BHAT, A. R., 2012. A new video quality metric for compressed video. Available from *OpenAIR@RGU*. [online]. Available from: http://openair.rgu.ac.uk

A New Video Quality Metric for Compressed Video

Abharana Ramdas Bhat

A thesis submitted as part of the requirements for the degree of Doctor of Philosophy awarded by the Robert Gordon University

February 2012

Abstract

Video compression enables multimedia applications such as mobile video messaging and streaming, video conferencing and more recently online social video interactions to be possible. Since most multimedia applications are meant for the human observer, measuring perceived video quality during the designing and testing of these applications is important. Performance of existing perceptual video quality measurement techniques is limited due to poor correlation with subjective quality and implementation complexity. Therefore, this thesis presents new techniques for measuring perceived quality of compressed multimedia video using computationally simple and efficient algorithms.

A new full reference perceptual video quality metric called the MOSp metric for measuring subjective quality of multimedia video sequences compressed using block-based video coding algorithms is developed. The metric predicts subjective quality of compressed video using the mean squared error between original and compressed sequences, and video content. Factors which influence the visibility of compression-induced distortion such as spatial texture masking, temporal masking and cognition, are considered for quantifying video content. The MOSp metric is simple to implement and can be integrated into block-based video coding algorithms for real time quality estimations. Performance results presented for a variety of multimedia content compressed to a large range of bitrates show that the metric has high correlation with subjective quality and performs better than popular video quality metrics.

As an application of the MOSp metric to perceptual video coding, a new MOSp-based mode selection algorithm for a H264/AVC video encoder is developed. Results show that, by integrating the MOSp metric into the mode selection process, it is possible to make coding decisions based on estimated visual quality rather than mathematical error measures and to achieve visual quality gain in content that is identified as visually important by the MOSp metric. The novel algorithms developed in this research work are particularly useful for integrating into block based video encoders such as the H264/AVC standard for making real time visual quality estimations and coding decisions based on estimated visual quality rather than the currently used mathematical error measures.

Acknowledgements

I would like to take this opportunity to thank everyone who has supported me in completing this research work.

My first acknowledgement is to Prof. Iain Richardson for giving me this unique opportunity to undertake this research work. I am very grateful for his constant guidance, support and encouragement throughout this project. I am particularly thankful for all the constructive discussions we have had and for taking time off his holiday in Canada to read my thesis chapters. Personally, I would like to thank him and Pat Ballantyne for cracking the whip and making me complete this thesis in time.

I would like to thank Dr. Yafan Zhao and Mr. Sampath Kannangara for being part of my supervisory team. I am particularly thankful to Dr. Yafan Zhao for her friendship, useful discussions and constant support throughout this project.

I would like to thank Prof. Maja Bystrom for her very constructive feedback and for critically reviewing my work on several occasions.

A lot of time has been spent in reading this thesis and giving me feedback. For this, I would like to thank Prof. Iain Richardson, Dr. Yafan Zhao, Dr. Laura Muir and James Philp.

I am very thankful to the countless number of volunteers, who were mostly staff and students at the Robert Gordon University, for their patience and kindness in taking time off their busy schedule to take part in the subjective tests.

Finally, I would like to take this unique opportunity to thank my family, who are always in my heart, for giving me endless encouragement and support. I would like to dedicate this work to them.

Contents

Part 1: Background

Part 2: Experimental Work

Part 3: Discussion And Conclusion

List of Figures

List of Tables

Abbreviations

AVC	Advanced Video Coding
CABAC	Context Adaptive Binary Arithmetic Coding
CAVLC	Context Adaptive Variable Length Coding
CI	Confidence Interval
CIF	Common Intermediate Format
CRT	Cathode Ray Tube
DCT	Discrete Cosine Transform
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DVD	Digital Versatile Disk
HDTV	High Definition television
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IIR	Infinite Impulse Response
IPTV	Internet Protocol television
ISDN	Integrated Services Digital Network
ISO/IEC	International Organisation of Standardization / International Electro-technical Commission
ITU-T	International Telecommunications Union, Telecommunications Standardization Sector
JVT	Joint Video Team
LCD	Liquid Crystal Display
LGN	Lateral Geniculate Nucleus
MAE	Mean Absolute Error
MOS	Mean Opinion Score
MOSp	Mean Opinion Score Predictor
MPEG	Motion Picture Experts Group
MSE	Mean Squared Error
NTIA	National Telecommunications and Information Administration
PC	Pair Comparison
PSNR	Peak Signal to Noise Ratio
PSTN	Public Switched Telephone Networ
QCIF	Quarter Common Intermediate Format

RDO	Rate Distortion Optimisation
RGB	Red, Green and Blue colour space
SAE	Sum of Absolute Error
SDTV	Standard Definition television
SSCQE	Single Stimulus Continuous Quality Evaluation
SSE	Sum of Squared Error
SSIS	Single Stimulus Impairment Scale
SVC	Scalable Video Coding
VCEG	Video Coding Experts Group
VGA	Visual Graphics Array
VLC	Variable Length Coding
VQEG	Video Quality Experts Group
VQM	Video Quality Metrics
YCbCr	Luminance and chrominance components

Part 1: Background

1 Introduction

1.1 The Research Problem

The rapid increase in computing power and communication speed, coupled with the availability of computer storage facilities, has led to a new age of multimedia applications. Multimedia has its presence in many applications such as online video databases, surveillance, mobile messaging, IPTV, video conferencing, interactive multimedia and more recently in multimedia based online social interaction. These new growing applications require storage of high-quality data, easy access to multimedia content, reliable transmission and delivery. Digital video compression has played a significant part in the realisation of these applications by bridging the gap between the demand for quality, performance and limitations of available storage and transmission capabilities.

The compression of digital video is accomplished by a video codec which consists of an encoder for compressing the original video signal into a suitable form for storage and transmission, and a decoder for reconstructing the compressed video signal for playback. In the past, video codecs were implemented on hardware platforms mostly due to the computational complexity of the process requiring a large amount of calculations. However, in recent years general purpose processors have significantly improved in performance, reliability and cost. Therefore, implementation of software only video codecs for real time applications such as video conferencing, video streaming and mobile video phones, has become feasible.

Advanced video compression algorithms such as the H.264/AVC video compression algorithm [1] can deliver significantly improved compression efficiency compared with previous video coding algorithms (up to 50% more) [2] by providing higher quality video over a wide range of bitrate channels. Due to its improved compression efficiency, error resilience features and increased flexibility in transmitting the coded data, H.264/AVC has enabled new multimedia video services such as mobile video messaging and multimedia streaming over wireless networks [3] which require compressed video to be transmitted across low bitrate channels. However, the quality of compressed video at such low bitrates is poor due to compression-induced distortions. Therefore, there is a need for video quality measurement techniques to be employed in the designing and testing of video compression algorithms.

Visual quality is a key factor in the performance of multimedia video applications because they are meant for the human observer. Although subjective measurement of mean opinion score (MOS) [4] is an accurate method of measuring visual quality, it is very expensive to perform and impractical in real time applications [5,6]. Therefore objective assessment methods have been developed to predict the subjective results based on video content and the characteristics of the human visual system. The video quality experts group (VQEG) have performed several evaluation tests to benchmark the performance of these quality metrics in context to multimedia sequences [7]. This has resulted in the standardisation of a few video quality metrics in the ITU-T Recommendation J.247 [8]. These metrics have varying degrees of success in predicting the subjective test scores, with reported correlations of 70% to 84% between each objective metric and the measured subjective quality scores indicating that better approaches are required to provide a more accurate prediction of subjective quality.

Although several objective measures have been developed in the literature, their application to real time video quality measurement of multimedia video sequences is limited due to implementation complexity and computational overload. Therefore there is a need for new video quality measurement methods which correlate well with subjective quality, are simple to implement and reasonably fast to run in real time multimedia video coding algorithms.

1.2 The Research Objective

The aim of this research work is to develop novel algorithms for effectively measuring perceived quality of multimedia video sequences compressed using block-based video coding algorithms. These algorithms should be computationally simple to implement and enable video coding algorithms to make accurate estimation of visual quality within reasonable computation time.

This research is particularly aimed at measuring visual quality of multimedia sequences with compression-induced artefacts because: (a) There is growing popularity for multimedia applications such as Internet and mobile video messaging and streaming, which require video compression for storage or transmission (b) Since these multimedia applications are meant for the human viewer, the visual quality of compressed video is an important factor when considering the performance of these applications.

The research aim is achieved through the following objectives which are structured into four stages. Each stage is briefly summarised as follows:

Stage 1:

1. Study existing subjective and objective video quality measurement techniques available in the literature to gain theoretical knowledge and identify the limitations of these techniques.
2. Evaluate video quality of compressed video sequences using these subjective and objective measurement techniques to investigate if there is a relationship between the two measurement techniques with a view to predicting subjective quality using objective measures.
3. Develop a new video quality measurement technique for predicting subjective quality of compressed video from objective measures.

Stage 2:

4. Develop techniques to automatically estimate the parameters of the new video quality metric from video content. Evaluate the performance of the developed metric to investigate if the predicted quality is in close agreement with subjective quality and whether the metric is computational simple and can be easily integrated into video coding algorithms for making real time quality estimations.

Stage 3:

5. Investigate other factors influencing visual quality of compressed video including cognition-based factors which attract viewer attention. Develop techniques to integrate these factors into the new video quality metric to further improve its prediction performance. Conduct experiments to investigate if the metric performance has improved.

Stage 4:

6. As an application of the developed video quality metric to perceptual video coding, develop a new mode selection algorithm for an H264/AVC encoder which will employ the metric in the mode selection process to improve the visual quality of compressed video sequences.

1.3 Novel contributions and Published material

This research aims to develop novel techniques for effectively measuring visual quality of multimedia video sequences compressed using block-based video coding algorithms. Key contributions of this research work to perceptual video quality measurement and perceptual video coding are listed below:

- Analysing the relationship between subjective quality (MOS) and objective quality (MSE) for a variety of multimedia content coded to a wide range of bitrates. Proposing the method of measuring MOS from MSE by exploiting the high correlation between the two measures.

- Developing a new full reference perceptual video quality metric called the MOSp metric for measuring subjective quality of multimedia video sequences compressed using block-based video coding algorithms. This development has led to two journal publications [9,10] and two conference papers [11,12].

- Investigating methods to quantify video content based on the visibility of compression-induced distortion using spatial texture and temporal change information.

- Developing algorithms to automatically derive the MOSp metric from MSE and video content. These algorithms have been presented in [9, 11, 12].

- Developing algorithms to extend the MOSp metric based on MSE and video content to incorporate cognition-based factors which attract viewer attention. This work has been published in [10]

- As an application of the MOSp metric to perceptual video coding, developing a new MOSp metric based mode selection algorithm for a H264/AVC encoder.

- Developing a new distortion measure based on the MOSp metric which can be used in other components of the video coding algorithm for making coding decisions.

- Developing an adaptive model for the Lagrange multiplier as a function of quantisation parameter (QP) and video content.

1.4 Organisation of this thesis

The thesis is organised as follows:

Chapters 2 and 3 are background chapters on the basic concepts of digital video coding and video quality measurement in context to multimedia applications. Chapter 4 explains the various experimental methods used in this research work. Chapters 5, 6 and 7 present algorithms for predicting visual quality of compressed video sequences using a new video quality metric called the MOSp metric. Application of the MOSp metric in perceptual video coding is investigated in chapter 8. Chapters 9 and 10 are the discussion and conclusion chapters. A detailed overview of each chapter is as follows:

Chapter 2 - Provides essential background knowledge on digital video representation and digital video compression with a particular focus on block-based video coding algorithms. Main functional blocks of a block-based video codec are briefly explained and the most widely used video coding standards including the H264/AVC coding standard are introduced.

Chapter 3 – Explores the basic concepts and approaches to measuring video quality. An overview of the various mechanisms involved in the processing of visual information and the limitations of human vision are presented. Approaches to objective video quality measurement are reviewed. Finally, the limitations of existing video quality measurement techniques and the need for a new video perceptual quality metric are discussed.

Chapter 4 – Is the experimental methodology chapter and it outlines the test material, equipment, experimental methods and data analysis techniques used in this research project.

Chapter 5 – Presents a video quality experiment conducted to investigate the relationship between subjective and objective video quality. Based on the experimental findings, a new full reference perceptual video quality metric called the MOSp metric is introduced.

Chapter 6 – Investigates techniques for calculating the parameters of the new MOSp metric from video content. Performance evaluations conducted to compare the MOSp metric with popular video quality metrics are presented and discussed.

Chapter 7 - Explores methods of extending the MOSp metric to incorporate cognition based factors which attract viewer attention with a view to further improve the metric performance. Experiments performed to investigate the metric performance are presented.

Chapter 8 – Investigates an application of the MOSp metric to perceptual video coding. A new mode selection algorithm for the H264/AVC encoder which uses the MOSp metric in the mode selection process to make coding decisions based on estimated visual quality is described. An experiment to evaluate the performance of the MOSp-based mode selection algorithm in comparison with the reference H264/AVC encoder is presented in this chapter.

Chapter 9 – Is the discussion chapter. A detailed summary of the main contributions of this research work is presented. The developed algorithms and experimental findings are critically analysed with emphasis to their benefits and limitations. The relevance of the main findings to addressing the research problem is also discussed in detail. Finally, possible directions for further developments and improvements in relation to the contributions of this research work are presented.

Chapter 10 – Is the conclusion chapter.

Appendix A – Contains list of publications related to this research work.

Appendix B – Contains training instructions given to viewers during subjective evaluations.

2 Digital Video Coding Fundamentals

2.1 Introduction

Digital representation of video signals requires large storage and transmission bandwidth. Multimedia video applications such as online broadcasting, mobile video streaming, internet video streaming and video on demand, require digital video in a form that is suitable for real-time transmission and storage. Therefore, video compression techniques are used to reduce the amount of data required to represent video in a digital form prior to transmission and storage. An overview of the basic concepts of digital video representation and video compression with a particular focus on block-based video coding algorithms is presented in sections 2.2 to 2.5.

The growing popularity in multimedia video applications has led academics and Industry to work together to standardise compression techniques in order to increase inter-operability between various applications and platforms. Several series of standards have been successfully developed by two organizations: International Organisation of Standardization, International Electro-technical Commission (ISO/IEC) and the International Telecommunications Union, Telecommunications Standardization Sector (ITU-T). These standards address a wide range of video applications in terms of bitrate, image quality, and complexity. A detailed summary of popular video coding standards is presented in section 2.6.

Data reductions caused during video compression to achieve bitrate savings have an effect on the video quality. Therefore video coding algorithms have to consider the trade-off between quality and rate when choosing optimum coding options. This is achieved by using rate-distortion optimisation techniques which are discussed in section 2.7. Finally, section 2.8 gives an overall summary of this chapter highlighting the advantages and limitations of existing video coding techniques in context to multimedia video applications.

2.2 Properties of digital video

Digital video is the visual representation of a real world scene in digital form, suitable for electronic storage and/or transmission. It is a 2-dimensional representation of the 3-dimensional real world scene. This section explains the main properties of digital video including resolution, frame rate, colour, video size, bitrate and frame representation.

2.2.1 Resolution

Video is captured using a camera and digitised into spatial and temporal samples. Spatial samples, often referred to as picture elements or pixels, are regularly spaced points on a 2-D rectangular grid to form a video frame as shown in Figure 2-1.

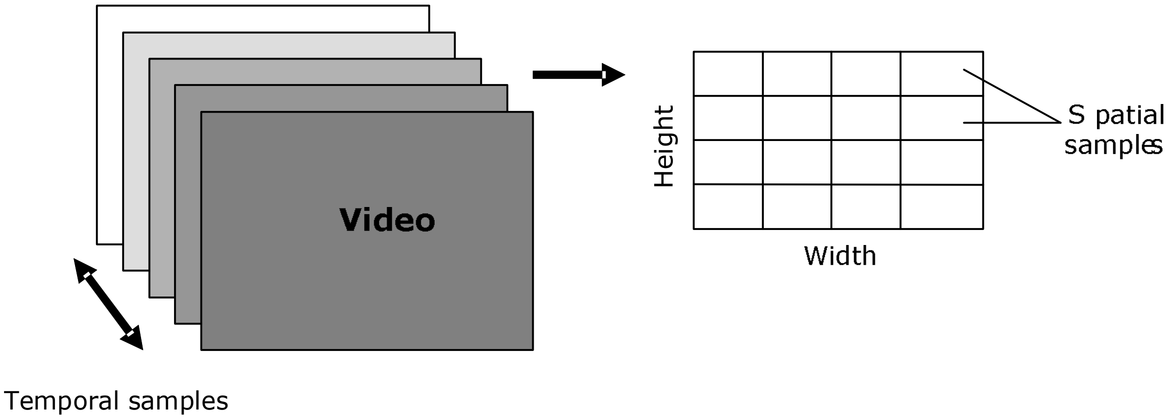


Figure 2-1: Video Sampling

The resolution of a video frame is the number of pixels in the frame. A larger number of pixels will produce a smooth and detailed visual representation of the scene. Resolution is expressed in terms of the number of pixels in the horizontal (width W) and vertical (height H) axes as W x H. Commonly used video resolution formats along with their applications are presented in Table 2-1.

Table 2-1: Common video resolutions and their applications

Format	Resolution (width x height)	Applications
Sub-quarter common intermediate format (SQCIF)	128x96	Mobile video
Quarter common intermediate format (QCIF)	176x144	Mobile video
Common intermediate format (CIF)	352x288	Multimedia applications and internet video streaming
4-common intermediate format (4CIF)	704x576	Standard definition television
High definition (HD)	1280x720 1920x1080	High definition television

2.2.2 Frame rate

Video is a sequence of video frames that are temporally sampled at a constant rate as shown in Figure 2-1. The number of temporal samples captured per second is the frame rate. A higher frame rate gives smoother representation of moving objects in the scene. The range of frame rates commonly used in video applications for a reasonably smooth display of video is between 20 – 30 frames per second (fps). Lower frame rates (below 10 fps) cause jerky appearance of motion in the video sequence [13].

2.2.3 Colour spaces

Pixels in a video frame contain colour information and are digitally represented using bits. For example: each pixel represented using 8-bits can have up to 256 (2^8) colour levels. More bits can represent more colour levels and hence more subtle variations in colour. There are two common colour spaces used for digital video representation: RGB (Red, green and blue) and YCbCr (Luminance, Red chrominance and Blue chrominance). In the RGB colour space, each pixel is represented by three numbers indicating the relative proportions of red, green and blue. Other colours of the visible light spectrum can be reproduced by combining varying proportions of the three primary colours. Typical RGB based video devices include television sets (LCD and plasma), mobile display screens, video projectors and digital video cameras. Although most video displays are driven by R, G and B signals, the RGB colour space is not the most efficient representation of video for storage and transmission because the R, G and B signals are correlated and cannot be separated into luminance and colour information. The human visual system is more sensitive to luminance information than colour information [4]. In colour images, detail perception is obtained from the luminance component of the pixels, because the human vision system is not well suited to detect structures defined by varying chrominance values. Hence by separating the luminance and colour data, video can be processed into perceptually relevant information.

