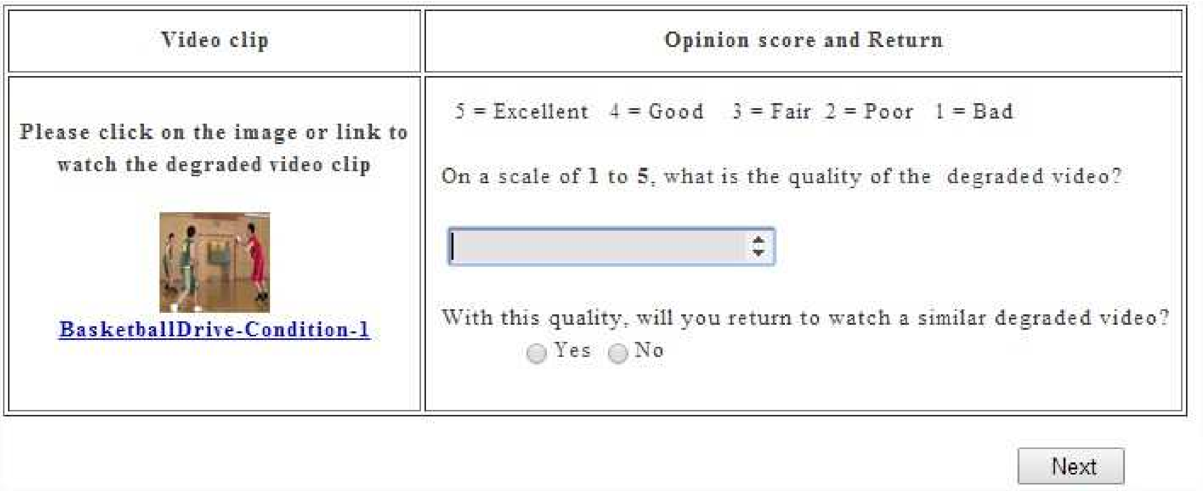
The degraded (impaired by encoder QP setting and network packet loss) together with unimpaired video sequences generated in chapter 3 are used for video quality assessment in this chapter.

5.2.1 The design of video quality evaluation platform

To carry out subjective testing, fourteen identical websites [23] are built. These websites are used to upload videos that subjective evaluators could evaluate. Each video was uploaded to its own webpage and rated independently on a discrete five level scale from “bad” (1) to “excellent” (5). To rate a video sequence after watching, participants were asked; “On a scale of 1 to 5, what is the quality of the degraded video?” where 1 indicates the worst quality (“bad quality”) and 5 indicated the best (“Excellent”). The individual ratings were mapped onto a Mean Opinion Score (MOS) scale from 1 to 5 where 1=“very bad”, 2=“poor”, 3=“fair”, 4=“good” and 5=“excellent” as recommended by ITU-R BT.500-13 [69]. To simplify numerical analysis and plotting of graphs, individual SSCQE ratings were averaged to obtain the MOS.

The time for subjective evaluation on each website was restricted to 10 minutes as recommended by ITU. The test sequences were randomly split (using a Latin square randomization technique [70]) into fourteen sets of fifteen videos each. Randomization reduces human memory effect [69].

Figure 5.1 shows the voting box (i.e. text box) of each website, the numerical value of this box was restricted to range between 1 and 5; evaluators whose scores were beyond this range were gently reminded with the following message “please grade the video with values between 1 and 5”.



Video clip	Opinion score and Return
Please click on the image or link to watch the degraded video clip BasketballDrive-Condition-1	5 = Excellent 4 = Good 3 = Fair 2 = Poor 1 = Bad On a scale of 1 to 5, what is the quality of the degraded video? With this quality, will you return to watch a similar degraded video? ○ Yes ○ No

Figure 5.1 Snapshot of voting box

The voting period was not restricted as evaluators had the option to view a video sequence many times before grading and confirming their scores through the “next” button presented on the testing website. On clicking the “next” button, the following activities and records are written to the backend database:

1. Grading or scores
2. Current date and time
3. Browser type and device information (device information extracted using the browser)
4. Network IP address information
5. Answers to questions
6. A reminder message to the evaluator to perform unfinished tasks
7. Load the next page containing another video sequence
8. Increase the progress bar

The ITU recommended 2 seconds waiting time between voting and loading the next video sequence in SSCQE subjective testing method was implemented by integrating a countdown timer between web pages.

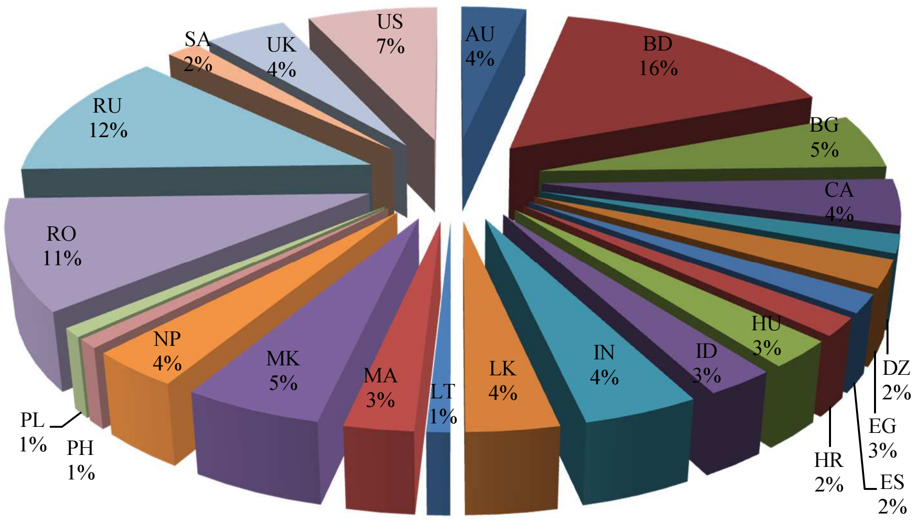
Considering that stalling and freezing may occur during quality evaluation because of limited Internet access bandwidth on the evaluator side which may as a consequence impact the manner in which videos are graded, the video sequences uploaded on each web page were downloadable i.e. evaluators had option to completely download the video sequence before starting the actual test when using Microsoft windows or Mac OS computer in order to overcome the impact of the network delay.

5.2.2 Subjective crowdsourcing testing

To capture the perceived quality in a crowdsourcing environment, ITU recommended SSCQE testing method [69] was used. This testing method shows the video being evaluated to evaluators without them having access to the original video. It mimics the perceptual response of typical video application consumers who have no access to the original video sequence.

The test sequences to be evaluated (including reference videos) were split and uploaded to fourteen test websites [23]. These test websites were then uploaded to Microworkers test platform once every week to control the time between testing. The test on Microworkers crowdsourcing testing platform was conducted over 4 months.

A total of 114 internet crowd evaluators from 23 different countries took part in the test with the majority of the evaluators coming from Bangladesh (18 evaluators), Russia (14), Romania (12), USA (8) and Bulgaria (6) respectively. The distribution of the evaluators from different countries is shown in Figure 5.2. The data collected in this thesis were only used for research purpose and were immediately destroyed afterward as agreed by the terms and conditions set by the ethical approval committee of Plymouth University.



AU	BD	BG	CA	DZ	EG	ES	HR	HU
Australia	Bangladesh	Bulgaria	Canada	Algeria	Egypt	Estonia	Croatia	Hungary

ID	IN	LK	LT	MA	MK	NP	PH	PL
Indonesia	India	Sri Lanka	Lithuania	Morocco	Macedonia	Nepal	Philippines	Poland

RO	RU	SA	UK	US
Romania	Russia	South Africa	United Kingdom	United States

Figure 5.2 Distribution of crowdsourcing evaluators by countries

5.2.3 Outlier detection

To determine if outliers existed in the crowdsourcing subjective test in terms of deviation from other scores, outlier detection is performed. This detection process was applied to all subjective test results. For example, MOS scores by evaluator j testing condition i with score sij were considered an outlier if: $sij \ < \ q1 - 1.5\ (q3 - q1)\ \&\ sij \ > \ q3 \ + \ 1.5\ (q3 - q1)$, where $q1$ and $q3$ are the 25th and 75th percentiles of the score distributions for the testing condition i respectively [94]. This interquartile range (IQR) [55] corresponds to 99.3% coverage of data that is normally distributed. Based on this statistical identification of outliers, crowdsourcing evaluators whose scores were outside this range were considered outliers. Results show that more than half of the score from crowdsourcing were outside the quartile range.

To get the ground truth of outlier existence on crowdsourcing results, a PCA based on correlation matrix is performed. Results of PCA are shown in Figure 5.3.

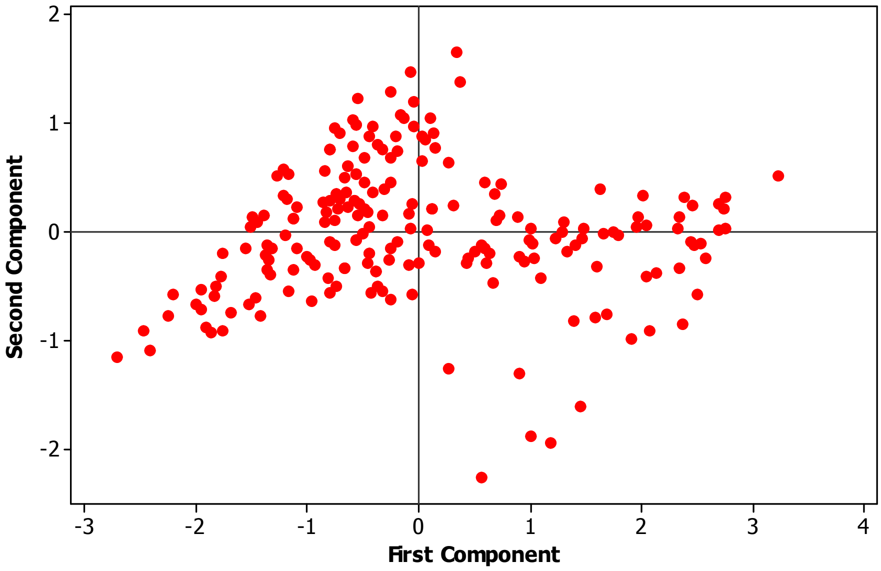


Figure 5.3 PCA of unscreened crowdsourcing subjective scores

Figure 5.3 shows that MOS scores from crowdsourcing evaluators are unevenly distributed along the first and second principal component indicating the existence of outliers. Based on the results, in the following sections, an algorithm that can be used to detect unreliable subjective test evaluators is proposed.

5.2.4 Screening unreliable crowdsourcing evaluators

Given the detection of outliers in the previous section, in this section, an algorithm for screening and identifying unreliable evaluators that may impact the quality of results obtained from the crowdsourcing platform is proposed. This algorithm takes into account some of the issues that have been identified to impact the quality of crowdsourcing results (especially those listed in section 5.1). Based on this, the decision to accept or reject test results from a crowdsourcing worker depends on how evaluator’s results agreed with predefined conditions not known to them. The decision making process is summarized in the form of a flow decision tree presented in Figure 5.4. It should be noted that condition “N” in the algorithm indicates any other condition that has not been used in the design of the current methodology. Therefore, this algorithm can be extended to include other conditions whenever possible.

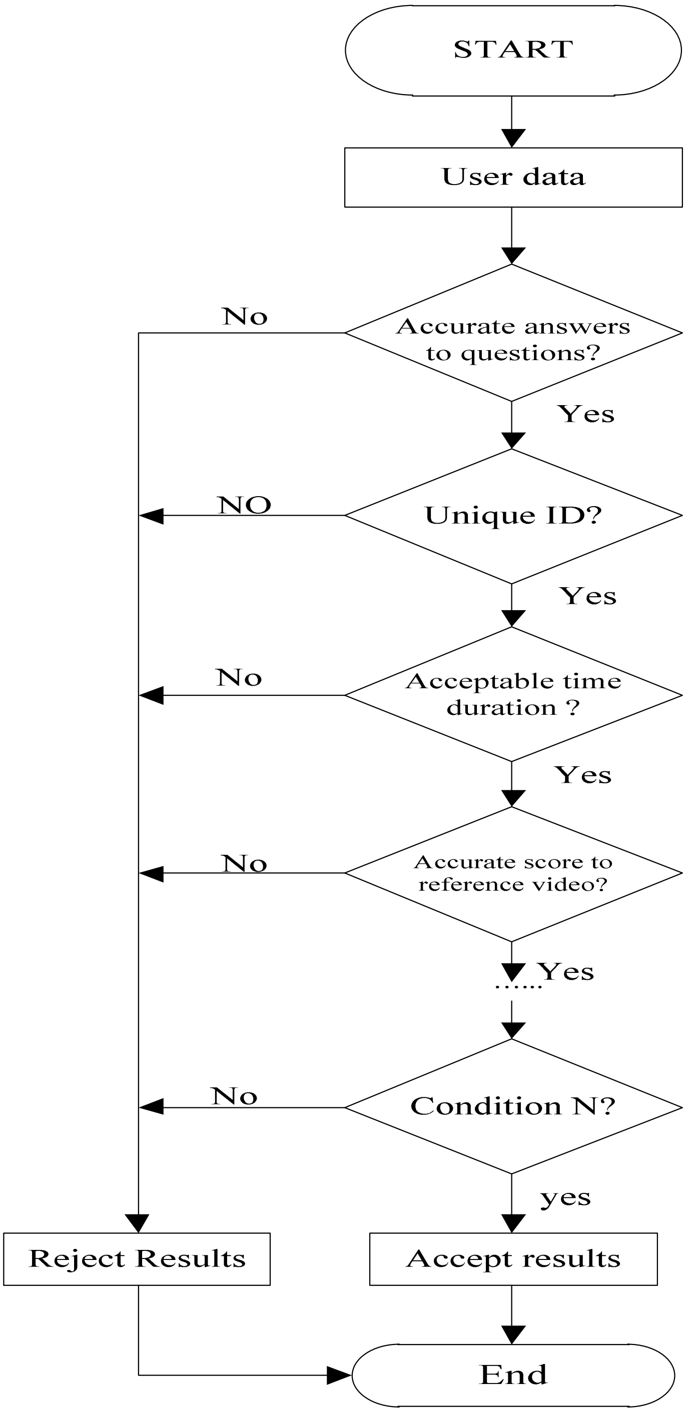


Figure 5.4 Crowdsourcing screening algorithm

To screen and determine the reliability of a crowdsourcing evaluator, a thorough analysis is performed on his/her data. This analysis takes the following conditions into account.

1. *Accurate answers to questions*: To determine the reliability of an evaluator, a set of questions is asked prior to the beginning of the test, for example, "what device are you using to take the test?", "Where are you taking the test?" Because the crowdsourcing platform is able to extract browser, device and IP address information, the answers to these questions can be obtained without the knowledge of the evaluator. By comparing the answers from the extracted data to those from the evaluator, it can be determined if the evaluator's MOS scores should be considered for further analysis or not.
2. *Determine if an evaluator has unique identification*: Given that crowdsourcing evaluators are anonymous in nature and unsupervised; this has continued to be a challenge to test organisers. Though a challenge, it also presents an opportunity to some crowd workers to maximize their income by creating multiple accounts for testing purposes. To mitigate this, the extracted evaluator's information (i.e. IP address, browser and device type) are compared to determine if multiple accounts have been used or not
3. *Determine if an evaluator actually performs the task (i.e. watch the video before scoring):* Test organisers rely on evaluators to perform the task in order to get paid, but that is not always the case as some of the evaluators are only interested in the remunerations. To determine if evaluators actually watch the processed video sequences, a hidden timer that calculates how many seconds an evaluator spent on a testing webpage that contained a video sequence is used. Following ITU recommendation, all video sequences were approximately 9 seconds long and because ITU recommends 10 seconds voting period, we expected reliable evaluators to spend on a range of 18 - 21 seconds on a web page before proceeding to the next video. Evaluators whose time on the website was below this range or on any of the web pages containing videos were automatically rejected.

4. *Accurate scores to hidden reference video*: hidden reference video sequences were incorporated on all testing websites. Because these sequences are not impaired, reliable evaluators (as recommended by ITU) who understood the task should give a MOS score of close to 4.5 or above 4.5 to these reference sequences. Evaluators who failed to score these sequences accurately were automatically rejected. It should be noted that accurate scores to hidden reference video is one of the applications of the common method of including "gold standard data" into crowdsourcing platform design to increase the trustworthiness of crowdsourcing evaluators [5].

5. *Evaluators not a member of a syndicate*: Given that syndicate members may know themselves, have common interests and live in a closed geographical location, the date and time of evaluators with a similar block of IP addresses were compared. Evaluators' information that shows minimum time difference and dates were further analysed to determine the existence of voting patterns. Furthermore, results from the different testing websites were compared to determine if similar trends existed.

As a result of applying the screening algorithm, 60 out of the 114 of the crowdsourcing workers were rejected. The breakdown of rejections for each condition is summarized in Figure 5.5.

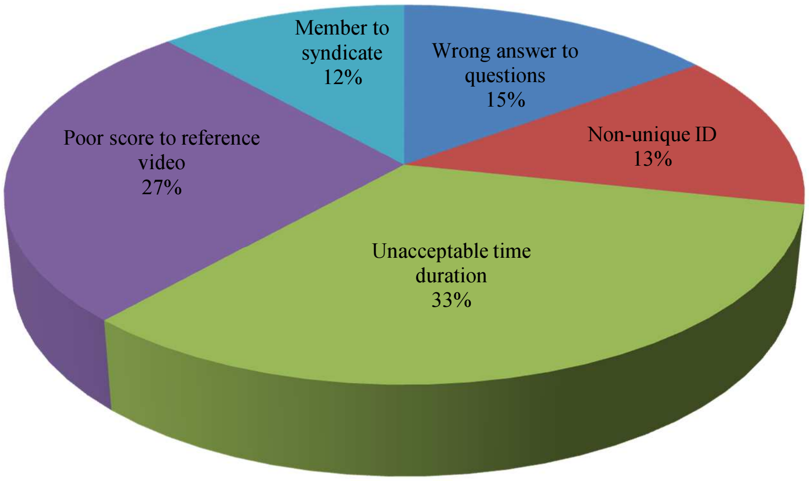


Figure 5.5 Crowdsourcing workers rejection pier chart

Results show that a majority (33%) of the rejection were due to unacceptable time spent on the websites (i.e. not watch the video sequences), 27% were rejected due to inappropriate MOS score submission (e.g. very low MOS) for the hidden reference videos. Results also show that 12 % of the evaluators were rejected because they show similarities in their voting and had the same IP address block (i.e. belonging to a syndicate).