The YCbCr colour space consists of the luminance component (Y) and two chrominance components (Cb and Cr). This colour space is popular in video processing algorithms such as video coding because the Y, Cb and Cr components are uncorrelated and therefore can be processed separately. The luminance component (Y) is a weighted average of red (R), green (G) and blue (B) and the chrominance components (Cb and Cr) are derived from Y, R and B [14] as shown below:

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ Cr &= R - Y \qquad (1) \\ Cb &= B - Y \end{aligned}$$

The separation of luminance and chrominance components means they can be stored or transmitted at different resolutions resulting in improved compression efficiency. Since the human visual system is more sensitive to luminance than colour, the luminance component can be stored or transmitted at higher resolutions and the chrominance components (Cb and Cr) can be sub-sampled to lower resolutions. There are three popular formats for sampling YCbCr components: 4:4:4, 4:2:2 and 4:2:0.

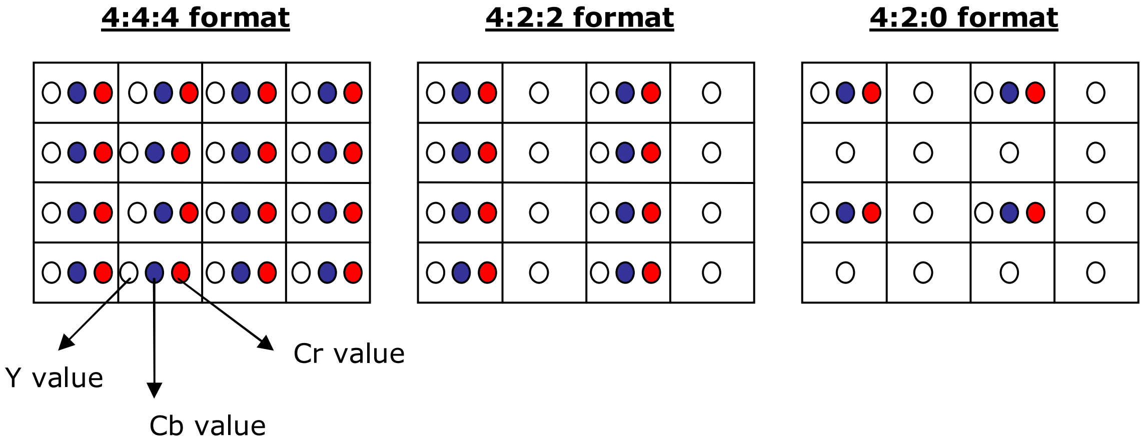


Figure 2-2: YCbCr sampling formats

In 4:4:4 format, the Y, Cb and Cr components are represented in the same resolution in both horizontal and vertical directions as shown in Figure 2-2. In 4:2:2

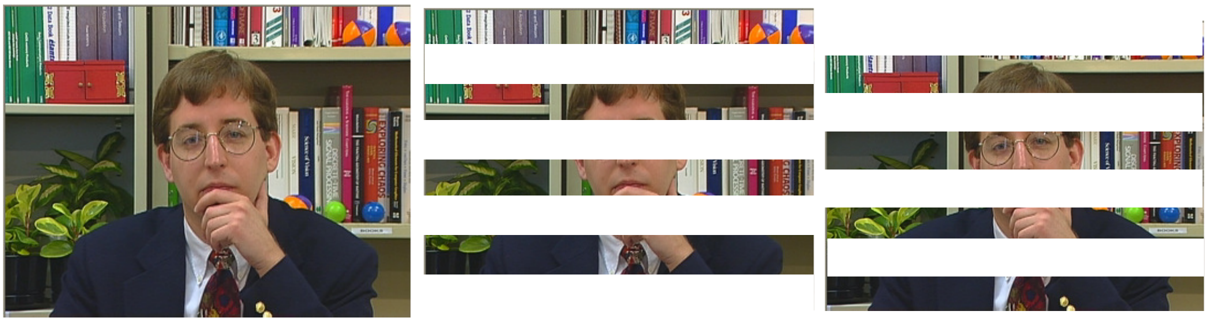
format, the Y component is represented in full resolution. However, the chrominance components (Cb and Cr) have the same vertical resolution but are sub-sampled to half the resolution as the Y component in the horizontal direction. 4:2:0 format means that the chrominance components are sub-sampled to half the resolution of the Y component in both horizontal and vertical directions. Since 4:2:0 video requires exactly half the number of samples as the 4:4:4 video, it is popular in video applications such as video conferencing and DVD storage.

2.2.4 Coded bitrate and video size

Coded bitrate, measured in bits per second (bps), is described as the average rate at which video data is transmitted in a given unit of time. When compressed video files are considered, bitrates may also be used to express the quality of video. Higher bitrate video has more bits to represent data and hence will have better quality compared to lower bitrate video. Coded video file size is the total number of bits used to store the video file and can be calculated as a product of coded bitrate and the duration of the video clip.

2.2.5 Progressive scan and interlaced video

There are two ways of rendering a video signal: interlaced scanning and progressive scanning. Interlaced scanning was developed for Cathode Ray tube (CRT) based television monitor displays and is used in most Standard Definition televisions (SDTV). Interlacing divides each video frame into odd and even lines stored and transmitted as two separate fields as shown in Figure 2-3. When displaying interlaced video, the display screen alternately refreshes the odd and even lines at 30 frames per second. This could sometimes lead to a “flickering” effect caused by delay in refresh rates between the two set of lines. Progressive scanning, as opposed to interlaced, scans the entire picture line by line from top to bottom and the video is transmitted as complete frames. This method is used in liquid crystal display screens (LCDS), plasma displays, DVDs and digital cameras. Since the frames are displayed at once, there is reduced flicker allowing for a greater range of motion for objects moving on screen. Video resolutions generally include “i” or “p”, such as 1080i and 720p, to denote either interlaced or progressive scanning.



Video frame **Top Field** **Bottom Field**

Figure 2-3: Progressive and interlaced scan.

2.3 Block-based video CODECs

Video is composed of a sequence of individual frames. In block-based video coding, video frames are broken down into individual blocks called macroblocks which contain 16x16 luminance samples and corresponding chrominance samples. For example, a picture from a video stream at CIF resolution (352x288) is divided into 396 (22x18) macroblocks. This practice simplifies the processing which needs to be done at each stage of compression. The macroblocks are individually compressed using a video codec. A video codec consists of an encoder for removing redundant information from the video signal and a decoder for re-inserting it. In video signals, two types of data redundancy can be identified:

- **Spatial and temporal redundancy:** Pixel values correlate with their neighbours both within the same frame and across frames. Therefore, pixel values may be predictable using the neighbouring pixel values.
- **Psychovisual redundancy:** The human eye has a limited response to fine spatial detail [15], and is less sensitive to detail near object edges or around scene changes. Consequently, data reduction in these regions may not be visible to a human observer.

The purpose and functioning of the encoder and decoder is discussed in the following two sections.

2.3.1 Video encoder

Figure 2.4 shows the block diagram of a block-based video encoder. A video encoder compresses video data by reducing spatial, temporal and psychovisual data redundancies. The main components of an encoder include predictive coding, transformation, quantisation and entropy encoding. Each video frame is individually encoded. The video frame is first divided into macroblocks. Predictive coding is performed on each macroblock to identify and eliminate spatial redundancies within a frame (*using intra prediction*) and temporal redundancies that may exist between individual video frames *(using inter prediction*). The prediction result is subtracted from the original data to form the residual. The resulting residual undergoes transformation from the spatial domain to the frequency domain in order to identify spatially correlated samples and reduce spatial redundancies.

The transformed coefficients are quantised to remove components that are unimportant to the visual presentation of the video frame leading to an irreversible data loss. The amount of compression can be controlled by varying the amount of quantisation. Entropy encoding is performed on the quantised transform coefficients to eliminate statistically redundant data. Entropy coded data also includes motion information. The encoded data forms the bit stream and is transmitted to the decoder through a transmission channel. Inverse quantisation and inverse transformation in the encoder perform the inverse operations of quantisation and transformation. The inverse operations are performed by the encoder to reconstruct the compressed video frames in order to facilitate motion estimation and compensation.

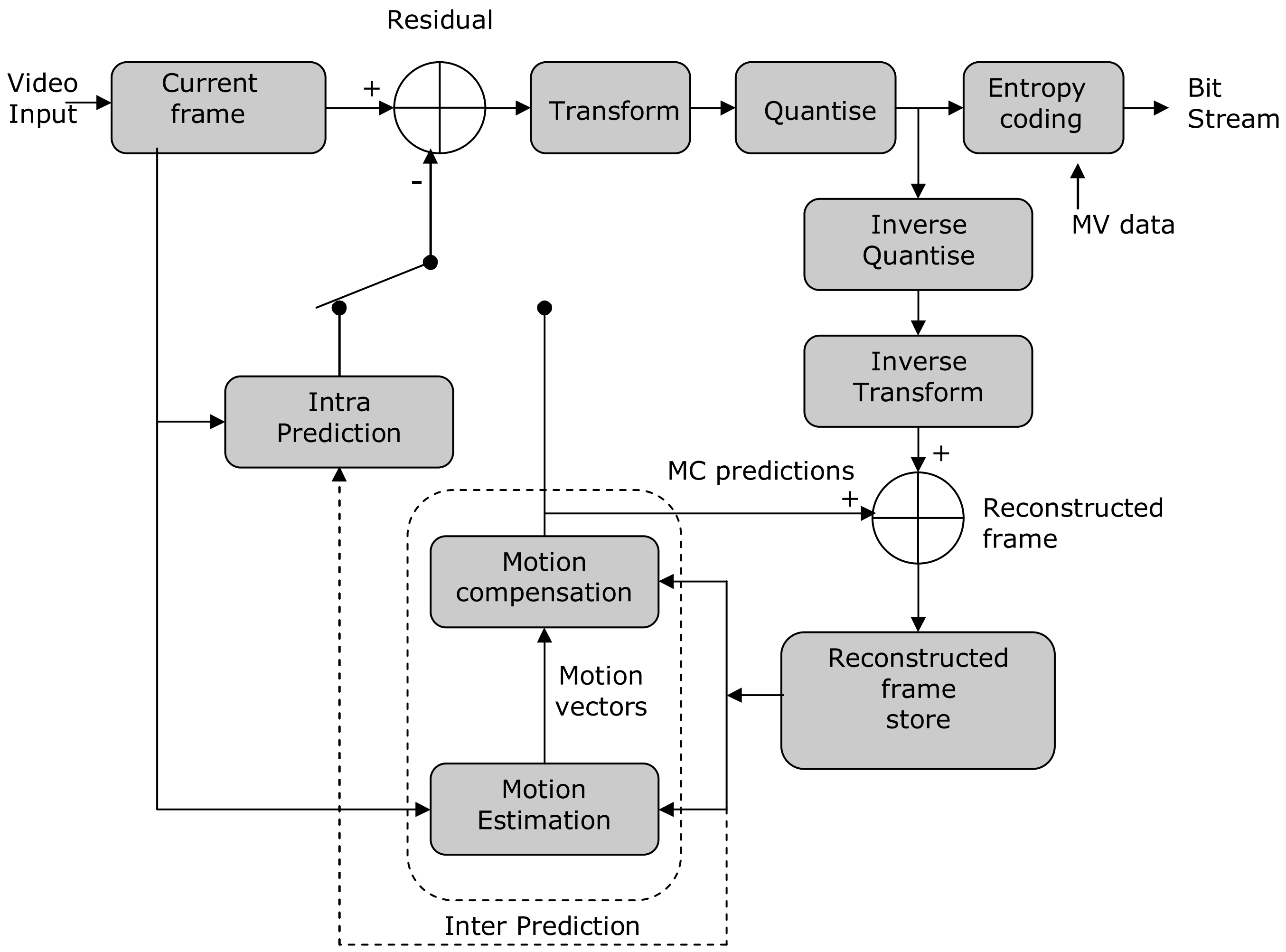


Figure 2-4: Block diagram of an encoder

2.3.2 Video decoder

Figure 2.5 shows the block diagram of a block-based video decoder which performs the inverse operation of the encoding process. The bit stream that is received at the decoder is entropy decoded, inverse quantised and inverse transformed to form the residual. The residual is added to the motion compensated prediction data of the previously decoded video frame to form the reconstructed frame. The reconstructed video frame at the decoder is not identical to the corresponding uncompressed frame at the encoder due to the permanent loss of data during the compression process (quantisation in particular).

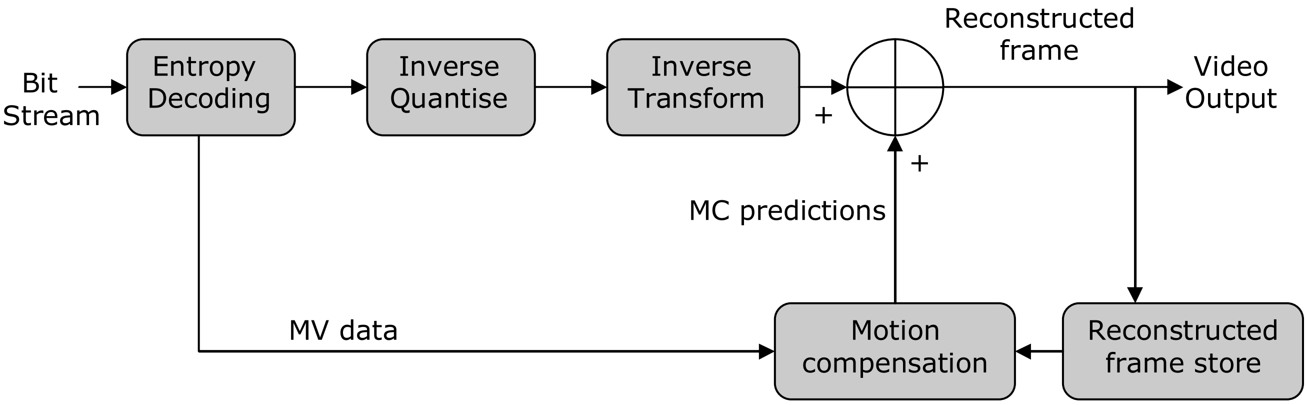


Figure 2-5: Block diagram of a decoder

2.4 Video structure

Figure 2-6 illustrates the video structure in block based coding algorithms. The reference video is coded as a stream of individual pictures. The basic coding unit of a video picture is a macroblock which contains 16x16 luminance samples and the corresponding chrominance samples depending on the YCbCr video format. Each picture consists of one or more slices. A slice is a group of macroblocks. Each slice is coded independently of the other slices in a picture in order to minimise the impact of data loss during transmission.

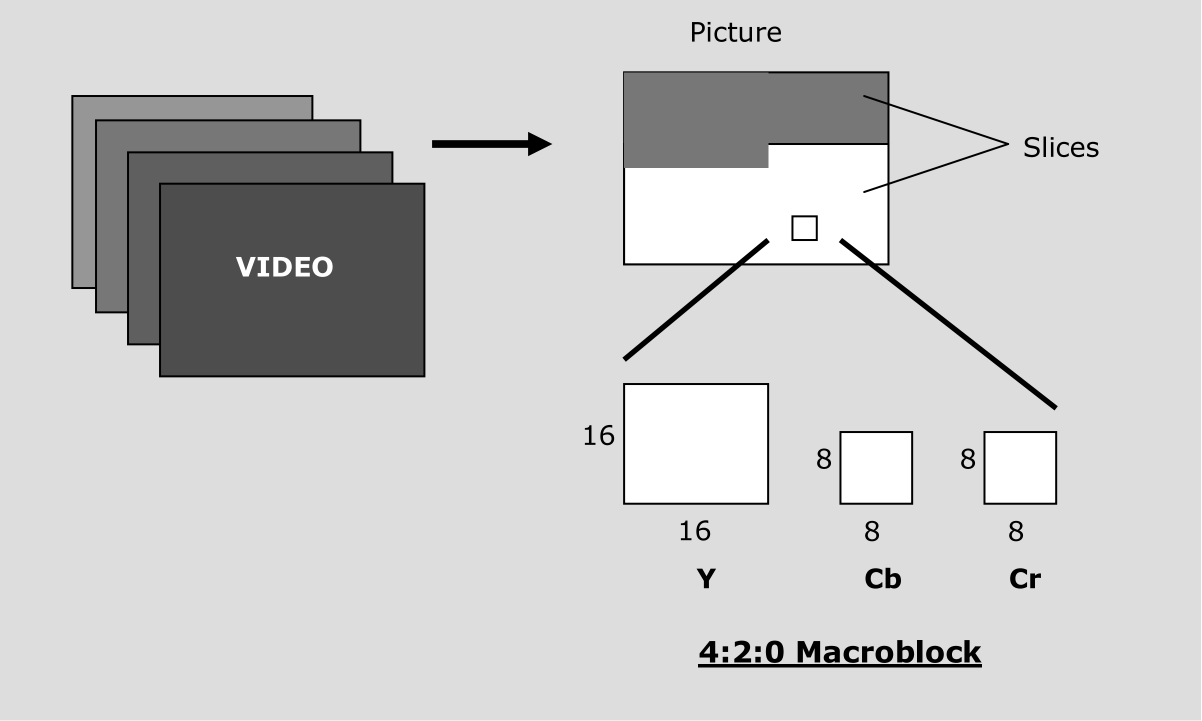


Figure 2-6: Video structure in block based video coding

2.5 Video coding techniques and tools

This section describes the various video coding techniques employed by video encoders including predictive coding for exploiting data redundancies, transform coding for converting data to a compactable form for further efficient compression, quantisation for performing lossy compression and entropy coding to remove statistical redundancy. Video coding tools including the de-blocking filter, which improve compression performance, are also explained in this section.

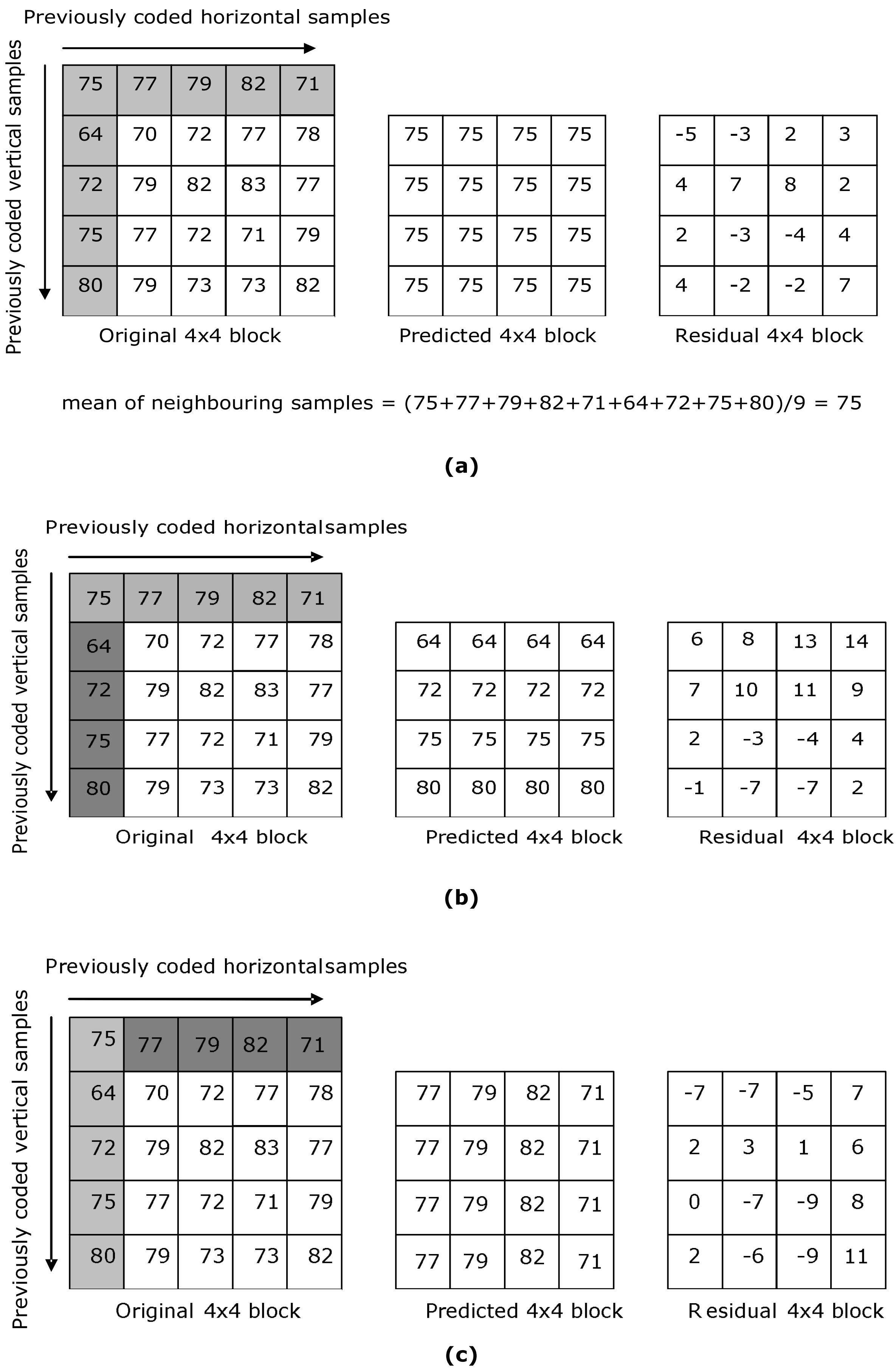
2.5.1 Predictive coding

Video frames may contain spatially and temporally redundant information. Coding efficiency may be improved by predicting current data from previously coded data and encoding the difference between the predicted and actual value. Predictive coding involves producing a prediction block by exploiting the spatial and temporal correlation between samples in the current block and previously coded samples either in the same video frame (intra prediction) or in previously coded video frame (inter prediction).

2.5.1.1 Intra Prediction

Intra prediction eliminates spatial redundancies. Intra prediction involves predicting the current block from previously coded neighbouring samples in adjacent blocks using a defined set of different directions. Luminance and chrominance samples are intra predicted separately. While the luminance macroblocks undergo partitioning into sub-blocks (i.e., 16x16, 8x8 or 4x4), the chrominance macroblocks are intra predicted without partitioning (i.e., 8x8 for 4:2:0 resolution format).

The purpose of using intra prediction is explained with the following example: Consider the intra prediction for a 4x4 block using three intra prediction techniques as shown in figure 2-7. Figure 2-7(a) illustrates intra prediction which uses the mean value of the neighbouring horizontal and vertical samples which have been previously coded. In figure 2-7(b), sample values of the 4x4 block are predicted using previously coded samples on the left of the block. In figure 2-7(c), the samples are predicted from previously coded upper neighbouring samples of the block. The difference between the predicted block and the actual block (i.e., the residual block) is then coded, which results in coding far fewer bits than would be the case for the original block.


Previously coded horizontal samples
Previously coded vertical samples

75	77	79	82	71
64	70	72	77	78
72	79	82	83	77
75	77	72	71	79
80	79	73	73	82

Original 4x4 block

75	75	75	75
75	75	75	75
75	75	75	75
75	75	75	75

Predicted 4x4 block

-5	-3	2	3
4	7	8	2
2	-3	-4	4
4	-2	-2	7

Residual 4x4 block

mean of neighbouring samples = (75+77+79+82+71+64+72+75+80)/9 = 75

(a)

Previously coded horizontalsamples
Previously coded vertical samples

75	77	79	82	71
64	70	72	77	78
72	79	82	83	77
75	77	72	71	79
80	79	73	73	82

Original 4x4 block

64	64	64	64
72	72	72	72
75	75	75	75
80	80	80	80

Predicted 4x4 block

6	8	13	14
7	10	11	9
2	-3	-4	4
-1	-7	-7	2

Residual 4x4 block

(b)

Previously coded horizontalsamples
Previously coded vertical samples

75	77	79	82	71
64	70	72	77	78
72	79	82	83	77
75	77	72	71	79
80	79	73	73	82

Original 4x4 block

77	79	82	71
77	79	82	71
77	79	82	71
77	79	82	71

Predicted 4x4 block

-7	-7	-5	7
2	3	1	6
0	-7	-9	8
2	-6	-9	11

R esidual 4x4 block

(c)

Figure 2-7: Example of intra prediction for a 4x4 block

Intra prediction which is available in H.263 and H.264/MPEG-4 AVC coding standards achieves better compression in smoother regions which prominently have spatial redundancies. Different directions for intra prediction exist to exploit inter-pixel redundancies within a video frame. Further details of the various directions for intra prediction are given in [16].

2.5.1.2 Inter Prediction

Consecutive video frames are typically very similar to each other and the differences usually arise due to moving objects in the video scene. By identifying and eliminating temporal redundancies, it may be possible to achieve higher compression. Inter prediction involves predicting the current block from a previously coded and reconstructed video frame using motion estimation and compensation processes. Motion estimation involves searching the previously coded video frame to obtain a good match for the current block. The result of motion estimation is a motion vector which represents the displacement between the locations of the current block and its best match in the previously reconstructed video frame. Motion compensation follows motion estimation and it involves finding the difference between the resulting best match and the current block to produce the residual block which is then coded. The main features of inter prediction are:

- Block matching to find the best match for the current block in a previously coded frame.
- Variable block size for improved motion estimation.
- Use of multiple reference frames to exploit re-occurring periodic motion.
- Motion estimation from past and future frames for improved compression efficiency.
- Sub-pixel motion estimation for increasing the precision of the motion vectors.

These features are described in detail in this section.

Block matching in motion estimation

Motion estimation is performed by searching an area in a previously coded video frame (the reference frame) to find a best match for the current block. A search area in the reference frame which is centred on the current block position is searched and the region within the search area that minimises a matching criterion is chosen as the best match for the current block as shown in figure 2-8. Commonly used matching criteria include the sum of absolute error (SAE) and mean absolute error (MAE) between the samples of the current block and the block-sized region in the search area of the reference frame as shown in equations (2) and (3).

$$SAE = \sum_{y=1}^{N} \sum_{x=1}^{N} \left| C(x, y) - R(x, y) \right| \tag{2}$$

$$MAE = \frac{1}{N^2} \sum_{y=1}^{N} \sum_{x=1}^{N} \left| C(x, y) - R(x, y) \right| \tag{3}$$

where, C(x,y) and R(x,y) are samples of the current and reference NxN blocks. These measures are popularly used for their computational simplicity.

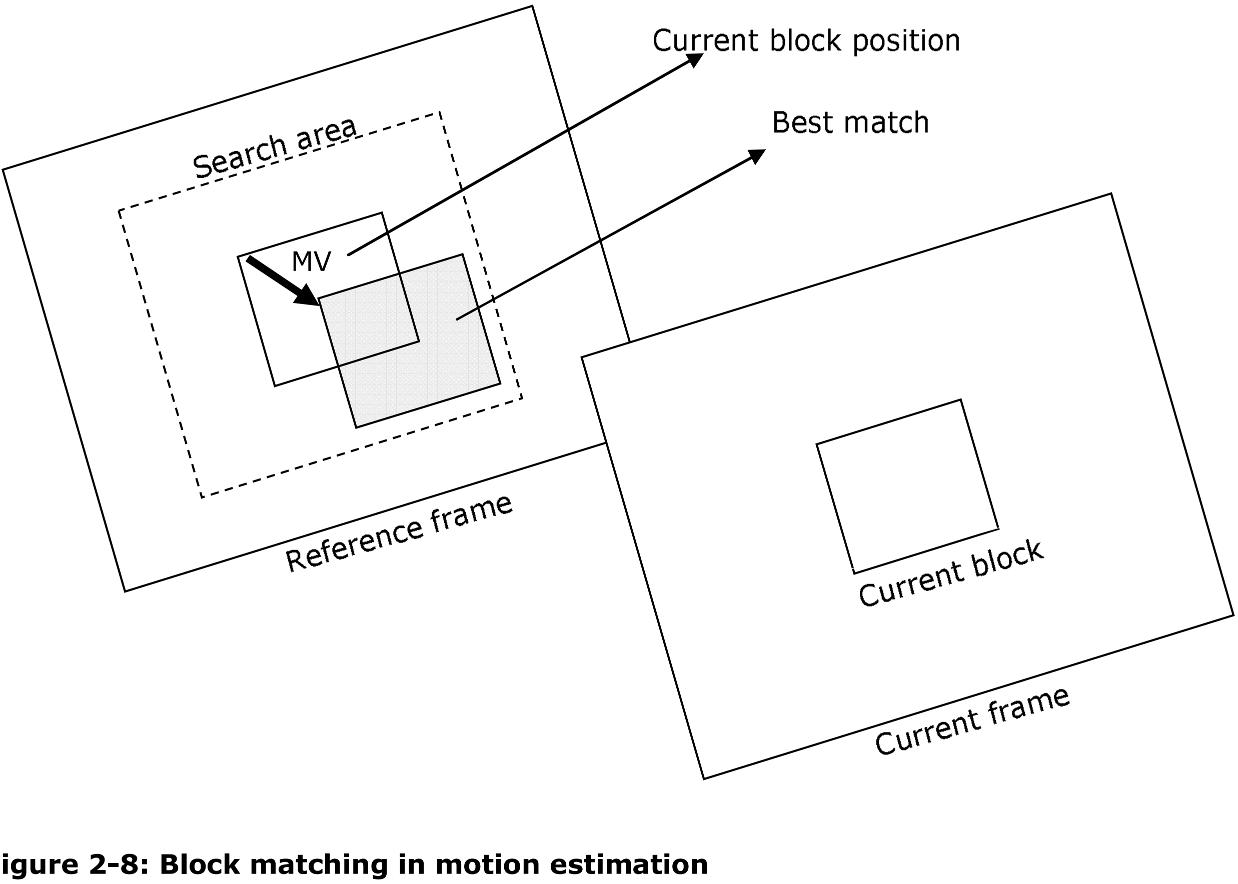


Figure 2-8: Block matching in motion estimation

Motion compensation follows motion estimation which involves obtaining the prediction block as the difference between the current block and its best matching block in the previously coded frame. The prediction block and the resulting motion vectors are coded and transmitted to the decoder.

Variable block size for improved motion estimation

A macroblock is the basic coding unit in block-based video coding algorithms. It contains 16x16 luminance samples and 8x8 chrominance samples (i.e., for 4:2:0 video format). To improve coding efficiency, video coding standards such as MPEG-4 and H.26x series support smaller block size motion estimation wherein the macroblock is broken down into smaller blocks, as shown in Figure 2-9, in an attempt to contain and isolate the motion. The resulting motion vectors from previous and/or future pictures are used to predict the current macroblock. Using smaller block sizes for motion estimation enables more accurate isolation of temporal changes within the macroblock resulting in a better prediction result.

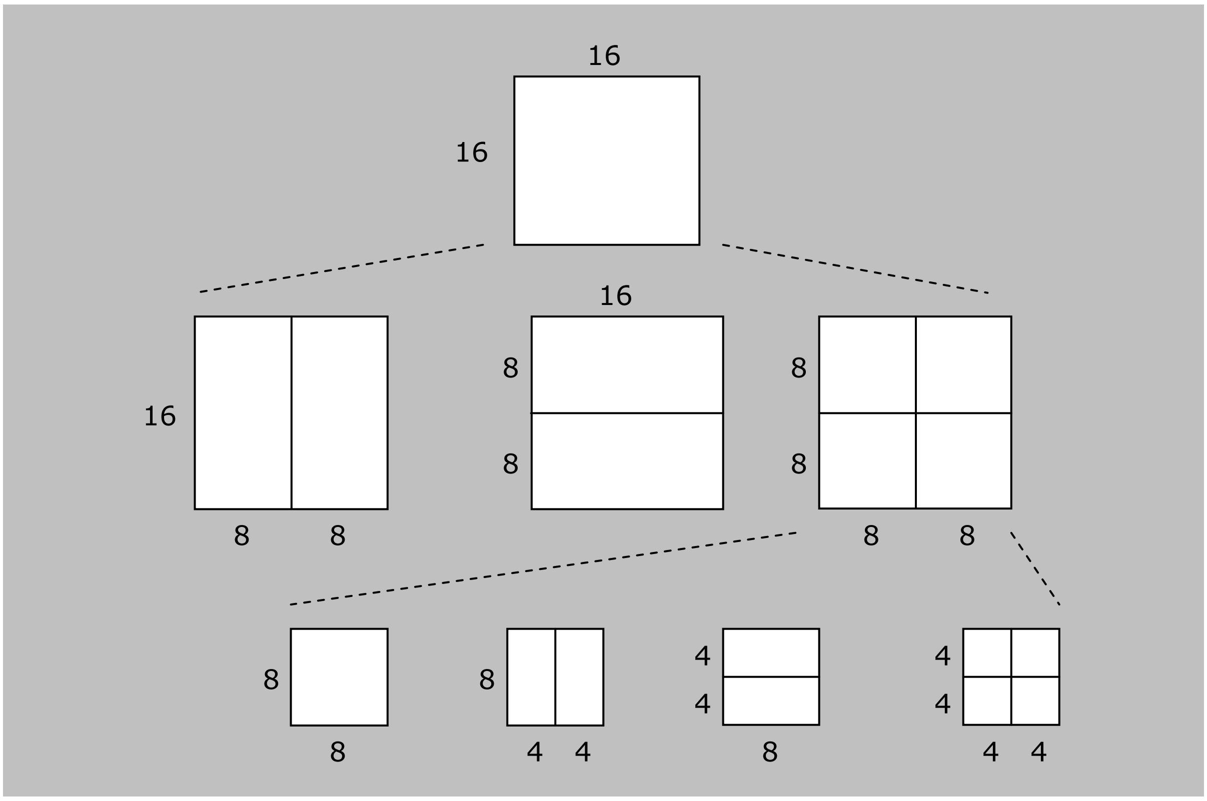


Figure 2-9: Sub-block partitions for motion estimation

The MPEG-2 coding standard supports only 16x16 block size leading to less accurate motion prediction but with fewer bits required to represent the motion data. MPEG-4 supports block sizes up to 8x8 offering moderate motion isolation. Advanced video coding standards such as H.264/MPEG-4 AVC introduce smaller block sizes (up to 4x4) for strong motion isolation, greater flexibility in block shapes, and greater precision in motion vectors in order to improve compression efficiency. However, the choice of predicting using 4x4 block size means an increase in motion data that has to be transmitted to the decoder for block reconstruction.

Use of multiple reference frames to exploit re-occurring periodic motion

Earlier coding standards such as MPEG-2 and MPEG-4 supported motion estimation of the current block from the immediate previously coded frame enabling low delay and minimal storage requirements. Advanced video coding standards such as H.264/AVC make it possible to find the best match for the current block from any of the previously coded reference frames. This feature is useful for dealing with:

1) Motion that is periodic in nature
2) Translating motion and occlusions
3) Alternating camera angles that switch back and forth between two different scenes

Although the use of multiple reference frames improves compression efficiency, it comes at an expense of increased computational cost at the encoder and increased storage requirement due to the need for storing previously coded reference frames.

Using I, P and B Macroblocks for improved compression efficiency

There are three types of macroblocks: I-macroblock, P-macroblock and B-macroblock as shown in Figure 2-10. An I-macroblock is coded using intra prediction, i.e. prediction from previously coded macroblocks in the same frame. A P-macroblock is inter predicted from a coded past picture and a B-macroblock is inter-predicted from two previously coded pictures which could be either from the past or future video frames. B-macroblocks give highest compression efficiency because motion data from past and future pictures are used in the prediction process but at a high computational expense.

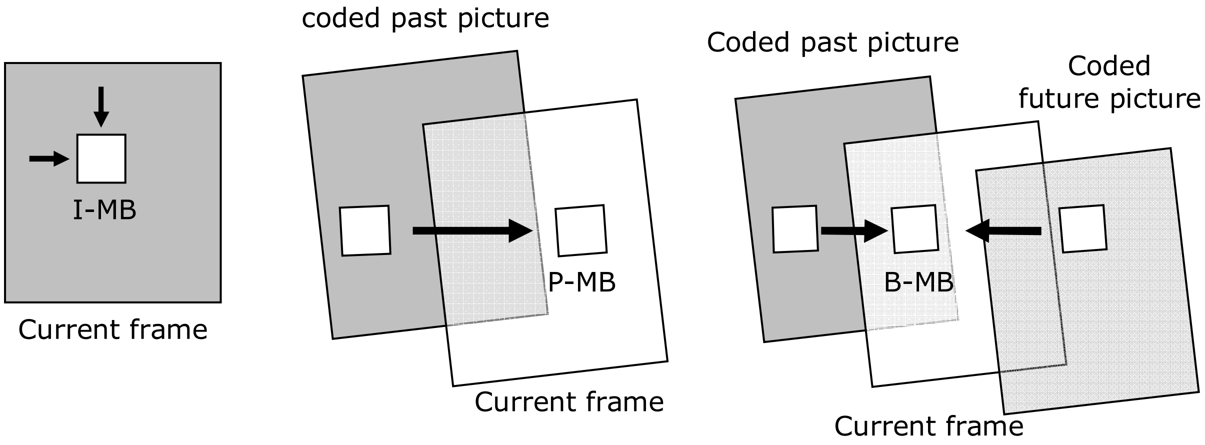


Figure 2-10: I, P and B macroblock types

Sub-pixel motion estimation

Motion estimation is performed to find the best matching block for the current block in a previously coded frame. The resulting motion vectors indicate the displacement between the positions of the current and best matching blocks. The efficiency of motion estimation depends on the accuracy of the motion vectors. Motion vectors can be obtained from full-pixel and sub-pixel locations. Sub-pixel options supported by video coding standards include: half-pixel, quarter-pixel and one-eighth pixel locations as shown in Figure 2-11. In order to obtain motion vectors from half-pixel locations, the luminance samples of the video frame are interpolated. Quarter-pixel and one-eighth pixel locations are obtained by interpolating half-pixel positions and quarter-pixel locations respectively.