5.3 Performance evaluation of screening algorithm

In this section, different approaches are used to determine if the proposed screening methodology of determining unreliable evaluators can lead to improvement in crowdsourcing results.

5.3.1 Laboratory based subjective testing

To determine the performance of the proposed screening algorithm, a lab based subjective testing exercise (controlled environment) is conducted using the same ITU recommended SSCQE testing methodology used in crowdsourcing based subjective testing environment. All test sequences, including reference videos (same as in the crowdsourcing based testing) were split and uploaded to fourteen test websites (same websites used in crowdsourcing based testing) [23]. It should be noted that the lab based subjective test is the same test conducted in chapter 3 section 3.6.2.

5.3.2 Lab based vs. unscreened crowdsourcing MOS

In this section, an evaluation is performed between the unscreened crowdsourcing results and lab based results in terms of R^2, Root Mean Square Error (RMSE) and empirical cumulative distribution function (ECDF). This evaluation is important because it helps determine if the screening algorithm is necessary or not.

The mutual differences between unscreened crowdsourcing and valid lab results is shown in Figure 5.6 where an ECDF plot of unscreened Microworkers MOS is compared with laboratory MOS results of the video sequences under test.

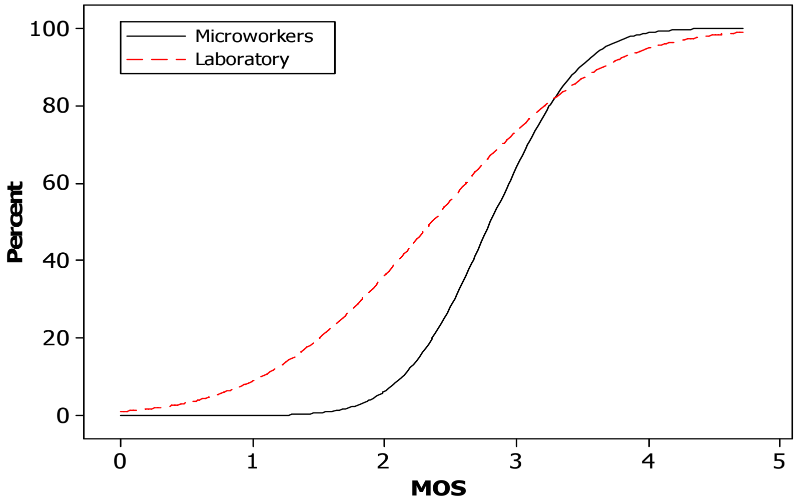


Figure 5.6 ECDF of unscreened Microworkers when compared with lab MOS

Results show that a high percentage (up to 80%) of the unscreened crowdsourcing evaluators underscored all video sequences with a MOS score of about 3.2 or less. These scores are in direct contrast with valid laboratory results which are used as a benchmark to determine the accuracy of crowdsourcing results. Additionally, ECDF results also show that approximately 20% of the unscreened crowdsourcing evaluators over graded the remaining video sequences with MOS scores of 4.8 or less.

To further determine how both results correlate in terms of R^2 and RMSE, a direct correlation between the two results is performed. The correlation coefficient R^2 and RMSE are summarized in Table 5.1.

Table 5.1 R^2 and RMSE for unscreened crowdsourcing and valid lab based MOS

Cor. Coefficient	BasketballDrive	Vidyo1	Johnny	BQterrace	Kimono1	ParkScene
R^2	0.38	0.41	0.43	0.35	0.37	0.31
RMSE	0.89	0.73	0.81	0.94	0.78	0.82

Results in Figure 5.6 and Table 5.1 clearly show that unscreened crowdsourcing results correlate poorly with lab based MOS. This is important because it has determined that it is necessary to apply the screening algorithm on crowdsourcing results.

5.3.3 Lab based vs. screened crowdsourcing MOS

Considering that the results in Figure 5.6 and Table 5.1 show very poor correlation between unscreened crowdsourcing and lab based results, in this section, the ECDF of MOS from the six video sequences is plotted to determine the consistencies between the screened crowdsourcing and valid lab results as shown in Figure 5.7.

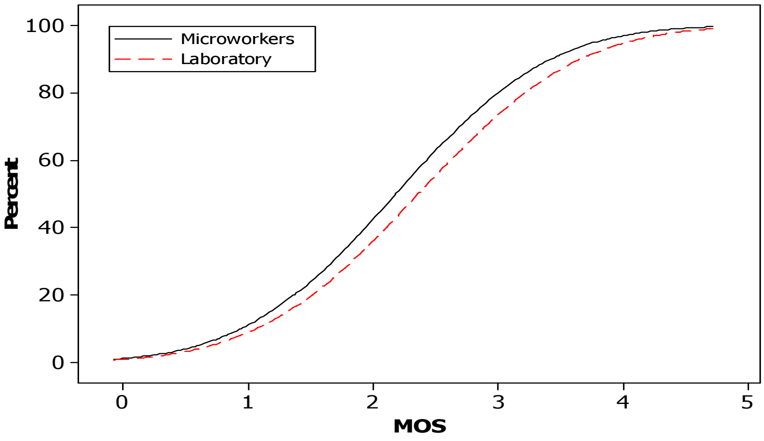


Figure 5.7 ECDF of screened Microworkers when compared with lab MOS

It can be observed in Figure 5.7 that screened crowdsourcing MOS are consistent with laboratory based MOS. This is important because it indicates that the filtering algorithm improves the quality of crowdsourcing results.

To quantify the quality improvement made on crowdsourcing results after the screening algorithm has been used to filter out unreliable evaluators, a direct correlation between the two MOS results in terms of R2 and RMSE is performed. The correlation results are summarized in Table 5.2.

Table 5.2 R^2 and RMSE for screened crowdsourcing and valid lab based MOS

Cor. Coefficient	BasketballDrive	Vidyo1	Johnny	BQterrace	Kimono1	ParkScene
R^2	0.97	0.98	0.95	0.93	0.96	0.98
RMSE	0.25	0.28	0.28	0.28	0.26	0.19

It can be observed in Table 5.2 that by applying the screening approach of determining genuine evaluators; the crowdsourcing results can be significantly improved by approximately 59% when compared to unfiltered (unscreened) results. This is important because crowdsourcing subjective testing may offer a cheaper alternative to lab based test which can be expensive and time consuming. However, due diligence must be taken to weed out untrustworthy test takers.

5.3.4 Objective quality measurement vs. screened crowdsourcing MOS

To further determine the performance of the proposed screening algorithm, the quality of each video sequence used in subjective quality testing is estimated using; 1) intrusive measurement based on PSNR metric, 2) convert the measured PSNR to MOS, and 3) performed an evaluation between the screened crowdsourcing results and objective results in terms of R^2, Root Mean Square Error (RMSE) and ECDF.

To convert intrusive PSNR values to MOS, the polynomial function PSNR to MOS conversion metric proposed in this project is used (see chapter 6 section 6.3.4).

Results of ECDF, R2 and Root Mean Square Error (RMSE) are shown in Fig.5.9 and Table 5.3.

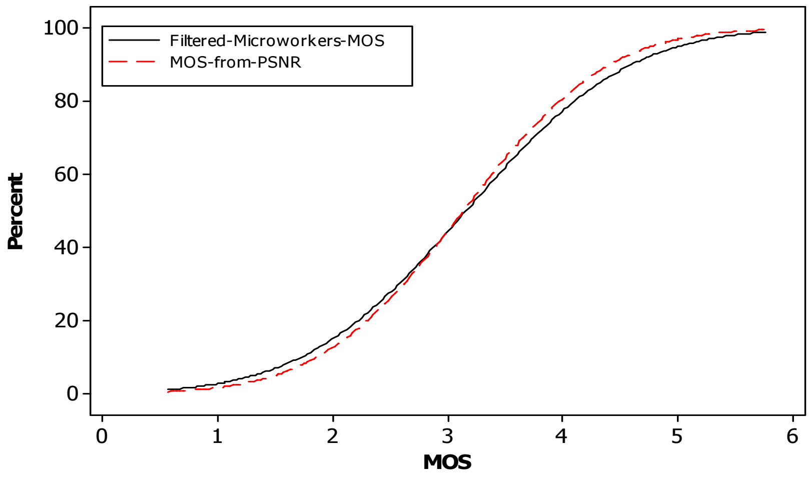


Figure 5.8 ECDF of screened Microworkers when compared with MOS from PSNR

Table 5.3 R^2 and RMSE for screened crowdsourcing and MOS from PSNR

Cor. Coefficient	BasketballDrive	Vidyo1	Johnny	BQterrace	Kimono1	ParkScene
R^2	0.97	0.97	0.98	0.93	0.97	0.96
RMSE	0.19	0.25	0.23	0.39	0.23	0.38

It can be observed in Figure 5.8 and Table 5.3 that the crowdsourcing screened results are consistent with objective based estimated MOS results. Additionally, the screened results show a good correlation with MOS from PSNR.

5.3.5 Limitations of the proposed algorithm

The proposed algorithm assumes the following:

- Evaluators' devices support the coding standard as it is not often feasible to provide uncompressed sequences to evaluators because of excessive bandwidth requirements.
- The laboratory test used as benchmarks to determine the validity of crowdsourcing results offers the most accurate reflection of the perceived quality.
- Given bandwidth constraints, evaluators will have to download video sequences before watching and evaluating to avoid additional network impairments.

Based on these assumptions, the proposed algorithm has some inevitable flaws that are beyond the assessor in terms of evaluators' fulfilling the five conditions. For example, the identification of evaluators' ID based on IP, device and browser type may not be 100% effective as an evaluator may use a different username, IP, device and browser type to maximize their income. Additionally, the exclusion of evaluators based on low score to the reference sequences may be limited because these scores may be due to surrounding disturbances (e.g. lights, low performance in wireless connectivity etc.). So excluding these users may be biased or unjustified. Excluding evaluator based on the time they spent on the web page containing a video sequence may also be limited because some evaluators may be quicker at watching and grading the videos than others. Finally, using the lab results as benchmarks to determine the validity of crowdsourcing results may be inaccurate as the results from the lab may be impacted by surrounding disturbances.

Although the proposed approach of screening subjective test may not be 100% perfect or accurate, by combining the predefined conditions, unreliable subjective evaluators of multimedia applications such as videos can be identified and weeded out of the subjective test data as shown by the results presented in this chapter.

5.4 Summary

In this chapter, some of the challenges of crowdsourcing subjective video testing have been addressed by designing and implementing a screening methodology that detects genuine evaluator based on predefined set of criteria. This approach takes into account an evaluator's own information extracted from the web browser. The extracted information includes IP information, device and browser type. Using the extracted information, a screening algorithm has been designed. This algorithm is able to screen and eliminates untrustworthy subjective evaluators and improved the crowdsourcing results by up to 59% when valid crowdsourcing results were compared with laboratory based results. Based on the screening algorithm, it was observed that a majority of the evaluators (33%) were rejected because they did not watch the video sequences before scoring. The results also show that, although crowdsourcing video testing makes it possible to reach a wider audience and in some cases a cheaper alternative to lab based testing, unfiltered crowdsourcing results can be misleading as they show a very poor correlation with controlled lab based results.

The work presented in this chapter is important because, it provides the basis for which crowdsourcing subjective video testing can be used to reduce the cost and time consumption that are associated with laboratory testing.

In the next chapter, the valid laboratory and crowdsourcing based subjective results will be used to develop video quality prediction models that can be used to predict the quality of different video sequences.

Chapter 6

6. Video Quality Prediction Models for HEVC Videos

6.1 Introduction

In chapter 2, the different objective (both intrusive and non-intrusive) and subjective video quality measurement methods were discussed. Intrusive methods are accurate and efficient; however, they are impractical in real time quality monitoring and estimation. Hence, non-intrusive methods are preferred to intrusive analysis as they are more suitable for real-time quality prediction. Additionally, because subjective testing approaches of quality measurement cannot be used in real-time, in this chapter, new models to predict the quality of different video sequences delivered over IP based network non-intrusively are developed. The prediction of video quality is based on a combination of parameters associated with the encoder parameter settings, IP network quality of service (NQoS) and video content type. The models predict video quality in terms of PSNR and MOS.

This chapter is organized as follows. Section 6.2 describes the prediction and performance evaluation of initial video quality prediction model. MOS based encoded and end-to-end video quality prediction models are presented in Section 6.3. Section 6.4 summarizes the chapter.

6.2 PSNR-based encoded video quality prediction

In this section, a reference free quality prediction model that can be used by content providers to predict the initial video quality of HEVC encoded videos is proposed.

Results in chapter 3 show that the initial video quality depends on the content type and the initial encoding process settings such as QP. As expected, lower QP results in higher bitrate (BR) which may lead to increased video quality. On the other hand, as the QP increases the BR reduces resulting in reduction of video quality. Meanwhile, different videos have different encoding bitrate requirements because of differences in spatiotemporal characteristics. Results also show that for the same QP setting, higher motion/complexity videos have a higher BR when compared to those with lower motion/complexity. Additionally, the results show that for all the encoded videos, the quality generally drops with increased compression (defined by QP). However, the drop in quality is steeper for videos with higher motion/complexity than those with lower motion.

Based on the conclusions arrived at in chapter 3, in this section, a content-based video quality prediction (CVQP) model is proposed. This model takes into account encoder parameter settings such as quantization parameter (QP) and video content type (metric) discussed in chapter 4. The initial encoding video quality model developed in this section is based on PSNR.

6.2.1 Experimental setup

Figure **6.1** shows a block diagram of the system that was designed and used to develop a non-intrusive model that is able to predict the quality of different HEVC encoded videos objectively. The system consists of encoding/decoding, video content type metric derivation, modelling and evaluation of the proposed video quality model.

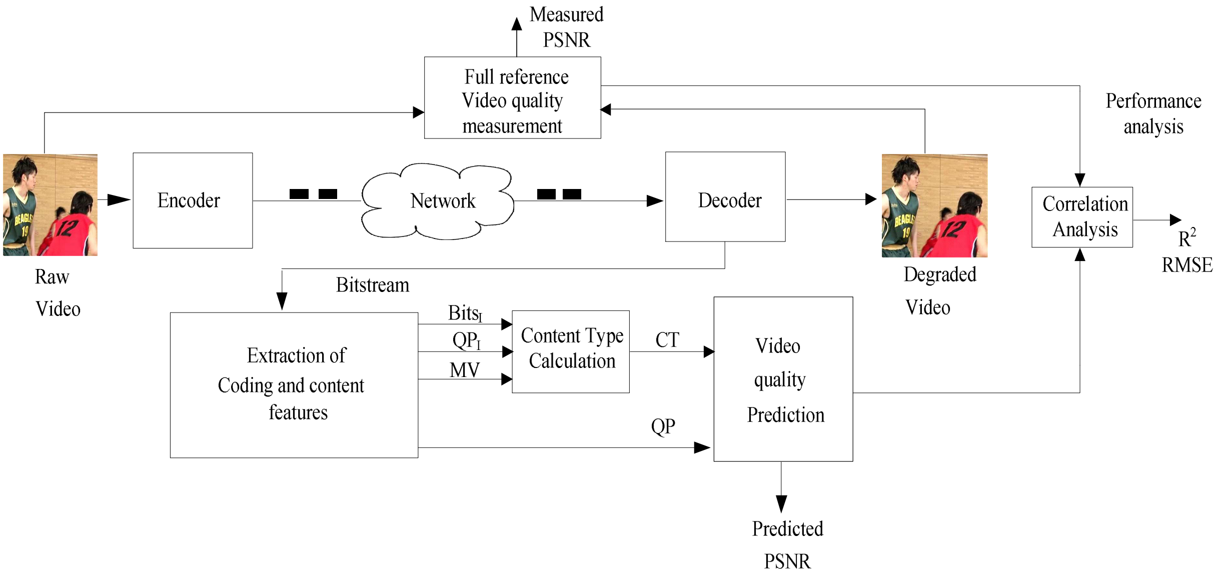


Figure 6.1 Block diagram of a system for initial encoded video quality prediction

- **Encoding and decoding**

Several video sequences from the recommended HEVC test sequences were encoded with different QP settings that ranged from 17 to 47 as recommended by JCT-VC. Additionally, different spatial and temporal resolutions were used throughout the encoding process. HEVC reference software HM-10.0 was used for encoding. The resultant encoded bitstreams were decoded using a modified versions HEVC decoder. The decoder was modified in order to extract the features/parameters that are needed to quantify video content type and quality prediction.

- **Statistical analysis of encoding and content type impact on video quality**

Results in chapter 3 show that encoded video sequences have different quality measurements even under the same encoder settings. This indicates that parameters other than the encoding settings impact the quality of encoded videos. Considering that, the sequences were encoded with the same spatial resolution and the same frame rate; the variation in quality may be due to the content type (CT) of the sequences.

To actually determine if the content type has any influence on video quality, several video sequences with different spatiotemporal characteristics were encoded, the spatial resolution and the frame rate were the same. Additionally, the sequences were encoded with different QP settings. The full reference PSNR and the CT of each sequence were computed, additionally, the QP values were also recorded. Using the QP, the computed CT and PSNR, a two-way Analysis of Variance (ANOVA) [95] is performed on the recorded QP, CT and PSNR of the sequences used. The results are presented in Table 6.1 where the Sum of Squares is represented in the first column of the ANOVA, second column is the Degrees of Freedom, the third column is the Mean Squares defined as the ratio of Sum of Squares to Degrees of Freedom. The F statistic is shown in the fourth column and the p-value is given in the fifth column. The p-value is determined from the cumulative distribution function (cdf) of F [95]. The P-values obtained from ANOVA indicates how QP and CT impact PSNR and also the combined impact of the two variables on PSNR. For example, a parameter with a p-value of less than 0.05 (p-value ≤ 0.05) will indicate that PSNR is significantly affected by such variable. On the other hand, a P-value of more than 0.05 will indicate that such variable has no significant impact on PSNR. The results in Table 6.1 show that both QP and CT impact quality i.e. the P-value for QP is 0 and the P-value for CT is 0.027. Results also show that the interaction between QP and CT does not significantly impact PSNR by having a p-value of 0.066. In order of ranking, results show that the QP has more influence on PSNR (with p-value of 0) than CT whose p-value is 0.027. The results are important because they provide an understanding on how statistically significant these parameters are and how they can be used in modelling video quality. Considering that CT varies with QP, to accurately determine the impact of CT on PSNR values, the CT and QP are used as continuous variables (covariate) [96] in ANOVA.

Table 6.1 ANOVA results for QP and CT impact on PSNR

Source	Sum of squares	Degree of Freedom	Mean Squares	F-Statistics	P-value
QP	1204.790	1	129.500	70.930	0.000
CT	56.990	1	9.700	5.320	0.027
QP*CT	6.540	1	6.540	3.580	0.066
Error	69.380	38	1.830		
Total	1337.710	41	129.500		

The findings of ANOVA can be summarized as follows:

1. Encoded video quality depends on content type as also indicated by authors in [33] and [48]. It is therefore important to measure and consider content type when designing a video quality prediction model.
2. The encoding QP setting determined the initial video quality and it is content dependent. To adequately predict video quality, the encoder setting parameters such as the QP should be taking into consideration.
3. The QP settings have more impact on video quality than the CT.

Considering that video content type significantly impact video quality as shown by the ANOVA. In the next section, the CT together with the encoder QP setting will be used to develop a non-intrusive video quality prediction model to estimate the initial encoding quality of different video sequences without the need for the original sequences. This approach of quality prediction addresses some of the limitations with content blind quality models. Additionally, the approach of using a single metric to quantify the content type of different video sequences also addresses the limitations with content type identification metrics that are based on grouping and classification of videos.

6.2.2 PSNR prediction

This section outlines the procedure for developing non-intrusive objective prediction model based on PSNR. To derive the initial video quality prediction model, the relationships between video quality and parameters needed for prediction i.e. QP and video content type are first established. These relationships are then used to develop a regression-based quality model that is able to accurately estimate the initial quality of different video sequences.

- **Relationship between QP and video quality**

To determine the relationship that exists between the encoder setting QP and quality, several video sequences with different spatiotemporal characteristics were encoded, using different QP settings. The frame rate and spatial resolution were kept constant. For each QP setting, the quality in terms of PSNR was computed and the QP recorded. A scatter plot of the recorded QP against PSNR is plotted. To determine the relationship between QP and PSNR, a trend line fitting based on exponential, linear and logarithmic relationships is performed. The result of each fitting in terms of R^2 is recorded as shown in Table 6.2. The results show that linear relationship has the highest R^2 thus indicating that QP and PSNR have a linear relationship as shown in Figure 6.2. Even though, the fixed QP setting is limited in that, the measured PSNR may poorly correlates with actual video quality, the relationship established between QP and PSNR is important because it provides an explanation on how QP impact video quality and how it should be used in designing video quality model.

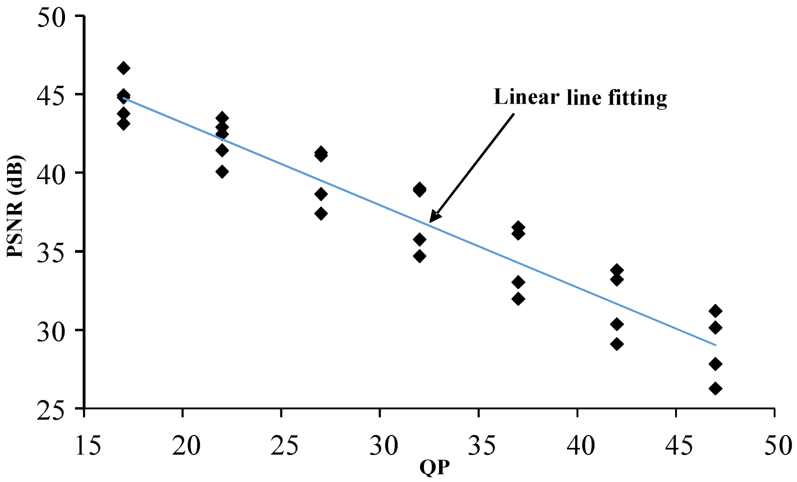


Figure 6.2 Relationships between QP and PSNR

Table 6.2 Relationship between QP and PSNR

Function	R^2
Exponential	0.89
Linear	0.90
Logarithmic	0.88

Relationship between content type and video quality

Results in chapter 3 show that the quality of an encoded video sequence varies with the spatiotemporal characteristics of the video. The level of video quality degradation as a result of variation in encoder settings is higher for high spatiotemporal content when compared to video sequences of lower spatiotemporal complexity. To determine and establish the relationship that exists between video content type (CT) and video quality in terms of PSNR, several videos are encoded with different spatiotemporal characteristics, different frame rate and different resolutions. The QP setting was kept constant at 22. The CT (computed using CT metric proposed in chapter 4) and the full reference PSNR of each sequence were computed. To determine the relationship between CT and quality, trend line fitting based on exponential, linear and logarithmic relationships are performed. The result of each fitting in terms of R^2 is recorded as shown in Table 6.3. The results show that logarithmic relationship has the highest R^2 thus indicating that CT and video quality have a logarithmic relationship as shown in Figure 6.3. The relationship is important because, it establishes how video content type relates to quality and how it can be used to model video quality.

Table 6.3 Relationship between CT and PSNR

Function	R2
Exponential	0.67
Linear	0.66
Logarithmic	0.80

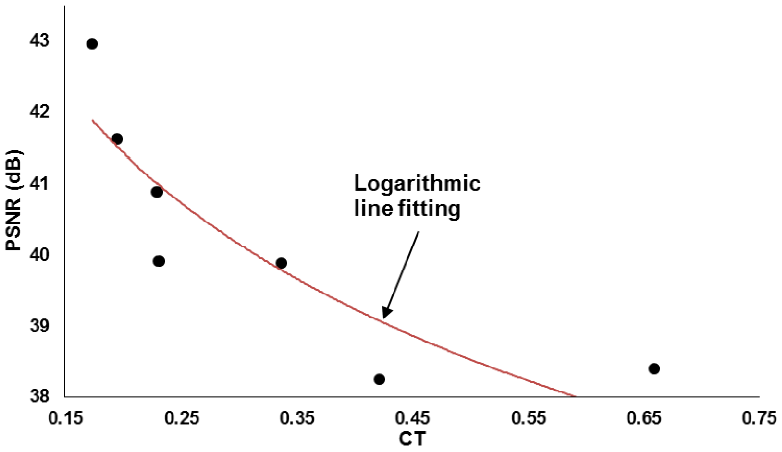


Figure 6.3 Relationship between CT and PSNR

- **Encoded video quality model derivation**

This section presents a content-based reference free quality prediction for HEVC encoded videos that can be used by content providers to predict the initial video quality in terms of PSNR metric.

Results in Figure 6.2 show that the PSNR values for each video sequence decreases as the QP increases. ANOVA on the other hand, shows that both QP and video content type significantly impact quality. Based on these results, the initial encoding quality (PSNR_e) of a video sequence was determined and formalised as,

$$PSNR_e = f(QP, CT) \tag{6.1}$$

Considering that quality has a linear relationship with QP and a logarithmic relationship with CT as shown in Figure 6.2 and Figure 6.3 respectively. It can be determined that the initial encoding quality (PSNR$_e$) of different video sequences can be estimated as,

$$PSNR_e = \alpha + \beta \times (QP) + \gamma \times \ln(CT) \tag{6.2}$$

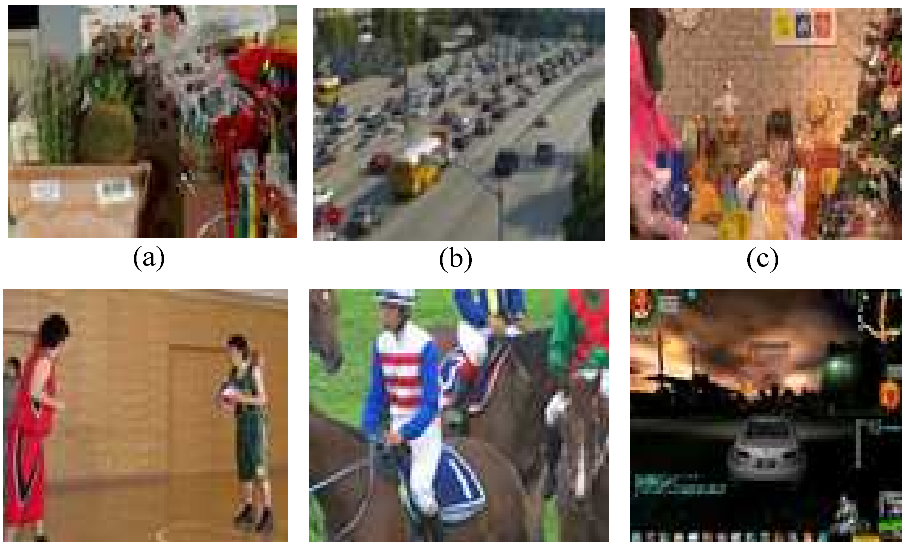
where $\alpha = 48.8$, $\beta = -0.856$ and $\gamma = -10.8$ are model parameters derived through regression. The derived coefficients hold for all video sequences.

6.2.4 Performance evaluation of proposed prediction model

This section outlines the different approaches that were taken to evaluate and determine the validity of the proposed initial encoded video quality model.

- **Performance evaluation with unseen objective data**

To evaluate the performance of the proposed model, several video sequences of different spatiotemporal characteristics were encoded. These sequences were not used in model derivation. They were encoded using different QP settings, different temporal resolution and spatial resolutions. The sequences include Cactus, Traffic, PartyScene, BasketballPass, RaceHorses and ChinaSpeed. A snapshot and encoder settings of the testing sequences are shown in Figure 6.4 and Table 6.4 respectively.



(a) (b) (c)

(d) (e) (f)

Figure 6.4 Snapshots of testing video sequences (a) Cactus, (b) Traffic (c) PartyScene (d) BasketballPasses (e) RaceHorses and (f) ChinaSpeed

Table 6.4 Encoder settings for testing sequences

Sequences	**Spatial resolution**	**Frame rate (fps)**
Cactus	1920 x 1080	50
Traffic	2560 x 1600	30
PartyScene	832 x 480	50
BasketballPass	416 x 240	50
RaceHorses	832 x 480	30
ChinaSpeed	1024 x 768	30

The PSNR of each test sequence encoded at a given QP was computed using the proposed model. The accuracy of the model is given in terms of correlation coefficient R^2 and Root Mean Square Error (RMSE) as summarized in Table 6.5 and

Table **6.6** and Figure 6.5 where a scatter plot of actual PSNR is plotted against model predicted PSNR ($PSNR_e$). Figure 6.6 and Figure 6.7 on the other hand, show graphs of model for the six testing and six training sequences respectively. A correlation coefficient of around 95% with the training video sequences and 94% with the testing sequences was achieved when compared with full reference PSNR measurements. There were a total of 42 training and 35 test data sets for model development and validation.

Table 6.5 Model prediction performance for training sequences

Cor. Coefficient	BasketballDrive	Vidyo1	Johnny	BQterrace	Kimono1	ParkScene
R^2	0.96	0.93	0.93	0.97	0.97	0.93
RMSE	0.79	1.29	1.39	0.91	0.49	1.87

Table 6.6 Model prediction performance for testing sequences

Cor. Coefficient	Cactus	Traffic	PartyScene	BasketballPass	RaceHorses	ChinaSpeed
R^2	0.93	0.96	0.94	0.94	0.95	0.94
RMSE	1.58	0.23	1.1	1.34	0.97	0.84

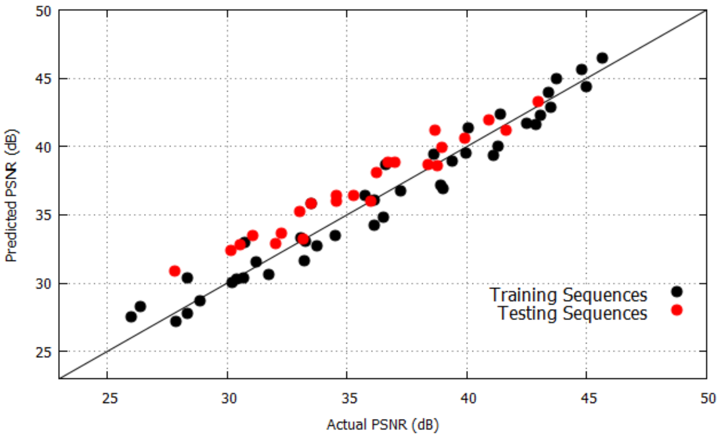


Figure 6.5 Actual vs. Predicted PSNR for training and testing sequences

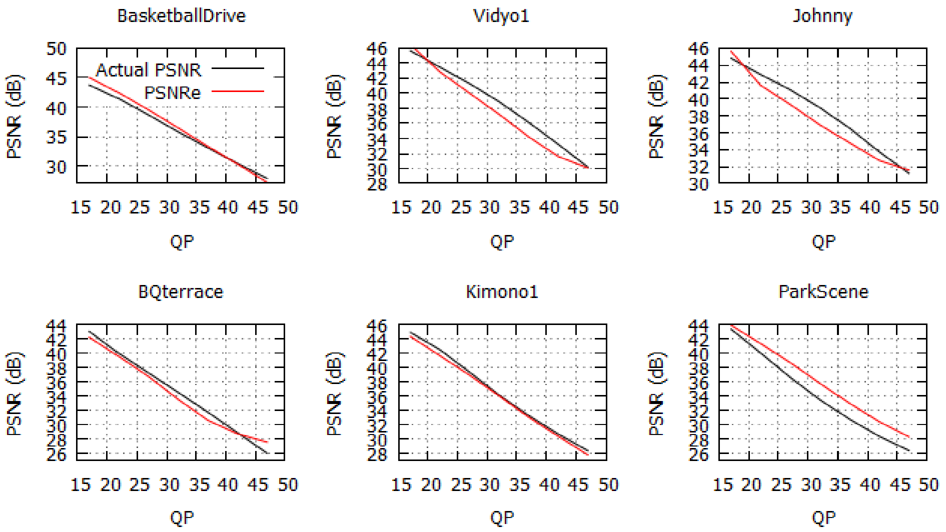


Figure 6.6 Impact of QP on Actual and Predicted PSNR for training sequences

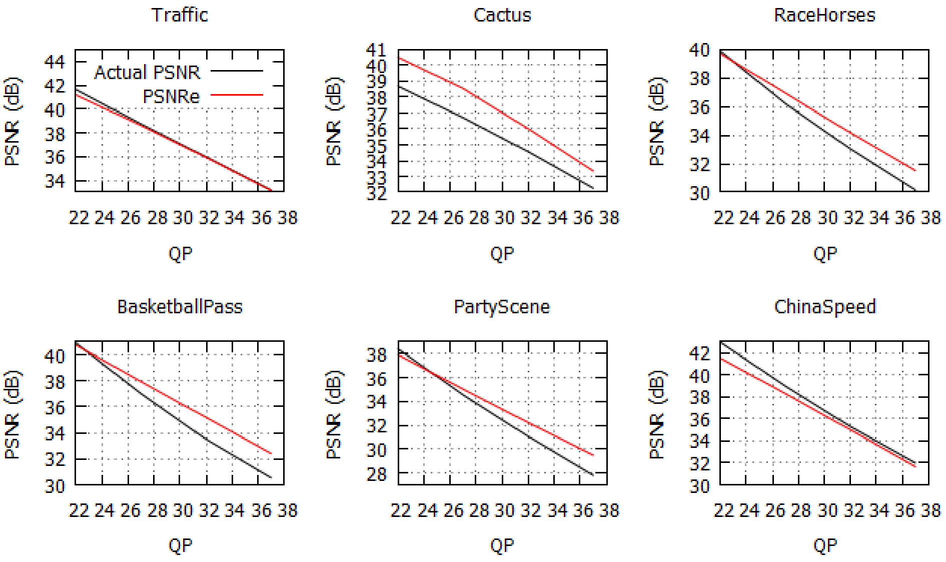


Figure 6.7 Impact of QP on Actual and Predicted PSNR for testing sequences

- **Performance evaluation with subjective data**

In this section, the performance of the proposed content-based initial video quality model is determined by comparing the degree of closeness between the predicted PSNR $(PSNR_e)$ with actual subjective ratings.

To carry out a subjective test that can be used to evaluate the performance of the proposed model, several video sequences of different spatiotemporal complexity were encoded. These are the same sequences that were used in model derivation. They were encoded using different

QP settings, the same temporal resolution and spatial resolutions. In essence, the same encoder settings used to encode sequences used in model derivation.

The subjective test plan used to collect MOS scores from end users followed the ITU recommendations for subjective video testing [69] [68] and was carried out using the same Double Stimulus Impairment Scale (DSIS) method used in chapter 3.

A total of 42 students from the School of Computing and Mathematics (SoCM) from Plymouth University took part in the test. This included 26 undergraduates and 16 graduate students and a mix of males and females with majority male. There were no monetary compensations for students who took the test. Although no attempt was made to measure the computer skills and familiarity with video testing of test takers, participants were all computer science students who had a high degree of computer literacy.