H.264/AVC supports quarter-pixel motion estimation for luminance samples and one-eighth pixel motion estimation for chrominance. The motion vectors are coded and transmitted along with the prediction data to the decoder. At the decoder, the corresponding reference frames are interpolated according to the precision of the motion vectors in order to reconstruct the current block.

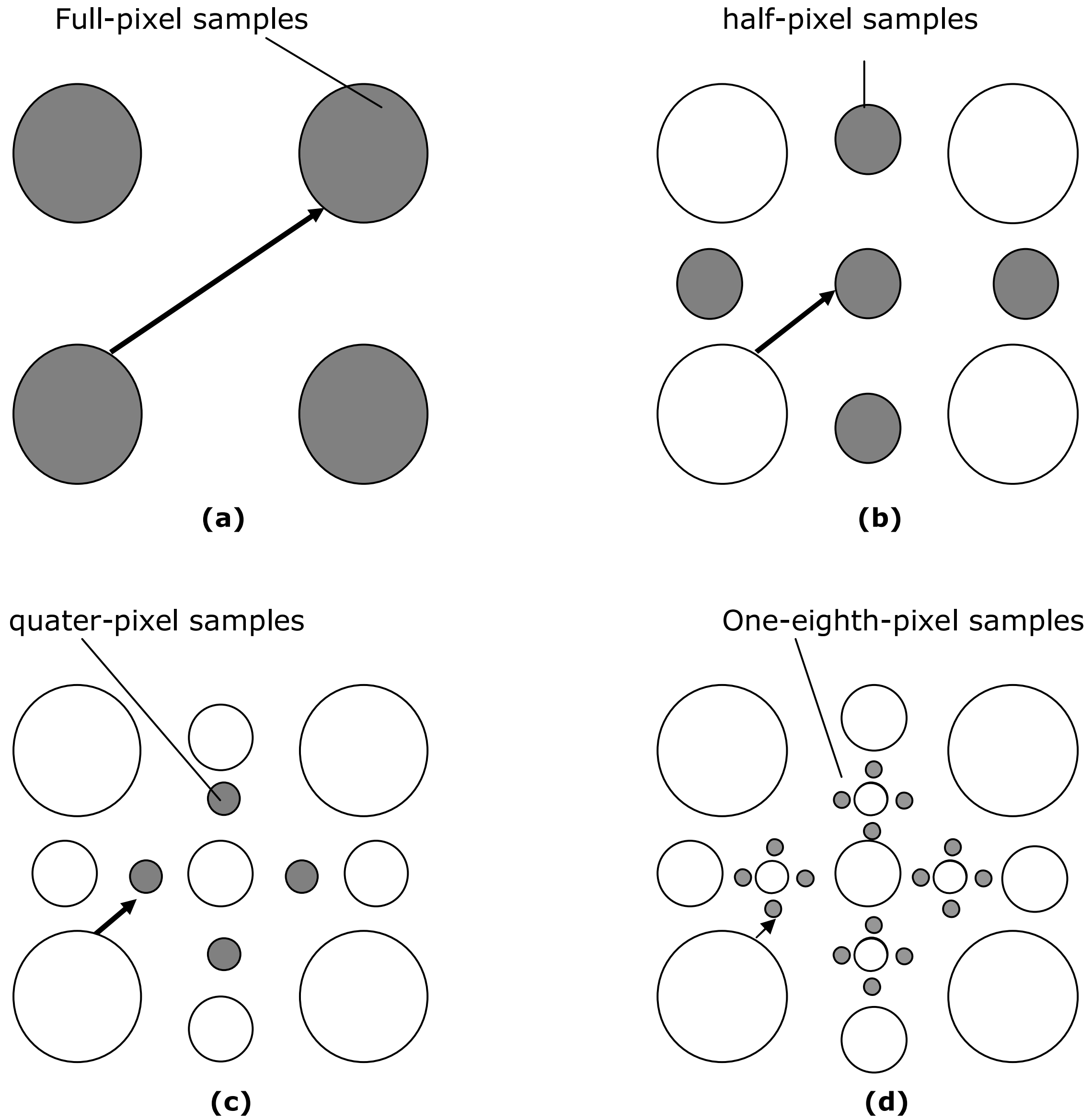


Figure 2-11: Sub-pixel motion estimation. (a) Full-pixel motion vector (1,1), (b)Half-pixel motion vector (0.5,0.5), (c) Quarter-pixel motion vector (0.25, 0.25) and (d) one-eighth pixel motion vector (0.125, 0.125) for chrominance samples.

2.5.2 Block-based transform coding

The residual blocks from motion estimation and compensation are transformed from spatial domain into frequency domain using block transform coding techniques. The residual block is converted into a block of transform coefficients which represent the magnitudes of spatial frequency components that make up the original residual block. Transformation does not lead to data loss and the process is completely reversible using inverse transformation. Transformation from spatial domain to frequency domain is performed for better energy compaction using a smaller number of larger coefficients and to de-correlate data by reducing inter-dependency. The human visual system is more sensitive to low/medium frequencies than high frequencies [17]. By separating image data in terms of frequencies, it may be possible to discard the higher frequencies without affecting the visual quality of the image block.

Discrete Cosine Transform (DCT) [17] is the most widely used block-based transform in video compression because it tends to concentrate the visually important contents of a block into a smaller number of coefficients for efficient encoding [18].

The two-dimensional DCT of an NxN block with pixels represented as f(i,j) and transform coefficients as F(u,v) is calculated using the following formula [17]:

$$F(u,v) = \frac{2}{N} C(u)C(v) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} F(i,j) \cos\left(\frac{(2i+1)u\partial}{2N}\right) \cos\left(\frac{(2j+1)v\partial}{2N}\right) \qquad (4)$$

where $C(n) = \begin{cases} \sqrt{\frac{1}{N}} & , n = 0 \\ \sqrt{\frac{2}{N}} & \text{otherwise} \end{cases}$

For an NxN block, the first DCT coefficient F(0,0) is called the 'DC coefficient' and is the average of all the samples in the block. It represents zero spatial frequency. The other DCT coefficients are called 'AC coefficients' which are arranged in order of increasing horizontal and vertical frequencies. The inverse DCT of an NxN block of coefficients will produce the original block in spatial domain with no data loss.

Video coding standards such as MPEG-2 employ a true DCT 8x8 transform as previously described that operates on floating-point coefficients. Since the computation of DCT can be computationally expensive and involves floating point operations, advanced video coding standards such as H.264/MPEG-4 AVC use a DCT-like 4x4 integer transform. The smaller block size of H.264/MPEG-4 AVC reduces blocking and ringing artefacts.

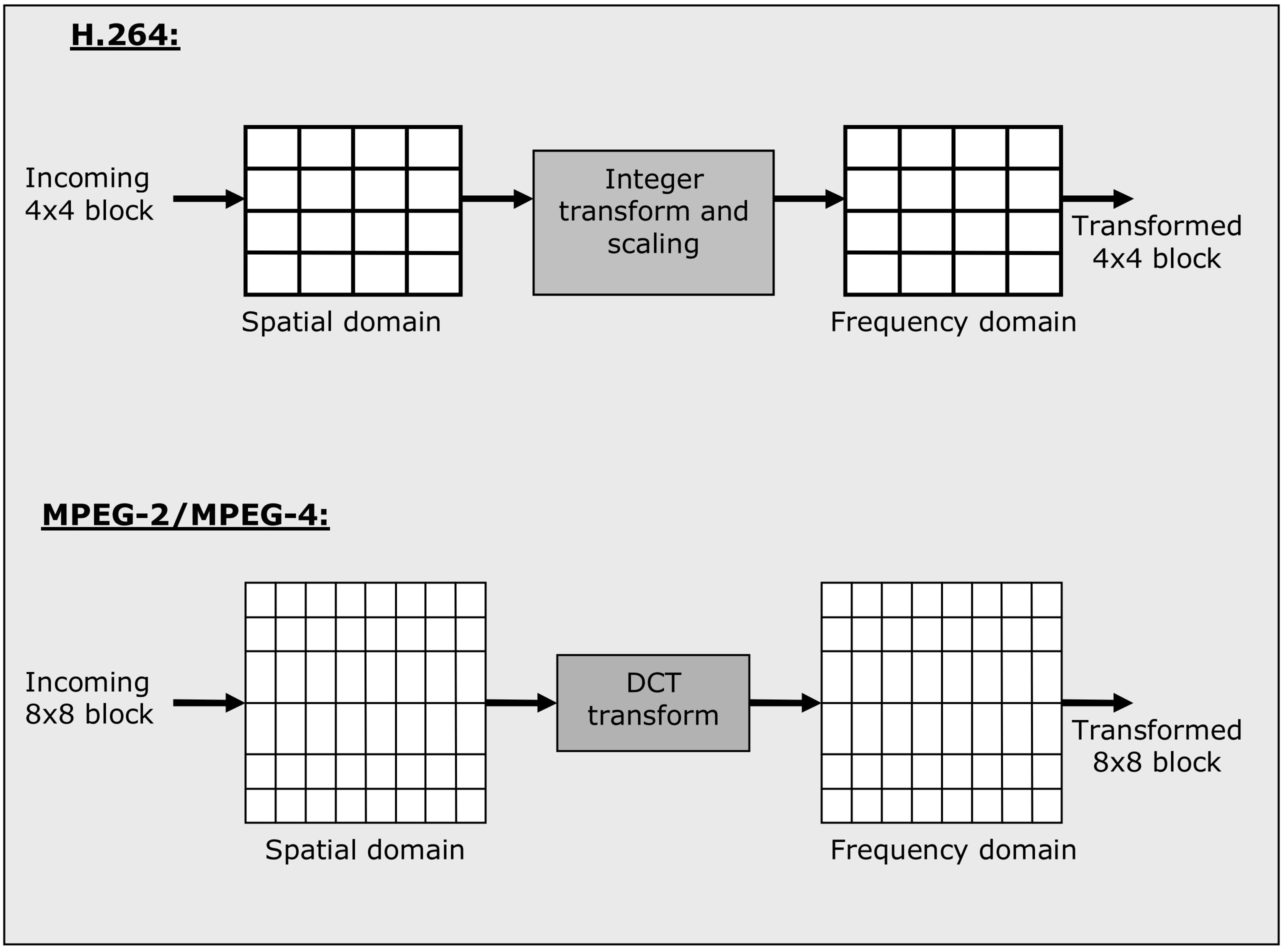


Figure 2-12: Transform coding in various video coding standards

2.5.3 Quantisation

In video compression, quantisation involves scaling the transform coefficient values of a block using a quantisation step and rounding to the nearest integer value. Information is lost during the rounding process. The amount of compression can be controlled by varying the size of the quantisation step. A larger quantisation step leads to a bigger rounding error and hence increased compression because the resulting quantised coefficients will prominently have small and zero coefficients. Inverse quantisation performed at the decoder involves rescaling the quantised coefficients with the quantisation step. However the rounding error caused during the quantisation process is irreversible leading to permanent data loss. The quantisation process eliminates high frequency coefficients. Since higher frequency coefficients contribute to image detail, using a large quantisation step will lead to elimination of higher frequencies resulting in blurring and blocking artefacts in the reconstructed block.

Lower bit rates can be achieved by increasing the levels of quantisation but at the expense of loss in image quality. Quantisation is also used for constant bit rate applications where it is varied to control the output bit rate. Figure 2-13 shows the effect of quantisation on a 4x4 DCT coefficient block. As can be seen, increasing the quantisation step causes an increase in the difference between the original DCT coefficients and the inverse quantised coefficients.

569	-6.7354	36.5	-34.553
85.949	-5.3406	4.0421	-17.489
52.5	-39.063	-114	-43.351
42.872	-139.49	-56.876	82.341

Original DCT coefficients

570	-10	40	-30
90	-10	0	-20
50	-40	-110	-40
40	-140	-60	80

Inverse Quantised DCT coefficients
(Quantisation Step=10)

560	0	40	-40
80	0	0	-20
60	-40	-120	-40
40	-140	-60	80

Inverse Quantised DCT coefficients
(Quantisation Step=20)

570	0	30	-30
90	0	0	-30
60	-30	-120	-30
30	-150	-60	90

Inverse Quantised DCT coefficients
(Quantisation Step=30)

Figure 2-13: Inverse Quantisation of 4x4 DCT coefficients using various quantisation steps

2.5.4 Entropy Coding

Statistical redundancies within the quantised DCT coefficient data can be exploited to gain further compression. The quantised DCT coefficients in a block contain fewer non-zero coefficients and a large number of zero coefficients. Entropy coding is a lossless process of converting encoded data into codes by exploiting its statistical redundancy. Before entropy coding can take place, the 4x4 or 8x8 quantized coefficient blocks must be serialised. Depending on whether these coefficients belong to a frame block or a field block, a different scan pattern is selected to create the serialised stream as shown in Figure 2-14. The scan pattern orders the coefficients from low frequency to high frequency. Then, since higher frequency quantized coefficients tend to be zero, run-length encoding is used to group the trailing zeros, resulting in more efficient entropy coding.

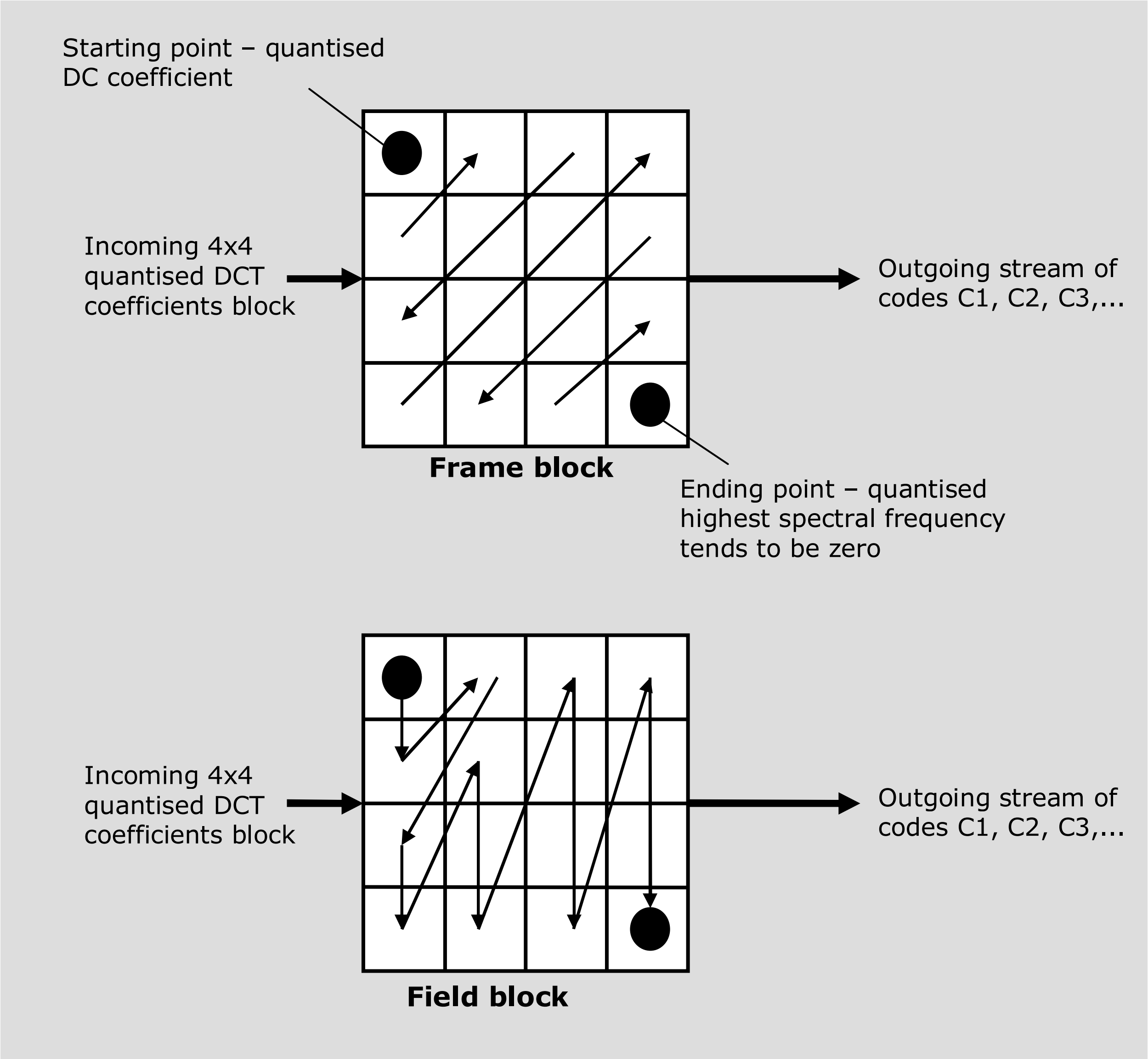

Starting point – quantised DC coefficient
Incoming 4x4 quantised DCT coefficients block
Outgoing stream of codes C1, C2, C3,...
Frame block
Ending point – quantised highest spectral frequency tends to be zero
Incoming 4x4 quantised DCT coefficients block
Outgoing stream of codes C1, C2, C3,...
Field block

Figure 2-14: Scanning order for a 4x4 block of quantised DCT coefficients

The entropy coding involves converting the quanitsed DCT co-efficients, video header information and motion vectors into bits. Entropy coding improves coding efficiency by assigning a smaller number of bits to frequently used symbols and a greater number of bits to less frequently used symbols.

There are three major types of entropy coding:
A: Variable Length Coding (VLC)
B: Context Adaptive Variable Length Coding (CAVLC)
C: Context Adaptive Binary Arithmetic Coding (CABAC).