The results were scanned for unreliability by using the interquartile range (IQR) and the subjective screening algorithm (discussed in chapter 5). Based on IQR and the screening algorithm, 7 out the 42 evaluators were rejected. The scores from valid evaluators were averaged to compute the overall MOS for each test condition. Figure 6.8 shows the histogram of subjective quality ratings. The mean quality in the experiment was determined to be 3.6.

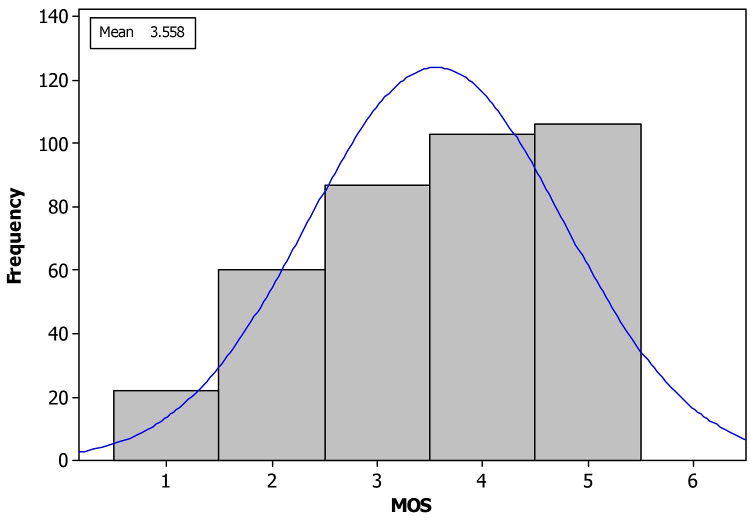


Figure 6.8 Histogram of subjective MOS (Mean 3.6)

- **Conversion of objective quality measurement metric to subjective metric**

Recently, novel metrics have been developed [54] [67] to convert objective video quality metric (e.g. PSNR, SSIM) to subjective metric such as MOS. This conversion, takes into account the non-linearity of human visual perception to quality [97] and in so doing adds human dimension to objective measurements. The conversion is typically used to determine the correlation between objective and subjective quality measurements. In this thesis, a PSNR to MOS conversion, is developed to determine the correlation between predicted PSNR and the actual MOS from DSIS subjective testing. To do this, a polynomial fitting between full reference PSNR and MOS values from subjective testing is performed. The metric for PSNR to MOS conversion is modelled as,

$$MOS_{PSNR} = d \times R^2 + u \times R + v \tag{6.3}$$

where $\mathrm{MOS_{PSNR}}$ denotes the MOS from PSNR to conversion, R represents the PSNR value, d = -0.0039, u = 0.4399 and v = -7.22 are modelling parameters that hold for all sequences.

- **Accuracy of the conversion metric**

To determine the accuracy of the PSNR to MOS mapping metric used, the full reference PSNR value of a sequence encoded at different QP was converted to MOS using three different metrics, i.e. the proposed metric, the popular Evalvid PSNR to MOS conversion metric [54] and the mapping metrics proposed in [67]. The derived MOS values were compared with those of actual MOS in terms of computing the R^2 and RMSE. Results of comparisons are shown in Table 6.7. The results show that, the proposed conversion metric outperformed the other metrics. Hence, in this project, the proposed PSNR to MOS conversion metric will be used for performance evaluation.

Table 6.7 PSNR to MOS metrics performance comparison

PSNR	Actual MOS	Proposed metric	Comparison with Evalvid metric [54]	Comparison with Dymarski metric [67]
43.12	4.5	4.50	5	4.2
40.07	4.4	4.14	5	3.9
37.39	4.2	3.78	5	3.6
34.70	3.3	3.35	5	3.4
31.97	2.8	2.86	4	3.1
29.10	1.9	2.28	3	2.9
26.27	1.7	1.64	3	2.6
R^2		0.96	0.90	0.94
RMSE		0.26	1.21	0.60

- **Comparison between MOS from PSNR and DSIS subjective MOS**

Recently, novel comparison models have been proposed to evaluate the performances of objective video prediction models. These approaches add human dimension to objective quality metric. For example, work presented in [98] performed a study of the performance of NTIA General Model [99] for HDTV video clips, the degree of accuracy was measured by comparing objective results with results of Single Stimulus Continuous Quality Evaluation (SSCQ) subjective quality ratings. Using a similar approach, the performance and accuracy of content-based Video Quality Prediction (CVQP) model was evaluated by comparing the predicted PSNR (converted to MOS) with MOS (obtained from DSIS subjective ratings) in terms of the correlation R^2 and RMSE. Results of performance evaluation are shown in Table 6.8. Additionally, a plot of actual vs. predicted MOS is presented in Figure 6.9.

Table 6.8 Comparison between predicted PSNR (MOS) and actual MOS

Cor. Coefficient	BasketballDrive	Vidyo1	Johnny	BQterrace	Kimono1	ParkScene
R^2	0.96	0.98	0.96	0.98	0.98	0.98
RMSE	0.35	0.26	0.42	0.26	0.23	0.34

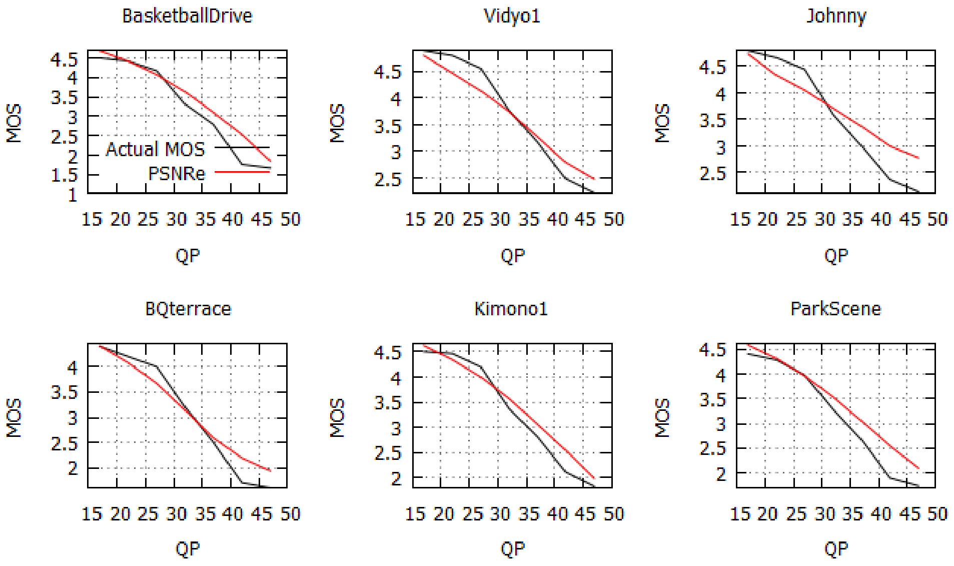


Figure 6.9 Performance comparisons between predicted PSNR and actual MOS

The comparison results indicate a good correlation between predicted MOS and actual MOS as perceived by end users. However, MOS from predicted PSNR show a higher degree of linearity than its subjective MOS counterpart. Results show that predicted PSNR underestimates the MOS score when QP is high. This highlights the limitations of PSNR-based metric which do not directly takes into account human perception of quality and indicates the importance of the work that will be presented in the upcoming sections which will focus on developing QoE model based on subjective results.

6.3 MOS based video quality prediction

In this section, we used subjective test results to develop video quality prediction models for HEVC encoded videos.

6.3.1 Experimental setup for MOS video quality prediction

Figure 6.10 and Figure 6.11 shows a simplified conceptual diagram for developing the non-intrusive prediction model for video quality over IP based networks. Figure 6.10 presents a block diagram of the video quality assessment model while Figure 6.11 shows the block diagram of the system that was designed and used to develop models that are able to predict the quality of videos delivered to end users. The prediction models developed in this section are based on MOS metric. The models are derived using both regression and exponential functions. Video quality is predicted from a combination of parameters associated network impairment (e.g. packet loss), codec related parameter (e.g. QP) and video content types. In real time, the parameters needed for quality prediction are extracted from received packets (e.g. RTP packets). It should be noted that, in this thesis, the parameters used to quantify video content type in terms of spatiotemporal characteristics included motion vectors, number of bits and QP of the coded I-frame. These parameters were extracted using a modified version of HEVC reference software.

The regression model presented in this section uses subjective MOS results obtained through subjective testing conducted in chapter 5. The MOS values are used for regression and exponential fitting. The approach of non-intrusive video quality prediction has the following benefits:

- It is based on end-to-end non-intrusive measurement of video quality. Thus, it can easily be applied to different IP based multimedia applications.
- Given the expensive and time consuming nature of subjective tests, the quality models proposed in this project, makes it possible for this cost to be avoided considering that no further subjective test need to be conducted to determine the quality of different video sequences.
- The models can be extended by adding new QoS parameters as desired.

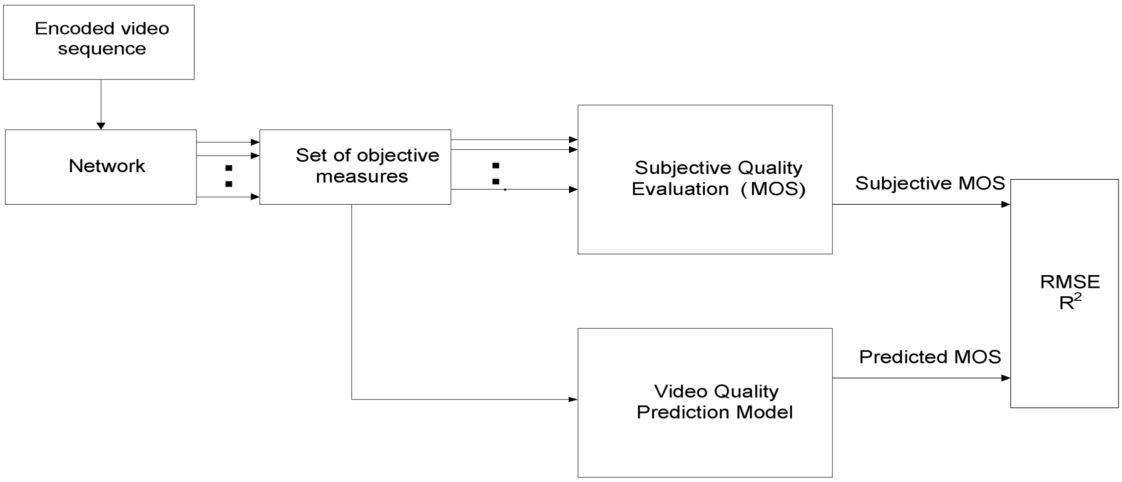

Encoded video sequence
Network
Set of objective measures
Subjective Quality Evaluation (MOS)
Subjective MOS
RMSE
R2
Predicted MOS
Video Quality Prediction Model

Figure 6.10 Block diagram of the proposed video quality prediction model

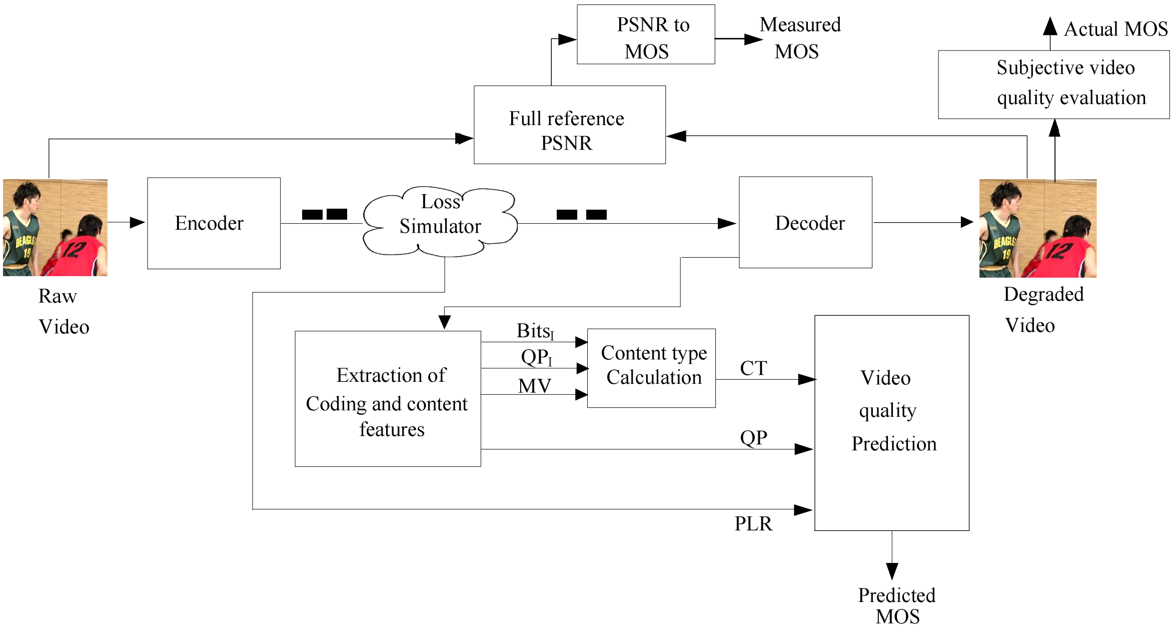


Figure 6.11 System block diagram of content based end-to-end video quality prediction

- **Analysis of valid crowdsourcing and lab based MOS**

In this section, the valid lab and crowdsourcing based results obtained from subjective testing (results are based on the subjective testing conducted in chapter 3 and 5) are analysed to establish the relationships between QP, PLR, CT and MOS. These relationships are needed to accurately model video quality.

- **Principal component analysis**

To investigate and understand how the variables for quality measurement (i.e. QP, CT, and PLR) contribute to the variability of the principal component (PC) i.e. MOS, the correlation matrix in PCA was used to determine the PC loadings. The results of PCA plot loadings for the QP, CT, PLR and MOS of the videos under test are shown in Figure 6.12 and Table 6.9.

From the results, it can be observed that both QP and PLR have positive loadings in PC1 while CT and MOS have negative loadings in PC1. This indicates that QP and PLR are negatively correlated to MOS (i.e. increasing QP and PLR leads to a decrease in MOS).

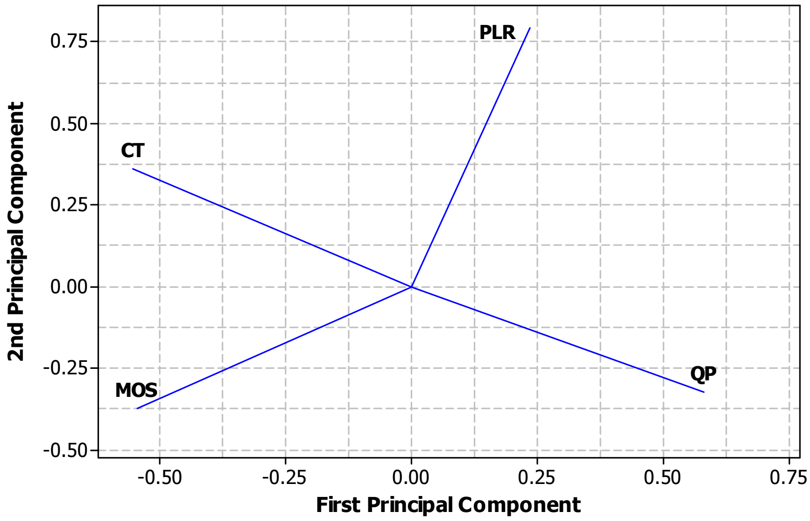


Figure 6.12 PCA of loadings for six video sequences

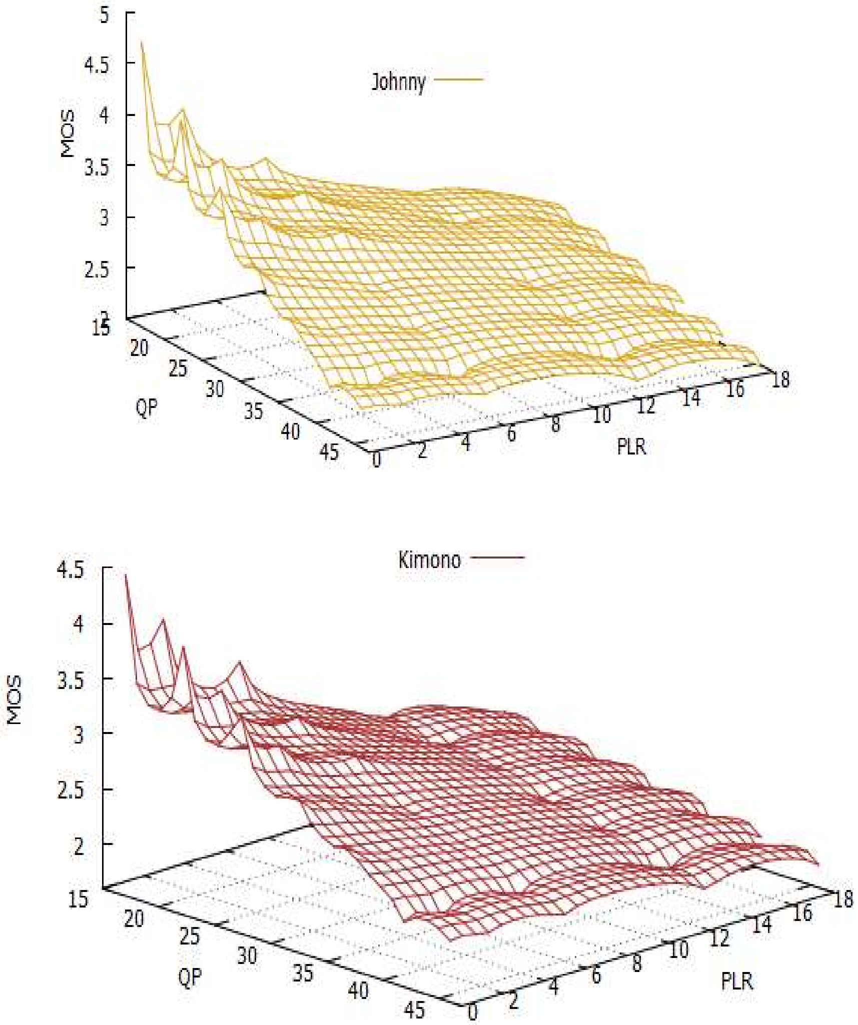
Table 6.9 PCA table of loadings

Variable	PC1	Squared loadings	PC2	Squared loadings
QP	0.58	0.34	-0.32	0.10
CT	-0.56	0.31	0.36	0.13
PLR	0.23	0.05	0.79	0.63
MOS	-0.55	0.30	-0.37	0.14
Sum		1		1

In terms of which of the variable has a stronger impact on quality, the results indicate that QP has the highest loading in PC1, this is followed by CT while PLR has the least value in PC1.

However, both results show that all the variables impact the principal component by not having a 0,0 cordination on the plot. PCA is important because it provides an understanding on how QP, CT and PLR impact MOS and how the variables should be used to appropriately model video quality quality.

To further understand the relationship between QP, PLR and MOS, 3-D graphs of QP vs. PLR vs. MOS are plotted as shown in Figure 6.13.



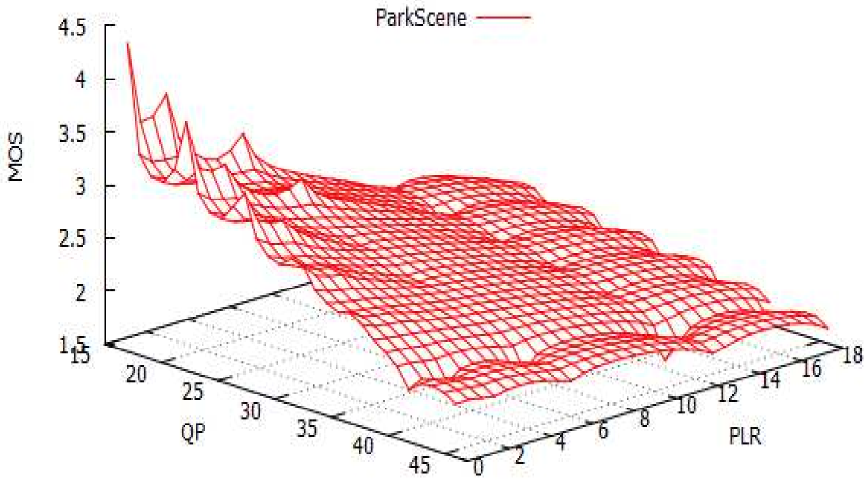


Figure 6.13 Relationship between QP, PLR and MOS

The results show that MOS values increases with decreased QP and PLR (consistent with PCA results). However, the increased or decreased in MOS as a result of variation in QP and PLR is content dependent as the three sequences show different quality (MOS) measurements under the same QP and PLR. For example, at QP 17 and PLR of 1.66 the MOS for Johnny is around 4.22, Kimono1 3.90 and ParkScene 3.70.

- **Relationship between packet loss rate and video quality**

In section 6.3 it was established that the initial video quality is determined by encoder settings such as the QP. However, as packet losses were introduced into the encoded bitstream, there was further degradation in video quality. To determine the relationship between packet loss and video quality, valid MOS values of sequences that were encoded with the same QP were extracted and plotted against the recorded PLR.

The relationship between PLR and video quality was established by using trend line fitting based on exponential, linear and logarithmic relationships. The result of each fitting in terms of R^2 is recorded. The results in Table 6.10 show that exponential relationship has the highest R^2 thus indicating that PLR and quality have an exponential relationship as shown in Figure

6.14. This relationship is important because it determines how PLR impact video quality and how it should be used in designing video quality model.

Table 6.10 Relationship between PLR and MOS

Function	R^2
Exponential	0.88
Linear	0.85
Logarithmic	0.83

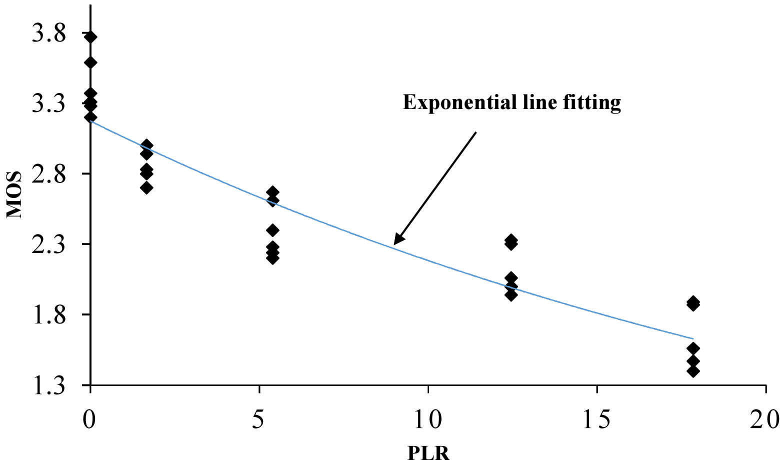


Figure 6.14 Relationship between PLR and MOS

To quantify the impact of packet loss on MOS scores, the ratio of MOS with packet loss impairment to the overall MOS (expressed as I_{PLR}) is calculated using Eq. 6.4. This is important because it establishes how MOS scores from subjective evaluators are impacted by packet loss impairment, and most essentially to further confirm the relation between packet loss rate and video quality.

$$I_{PLR} = \frac{MOS_{PLR}}{MOS_{zeroPLR}} \qquad (6.4)$$

where MOS_{PLR} denotes the MOS scores from video with PLR impairment and $MOS_{zeroPLR}$ represent MOS scores with no packet loss impairment.

The following can be deduced from Eq. 6.4:

1. Sequences that subjective test evaluators give high $\mathrm{MOS}_{\mathrm{PLR}}$ will result to high $\mathrm{I}_{\mathrm{PLR}}$ values (closer to 1) indicating MOS is insignificantly impacted by PLR.
2. Sequences that subjective test evaluators give lower $\mathrm{MOS}_{\mathrm{PLR}}$ will result to lower $\mathrm{I}_{\mathrm{PLR}}$ values (further away from 1) indicating that packet loss has a high impact on the overall MOS values.
3. An $\mathrm{I}_{\mathrm{PLR}}$ of 1 will indicates that $\mathrm{MOS}_{\mathrm{PLR}}$ is equal to $\mathrm{MOS}_{\mathrm{zeroPLR}}$ and packet loss has no impact on the overall MOS score.

Results of $\mathrm{I}_{\mathrm{PLR}}$ calculations for the sequences under test are shown in Figure 6.15. The results show that $\mathrm{I}_{\mathrm{PLR}}$ decreases (in terms of ratio) with increased packet loss and show approximately an exponential relationship for all video sequences.

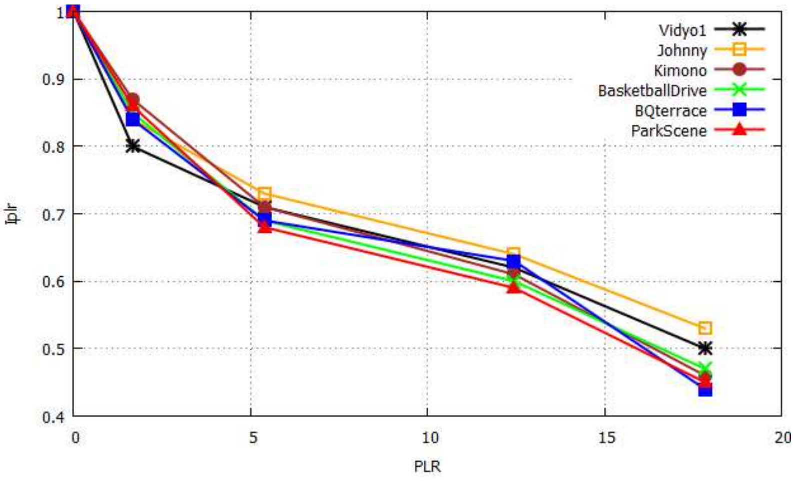


Figure 6.15 Relationship between PLR and $\mathbf{I}_{\mathbf{plr}}$

Based on the relationship that exists between packet loss and MOS, the impact of packet loss on quality $(\mathrm{I}_{\mathrm{PLR_Q}})$ can therefore be estimated non-intrusively using an exponential function, formalised as,

$$I_{PLR_Q} = \exp(\delta \times PLR) \qquad \mathrm{I}_{\mathrm{PLR_Q}} \in [0,1] \tag{6.5}$$

where δ (-0.045) is the modelling parameter obtained through exponential fitting.

6.3.2 MOS prediction

This section outlines the procedure for developing non-intrusive objective prediction model. To do this, the relationships established between QP, CT and video quality in Section 6.3 are used to develop a regression based model that is able to predict the initial video quality. Furthermore, both regression and exponential function based equations are combined to derive a single model that is able to predict the end-to-end video quality of different video sequences.

- **MOS based initial video quality prediction**

This section presents a MOS based reference free quality prediction for HEVC encoded videos that can be used by content providers to predict the initial video quality.

Results in Figure 6.13 show that the MOS values for each video sequence decreases as the QP increases. PCA on the other hand, shows that both QP and video content type significantly impact quality. Based on these results, the encoding quality $(\mathrm{MOS_e})$ of video sequences can be formalised as,

$$MOS_e = f(QP, CT) \tag{6.6}$$

where QP is the encoder setting QP and CT measured in terms of picture complexity and video motion activity.

Considering that quality has a linear relationship with QP and a logarithmic relationship with CT as shown in Figure 6.2 and Figure 6.3 respectively (Section 6.3). It is further determined that the initial encoding quality $(\mathrm{MOS_e})$ of different video sequences can be predicted as,

$$MOS_e = \alpha + \beta \times (QP) + \gamma \times \ln(CT) \tag{6.7}$$

where α = 6.02, β = -0.145 and γ = -1.43 are the modelling parameters obtained through regression fitting.

- **Performance evaluation of MOS based initial quality model**

This section outlines the steps used to evaluate the performance and the validity of the proposed initial video quality model.

- **Performance evaluation with unseen subjective data**

To evaluate the performance of the proposed initial encoded quality model, three video sequences not used in model derivation were used. These sequences include BasketballDrive, Vidyo1 and BQTerrace. The MOS of each sequence encoded at a given QP value is computed using the proposed model. The accuracy of the model is determined by comparing MOS predicted values with MOS values obtained from subjective testing in terms of correlation coefficient R^2 and Root Mean Square Error (RMSE) as summarized in Table 6.11 and Table 6.12.

Table 6.11 Model performance evaluation for training sequences

Cor. Coef.	Johnny	Kimono1	ParkScene
R^2	0.97	0.95	0.95
RMSE	0.39	0.41	0.46

Table 6.12 Model performance evaluation for testing sequences

Cor. Coef.	BasketballDrive	Vidyo1	BQterrace
R^2	0.94	0.94	0.95
RMSE	0.46	0.45	0.47

Furthermore, Figure 6.16 shows the scatter plot between the actual and model predicted MOS. A correlation coefficient of around 94% with the training sequences and 92% for testing sequences was achieved. There were 21 training and 21 testing datasets for developing and validating the model.

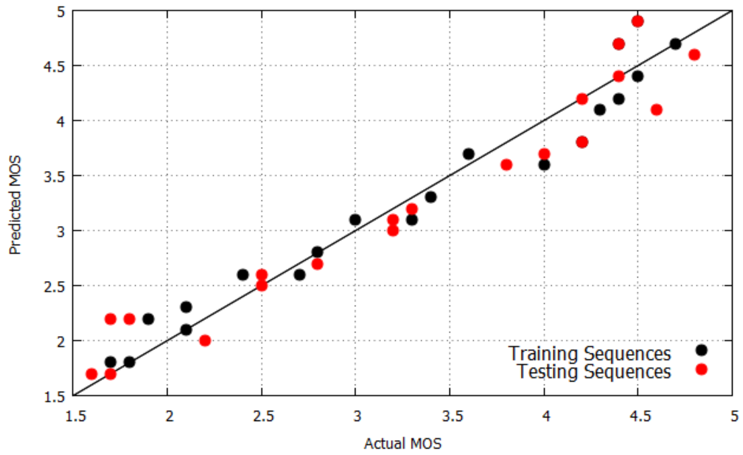


Figure 6.16 Scatterplot of Actual vs. Predicted MOS

- **Performance evaluation with objective data**

To further determine the performance of the proposed quality model, the same testing sequences used in section 6.3 were used. The quality of each encoded sequence was estimated by performing an intrusive measurement based on PSNR. The measured PSNR values are then converted to MOS using Eq 6.3. To determine the validity of the model, a further non-intrusive measurement is performed on each sequence encoded at a given QP value using the proposed model. The accuracy of computations is given in terms of correlation coefficient R^2 and Root Mean Square Error (RMSE) as summarized in Table 6.13. Furthermore, Figure 6.17 shows the scatter plot between the actual (PSNR to MOS conversion) and the model predicted MOS. A correlation coefficient of around 96% was achieved when MOS values from PSNR are compared with predicted MOS values.

Table 6.13 Model prediction performance for testing sequences

Cor. Coefficient	Cactus	Traffic	PartyScene	BasketballPass	RaceHorses	ChinaSpeed
R^2	0.96	0.96	0.93	0.97	0.97	0.98
RMSE	0.20	0.20	0.29	0.15	0.15	0.09

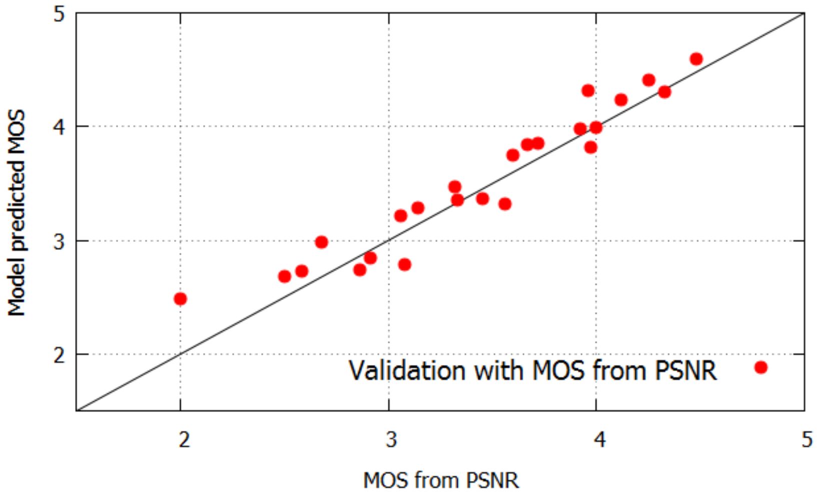


Figure 6.17 Scatter plot of MOS from PSNR vs. Model predicted MOS

Results show that the predicted MOS values are consistent with those of MOS from PSNR indicating that the proposed model is valid and can predict the quality of different video sequences with good accuracy.

- **MOS based end-to-end video quality prediction**

In this section, a non-intrusive end-to-end video quality prediction model based on the combination of video content type, encoder QP setting and packet loss rate is developed. Considering that video quality is impacted by the content type, the initial encoder settings and the impairment caused by the network, the end-to-end quality of video ($\mathrm{MOS}_{\mathrm{e2e}}$) of a sequence can be modelled as,

$$MOS_{e2e} = f(QP, PLR, CT) \tag{6.8}$$

PCA results show that MOS values decreases when both QP and PLR increases and vice versa. This therefore indicates that QP and PLR are directly proportional in terms of MOS estimation. Based on this proportional relationship, the $\mathrm{MOS}_{\mathrm{e2e}}$ of a video sequence can be predicted as,

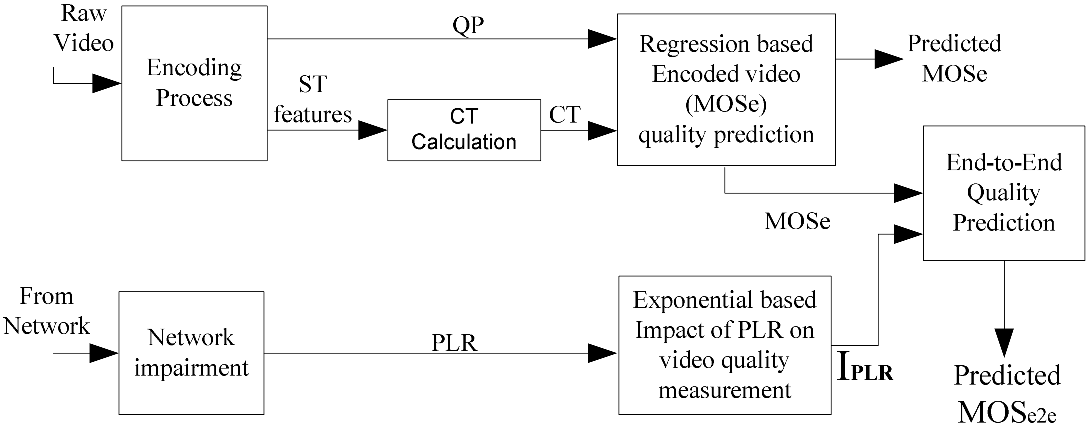
$$MOS_{e2e} = MOS_e \times I_{PLR_Q} \tag{6.9}$$

where MOS_e and I_{PLR_Q} are the encoding quality and network impact respectively.

The final $\mathrm{MOS}_{\mathrm{e2e}}$ equation will be as follows when the encoding quality prediction model is combined with non-intrusive measurement of the impact of packet loss on quality.

$$MOS_{e2e} = [\alpha + \beta \times (QP) + \gamma \times \ln(CT)] \times \exp(\delta \times PLR) \quad (6.10)$$

where model coefficients α, β, γ and δ are defined in Eq.6.5 and 6.7. The block diagram for end-to-end quality prediction is also presented in Figure 6.18.