Variable length coding involves assigning codes to the non-zero quantised DCT coefficients in a block based on the frequency of occurrence. Short codes are assigned to more frequently occurring coefficients and longer codes are assigned to less frequently occurring values. The zero coefficients are encoded using run-length encoding which involves transmitting a number to represent the length of the current ‛run’ of zeros. VLC is used in MPEG-2 standard.

Context Adaptive Variable Length Coding (CAVLC) offers superior coding efficiency compared to VLC. Context adaptive means that different code tables are used according to the content of local statistics of the block in order to achieve better coding efficiency. The disadvantage of CAVLC is that it can only encode residual coefficients context adaptively. This encoding technique is adopted by H.264/AVC.

Context Adaptive Binary Arithmetic Coding offers superior coding efficiency over VLC and CAVLC by adapting to the changing probability distribution of symbols, by exploiting correlation between symbols, and by adaptively exploiting bit correlations using arithmetic coding. Unlike VLC and CAVLC, Arithmetic coding generates non-integer codes for higher efficiency. CABAC is not only limited to encoding residual coefficients but also syntax elements such as motion information, encoding parameter sets and header data. H.264/AVC supports CABAC encoding technique.

H.264 also supports Context Adaptive Variable Length Coding (CAVLC) which offers superior entropy coding over VLC without the full cost of CABAC. However, CABAC has been reported to achieve 9%-14% higher compression efficiency compared to CAVLC [19].

2.5.5 De-blocking filter

Advanced video compression standards, such as H.264/MPEG-4 AVC have an in-loop de-blocking filter that operates on both 16x16 macroblocks and 4x4 block boundaries. The aim of this filter is to smooth the blocking edges around the boundary of each macroblock without affecting the sharpness of the picture. The reconstructed pictures at the encoder are (optionally) filtered using the de-blocking filter before being used as reference pictures for inter prediction of future frames. In the case of macroblocks, the filter is intended to remove artefacts that may result from adjacent macroblocks having different estimation types (e.g. motion vs.

intra estimation), and/or different quantisation scales. In the case of blocks, the filter is intended to remove artefacts that may be caused by transformation/quantisation and from motion vector differences between adjacent blocks. The loop filter typically modifies the two pixels on either side of the macroblock/block boundary using a content adaptive non-linear filter. The filter strengths depend on the level of quantisation used. A detailed explanation of the de-blocking filter is given in [20].

2.6 Video coding standards

Standardising video coding techniques enables improved encoding and decoding strategies to be employed in a standard-compatible manner in order to encourage interoperability between video communication systems developed by different manufacturers. There are two international bodies that are responsible for standardising video codecs and helping shape the video communications Industry:

A: Video Coding Experts Group (VCEG) formed by the Telecommunications sector of the International Telecommunications Union (ITU-T).

B: Motion Pictures Experts Group (MPEG) formed by the International Standardisation Organisation (ISO)

Each video coding standard specifies the *syntax* of the bit stream and the decoding process (example: use inverse discrete cosine transform (IDCT), but not how to implement IDCT) as shown in Figure 2-15. Standards do not specify the encoder or decoder specifications.

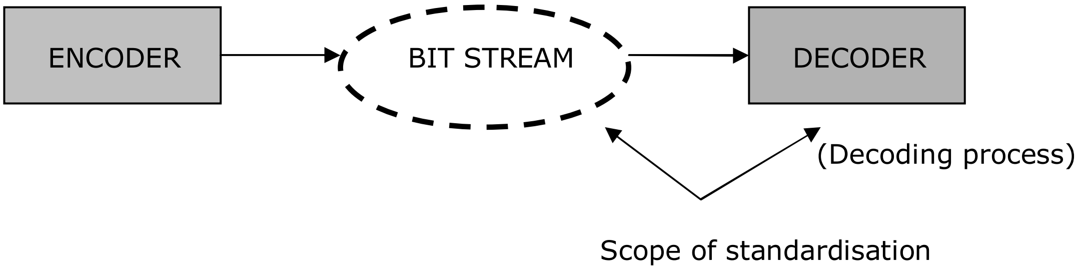


Figure 2-15: Scope of standardisation

The standards released by the ITU-T include the H.26x series and the ISO has released the MPEG series. This section gives an overview of popular video coding standards released to date including their features and applications.

2.6.1 MPEG-1 [21]

The draft MPEG-1 standard was released in 1993. The main features of this standard include support for progressively scanned video with bitrate up to 1.5Mbps, support for flexible picture types such as I, P and B pictures to provide improved compression efficiency, half-pixel motion compensation and real-time playback. The standard was primarily developed for storage of video and audio on digital media such as CD-ROM.

2.6.2 MPEG-2 [22]

MPEG-2 is based on MPEG-1 and was developed in 1995 to support a wider range of resolutions and bitrate. It was aimed at applications including digital television, high definition television (HDTV) and satellite television broadcasting. Video coding tools in the standard include support for interlaced video and scalable video coding. This standard introduced the concept of "profiles and levels" to specify a set of tools and capabilities required by the decoder to support different applications, resolutions, and bitrate, and provide inter-operability between different decoders.

2.6.3 H.261 [23]

Released in 1990, this standard was aimed at low bitrate video coding applications including video conferencing and videophone over Integrated Services Digital Network (ISDN) channels. The standard utilises hybrid video coding which consists of block-based motion estimation and DCT transform coding, and supports only CIF and QCIF resolutions of non-interlaced video.

2.6.4 H.263 [24]

H.263 was standardised by ITU-T in 1993 for low bit rate video communication over Public Switched Telephone Network (PSTN) and mobile networks with transmission bitrates of around 10-24kbps or above. The core algorithm of H.263 is based on H.261 but it supports a bigger range of resolution formats and coding tools including:

- Half-pixel motion estimation where the motion vector accuracy is up to half a pixel.
- Unrestricted motion vector mode where the motion vector is allowed to point outside the boundaries of the video frame.
- Predictive coding of motion vectors where the current macroblock is predicted using previously coded macroblocks either in the same video frame or previous video frames.
- Advanced prediction where a macroblock is divided into four 8x8 blocks and each block is individually motion compensated to yield four motion vectors for each macroblock. This method of prediction results in higher compression efficiency and flexibility as it is able to represent motion within a macroblock with better accuracy.

2.6.5 H.264/MPEG-4 Part 10: Advanced Video Coding [1,25]

The H.264/AVC standard was developed by the joint video team (JVT) for a variety of applications such as internet video streaming, mobile video, high definition television and DVD. The H.264/AVC is capable of achieving significantly improved compression performance and flexibility compared to previous video coding standards.

The core algorithm is similar to H.263 but it includes improvements in coding techniques such as:

- **Enhanced motion estimation and compensation** using varying block sizes from 16x16 pixel to 4x4 pixel sized macroblock sub-partitions. Smaller block sizes provide more accurate motion vectors and hence better motion compensation. H.264 also supports quarter pixel motion estimation and multiple reference frames to provide improved motion vector accuracy.
- Unlike previous coding standards, H.264 uses a 4×4 **integer block transform** [26,27], which is based on the DCT transform, operating on every 4×4 residual blocks. Compared to the conventional DCT transform, the integer transform does not produce any loss of data as it is defined exactly by the integer arithmetic operation, so that inverse transform mismatch is avoided.
- H.264 uses an **improved in-loop deblocking filter** [20,28] to smooth the blocking around the boundary of each macroblock without affecting the

sharpness of the picture. Therefore, subjective video quality is dramatically improved. Motion estimation predicted from filtered macroblocks has been shown to produce better results compared with non-filtered macroblocks.

- H.264 supports two **advanced entropy coding** techniques, CABAC and CALVC, depending on the coding modes.

The above mentioned features have enabled H.264/AVC to achieve an average bitrate saving of up to 50% compared to previous video standards [2]. The AVC/H.264 standard defines four different Profiles: Baseline, Main, Extended and High Profile to provide support for a variety of applications, bitrate and resolutions:

- **Baseline Profile**: provides support for I and P frames, progressive video and CAVLC only entropy encoding.
- **Extended Profile**: supports I, P, B, SP and SI frames, progressive video and CAVLC only entropy encoding
- **Main Profile**: supports I, P and B frames, progressive and interlaced video, and offers both CAVLC and CABAC.
- **High Profile** adds to the Main Profile: 8x8 intra prediction, lossless video coding, support for more video formats including 4:0:0, 4:2:0, 4:2:2 and 4:4:4.

2.6.6 Annex G of H.264/AVC: Scalable video coding [29]

Video is currently used in increasingly diverse applications on many client devices from IPTV to mobile devices and the video streams for these devices are different in terms of resolutions, framerate and available bandwidth. To be made more compatible with a specific viewing device and channel bandwidth, the video stream must be encoded many times with different settings. Each combination of settings must yield a stream that targets the bandwidth of the channel carrying the stream to the consumer as well as the decoding capability of the viewing device.

The scalable video coding extension to the H.264 standard (H.264 SVC) is designed to address this problem. It is based on the H.264 advanced video codec standard (H.264 AVC) but the encoded stream it generates is scalable spatially, temporally and in terms of video quality. Therefore, the decoded video can be rendered at different frame rates, resolutions, or quality levels to suit the requirements of the transmission channel and the viewing device. Unlike the original H.264/AVC, the SVC extension introduces layers within the encoded stream:

- A base layer containing the lowest temporal, spatial, and quality representation of the encoded video stream.
- Enhancement layers containing additional information required to reconstruct higher quality, resolution, or temporal versions of the video during the decoding process.

This layered approach allows the generation of an encoded stream that can be truncated to meet the computational requirements of the decoder. The decoder can simply extract the required layers from the encoded video stream with no additional processing on the stream. This process can even be performed "in the network". The video stream transitions from a high bandwidth to a lower bandwidth network could be made to suit the available bandwidth and the decode capabilities of the handheld device. Further information on the technicalities of the H.264/SVC can be found in [29].

2.6.7 High Efficiency Video Coding / HEVC / H.265 [30]

High efficiency video coding is a draft standard and a successor of the H.264/AVC. It is currently under development by the ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG). HEVC is aimed at improving the coding efficiency of the H.264/AVC high profile in terms of bitrate reductions, robustness to errors, computational complexity and processing delay time. HEVC targets next generation HDTV and support for a wide range of resolutions from QCIF to ultra high definition video (7680x4320).

Main features of the draft standard include:

- Extended block sizes for the coding unit from 8x8 to 64x64.
- Larger transform block sizes which are non-square, quad-tree structured with sizes from 4x4 to 32x32 samples.
- Larger number of Intra prediction directions (up to 34).
- Adaptive motion vector prediction
- Entropy coding using CABAC or low complexity entropy coding.
- Advanced de-blocking loop filter (ALF).
- High-accuracy interpolation using 6- or 12-tap interpolation filter.

Investigations are underway to evaluate the performance of these features and the final draft of this standard is expected to be completed in January 2013 [31].

2.7 Rate-Distortion optimised video coding

Rate distortion optimisation is the process of minimising distortion for a given bitrate. In video compression, rate distortion optimisation is a technique for selecting the best coding option to encode each coding unit that will minimise distortion for a target bitrate [32,33]. Research [33] has shown that by optimising this selection process, the overall performance of video coding increases. Advanced block-based video encoders such as the H.264/AVC offer a large number of coding modes to encode each coding unit to suit a variety of spatial and temporal content. For example, relatively dormant (stationary) regions of the video scene could simply be copied from previously decoded frames into the current frame using the SKIP mode. New areas in the video scene may be effectively coded directly using INTRA modes. On the other hand, key changing regions could be coded using block-based motion compensation followed by encoding of the prediction residual using INTER modes. Hence, it is a challenging task for the encoder to choose the best mode for each coding unit from a very large set of mode choices.

Rate distortion optimization (RDO) can be applied in the video encoder for optimizing motion estimation, rate control and mode decision processes [34]. This section explores the practical implementations of these processes for the H264/AVC encoder.

2.7.1 Rate Distortion optimised motion estimation

Block-based motion estimation involves finding a motion vector which represents the displacement between the location of the current coding unit and its best match in the previously reconstructed frame. This is followed by obtaining the motion-compensated residual block as the difference between the two blocks. Optimising the motion estimation process in the rate-distortion sense would ideally involve coding the residual data for each possible motion vector, decoding and reconstructing the macroblock in order to measure the corresponding bit usage and distortion. However, due to the very large number of possible motion vectors to choose from, the computational overload of coding every residual difference signal is very large in a practical video encoder. Hence, rate-distortion optimised motion estimation involves finding a motion vector that minimises the motion vector cost J_{mv} which is calculated as follows:

$$J_{mv} = D_{DFD} + \lambda_{motion} R_{mv} \qquad (5)$$

Where D_{DFD} is the pixel differences between the current macroblock and the motion compensated (displaced) macroblock for the corresponding motion vector. It is usually measured as the sum of absolute differences (SAD) or the sum of squared differences (SSD). R_{mv} is the number of bits used to transmit the motion vector and λ_{motion} is the Lagrangian multiplier for the motion estimation process.

It must be noted that in the context of video coding, rate-distortion optimisation usually involves rate and distortion measures obtained using the original macroblock and the reconstructed macroblock which is obtained as a result of coding and decoding process. However, D_{DFD} and R_{mv} measures used in equation (5) are estimations used instead of the 'actual' distortion and rate values of the macroblock. Therefore, optimisation of motion estimation involves minimising the block difference subject to a constraint on the motion vector bits rather than the actual rate-distortion optimisation process.

2.7.2 Rate distortion optimised mode selection

Motion estimation on a macroblock results in the selection of the most appropriate motion vector for each available inter prediction mode. This is followed by the mode selection process which involves choosing the best mode to encode a macroblock from the available modes. In case of H.264, the available modes include 7 different coding modes: SKIP, INTER 16x16, INTER 16x8, INTER 8X16, INTER 8X8, INTRA 16X16 and INTRA 4X4 so that spatial and temporal detail in a macroblock can be best presented.

The goal of the rate-distortion optimised mode selection algorithm is to find the best coding mode from a set of available modes that minimises the distortion for a rate constraint. The process involves encoding each macroblock using a certain mode followed by decoding and reconstruction to obtain the actual rate and distortion measures. The optimisation is carried out by minimising the following mode selection cost function J_{md} :

$$J_{md} = D_{\mathrm{Re}c} + \lambda_{\mathrm{mod}e} R_{\mathrm{Re}c} \tag{6}$$

Where $D_{\mathrm{Re}c}$ is the actual distortion between the original and reconstructed macroblocks obtained by calculating the sum of squared difference (SSD) between the two macroblock pixels. $R_{\mathrm{Re}c}$ is the number of bits spent for coding the entire macroblock (including transformed residual coefficients and motion vectors) and $\lambda_{\mathrm{mod}e}$ is the Lagrange multiplier for mode selection. Figure 2-16 shows the process of calculating the rate-distortion cost function during the mode selection process in the H.264 video encoder.

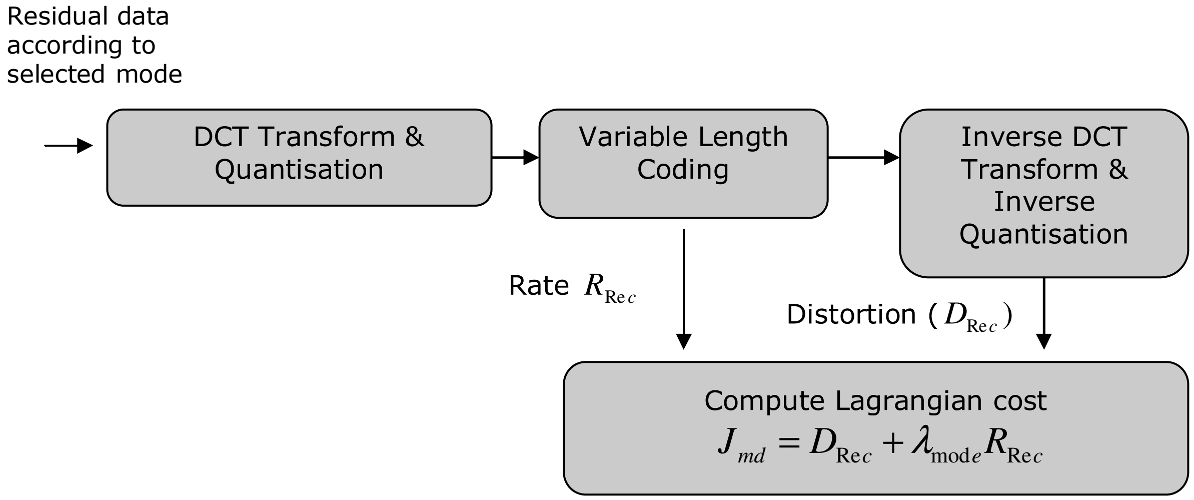

Residual data according to selected mode
DCT Transform & Quantisation
Variable Length Coding
Inverse DCT Transform & Inverse Quantisation
Rate $R_{\mathrm{Re}c}$
Distortion ($D_{\mathrm{Re}c}$)
Compute Lagrangian cost
$J_{md} = D_{\mathrm{Re}c} + \lambda_{\mathrm{mod}e} R_{\mathrm{Re}c}$

Figure 2-16: The process of calculating mode selection Lagrangian cost in H.264

2.7.3 Choice of λ for mode selection and motion estimation

Lagrangian R-D optimisation involves finding a mode that minimises the rate-distortion cost function for an appropriate Lagrange multiplier that satisfies a certain rate constraint. The quantisation parameter (QP), which controls the amount of compression, also plays a key role in achieving this rate target. Hence different QP values should also be evaluated along with the available macroblock modes for minimising the Lagrangian R-D cost function.

Finding an appropriate Lagrange multiplier for a certain rate constraint is an iterative process which involves:

- Choosing a particular value for λ = λ*
- Minimising the R-D cost function for all available QP and mode combinations.
- Selecting the best mode and QP combination (QP*, mode*) for the given λ* value.

The above iterative process, though optimal, is impractical for real-time video coding scenarios due to the large computational load. Hence, modelling the Lagrange multiplier is essential to determine the λ value prior to encoding the macroblock.

Sullivan et al [33] conducted experiments to model the Lagrange multiplier using the H.263 test model [35]. Lagrangian optimisation of mode selection was performed on various sequences using different QP values along with available macroblock modes for a selected set of λ values. By plotting the $\lambda_{\mathrm{mod}e}$ versus average macroblock QP values, they determined the approximation of the functional relationship between $\lambda_{\mathrm{mod}e}$ and average QP as:

$$\lambda_{\mathrm{mod}e} = 0.85 * QP^2 \qquad (7)$$

This equation is a one-to-one relationship between λ and QP which implies that for a certain $\lambda_{\mathrm{mod}e}$ value, a specific QP can be selected. This relation also means that, if a certain QP is selected prior to encoding (by a rate control algorithm), the corresponding $\lambda_{\mathrm{mod}e}$ value can be calculated using equation (7) to optimise the mode selection process. Sullivan et al also found that equation (7) holds for sequences with widely varying content indicating that sequence statistics have insignificant impact on the $\lambda_{\mathrm{mod}e}$ - QP relationship. However, it must be noted that ignoring variations in sequence statistics may lead to sub-optimal solutions.

The following Lagrange multiplier models for motion estimation were also proposed by Sullivan et al through experimental evaluations:

1. When the distortion D_{DFD} is calculated using sum of squared differences (SSD), the Lagrange multiplier can be calculated as:

$$\lambda_{motion} = \lambda_{\mathrm{mod}e} \qquad (8)$$

2. When the distortion D_{DFD} is calculated using sum of absolute differences (SAD), the Lagrange multiplier can be calculated as:

$$\lambda_{motion} = \sqrt{\lambda_{\mathrm{mod}e}} \qquad (9)$$

It was shown by the Sullivan et al that the overall performance gain of the rate-distortion optimised mode selection and motion estimation process was around 10% reduction in bit rate and around 0.5dB in PSNR for fixed output picture quality. Therefore, this algorithm was adopted in the H.263 reference encoder TMN10 [36].

Following the experiments carried out in [33], Weigand et al presented a new model for the Lagrange multiplier as a function of QP for the H.264/AVC reference encoder [37]. Unlike the model described in [33], the Lagrange multiplier model for mode selection is selected based on the frame type.

For I and P frames:

$$\lambda_{\mathrm{mod}e,P} = 0.85 * 2^{QP/3} \qquad (10)$$

For B frames:

$$\lambda_{\mathrm{mod}e,B} = \max\left(2, \min\left(4, \frac{QP}{6}\right)\right) * \lambda_{\mathrm{mod}e,P} \qquad (11)$$

The Lagrange multiplier for motion estimation is calculated as:

$$\lambda_{motion} = \sqrt{\lambda_{\mathrm{mod}e}} \qquad (12)$$

This is similar to H.263 (9) but the distortion D_{DFD} is calculated using sum of (transformed) absolute differences (SATD) as opposed to sum of absolute differences.

A graphical interpretation of the Lagrange multiplier model for I and P frames is given in Figure 2-17.

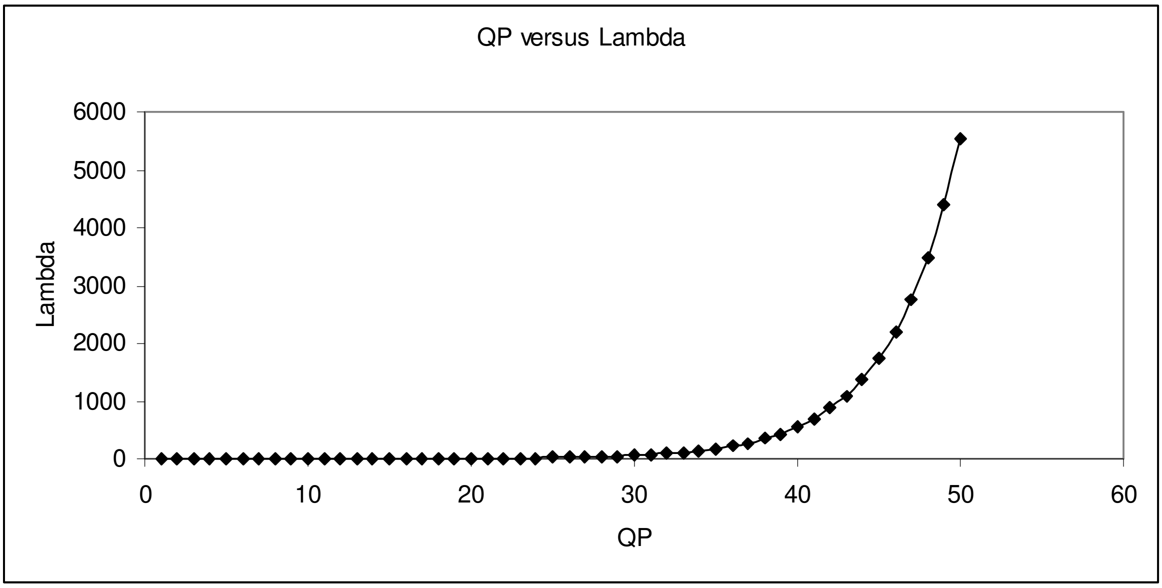


Figure 2-17: Relationship between QP and the Lagrange multiplier in H264/AVC

A large QP value will result in a large Lambda value λ and the Lagrangian cost J=D+λR weigh more on coded bits R making it a dominating factor in the mode selection process. Therefore, modes that produce lower coded bits will have a higher probability of being selected. Similarly, a smaller QP yields a smaller λ and the Lagrangian cost J=D+λR weigh more on the distortion parameter D making. Hence, modes (such as intra prediction) resulting in a lower distortion may then be chosen.

In summary, the Lagrange multiplier models described above are very simple models which can be easily incorporated into any block-based encoder. The models are based on experimental results, assumptions and approximations. Since these models do not incorporate changing sequence statistics, the relative R-D performance gains may vary depending on the sequence under test. More details about the Lagrange multiplier models for H.264/AVC can be found in [38].

2.7.4 Rate Distortion optimisation and rate control

The video encoder uses parameters such as quantisation parameter and motion estimation search area, to control the outcome of an encoding process. If these control parameters are kept constant, then the number of bits produced for each macroblock will change depending on its content, causing the bit rate of the encoder output (measured in bits per second) to vary. An encoder with constant control parameters will typically produce more bits for high motion and/or detail content and fewer bits for low motion and/or detail. Figure 2-18 shows the variation of mean coded bitrate across consecutive frames in the Foreman CIF sequence compressed using H.264/AVC reference codec at fixed QP=26.

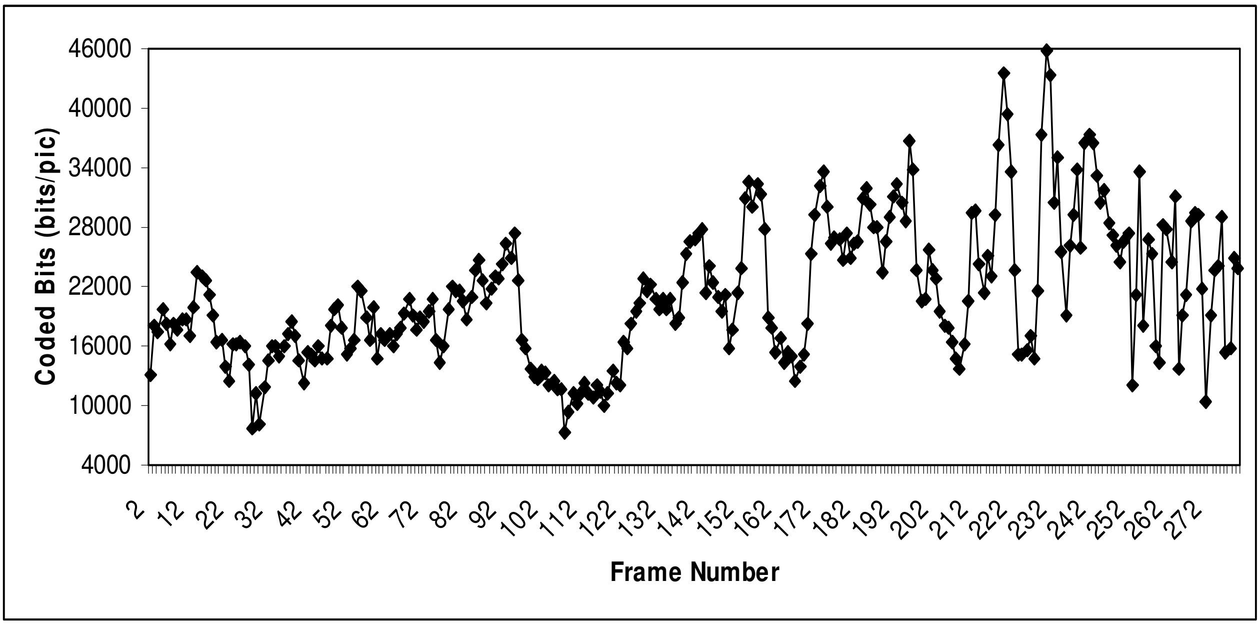


Figure 2-18: Variation of coded bits between consecutive frames of Foreman sequence when coded using fixed QP.

This variation in bitrate can be a problem for practical video communication applications such as video transmission across a constant bit rate or a congested transmission channel. In these cases, it is necessary to adapt or control the bit rate produced by a video encoder to match the available bitrate.

Rate control involves modifying the encoder control parameters in order to maintain a target bit rate. Rate control is not a part of video coding standards, but the standards group has issued non-normative guidance to aid in implementation. A common approach to rate control is to modify the quantisation parameter QP to achieve a target bit rate. Increasing QP reduces coded bitrate (at the expense of lower decoded quality) and decreasing QP increases coded bit rate. A rate control mechanism recommended for H.264/AVC is described in [38,39,40]. This mechanism makes use of a quantitative model that adapts to changing macroblock statistics in order to determine the relationship between QP and bit rate. Models are necessary to avoid iterative encoding using different QP values in order to achieve the target bit rate for the macroblock. However, models lead to approximate results. Therefore the required rates may not be tightly achieved for each macroblock. In any case, multiple encoding is avoided to minimise the computational complexity and overall delay.

2.8 Summary

This chapter explained the fundamental concepts of digital video representation and video coding. These are summarised below:

- The high bandwidth requirement of digital video means that digital video coding is a necessary part of multimedia video communication applications where transmission bandwidth and storage capacities are limited.
- Block-based video coding algorithms employ various techniques including predictive coding for exploiting data redundancies, transform coding for converting data to a compactable form for efficient compression, quantisation for performing lossy compression, entropy coding to remove statistical redundancy and coding tools such as the de-blocking filter to improve the quality of compressed video.
- In order to increase the inter-operability of video coding algorithms on various platforms and applications, video coding standards have been introduced.
- H.264/AVC is the most efficient of all existing video coding standards in terms of bitrate reductions and flexibility. This high efficiency is achieved due to the use of advance coding tools and the efficient rate-distortion optimisation techniques discussed in this chapter.
- However, with bit rate savings comes quality reduction and video quality is an important factor in multimedia applications. Therefore, maintaining good quality whilst achieving high compression efficiency is a challenge to existing video compression algorithms. The next chapter will look at various techniques for measuring and evaluating the quality of video sequences compressed using block-based video coding algorithms.

3 Video Quality Measurement

3.1 Introduction

Multimedia video data may be subjected to various forms of distortion during video capture, transmission and storage. Video capture devices may introduce distortion, such as aliasing, in the video during the digitisation process. Video coding algorithms used to reduce transmission bandwidth and storage requirements of video data may cause degradations in video quality during compression. Error-prone communication networks such as the Internet may cause data loss or delay during video data transmission. All these imperfections may cause degradations in video quality. Therefore, video quality measurement techniques are necessary in multimedia video communication systems for evaluating and quantifying degradations in video quality so that they can be monitored, managed and possibly reduced.

Video quality metrics can be employed in various stages of the video communication process. Video acquisition systems such as digital cameras may use quality metrics to monitor and automatically adjust their settings to acquire best quality video. These metrics may be embedded into video coding algorithms for optimising encoding parameters. Several video processing algorithms available for a specific task could be benchmarked using video quality metrics to determine the optimum choice. Network video servers can examine and control the quality of video transmitted through the network.

This chapter will focus on the basic concepts and approaches of existing video quality assessment techniques in order to understand their advantages and limitations. Most multimedia video applications are meant for the human observer. Therefore understanding the various mechanisms involved in the processing of visual information by the human visual system is important. Section 3.2 gives an overview of these mechanisms including the limitations of human vision. In section 3.3, various types of compressed-induced distortion are discussed. Approaches to measure video quality are reviewed in Section 3.4. Section 3.5 introduces the current status of objective video quality measurement research with emphasis on full reference objective metrics. In section 3.6, issues related to the validating and standardising of objective video quality measurement techniques are discussed. Techniques for evaluating the performance of video quality metrics are mentioned in section 3.7. Finally, section 3.8 makes some concluding remarks highlighting the limitations of the existing video quality assessment techniques and the need for a new video perceptual quality metric.

3.2 The Human Visual System (HVS)

Video displays and video compression algorithms are evolving to meet the requirements demanded by the human visual system. Hence, there is a need to understand the fundamentals of human vision in order to determine what is essential to process and display video in a way that is relevant to the human observer. This understanding involves looking at the various components of the human visual system, the mechanisms of processing of visual information and the limitations of visual perception.

3.2.1 Structure of the human eye

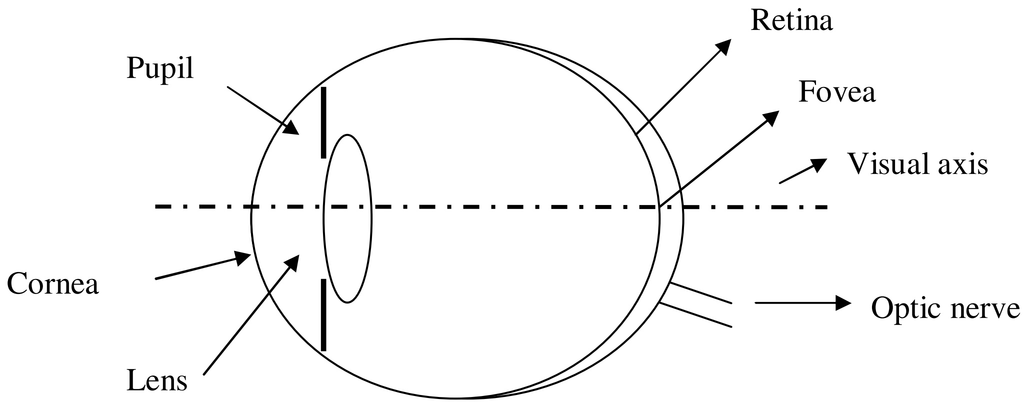


Figure 3-1: Diagram showing the structure of the eye

Figure 3-1 shows the main components of the human eye. Light coming from an object first encounters the eye at the cornea, the main refractive surface of the eye. The light then enters the eye through the pupil, the hole in the centre of the pigmented iris. The pupil is able to change its diameter in order to control the amount of light entering the eye, hence contributing to the eye's ability to adapt to a wide range of illuminations. The light then passes through the lens of the eye, a transparent flexible structure that changes its shape to focus the image onto the back of the eye. This flexible nature of the lens makes it possible to see near and distant objects. The fluids that fill the eye (vitreous and aqueous humour) help maintain its shape.

When the eye is properly focused, light from an object is imaged onto the back of the eye. Lining the back of the eye is the retina where light sensitive neurons called photoreceptors work as transducers to convert light energy into electro-chemical signals used by the nervous system for interpretation of visual data. There are two classes of photoreceptors: rods and cones. Rods facilitate vision in low levels of illumination. They serve to give an overall picture of the field of view and do not contribute to colour vision. The cones operate at higher levels of illumination and contribute to colour vision. The fovea of the eye is the central region of the retina and has the highest concentration of cones for high resolution vision. There are three types of cones depending on the sensitivity to the various wavelengths of visible light (400nm - 700nm). These are the short wavelength sensitive cones (S-cones), middle wavelength sensitive cones (M-cones) and long wavelength sensitive cones (L-cones). The three cone types are responsible for splitting the image projected onto the retina into three visual streams which can be thought of as the Red, Green and Blue colour components. The signals from the photoreceptors are transmitted from the eye to the brain through interconnecting nerve cells in the optic nerve called the ganglion cells. Further information on the structure of the human eye can be found in chapter 2 of [4].

3.2.2 The Visual Pathway

Visual pathway in the brain includes the eyes, the optical chiasms, the lateral geniculate nucleus (LGN) and the primary visual cortex [4]. Visual signals from the eye reach the brain for interpretation through the optic nerve. The optic nerves from both eyes meet and cross at the optic chiasm present at the base of the brain. This is where information coming from both eyes is combined and then distributed according to the visual field. The right half of the field of view is sent to the left side of the primary visual cortex and the left half of the field of view is sent to the right side of the primary visual cortex for processing. A small region in the centre of the field of view is processed redundantly by both halves of the primary visual cortex. The neurons in the visual cortex are known to be tuned to various aspects of the incoming visual streams such as spatial and temporal frequencies, orientation and motion. The visual streams undergo higher levels of processing in the brain for interpretation of data input by the human visual system. The lateral geniculate nucleus is a "bent knee" like structure found in the thalamus of the brain and is present on both sides of the brain. It is the primary relay centre for visual information coming from the retina and relays this information to the primary visual cortex for further processing.

3.2.3 Mechanisms of the Human Visual System

Various mechanisms of the human visual system facilitate the processing of the visual information and correlate with perceptual image and video quality [41]. These mechanisms include:

A: Light adaptation

B: Contrast sensitivity

C: Colour processing

D: Masking effects

E: Pooling of multi-channel information

Light adaptation is the ability of the human visual system to adapt to a wide range of light intensities ranging from scotopic (very low light) vision to photopic (bright/daylight) vision. Light adaptation by the human visual system is possible due to the controlling of the amount of light entering the eye through the pupil and the adaptation mechanisms of the photoreceptors which increase or decrease the signal output of the photoreceptors depending on the changing light intensities.

Contrast sensitivity is the sensitivity of the human visual system to relative variations in luminance over a wide range of background light intensity. The phenomenon that maintains contrast sensitivity over a wide range of intensities (ranging from faint lighting to daylight) due to the adaptation capabilities of the human visual system is called *Weber-Fechner's law*.

Colour processing by the human visual system is facilitated by the three types of cones in the retina: L-, M- and S- cones. Colour as perceived by the human visual system has five attributes: brightness - which is the intensity of the colour, lightness - which is relative to white colour, colourfulness - which is the chromaticity of the colour, chroma – which is chromaticity relative to white colour and hue – which is the attribute of a colour.

Masking is a phenomenon that explains why a stimulus which is visible by itself may not be visible in the presence of another stimulus. The masking effect reduces the visibility of the stimulus under test. Spatial masking occurs when a spatial component (such as texture or a strong edge) that is visible by itself may not be detected in the presence of another spatial component. Spatial masking is strongest when the image component under test and the surrounding image components have similar frequencies, colour and orientation. On the other hand, temporal masking is an elevation of visibility thresholds due to temporal changes in intensity. An increase in the amount of temporal change causes an increase in the masking effect.