Raw Video
Encoding Process
QP
ST features
CT Calculation
CT
Regression based Encoded video (MOSe) quality prediction
Predicted MOSe
End-to-End Quality Prediction
MOSe
From Network
Network impairment
PLR
Exponential based Impact of PLR on video quality measurement
IPLR
Predicted MOSe2e

Figure 6.18 Functional blocks of End-to-End video quality prediction model

6.4 Summary

In this chapter, the content type metric proposed in chapter 4 has been used to develop objective, reference free video quality models based on PSNR and MOS to predict the initial encoding quality of different video sequences. The MOS based objective quality metric was evaluated using subjective data and MOS derived from PSNR to MOS conversion. Additionally, an end-to-end video quality prediction model that estimates the quality of different video sequences delivered over IP based network was also developed. This new content-aware model takes into account three parameters (i.e. content type, QP and PLR) that impact the quality of different video sequences.

The work presented in this chapter is important because it provides the basis for which video content type can be used as a parameter to design video quality models which will enable effective monitoring and provisioning of videos with acceptable quality.

The upcoming chapter will focus on evaluating and determining the validity of the proposed end-to-end video quality model.

Chapter 7

7. Performance Evaluation of End-to-End Video Quality Model

7.1 Introduction

In chapter 6 regression based models were developed to predict the initial encoding and end-to-end video quality non-intrusively. This chapter presents the steps taken to evaluate and validate the proposed quality models.

To evaluate the proposed prediction model, subjective data not used in model derivation were used to compare model predicted values with those of subjective video evaluation. Furthermore, a standalone video quality evaluation application that enables the evaluation of the prediction model under different encoder parameter settings and network quality of service (NQoS) was also developed. This tool is the frontend implementation of the proposed prediction model.

The chapter is organized as follows. Section 7.2 presents the evaluation of the proposed end-to-end video quality model with subjective data. Section 7.3 presents a standalone video quality evaluation tool that enables the estimation of video quality over a time period. Section 7.4 summarizes the chapter.

7.2 Video quality prediction model evaluation

This section outlines the steps taken to evaluate the performance and the validity of the proposed model.

The data set used for evaluations are based on subjective testing (chapter 3 and 5), the encoding and NS-3 simulation platform proposed in chapter 3.

7.2.1 Performance evaluation with unseen subjective data

To evaluate the performance of the proposed end-to-end model, three testing video sequences not used in model derivation are used. These sequences include BasketballDrive, Vidyo1 and BQterrace. The MOS of each sequence encoded at a given QP and impaired with different levels of PLR was computed using the proposed model.

The accuracy of the model is given in terms of correlation coefficient R^2 and RMSE as summarized in Table 7.1 and Table 7.2. Figure 7.1 shows the scatter plot of actual MOS against model predicted MOS for training and testing sequences. A correlation of around 94% and 93% was achieved with the training and testing sequences respectively. There were 105 training and 105 testing datasets test conditions for model development and validation.

Table 7.1 Model performance evaluation with training sequences

Cor. Coef.	Johnny	Kimono1	ParkScene
R^2	0.93	0.95	0.95
RMSE	0.29	0.26	0.29

Table 7.2 Model performance evaluation for testing sequences

Cor. Coef.	BasketballDrive	Vidyo1	BQterrace
R^2	0.94	0.94	0.92
RMSE	0.32	0.35	0.29

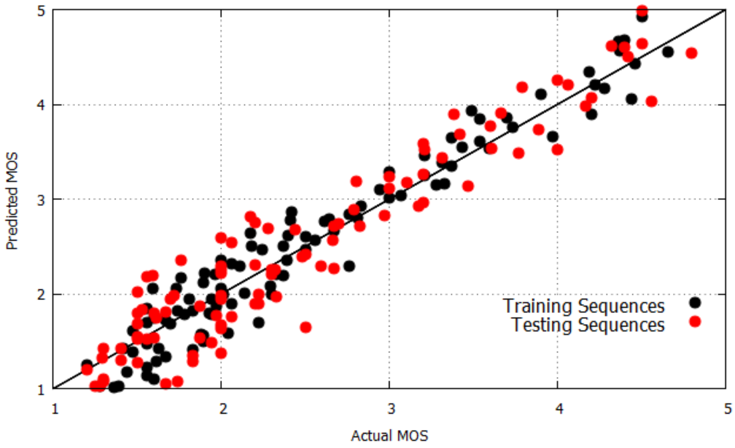


Figure 7.1 Scatter plot of Actual MOS vs. Predicted MOS

7.2.2 Performance comparison with content-blind model

To determine if by using content type (CT) as an additional variable improves the accuracy of video quality prediction, the CT-based prediction model is compared with content-blind (CT-blind) approach of quality prediction (i.e. encoder parameter settings and network impairments based method of quality prediction), to do this, a different model that did not take into account video content type was developed using the same dataset used in deriving the content-aware model.

Considering that QP and PLR have linear and exponential relationships with MOS respectively, we determined that the CT-blind end-to-end video quality (MOS_{e2e_b}) can be can be formalised as,

$$MOS_{e2e_b} = \gamma \times QP + \chi \times \exp(\rho \times PLR) \tag{7.1}$$

where γ= -1001, χ= 6.5614 and ρ= -0.046 are modelling parameters obtained by minimising the mean square error of all sequences. Comparison results are shown in Table 7.3 and Table 7.4.

Table 7.3 Comparing CT-based and CT-blind models using training sequences

	Johnny		Kimono1		ParkScene	
Cor. Coef.	CT-based	CT-blind	CT-based	CT-blind	CT-based	CT-blind
R2	0.93	0.84	0.95	0.84	0.95	0.84
RMSE	0.29	1.39	0.26	1.24	0.29	1.13

Table 7.4 Comparing CT-based and CT-blind models using testing sequences

	BasketballDrive		Vidyo1		BQTerrace	
Cor. Coef.	CT-based	CT-blind	CT-based	CT-blind	CT-based	CT-blind
R2	0.94	0.81	0.94	0.83	0.95	0.80
RMSE	0.32	1.16	0.35	1.47	0.29	1.11

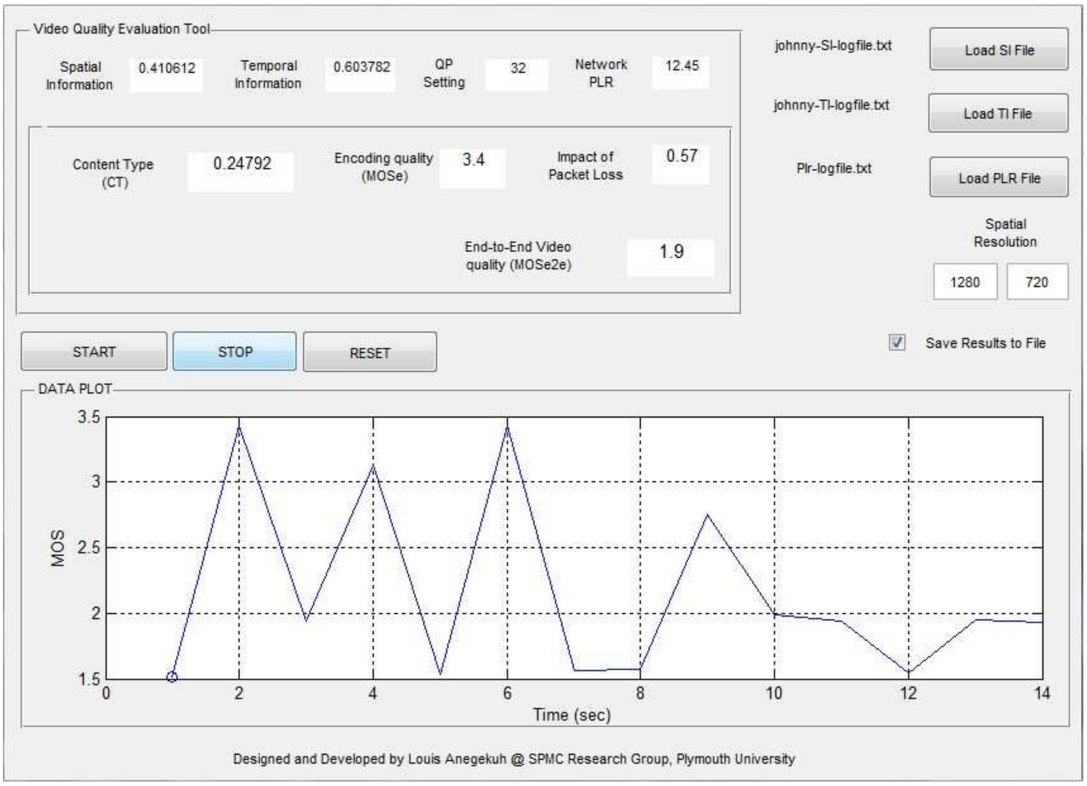
The results show that CT-based model provides improved prediction accuracy when compared with CT-blind method.

7.3 Model evaluation with video quality evaluation tool

In the absence of an open source real-time HEVC encoder that can be used to stream and evaluate the proposed model in real-time, in this section, a standalone video quality evaluation application that enables the evaluation of the proposed model is developed.

7.3.1 Video quality evaluation application

A snapshot of the proposed video quality evaluation tool is shown in Figure 7.2. This application computes the content type (based on spatial and temporal information), the initial encoding quality, the network quality of service (NQoS) and the end-to-end video quality. Additionally, the application also saves the results of computation to file for future analysis. The input data into the application is based on extracted features from the encoded bitstream.


Video Quality Evaluation Tool

Spatial Information: 0.410612

Temporal Information: 0.603782

QP Setting: 32

Network PLR: 12.45

Content Type (CT): 0.24792

Encoding quality (MOSe): 3.4

Impact of Packet Loss: 0.57

End-to-End Video quality (MOSe2e): 1.9

johnny-SI-logfile.txt

johnny-TI-logfile.txt

Plr-logfile.txt

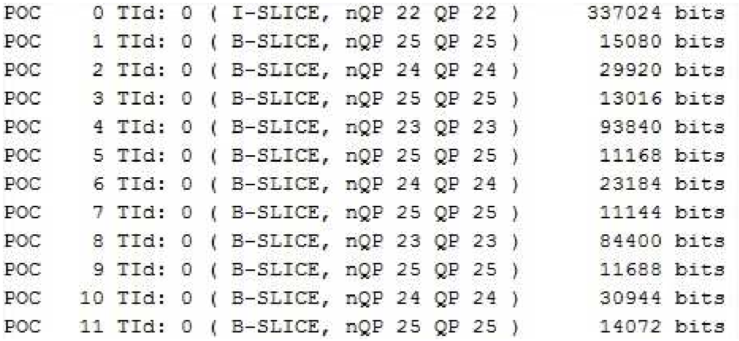
Spatial Resolution: 1280 720

☑ Save Results to File

Designed and Developed by Louis Anegekuh @ SPMC Research Group, Plymouth University

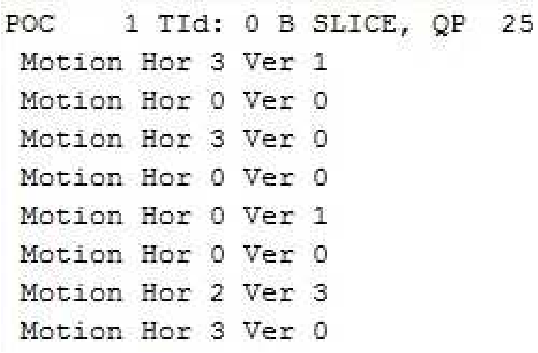
Figure 7.2 Snapshot of video quality evaluation tool

The input data needed for quality evaluation includes the number of bits and QP of coded I-frame, motion vectors, encoder QP settings and the network impairment (i.e. packet loss rate). To generate the input data, several video sequences of different spatiotemporal characteristics were encoded using varied encoder QP settings, spatial and temporal resolution. Based on the encoded videos, content and coding features were extracted and save to a log file. A sample SI, TI and PLR log files are shown in Figure 7.3, Figure 7.4 and Figure 7.5.



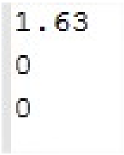
```
POC    0 TId: 0 ( I-SLICE, nQP 22 QP 22 )     337024 bits
POC    1 TId: 0 ( B-SLICE, nQP 25 QP 25 )      15080 bits
POC    2 TId: 0 ( B-SLICE, nQP 24 QP 24 )      29920 bits
POC    3 TId: 0 ( B-SLICE, nQP 25 QP 25 )      13016 bits
POC    4 TId: 0 ( B-SLICE, nQP 23 QP 23 )      93840 bits
POC    5 TId: 0 ( B-SLICE, nQP 25 QP 25 )      11168 bits
POC    6 TId: 0 ( B-SLICE, nQP 24 QP 24 )      23184 bits
POC    7 TId: 0 ( B-SLICE, nQP 25 QP 25 )      11144 bits
POC    8 TId: 0 ( B-SLICE, nQP 23 QP 23 )      84400 bits
POC    9 TId: 0 ( B-SLICE, nQP 25 QP 25 )      11688 bits
POC   10 TId: 0 ( B-SLICE, nQP 24 QP 24 )      30944 bits
POC   11 TId: 0 ( B-SLICE, nQP 25 QP 25 )      14072 bits
```

Figure 7.3 Snapshot of SI log file



```
POC    1 TId: 0 B SLICE, QP  25
 Motion Hor 3 Ver 1
 Motion Hor 0 Ver 0
 Motion Hor 3 Ver 0
 Motion Hor 0 Ver 0
 Motion Hor 0 Ver 1
 Motion Hor 0 Ver 0
 Motion Hor 2 Ver 3
 Motion Hor 3 Ver 0
```

Figure 7.4 Snapshot of TI log file



```
1.63
0
0
```

Figure 7.5 Snapshot of PLR log file

To generate the input data of packet loss, several video sequences of different spatiotemporal characteristics were encoded. They were encoded using exactly the same spatial and temporal resolutions and different QP settings. Each encoded bitstream was packetized and streamed from the UDP server to the UDP client over different bandwidth using the NS-3 streaming framework proposed in chapter 3 (i.e. Figure 3.3). The QP, bandwidth and the number of TCP connections were varied for each sequence because of the differences in spatiotemporal characteristic, i.e. sequences with high complexity and motion characteristics require more bandwidth and higher QP values than sequences with low complexity and motion for almost the same quality. The resulting packet loss rate and the computed MOS are shown in Table 7.5.

Table 7.5 NS-3 generated PLR and quality estimation

BasketballDrive				
QP	Bandwidth (Mbps)	Congestion	PLR	MOS
17	10	6	3.48	4.26
22	5	5	0.13	4.45
32	1	2	1.37	3.17
32	1	4	3.24	2.91

Vidyo1				
QP	Bandwidth (Mbps)	Congestion	PLR	MOS
17	5	6	0.29	5
22	1	2	1.96	4.2

Johnny				
QP	Bandwidth (Mbps)	Congestion	PLR	MOS
17	5	6	0.82	4.88
22	1	2	1.49	4.09

BQTerrace				
QP	Bandwidth (Mbps)	Congestion	PLR	MOS
22	5	4	1.98	3.75
22	5	6	2.88	3.60
27	1	2	13.95	1.90
27	1	4	15.5	1.77

Kimono1				
QP	Bandwidth (Mbps)	Congestion	PLR	MOS
32	1	4	1.1	3.16
32	1	6	1.38	3.12

ParkScene				
QP	Bandwidth (Mbps)	Congestion	PLR	MOS
17	10	4	5.98	3.7
17	10	6	6.16	3.67
22	5	4	1.11	4.13
22	5	6	2.53	3.87

7.3.2 Video quality evaluation over time

To demonstrate how the proposed quality evaluation tool can be used to estimate the quality of videos streamed over IP networks over time, the data set generated above is used to estimate the quality of each sequence after every 15 frames. This number of frames (window size) correspond to a 0.5 second time period for a video encoded at 30fps temporal resolution and 0.3 seconds for a video encoded at 50 frames per second. The results output file for each sequence is shown in Table 7.6.

Table 7.6 Video quality prediction using quality evaluation tool

BasketballDrive							
SI	TI	QP	PLR	CT	MOSe	Iplr	MOSe2e
0.42	0.64	32	1.37	0.27	3.26	0.94	3.07
0.42	0.61	32	1.37	0.26	3.31	0.94	3.12
0.42	0.61	32	3.24	0.26	3.32	0.86	2.87
0.42	0.59	32	1.37	0.25	3.36	0.94	3.16
0.42	0.58	32	1.37	0.25	3.38	0.94	3.18
0.42	0.60	32	1.37	0.25	3.34	0.94	3.14
0.42	0.62	32	3.24	0.26	3.30	0.86	2.86
0.42	0.60	32	3.24	0.25	3.34	0.86	2.88
0.42	0.60	32	1.37	0.25	3.35	0.94	3.15
0.42	0.60	32	1.37	0.25	3.35	0.94	3.15
0.42	0.59	32	3.24	0.25	3.38	0.86	2.92
0.42	0.59	32	1.37	0.25	3.38	0.94	3.17
0.42	0.61	32	1.37	0.26	3.33	0.94	3.13
0.42	0.60	32	3.24	0.25	3.34	0.86	2.89
0.42	0.60	32	3.24	0.25	3.34	0.86	2.88

Vidyo1							
SI	TI	QP	PLR	CT	MOSe	Iplr	MOSe2e
0.49	0.60	22	1.96	0.29	4.58	0.92	4.19
0.49	0.58	22	1.96	0.29	4.62	0.92	4.23
0.49	0.56	22	1.96	0.27	4.68	0.92	4.28
0.49	0.57	22	1.96	0.28	4.65	0.92	4.26
0.49	0.59	22	1.96	0.29	4.58	0.92	4.20
0.49	0.63	22	1.96	0.31	4.50	0.92	4.12
0.49	0.65	22	1.96	0.32	4.45	0.92	4.07
0.49	0.66	22	1.96	0.33	4.42	0.92	4.05
0.49	0.60	22	1.96	0.29	4.58	0.92	4.19
0.49	0.63	22	1.96	0.31	4.50	0.92	4.12
0.49	0.63	22	1.96	0.31	4.50	0.92	4.12
0.49	0.59	22	1.96	0.29	4.59	0.92	4.21
0.49	0.59	22	1.96	0.29	4.60	0.92	4.21
0.49	0.62	22	1.96	0.31	4.52	0.92	4.14
0.49	0.62	22	1.96	0.31	4.52	0.92	4.14

Johnny							
SI	TI	QP	PLR	CT	MOSe	Iplr	MOSe2e
0.48	0.67	22	1.49	0.32	4.44	0.94	4.15
0.48	0.68	22	1.49	0.33	4.41	0.94	4.12
0.48	0.67	22	1.49	0.33	4.43	0.94	4.14
0.48	0.73	22	1.49	0.35	4.32	0.94	4.04
0.48	0.69	22	1.49	0.33	4.40	0.94	4.12
0.48	0.72	22	1.49	0.35	4.33	0.94	4.05
0.48	0.68	22	1.49	0.33	4.42	0.94	4.13
0.48	0.67	22	1.49	0.32	4.44	0.94	4.15
0.48	0.65	22	1.49	0.32	4.48	0.94	4.19
0.48	0.68	22	1.49	0.33	4.42	0.94	4.13
0.48	0.70	22	1.49	0.34	4.37	0.94	4.09
0.48	0.67	22	1.49	0.33	4.43	0.94	4.14
0.48	0.70	22	1.49	0.34	4.37	0.94	4.09
0.48	0.71	22	1.49	0.35	4.35	0.94	4.06
0.48	0.73	22	1.49	0.35	4.32	0.94	4.04

BQTerrace							
SI	TI	QP	PLR	CT	MOSe	Iplr	MOSe2e
0.56	0.73	22	2.88	0.41	4.10	0.88	3.60
0.56	0.76	22	2.88	0.42	4.06	0.88	3.56
0.56	0.76	22	1.98	0.43	4.05	0.91	3.71
0.56	0.77	22	2.88	0.43	4.04	0.88	3.55
0.56	0.77	22	2.88	0.43	4.03	0.88	3.54
0.56	0.77	22	2.88	0.43	4.03	0.88	3.54
0.56	0.77	22	1.98	0.43	4.02	0.91	3.68
0.56	0.75	22	1.98	0.42	4.06	0.91	3.71
0.56	0.78	22	1.98	0.44	4.01	0.91	3.67
0.56	0.78	22	1.98	0.44	4.00	0.91	3.66
0.56	0.77	22	1.98	0.43	4.03	0.91	3.68
0.56	0.78	22	2.88	0.44	4.01	0.88	3.52
0.56	0.78	22	1.98	0.44	4.02	0.91	3.68
0.56	0.75	22	2.88	0.42	4.06	0.88	3.57
0.56	0.75	22	2.88	0.42	4.07	0.88	3.57

Kimono1							
SI	TI	QP	PLR	CT	MOSe	Iplr	MOSe2e
0.43	0.64	32	1.38	0.28	3.22	0.94	3.03
0.43	0.63	32	1.38	0.28	3.23	0.94	3.03
0.43	0.63	32	1.38	0.27	3.23	0.94	3.04
0.43	0.63	32	1.38	0.27	3.23	0.94	3.04
0.43	0.62	32	1.38	0.27	3.25	0.94	3.06
0.43	0.64	32	1.38	0.28	3.22	0.94	3.03
0.43	0.63	32	1.38	0.27	3.24	0.94	3.04
0.43	0.6	32	1.38	0.26	3.29	0.94	3.1
0.43	0.63	32	1.38	0.27	3.23	0.94	3.04
0.43	0.63	32	1.38	0.27	3.24	0.94	3.05
0.43	0.6	32	1.38	0.26	3.31	0.94	3.11
0.43	0.56	32	1.38	0.24	3.41	0.94	3.2
0.43	0.55	32	1.38	0.24	3.43	0.94	3.22
0.43	0.56	32	1.38	0.24	3.41	0.94	3.2
0.43	0.55	32	1.38	0.24	3.42	0.94	3.21

ParkScene							
SI	TI	QP	PLR	CT	MOSe	Iplr	MOSe2e
0.60	0.70	17	6.16	0.42	4.8	0.76	3.64
0.60	0.70	17	5.98	0.42	4.79	0.76	3.66
0.60	0.71	17	6.16	0.43	4.77	0.76	3.62
0.60	0.70	17	6.16	0.42	4.79	0.76	3.63
0.60	0.70	17	5.98	0.42	4.79	0.76	3.66
0.60	0.70	17	5.98	0.42	4.79	0.76	3.66
0.60	0.71	17	5.98	0.43	4.77	0.76	3.64
0.60	0.71	17	5.98	0.43	4.77	0.76	3.65
0.60	0.71	17	5.98	0.43	4.78	0.76	3.65
0.60	0.72	17	5.98	0.43	4.76	0.76	3.64
0.60	0.70	17	6.16	0.42	4.78	0.76	3.63
0.60	0.71	17	5.98	0.43	4.77	0.76	3.64
0.60	0.71	17	5.98	0.43	4.77	0.76	3.64
0.60	0.71	17	5.98	0.43	4.78	0.76	3.65
0.60	0.71	17	5.98	0.42	4.78	0.76	3.65

7.4 Summary

In this chapter, the proposed end-to-end quality model has been evaluated by using subjective data not used in model derivation. The evaluation shows that the end-to-end quality model is able to achieve around 94% and 93% accuracy for training and testing sequences respectively.

To determine if the proposed approach of quality estimation improves the accuracy of prediction, a content-blind model that is based on only encoder parameter settings (i.e. QP) and network impairment (i.e. PLR) was further developed. When prediction quality values from the two models are compared, results show that the content-based approach of quality prediction significantly improves the prediction accuracy of video quality. This is important because, the results underpin the significance of the work presented in this project.

Considering that no open source real-time HEVC encoder exists to enable the streaming and evaluation of the proposed end-to-end model in a real-time environment, a standalone video quality evaluation tool that can be used to estimate the quality of different video sequences is developed. Using the proposed tool, the quality of different videos streamed over IP based network (simulated using NS-3) was estimated. Additionally, the tool was also used to demonstrate how the content-based quality measurement approach can be implemented in a real-time environment where the quality of a sequence is estimated after a number of frames.

The work presented in this chapter is important because it confirms the validity of the proposed quality model and also provides the basis for which video content type can be quantified and used as a parameter to design video quality models which will enable effective monitoring and provisioning of videos with acceptable quality.

Chapter 8

8. Discussion, Future Work and Conclusions

8.1 Introduction

The increased computational power of multimedia devices, compression efficiency and advances made in communication networks has increased the prevalence and diversity of video applications. However, the success of current and future applications will depend on the perceivable Quality of Experience (QoE) of end users. In general, the delivery of video applications to end users' devices with acceptable QoE, i.e. the minimum QoE which a customer could accept the service, requires a joint consideration of two fundamental processes of compression and bandwidth constraints. Considering that, increasing compression to meet bandwidth constraints and transmitting over a bandwidth limited network may result in poor quality which in turn leads to reduced usage of the application/services and hence reduced revenues. It has therefore become an inevitable task for both service and network providers to measure, monitor or predict the QoE especially on end users' devices.

8.2 Contribution to Knowledge

The main contributions of this thesis are:

(1) **Objectively quantified the content type of videos using spatial and temporal features/parameters of video sequences**

The work has contributed a metric for quantifying and using video content type to develop models for video quality prediction. Considering that, video content type was found to be a significant parameter that determines how video sequences are impacted by both encoding and networking processes; a new metric was developed to objectively quantify the content type of different video sequences. This new metric is based on the extraction of spatial and temporal features from the encoded bitstream using modified decoders (i.e. decoder of HEVC standards). The developed metric for quantifying video content type was then used together with encoding and network parameters to develop non-intrusive video quality prediction models.

The contribution in this area has been made public to the research community through the following publication-

[22] [100]. The work is described in chapter 4.

(2) Developed an algorithm to screen unreliable laboratory and crowdsourcing subjective evaluators.

This work has contributed platforms and an algorithm for screening unreliable crowdsourcing and laboratory subjective evaluators.

Considering the costs and time demands posed by laboratory subjective tests, recently, researchers have proposed alternative cheaper and less time consuming approaches such as crowdsourcing subjective test method that can produce results that are similar to those of lab based testing. An example of such method includes crowdsourcing; however, results from the crowdsourcing subjective evaluation are inherently marred by poor quality because of the unsupervised nature of the test. To weed out the unreliable evaluators, a platform was developed to rapidly carry out subjective quality evaluation. This platform is able to extract information such as IP addresses, device and browser type, time spent on a website and scores to hidden reference sequences. Based on the extracted information, a screening algorithm that is able to identify unreliable testing takes is developed.

The valid subjective test results have been made publicly available at **http://www.tech.plymouth.ac.uk/spmc/staff/laanegekuh/** [27] to the research community as currently there is a shortage of video quality assessment database available that combines distortions caused by the encoder and IP network for different types of video content, especially for the newly released HEVC codec.

The contribution in this area has been made public to the research community through the following publication-

[24]. The work is described in chapter 5.

(3) New models to predict non-intrusively the initial video quality and the quality of videos delivered over IP based networks.

This work has contributed new models to non-intrusively predict the initial video quality (based on quantization parameter and video content type) and the quality of videos delivered to end users' devices over IP-based networks i.e. end-to-end video quality. The models are based on a combination of parameters associated with the encoder, the IP network and video content type. The initial video quality models are based on regression and used objective metric (PSNR) and MOS obtained through subjective test. The end-to-end video quality model on the other hand, is based on regression and exponential function, this model uses MOS obtained from laboratory and crowdsourcing subjective testing. The initial prediction model prediction model has an accuracy of 95% when the model predicted PSNR values from video sequences not used in model derivation are compared with those of full reference PSNR. The end-to-end quality prediction model has an accuracy of 93% when the model predicted MOS values from video sequences not used in model derivation are compared with those of subjective MOS.

The contribution in this area has been made public to the research community through the following publication-

[25]. The work is described in chapter 6.

8.3 Limitations of the current work and discussions

The work carried out in this thesis has a number of limitations that should be addressed in future studies.

(1) Subjective video test screening algorithm

The identification of unreliable subjective evaluators using the proposed screening algorithm assumes that the evaluators' devices support the coding standard and have no influence on subjective evaluations. The algorithm also assumes that, evaluators will have to download video sequences before watching and evaluating to avoid additional network impairments. Considering that, research has also shown that the perception of video quality by end users is influenced by their location and the devices in which the applications are being consumed with [101], by not taking into account these other variables (i.e. devices and location), the exclusion of evaluators based on only their five points MOS scores may be limited. The assumption that evaluators will have to download the application before watching/evaluating to cater for bandwidth constraints (which may result to stalling and buffering) only apply to desktop PC and laptops. This is limited because not all multimedia devices allow applications to be downloaded. Furthermore, the exclusion of evaluators based on the lowest score to the reference sequences may be limited because these scores may be due to surrounding disturbances (e.g. lights, low performance in wireless connectivity etc.). So excluding these evaluators may be biased or unjustified. Furthermore, excluding evaluator based on the time they spent on the webpage containing a video sequence may be limited because some evaluators may be quick at watching and grading the videos than others. Based on these limitations, the subjective results used for model derivation and performance evaluation may not fully reflect the user perception of quality.

(2) Simulation based performance evaluation

The impact of network impairments on HEVC video quality is assessed in simulated networks using mainly NS-3 simulator and AHG loss simulator. This approach has the benefits of being fast, repeatable, easy to configure and customize. In a simulated network, many parameters such as packet loss, bottleneck link and bandwidth are controllable. Simulation based tests are also much more economical than those based on emulation or physical implementation, which involves computing devices and communications interfaces. However, the reliability and consistency of the simulation based tests depend on the quality and accuracy of the simulation models used. In real networks the network conditions are unpredictable.

(3) Limited consideration of end-to-end impairments

In this thesis, the IP network impairment considered for video quality modelling is mainly based on packet loss. However, in reality IP network suffers from other impairments such as delay, jitter, bandwidth etc. These other impairments may have a significant impact on end-to-end video quality.

(4) Limited consideration of video content types

In this thesis, twelve different video sequences were considered (mainly JCT-VC recommended sequences). These sequences broadly covered slow moving (head and shoulder) to fast moving (sports type). However, cartoon clips and movies were not considered. The spatiotemporal features of cartoon and movie clips may have an impact on content type metric and consequently the accuracy of the end-to-end quality model proposed.

(5) Limited validation of the work

Although, the models developed in this work, have been evaluated with subjective data not used in model derivation, nevertheless, validations with external database and cross validations are still needed.

8.4 Suggestions for Future Work

General directions of future work

Considering the limitations of the work presented in previous sections, the main aspects of the research that can be improved and extended in future work includes:

1. **Enhance subjective testing platform and the algorithm for screening unreliable evaluators**

The proposed screening algorithm for weeding out unreliable evaluators depend on evaluator's own data (i.e. IP addresses, device and browser type, time spent on a website and scores to hidden reference sequences) extracted from the web-based subjective test platform. To determine if the location of an evaluator impact video quality, the testing platform can be incorporated with user's location services, for example, GPS. The user's location can then be used as an additional parameter/variable in the development of video quality models. Furthermore, information about devices used for evaluation can also be extracted and use to evaluate video quality.

To determine the validity of the algorithm, further experiments can be conducted with the introduction of deliberate uncontrolled "malicious" evaluators to check whether those malicious evaluators are indeed captured by the proposed algorithm.

2. **Using real network scenarios to evaluate models**

Although the quality models developed in this work give good accuracy and outperformed content-blind models, more work may be required to test the models rigorously. In addition, the simulation based impact assessment can be addressed in the future by emulation. Emulation brings in some aspects of reality while keeping a certain degree of repeatability, configurability and other advantages of simulation. Emulation can be considered as a compromise solution between simulation and physical implementation in a test bed. By using emulation experiments are performed in a semi realistic environment, i.e. using real operating systems that are operating on real devices and running real applications. The models developed in this project can be tested in an emulation environment to get more convincing results and to observe their performance when being implemented in real devices.

In order to study the behaviour of developed technologies in reality and to validate the simulation/emulation results, their physical implementations in real systems are eventually required as the research matures. The actual performance of the models as well as accuracy of the simulation/emulation experiments can be fully tested as real functioning software. Furthermore, the collected results can be used to prove their usefulness for commercial applications.

8.5 Conclusions

Inspired by advanced innovations in multimedia devices and the exponential growth of different video applications delivered over IP based networks, this project was initiated to investigate the impact of different QoS parameters on video quality and to develop video quality models that have less computational overheads and can accurately and efficiently estimate quality in a non-intrusive manner without the need of the original sequence and subjective evaluation which could be expensive and time consuming.

In this thesis, an investigated has been carried out to determine the impact of video content type, encoder Quantization Parameter (QP) settings and packet loss rate (PLR) on the quality of videos encoded with the newly released HEVC codec. Based on the results and analysis, it was determined that the content type of a video sequence has a significant impact on video quality and in doing so, a new metric called content type (CT) was proposed to quantify the content types of different video sequences. This new metric is based on temporal information (TI) and spatial information (SI) defined by the ratio of non-zero motion vectors to total number of motion vectors and the complexity of video sequences respectively. Based on the proposed metric, it was observed that videos with high CT have high number of bits of coded I-frame and bitrate than those with lower CT. It was also observed that although higher CT videos have high number of bits of coded I-frame, videos with lower CT values have higher PSNR values than those with higher CT under the same encoding and network impairment.

Based on the proposed content type metric, different video quality prediction models that are able to estimate the quality of different video sequences were developed. These new content-aware models, takes into account three parameters (i.e. CT, QP and PLR) that impact the quality of different video sequences.

To evaluate these new models, the model predicted values were compared with quality values obtained from laboratory and crowdsourcing subjective testing environments.

The work presented in this thesis is important because it provides the basis for which valid subjective data can be collected from internet for video quality evaluation. Additionally, the work also provides the foundation for which video content type can be quantified and used as a parameter to design video quality models which will enable effective monitoring and provisioning of videos with acceptable quality.

The outcomes of this work can be used as building blocks for future work in this area. However, the performances of the proposed models need to be further confirmed with physical systems, real test beds and large scale networks and for other applications before they are made available for commercial applications and before a realistic implementation can be made for QoS sensitive applications.

References

[1] K. Hiramatsu, S. Nakao, M. Hoshino, and D. Imamura, “Technology Evolutions in LTE / LTE-Advanced and Its Applications,” IEEE Int. Conf. Commun. Syst., pp. 161–165, 2010.

[2] H. Choi, J. Nam, D. Sim, and I. V Bajiü, “Scalable Video Coding Based on High Efficiency Video Coding (HEVC),” IEEE Pacific Rim Conf. Commun. Comput. Signal Process., pp. 346–351, 2011.

[3] A. Takahashi, D. Hands, and V. Barriac, “Standardization Activities in the ITU for a QoE Assessment of IPTV,” IEEE Commun. Mag., vol. 46, no. 2, pp. 78–84, 2008.

[4] C. Wu, K. Chen, Y. Chang, and C. Lei, “Crowdsourcing Multimedia QoE Evaluation : A Trusted Framework,” IEEE Trans. Multimedia, no. August, pp. 1121–1137, 2013.

[5] T. Hoßfeld, Seufert, and R. Schatz, “Quantification of YouTube QoE via Crowdsourcing,” IEEE Int. Work. Multimedia Qual. Exp., 2011.

[6] D. C. Brabham, “Crowdsourcing as a Model for Problem Solving: An Introduction and Cases,” Converg. Int. J. Res. into New Media Technol., vol. 14, no. 1, pp. 75–90, Feb. 2008.

[7] S. Winkler and P. Mohandas, “The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics,” IEEE Trans. Broadcast., vol. 54, no. 3, pp. 660–668, Sep. 2008.