Pooling of multi-channel information involves the integration of various types of visual information such as colour, texture, contrast, motion, shape, orientation and masking effects in order to make an interpretation. It is not quite understood how the human visual system performs pooling but it is known to involve cognition.

Apart from the above mentioned visual mechanisms, compression induced distortion also has influence on the visual quality of multimedia video sequences. This is described in the next section.

3.3 Compression induced visual distortion in video

Block-based compression algorithms rely on motion estimation and compensation, block-based Discrete Cosine Transform (DCT) and quantisation processes in order to compress video. In such coding schemes, visual distortions are mainly caused due to the quantisation of transform coefficients. Other factors contributing to visual distortion include motion prediction error and sampling of video data which causes aliasing. This section introduces prominent compression induced visual distortion such as blocking effect, blurring, colour bleeding, ringing, staircase effect, block motion prediction error induced artefacts, false edges and other temporal artefacts. A detailed description of compression induced distortion can be found in chapter 3 of [42].

3.3.1 Blocking effect

Blocking effect or blockiness is the most prominent visual distortion in video compressed using block-based compression algorithms. It refers to the block pattern in compressed video characterised by discontinuities between adjacent blocks in a video frame. It is due to coarse quantisation of the DCT coefficients in individual blocks. The visibility of blockiness depends on the content of the blocks as well as the masking effects of the HVS. The blocking effect is usually prominent in smoothly textured regions and around moving objects as a result of poor motion compensation and block mismatch.

3.3.2 Blurring

Blurring in compressed video images occurs due to the loss of detail from moderate to high spatial activity regions such as roughly textured areas and around scene object edges. In intra-frame coded macroblocks, blurring is related to coarse quantisation of higher order AC DCT coefficients. In inter-frame coded macroblocks, blurring is mainly a consequence of the quantisation process and prediction from previously coded macroblocks which lack spatial detail.

3.3.3 Colour bleeding

Luminance and chrominance video data are processed separately in block-based compression algorithms. The loss of detail in luminance information results in blurring. The corresponding effect in chrominance information results in smearing of colours between areas of strongly contrasting chrominance due to coarse quantisation of higher order AC coefficients in the chrominance blocks.

3.3.4 Ringing

Ringing effects occur along high contrast edges in areas of generally smooth texture. It appears as a wave like transition or rippling moving outwards from the edges. The higher the contrast of the edge, the level of peaks and troughs of the rippling will be greater.

3.3.5 Staircase effect

The Staircase effect appears in diagonal edges that are represented within a string of blocks. Coarse quantisation of these blocks leads to discontinuities around block boundaries. Figure 3-2 shows the staircase effect along the edge of the building in the image.

3.3.6 Block MC mismatch induced artefacts

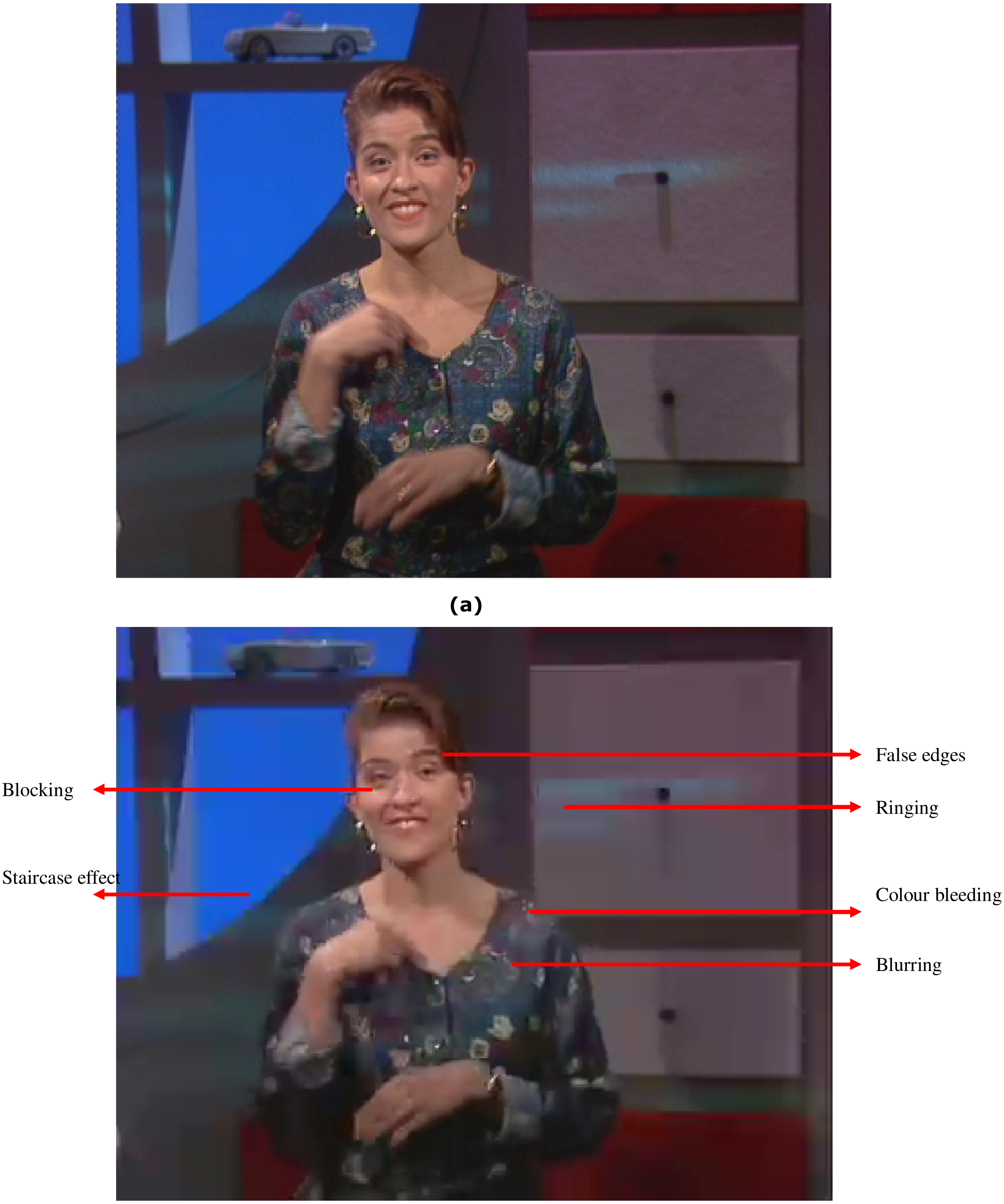
Block motion compensated mismatch occurs when objects overlap in macroblocks. Subsequent motion compensation will be unable to find a satisfactory match for the overlapping objects resulting in high prediction errors. These errors are ineffectively coded due to quantisation leading to higher visibility of the mismatch.

3.3.7 False edges

False edges mainly occur in inter-frame coded macroblocks which have been predicted using macroblocks which contain blocking artefacts. These artefacts are more prominent in smooth areas and object boundaries.

Figure 3-2 illustrates various compression induced distortions using the original and compressed video frames of the "Sign Irene" sequence. These include: false edges around the eyebrow area, ringing (or rippling effect) at object boundaries in the background, colour bleeding or smearing between the maroon and turquoise colours on the shirt, blurring of detail on the shirt area, staircase effect on the edge of the window and prominent block effect in the facial region.

Other temporal artefacts include: jerkiness and temporal fluctuations in stationary areas resulting in flickering effect caused due to quantisation of prediction errors. Suppression of the compression induced visual distortions is a priority to video compression algorithms. Hence video quality assessment techniques have been extensively researched to develop and evaluate new techniques which can help identify and manage visual distortions. The following sections in this chapter will focus on the various state-of-the-art approaches to video quality measurement, their applications and limitations.



(a)

(b)

Figure 3-2: Compression artefacts. (a) Original frame (b) Compressed frame

3.4 Classification of video quality measurement techniques

There are two approaches to measuring video quality: subjective assessment and objective measurement. Subjective assessment involves utilising human observers to assess video quality and express their opinion on a specific rating scale. The average quality of the degraded video is the mean opinion score (MOS). Subjective assessment is an accurate way of measuring perceived quality. However, it is expensive in terms of time and complexity, and cannot be easily implemented in real-time video applications.

Hence objective measurement techniques have been developed to predict subjective quality without human input. These techniques are automatic and are based on the physical aspects of the video signal and characteristics of the HVS. The performance of an objective quality measure depends on how closely it correlates with subjective results. Although existing objective measures do not completely reproduce the subjective assessment result, they are widely employed in video communication systems due to their repeatability, speed and simplicity.

3.4.1 Subjective quality measurement

Subjective measurement involves evaluating, comparing and assessing the picture quality of a video sequence under test using human observers. The outcome of subjective quality tests depend on many factors such as: selection of test material, selection of participants, experimental setup and following standardised testing methods.

Selection of test material is an important factor during subjective evaluation and is application specific depending on the video communication system under test. For example, if video sequences are being tested for a multimedia video communication system then the test video bit rate, frame rate and resolution should be within the range suitable for multimedia applications. Apart from video specifications, video content also plays an important role in the outcome of the subjective testing process and is dependent on the application. For testing videoconferencing systems, the test material would have sequences with “head and shoulder” shots and little motion. Similarly test material for assessing surveillance video systems will contain both indoor and outdoor video clips with changing backgrounds and large motion. The duration of each test video clip is also important. Since a number of test sequences are evaluated in a single test, the duration of each test sequence must be long enough to rate overall quality and short enough to keep the time taken to complete the test within the limited time specified in the ITU-T Recommendation P.910 [6] for multimedia applications. Generally, the duration of video clips for subjective quality assessments is around 10 seconds.

Participants for a subjective test could either be expert or non-expert. Experts in video communications have experience in designing and evaluating video communication systems. The advantage of using experts is the test process is quicker, they know what they are looking for and their feedback may be valuable in improving the video communication system under test. The disadvantage of using experts is the results may not be representative of the average consumer. Non-experts, on the other hand, represent the general public or the average consumer with no pre-determined way of looking at a video sequence. The ideal number of participants for a subjective test depends on the standard deviation of the subjective ratings for each video sequence and the 95% confidence interval due to the fact that the video has been rated by a limited number of the population.

Experimental setup includes environmental factors that need to be taken into account while conducting a subjective test. These include: the number of test sequences, duration of the test, the video display device and the test room conditions such as ambient noise and lighting. Standard environmental setup parameters have been defined in ITU-R Recommendation BT.500-11 [5] for subjective assessment of digital television pictures and in ITU-T Recommendation P.910 [6] for multimedia applications.

There are several methodologies for conducting subjective assessment. Standardised methodologies [5,6] that are internationally accepted include:
A: Single stimulus continuous quality evaluation (SSCQE)
B: Double stimulus continuous quality scale (DSCQS)
C: Double stimulus impairment scale (DSIS)
D: Pair comparison (PC)

Single stimulus continuous quality evaluation (SSCQE) method involves assessing the picture quality of video sequences independent of one other. The viewers rate each video sequence using a five grade rating scale: Excellent (=100), Good (=75), Fair (=50), Poor (=25) and Bad (=0). SSCQE is a continuous evaluation method where video sequences are presented to the viewers one after the other as shown in Figure 3-3. The reference video is not shown. This method is used to assess quality of video sequences that are scene dependent and time-varying. The analysis is based on calculating the mean opinion score for the ratings of each test video sequence.

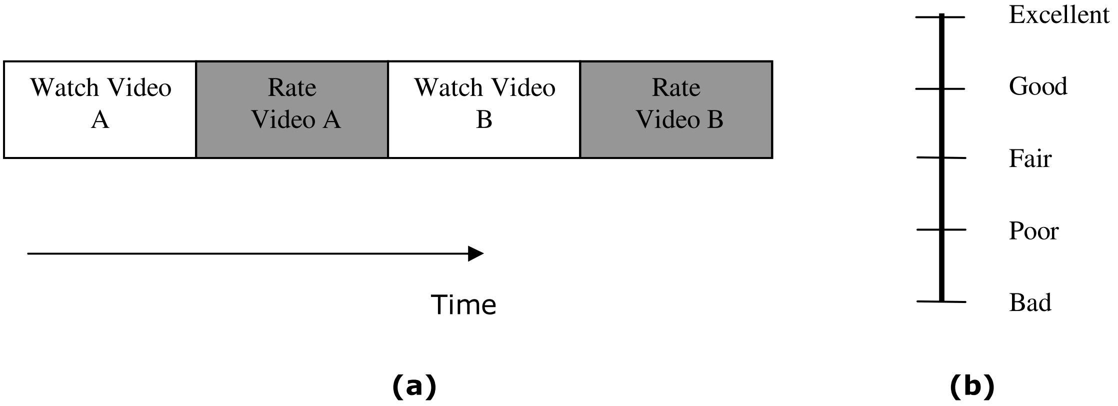


Figure 3-3: The SSCQE method: (a) Order of presenting and rating video. (b) Five-grade rating scale.

In the Double stimulus continuous quality scale (DSCQS) method viewers are presented with the reference video and the video under test twice in an alternating fashion. The order of the two sequences is displayed randomly and picture quality is rated using a five-grade rating for each sequence separately as shown in Figure 3-4. The analysis is based on the difference in rating for each pair. This method is preferred when the differences in picture quality of the reference and degraded video are small.

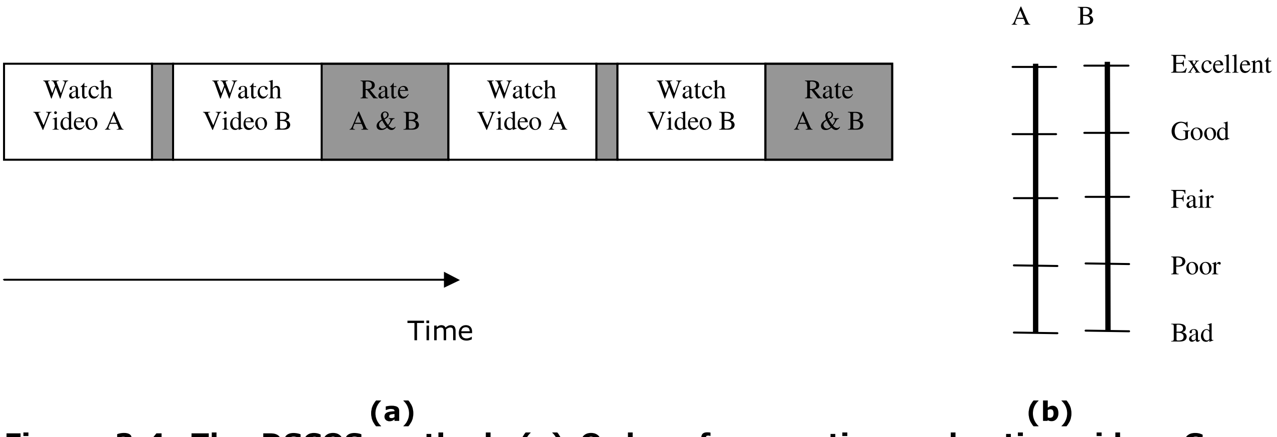
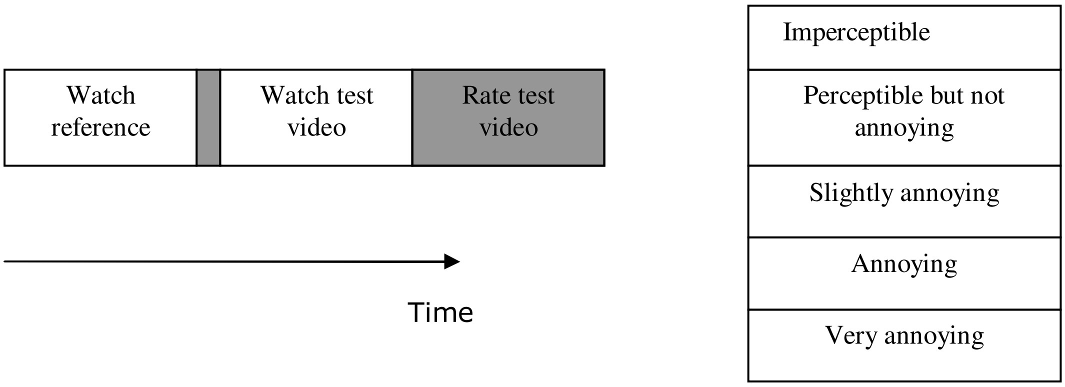


Figure 3-4: The DSCQS method: (a) Order of presenting and rating video. Grey area represents delay in viewing (b) Rating scale for videos A and B.

Double stimulus impairment scale (DSIS) involves presenting the reference video once followed by the test video sequence. Viewers rate the amount of impairment in the test video sequence in comparison with the reference video using a five grade rating scale as shown in Figure 3-5.



(a) **(b)**

Figure 3-5: The DSIS method: (a) Order of presenting and rating video. Gray area represents delay in viewing (b) Rating scale for test video.

Pair comparison (PC) involves displaying the two video sequences under test at the same time on the same screen to make a preference judgement based on picture quality. This method is useful for when there is very fine discrimination between the two video clips.

The analysis of the above mentioned subjective test ratings is performed by averaging the ratings from all observers for each test sequence into a mean opinion score (MOS) to represent the subjective quality of the corresponding test video.

<u>Limitations of subjective assessment methods:</u>

- The subjective results can vary significantly depending on the assessor and also on the video sequence under test.
- Repetitions of sequences may lead to the viewers becoming familiar with the degradations and materials under test.
- Longer sequences (over 30 seconds in duration) are more representative of the actual video broadcasting. Implementing the subjective assessment methods using longer sequences would be difficult and time consuming.
- Subjective assessment gives an overall rating of the video sequence. Hence it is difficult to pinpoint severity of the impact of individual degradations.

- Subjective assessment results may not be suitable for long sequences due to the recency effect which means that the judgement of the overall video quality may heavily depend on the last 5-10 seconds of the video sequence [43].

Subjective assessment is used to measure perceived video quality. However, the above mentioned limitations make it expensive in terms of time and resources, and not suitable for real-time video applications. Hence, objective quality measures have been developed to predict the subjective results automatically based on the video content and by modelling the characteristics of the human visual system.

3.4.2 Objective quality measurement

There are three approaches to objective measurement of perceived video quality depending on the availability of a reference video: full-reference, reduced reference and no reference. An overview of the approaches is provided in the remainder of this section.

Full-reference video quality measurement makes an assessment of the quality of the degraded video sequence by making a comparison with the reference video sequence as shown in Figure 3-6. This approach provides the highest quality measurement accuracy amongst the objective measurement approaches because it has access to the reference data. Full reference quality measures are typically employed in designing and benchmarking new video communication algorithms where the availability of the reference video and computational complexity is not an issue. Several full reference models have been proposed in the literature. An overview of existing full reference models is given in section 3.5 of this chapter.