[8] O. Verscheure, P. Frossard, and M. Hamdi, “MPEG-2 Video Services over Packet Networks : Joint Effect of Encoding Rate and Data Loss on User-Oriented QoS,” 8th Int. Work. Netw. Oper. Syst. Support Digit. Audio Video (NOSSDAV 98), pp. 257–264, 1998.

[9] A. Khan, L. Sun, and E. Ifeachor, “Content-Based Video Quality Prediction for MPEG4 Video Streaming over Wireless Networks,” J. Multimedia, vol. 4, pp. 228–239, 2009.

[10] B. Lee and M. Kim, “No-Reference PSNR Estimation for HEVC Encoded Video,” IEEE Trans. Broadcast., vol. 59, no. 1, pp. 20–27, Mar. 2013.

[11] A. Eden, “No-Reference Estimation of the Coding PSNR for H.264-Coded Sequences,” IEEE Trans. Consum. Electron., vol. 53, no. 2, pp. 667–674, 2007.

[12] M. Naccari, M. Tagliasacchi, and S. Tubaro, “No-Reference Video Quality Monitoring for H . 264/AVC Coded Video,” IEEE Trans. Multimedia, vol. 11, no. 5, pp. 932–946, 2009.

[13] P. Calyam, E. Ekici, M. Haffner, and N. Howes, “A ‘GAP-model’ based framework for online VVoIP QoE measurement,” J. Commun. Networks, vol. 9, no. 4, pp. 446–456, Dec. 2007.

[14] S. Mohamed and G. Rubino, “A Study of Real-Time Packet Video Quality using Random Neural Networks,” Circuits Syst. Video Technol. IEEE Trans., vol. 12, pp. 1071–1083, 2002.

[15] J. Apostolopoulos, R. A. Guérin, S. Tao, and R. Gu, “Real-Time Monitoring of Video Quality in IP Networks,” in Proc. of ACM NOSSDAV, 2005, vol. 16, no. 6.

[16] A. Khan, L. Sun, and E. Ifeachor, “Content Clustering Based Video Quality Prediction Model for MPEG4 Video Streaming over Wireless Networks,” in IEEE International Conference on Communications - ICC, 2009, pp. 1–5.

[17] M. Ries, O. Nemethova, and M. Rupp, “Motion Based Reference-Free Quality Estimation for H.264/AVC Video Streaming,” 2nd Int. Symp. Wirel. Pervasive Comput., 2007.

[18] G. Van Wallendael, N. Staelens, and L. Janowski, "No-reference bitstream-based impairment detection for high efficiency video coding," Fouth Int. Work. Qual. Multimedia Exp., pp. 7–12, 2012.

[19] G. J. Sullivan and J. Ohm, "Recent developments in standardization of high efficiency video coding (HEVC)," SPIE Opt. Eng. + Appl. Int. Soc. Opt. Photonics, vol. 7798, no. 7798, 2010.

[20] Frank Bossen, "Common test conditions and software reference configurations," JCT-VC Doc. JCTVC-G1200, 2010.

[21] Stephan Wenger, "'NAL Unit Loss Software,'" JCT-VC Doc. JCTVCH0072, 2012.

[22] L. Anegekuh, L. Sun, and E. Ifeachor, "Encoded Bitstream based Video Content Type Definition for HEVC Video Quality Prediction," in IEEE ICC conference, Sydney, Australia, 2014.

[23] "Plymouth University- SPMC Subjective Video Testing." [Online]. Available: http://www.tech.plymouth.ac.uk/spmc/staff/laanegekuh/videotesting1. [Accessed: 12-Mar-2014].

[24] L. Anegekuh, L. Sun, and E. Ifeachor, "A Screening Methodology for Crowdsourcing Video QoE Evaluation," in IEEE Globecom conference, Austin, TX USA, 2014.

[25] L. Anegekuh, L. Sun, and E. Ifeachor, "Encoding and video content based HEVC video quality prediction," Multimedia Tools Appl., Dec. 2013.

[26] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos.," IEEE Trans. image Process., vol. 19, no. 2, pp. 335–350.

[27] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, "Modeling Packet-Loss Visibility in MPEG-2 Video," IEEE Trans. Multimedia, vol. 8, no. 2, pp. 341–355, 2006.

[28] A. R. Reibman, V. A. Vaishampayan, and Y. Sermadevi, "Quality Monitoring of Video Over a Packet Network," IEEE Trans. Multimedia, vol. 6, no. 2, pp. 327–334, 2004.

[29] G. Valenzise, S. Magni, M. Tagliasacchi, and S. Tubaro, "Estimating channel-induced distortion in H.264/AVC video with- out bitstream information," in IEEE Quality of Multimedia Experience (QoMEX), 2010, pp. 100–105.

[30] J. Nightingale, Q. Wang, and C. Grecos, "HEVStream : A Framework for Streaming and Evaluation of High Efficiency Video Coding (HEVC) Content in Loss-prone Networks," IEEE Trans. Consum. Electron., vol. 58, no. 2, pp. 404–412, 2012.

[31] T. Liu, X. Feng, A. Reibman, and Y. Wang, "Saliency Inspired Modeling of Packet-loss Visibility in Decoded Videos," in Int. Workshop on Video Perceptual Quality Metric (VPQM), 2009, pp. 1–5.

[32] U. Engelke, M. Barkowsky, and P. Le Callet, "Modelling saliency awareness for objective video quality assessment," in Workshop on Quality of Multimedia Experience (QoMEx, 2010, pp. 212–217.

[33] G. Zhai, J. Cai, and W. Lin, "Cross-dimensional Perceptual Quality Assessment for Low Bitrate Videos," IEEE Trans. Multimedia, vol. 10, no. 7, pp. 1316–1324.

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

[35] M. Welling, "Support Vector Regression," Dep. Comput. Sci. Univ. Toronto, 2004.

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

[37] K. Seshadrinathan and A. C. Bovik, “Motion-based Perceptual Quality Assessment of Video,” IS&T/SPIE Electron. Imaging. Int. Soc. Opt. Photonics, Feb. 2009.

[38] N. Kanwisher and E. Wojciulik, “Visual attention: insights from brain imaging,” Nat. Rev. Neurosci., vol. 1, pp. 1–10, 2000.

[39] E. Huang, H. Zhang, and D. C. Parkes, “Toward Automatic Task Design : A Progress Report,” Proc. ACM SIGKDD Work. Hum. Comput., pp. 77–85, 2010.

[40] S. Wenger, “H.264/AVC over IP,” IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 645–656, Jul. 2003.

[41] T. Oelbaum and K. Diepold, “Building a Reduced Reference Video Quality Metric with Very Low Overhead using Multivariate Data Analysis,” in The 4th International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA’07), 2007, vol. 6, no. 5.

[42] A. Khan, L. Sun, E. Ifeachor, J. O. Fajardo, and F. Liberal, “Video Quality Prediction Model for H.264 Video over UMTS Networks and Their Application in Mobile Video Streaming,” IEEE Int. Conf. Commun. Cape Town, South Africa, pp. 1–5, May 2010.

[43] M. Ries, O. Nemethova, and M. Rupp, “Video Quality Estimation for Mobile H . 264 / AVC Video Streaming,” J. Commun., vol. 3, no. 1, pp. 41–50, 2008.

[44] T. Shanableh, “Prediction of Structural Similarity Index of Compressed Video at a Macroblock Level,” IEEE Signal Process. Lett., vol. 18, no. 5, pp. 335–338, May 2011.

[45] H. Koumaras, A. Kourtis, C. Lin, and C. Shieh, “A Theoretical Framework for End-to-End Video Quality Prediction of MPEG-based Sequences,” Third Int. Conf. Netw. Serv. ICNS., 2007.

[46] R. Feghali, F. Speranza, D. Wang, and A. Vincent, “Video Quality Metric for Bit Rate Control via Joint Adjustment of Quantization and Frame Rate,” IEEE Trans. Broadcast., vol. 53, no. 1, pp. 441–446, Mar. 2007.

[47] A. G. Davis, D. Bayart, and D. S. Hands, “Hybrid no-reference video quality prediction,” 2009 IEEE Int. Symp. Broadband Multimedia Syst. Broadcast., pp. 1–6, May 2009.

[48] A. Khan, L. Sun, and E. Ifeachor, “QoE Prediction Model and its Application in Video Quality Adaptation Over UMTS Networks,” IEEE Trans. Multimedia, vol. 14, no. 2, pp. 431–442, Apr. 2012.

[49] G. Zhai, J. Cai, and W. Lin, “Cross-Dimensional Perceptual Quality Assessment for Low Bit-Rate Videos,” EEE Trans. Multimedia, vol. 10, no. 7, pp. 1316–1324, 2008.

[50] “A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, LIVEWireless Video Quality Assessment Database.” [Online]. Available: http://live.ece.utexas.edu/research/quality/live_wireless_video.html. [Accessed: 14-Jul-2014].

[51] T. Liu, H. Yang, A. Stein, and Y. Wang, “Perceptual quality measurement of video frames affected by both packet losses and coding artifacts,” Proc. QoMEX 2009.

[52] M. Goudarzi, L. Sun, and E. Ifeachor, “Audiovisual Quality Estimation for Video Calls in Wireless Applications,” IEEE Globecom 2010 proceedings., pp. 1–4, 2010.

[53] C.-H. Ke, C.-K. Shieh, W.-S. Hwang, and A. Ziviani, “An Evaluation Framework for More Realistic Simulations of MPEG Video Transmission,” J. Inf. Sci. Eng., vol. 24, no. 2, pp. 425–440, 2008.

[54] J. Klaue, B. Rathke, and A. Wolisz, “EvalVid - A Framework for Video Transmission and Quality Evaluation,” in 13th international conference on Modelling Techniques and Tools Computer Performance Evaluation, 2003, pp. 255–272.

[55] D. Saladino, a. Paganelli, and M. Casoni, “A tool for multimedia quality assessment in NS3: QoE Monitor,” Simul. Model. Pract. Theory, vol. 32, pp. 30–41, Mar. 2013.

[56] "Network Simulator 3." [Online]. Available: http://www.nsnam.org/. [Accessed: 22-Jun-2014].

[57] "High efficiency video coding," Recomm. ITU-T H.265, 2013.

[58] G. Sullivan, J. Ohm, W. Han, T. Wiegand, A. High, E. Video, and C. Hevc, "Overview of the High Efficiency Video Coding," IEEE Trans. circuits Syst. video Technol., vol. 22, no. 12, pp. 1649–1668, 2012.

[59] ITU-T, "Audiovisual and multimedia systems, Infrastructure of audiovisual services — coding of moving video 'Video coding for low bit rate communication,'" vol. 263, 2005.

[60] I. 14496-2:2001, "Information technology – Coding of audio-visual objects – Part 2: Visual," vol. 2001, 2001.

[61] S. H. ITU-T H.264, "Audiovisual and multimedia systems, Infrastructure of audiovisual services—coding of moving video, "Advanced video coding for generic audiovisual services," International Telecommunication Union," 2005.

[62] I. Richardson, H.264 and MPEG-4 Video Compression Video Coding for Next-generation Multimedia. 2003.

[63] Y. Ou, Z. Ma, and Y. Wang, "A Novel Quality Metric for Compressed Video Considering both Frame Rate and Quantization Artifacts," Int. Work. Image Process. Qual. Metrics Consum., 2008.

[64] Y. Sanchez, S. Wenger, T. Schierl, M. Hannuksela, and Y.-K. Wang, "RTP Payload Format for High Efficiency Video Coding."

[65] J. Postel, "Internet Protocol" RFC 791, 1981." [Online]. Available: http://www.ietf.org/rfc/rfc791.txt. [Accessed: 13-May-2014].

[66] "RTP: A Transport Protocol for Real-Time Applications RFC 3550."

[67] P. Dymarski, S. Kula, and T. N. Huy, "QoS Conditions for VoIP and VoD," J. Telecommun. Inf. Technol., pp. 29–37, 2011.

[68] "Subjective video quality assessment methods for multimedia applications," ITU-T Recomm. P.910, 1999.

[69] "Methodology for the subjective assessment of the quality of television pictures," ITU-T Recomm. BT.500-13.

[70] D. S. Hands, "A Basic Multimedia Quality Model," IEEE Trans. Multimedia, vol. 6, no. 6, pp. 806–816, Dec. 2004.

[71] "VQEG Hybrid Perceptual/Bitstream (HBS) group." [Online]. Available: http://vqegstl.ugent.be/?q=node/16. [Accessed: 01-May-2014].

[72] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," Signal Process. Image Commun., vol. 19, no. 2, pp. 121–132, Feb. 2004.

[73] M. H. Pinson and S. Wolf, "A New Standardized Method for Objectively Measuring Video Quality," IEEE Trans. Broadcast., vol. 50, no. 3, pp. 312–322, 2004.

[74] Y. Wang, "Survey of Objective Video Quality Measurements," 2006.

[75] "P.NBAMS Terms of Reference," ITU-T Q14/12, 2011.

[76] "Parametric non-intrusive bitstream assessment of video media streaming quality – higher resolution application area," ITU-T P.1202.2, 2013.

[77] Cisco Inc, "Cisco Visual Networking Index : Global Mobile Data Traffic Forecast Update , 2012 – 2017," Cisco Tech. White Pap., 2013.

[78] ITU-T, "G.1050 Network model for evaluating multimedia transmission performance over Internet Protocol," Revis. E, 2011.

[79] TIA, "TIA-921 Network model for evaluating multimedia transmission performance over Internet Protocol," Revis. B, 2010.

[80] Y. Wang and M. M. Hannuksela, "Common conditions for SVC error resilience testing," ISO/IEC JTC1/SC29/WG11 ITU-T SG16 Q.6, pp. 7–10, 2005.

[81] J. Nightingale, Q. Wang, and C. Grecos, “Benchmarking Real-Time HEVC Streaming James,” SPIE Photonics Eur. Int. Soc. Opt. Photonics, no. 8437 84370D-6, p. 84370D–84370D–14, Jun. 2012.

[82] D. Ammar, T. Begin, and I. Guerin-Lassous, “A new tool for generating realistic Internet traffic in NS-3,” in 4th International ICST Conference on Simulation Tools and Techniques., 2011, pp. 81–83.

[83] D. Wischik and N. Mckeown, “Part I : Buffer Sizes for Core Routers,” ACM SIGCOMM Comput. Commun. Rev., vol. 35, no. 2, pp. 75–78, 2005.

[84] M. H. Pinson and S. Wolf, “Comparing subjective video quality testing methodologies,” SPIE Proc, vol. 5150, no. 3, pp. 573–582, Jun. 2003.

[85] “University of Plymouth - SPMC Subjective Video Test.” [Online]. Available: http://www.tech.plymouth.ac.uk/spmc/staff/laanegekuh/subjective1/. [Accessed: 11-Jul-2014].

[86] W. J. Krzanowski, “Principles of multivariate analysis,” Oxford Univ. Press, 2000.

[87] Suhr and D. D, “Principal component analysis vs. exploratory factor analysis,” in SUGI 30 Proceedings, 2005, pp. 203–230.

[88] J. Hu and H. Wildfeuer, “Use of content complexity factors in video over IP quality monitoring,” Int. Work. Qual. Multimedia Exp., pp. 216–221, Jul. 2009.

[89] L. Yufei, F. Xiubo, and W. Qin, “A High-Performance Low Cost SAD Architecture for Video Coding,” IEEE Trans. Consum. Electron., vol. 53, no. 2, pp. 535–541, 2007.

[90] H. Bhaskar, R. L. Kingsland, and S. Singh, “Multi-resolution based motion estimation for object tracking using genetic algorithm,” IET Int. Conf. Vis. Inf. Eng. (VIE 2006), pp. 583–588, 2006.

[91] J. Zhai, K. Yu, J. Li, and S. Li, “A Low Complexity Motion Compensated Frame Interpolation Method,” IEEE Int. Symp. Circuits Syst. - ISCAS, pp. 4927–4930, 2005.

[92] E. Rosdiana and M. Ghanbari, “Picture complexity based rate allocation algorithm for transcoded video over ABR networks,” Electron. Lett. 36.6, vol. 36, no. 6, pp. 521–522, 2000.

[93] F. Jurie and M. Dhome, “Real Time Robust Template Matching,” BMVC2002 - Br. Mach. Vis. Conf., pp. 123–132, 2002.

[94] P. Hanhart, M. Rerabek, F. De Simone, and T. Ebrahimi, “Subjective quality evaluation of the upcoming HEVC video compression standard,” SPIE Opt. Eng. Appl. Int. Soc. Opt. Photonics, 2012.

[95] George Waddell Snedecor and William Gemmell Cochran, Statistical Methods, 8th ed. Ames: Iowa state univ press, 1989.

[96] J.-O. Kim and F. J. Kohout, “Analysis of variance and covariance: subprograms ANOVA and ONEWAY,” Stat. Packag. Soc. Sci., vol. 2, pp. 398–433, 1975.

[97] E. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, and F. Moschetti, “Low bit rate quality assessment based on perceptual characteristics,” in IEEE International Conference on Image Processing- ICIP 2003, 2003, no. 1, pp. 3–5.

[98] S. Wolf and M. Pinson, “Application of the NTIA general video quality metric (VQM) to HDTV quality monitoring,” in Proceedings of The Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), 2007, pp. 4–8.

[99] M. H. Pinson and S. Wolf, “A New Standardized Method for Objectively Measuring Video Quality,” IEEE Trans. Broadcast., vol. 50, no. 3, pp. 312–322, Sep. 2004.

[100] L. Anegekuh, L. Sun, E. Jammeh, I. Mkwawa, and E. Ifeachor, “Content based Video Quality Prediction for HEVC Encoded Videos Streamed over Packet Networks - IN PREPARATION,” IEEE Trans. Multimedia, pp. 1–19.

[101] M. G. Manzato and R. Goularte, “Live Video Adaptation : A Context-Aware Approach,” in 11th Brazilian Symposium on Multimedia and the web. ACM, 2005, pp. 1 – 8.