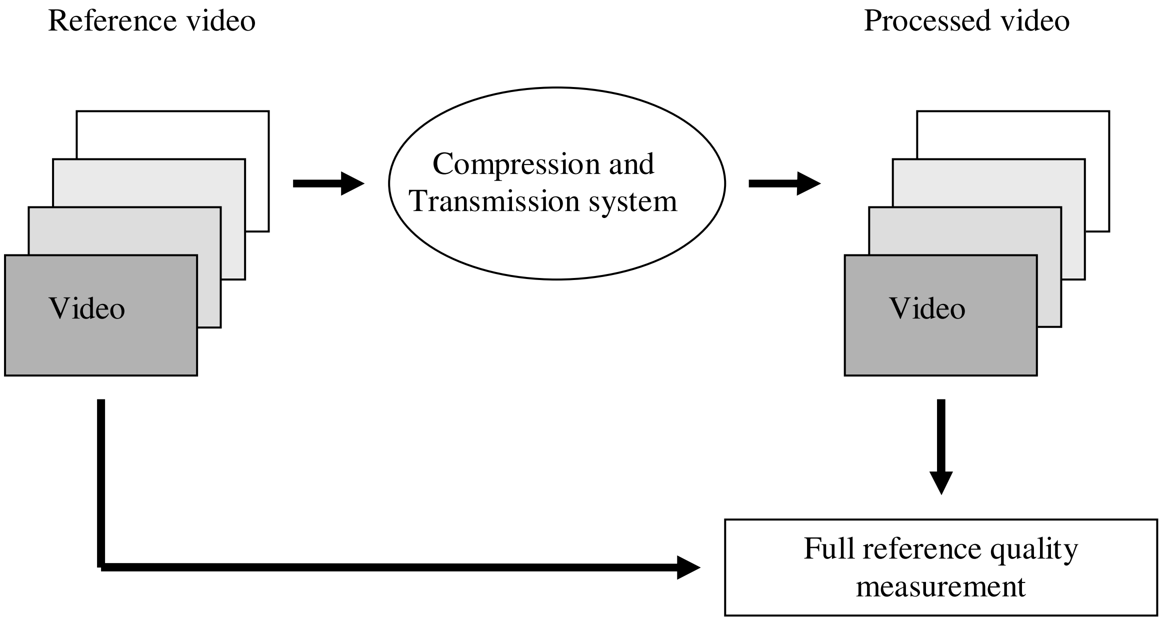

Reference video
Processed video
Compression and Transmission system
Video
Video
Full reference quality measurement

Figure 3-6: Full reference video quality measurement

In **reduced reference video quality measurement**, specific features are extracted from both the reference and processed video sequences as shown in Figure 3-7. These features could include spatial, temporal, blockiness and blurriness information. Features extracted from the reference video are transmitted to the receiving system through a side channel for quality estimation of the processed video sequence. Reduced reference approach is not as computationally expensive as the full reference approach. However, it requires a side-channel for transmission of reference feature information and this channel must be error-free.

Several reduced reference techniques have been proposed in the literature. Wolf and Pinson [44] have developed a reduced reference model for in-service quality measurement of standard television video sequences. The model uses low-level features extracted from the spatio-temporal regions of the reference and degraded video sequences with region size based on the side-channel bandwidth and the accuracy requirement of the system under test. Spatial and temporal features based on the visibility and masking of artefacts are obtained from both the degraded and reference video sequences and processed through comparison functions in order to obtain an overall quality measure of the degraded video sequence.

In [45] a reduced reference quality assessment model for standard definition compressed video sequences is proposed. The model measures local harmonic gain/loss feature which is derived from image spatial gradients. The harmonic gain/loss feature is used to identify blockiness and blurriness in a video frame. If the gain/loss measure indicates energy gain then it represents blockiness and conversely, energy loss would indicate blurriness in the video frame. A motion correction factor has been incorporated into the harmonic gain/loss feature to deal with temporal changes in the video sequence.

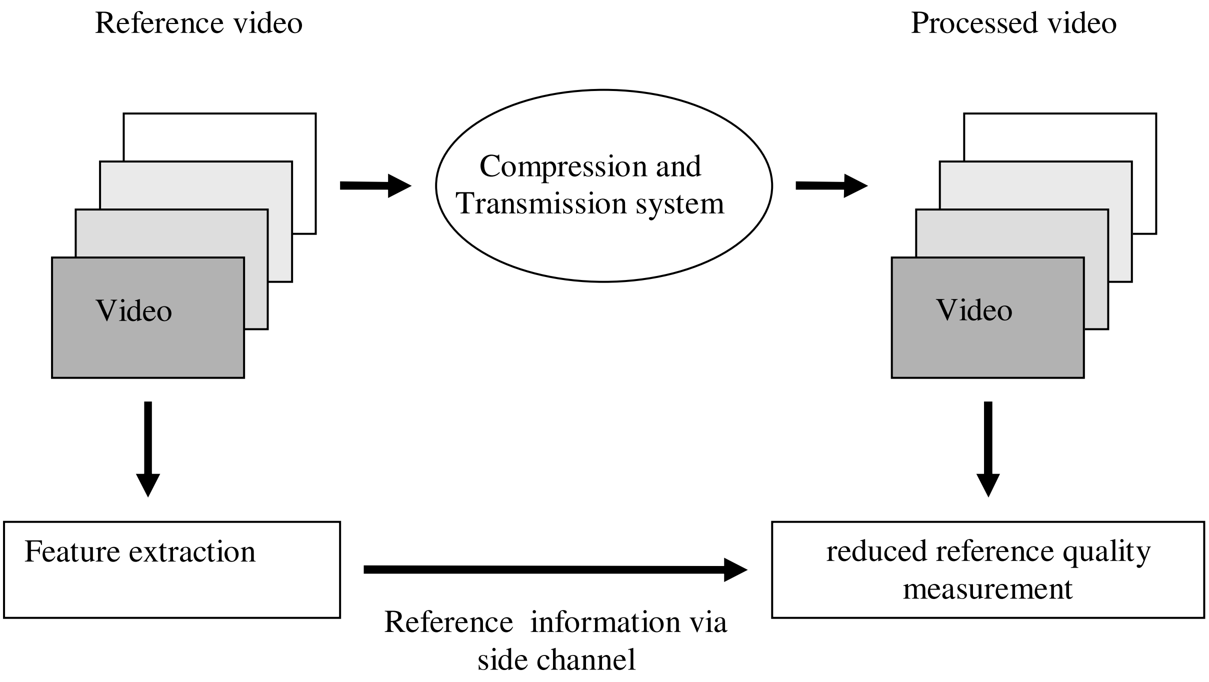

Reference video
Processed video
Compression and Transmission system
Video
Video
Feature extraction
reduced reference quality measurement
Reference information via side channel

Figure 3-7: Reduced reference video quality measurement

No-reference video quality measurement methods are employed in scenarios where access to the reference video sequence is not possible. The quality measurement is made based only on the analysis of content of the degraded video sequence as shown in Figure 3-8. The lack of reference video means that this measurement technique has prediction accuracy lower than the full-reference and reduced-reference approaches. No-reference methods are based on models of visual distortions built using training data sets. These methods are popularly used to measure the impact of transmission errors on video quality.

In [46], a no-reference metric for measuring blockiness in reconstructed video sequences caused due to packet loss is described. The metric is based on measuring the activity around block edges and counting the number of blocks that contribute to the overall perception of blockiness in the video image. Standard deviation and gradients are computed for each block to identify blockiness. Counting the number of blocks with blockiness artefacts enables the determination of the extent of packet loss per video frame.

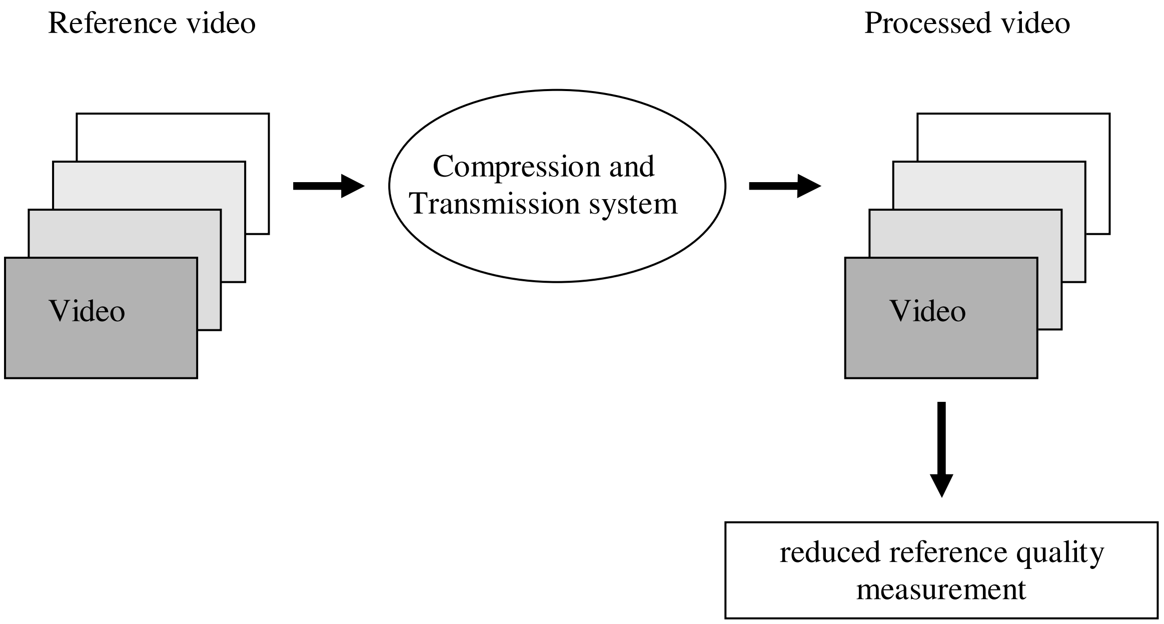

Reference video
Processed video
Compression and Transmission system
Video
Video
reduced reference quality measurement

Figure 3-8: No reference video quality measurement

3.5 Full reference objective quality measurement techniques – a review

Video compression algorithms employ full reference measures such as MSE and PSNR to make optimum compression decisions. Several approaches to full reference quality measurement have been proposed in the literature. In this section, an overview of popular full reference video quality measures such as: pixel-based measures, HVS-based models, standardised full reference models and visual masking based models are presented.

3.5.1 Pixel-based quality measures

Pixel-based measures are based on a pixel-by-pixel comparison of two video sequences. Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR) between the reference and distorted video data are simplistic but widely used pixel-based difference measures in video compression algorithms [32]. MSE is the mean of the squared differences between the samples of the reference video sequence (I) and degraded video sequence (I_c) with picture size MxN and T frames per sequence as follows:

$$MSE = \frac{1}{M*N*T}\sum_{t=1}^{T}\sum_{y=1}^{N}\sum_{x=1}^{M}\left[I(x,y,t) - I_C(x,y,t)\right]^2 \qquad (13)$$

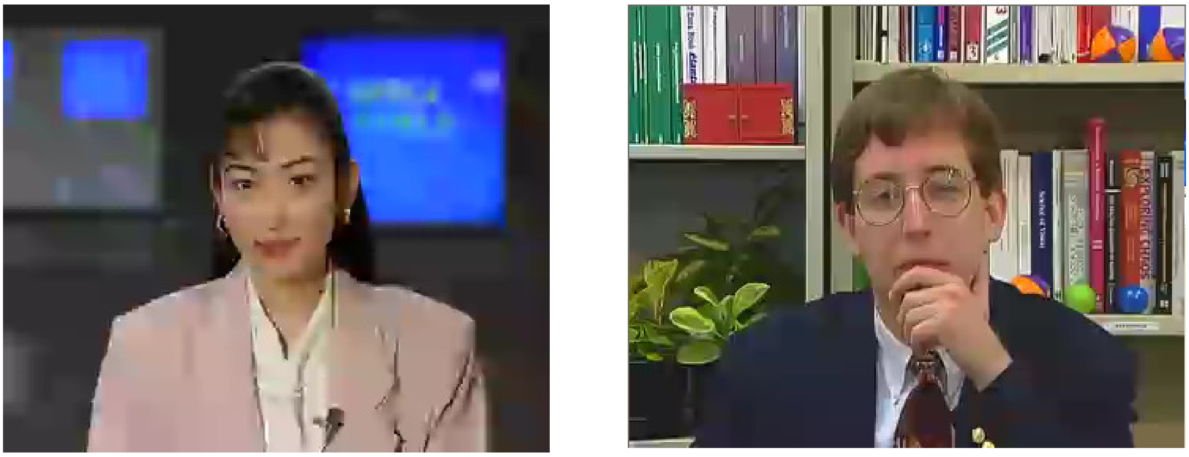
While MSE measures the mean difference between two video sequences, PSNR gives a measure of fidelity i.e., how closely a video sequence resembles the reference video. PSNR is measured in decibels on a logarithmic scale using MSE between the reference and degraded video sequences and the square of the highest possible sample value in video image (i.e., 255 for an 8-bit image) [13]. PSNR is calculated for an n-bit image as:

$$PSNR = 10\log_{10}\frac{(2^n - 1)^2}{MSE} \qquad (14)$$

The popularity of the two metrics is due to the fact that minimizing MSE and/or maximising PSNR is well understood from a mathematical point of view. Besides, the computational time for the calculation of MSE and PSNR is very small and their implementation is relatively simpler.

Pixel-based measures such as MSE and PSNR make a comparison on a pixel-by-pixel basis. Hence they may not have good correlation with the distortion perceived by the human observer and therefore are not accurate measures of perceived quality for compressed video sequences [47,48,49].

Visibility of distortions depends on factors such as video content, task in hand and viewer interest. Therefore, perceived quality of two video images with very similar MSE or PSNR may be very different. This is illustrated in Figure 3-9 which shows video frames from two sequences: (a)Akiyo, with average frame MSE = 74.46 and (b)Deadline with average frame MSE = 74.61. It can be seen that although the average MSE of both video frames are similar, the visual quality of video frame (b) may be ranked as better than video frame (a). The detail in the background of frame (b) and the facial features are better than frame (a).



(a) **(b)**

Figure 3-9: Video frames of two test sequences (a) Akiyo, average frame MSE = 74.46 and (b) Deadline, average frame MSE = 74.61

Hence the above example demonstrates that MSE does not necessarily correlate with perceived quality. Therefore, several full reference objective video quality measures have been proposed and analysed as alternatives to MSE and PSNR. These measures focus on modelling the known psycho-visual properties of the human visual system. To date, attempts to use these objective metrics to measure real-time video quality have been limited by their accuracy and computational complexity.

3.5.2 Human visual system based models

To overcome the limitations of pixel-based quality measures, several human visual system based models have been proposed. HVS-based metrics consider the distorted signal to be the sum of the reference signal and an error signal. A quality assessment of the degraded signal is made by evaluating the visibility of the error signal based on the physiological and psychophysical characteristics of the human visual system. The general framework of HVS-based metrics [50] is given in Figure 3-10.

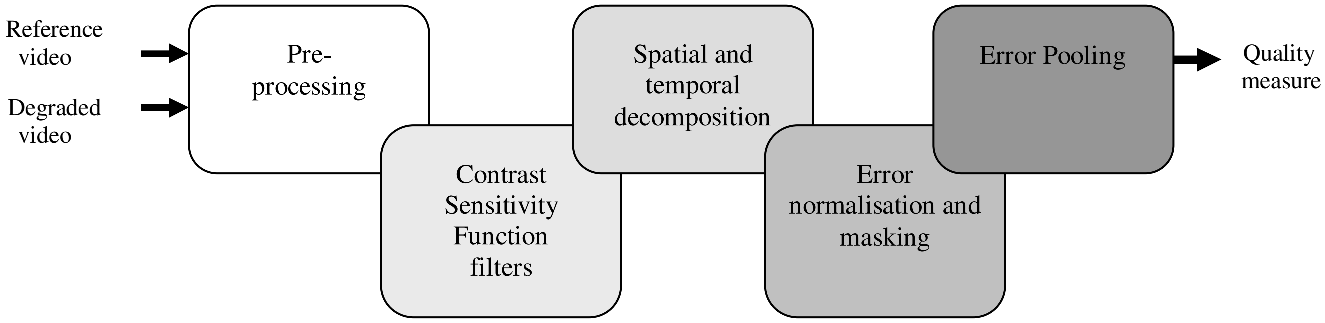

Reference video
Degraded video
Pre-processing
Contrast Sensitivity Function filters
Spatial and temporal decomposition
Error normalisation and masking
Error Pooling
Quality measure

Figure 3-10: General Framework for HVS-based video quality metrics

Pre-processing stage:

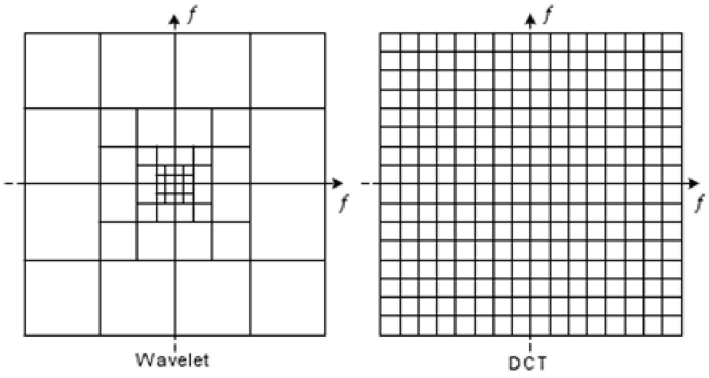
Both the reference and degraded video sequences undergo pre-processing operations which may involve [51,52]: spatial and temporal alignment to align samples between the two video sequences, colour transformation to a colour space that conforms better with the human visual system (such as CIE L*a*b*), low pass filtering to simulate the point spread function of the eye and light adaptation to exploit the non-linear perception of luminance by the human visual system.

Contrast sensitivity function filtering:

An important characteristic of the HVS concerns the decreasing sensitivity to higher spatial and temporal frequencies. This phenomenon is parameterised by the contrast sensitivity function. Linear filters are generally used to approximate the spatial frequency response of the HVS while infinite impulse response (IIR) filters are used to model the temporal frequency responses [48,51,53]. Each colour channel resulting from the colour space conversion is separately processed using CSF filtering.

Spatial and temporal decomposition:

Spatial and temporal decomposition involves separating the various colour channels into different spatial and temporal frequencies (sub-bands). This may be accomplished using several methods such as: block-based discrete cosine transform (DCT) [50], Gaussian/Laplacian pyramids [51] to perform multi-scale band-pass decomposition and separable wavelet transforms [54] to decompose the signals into logarithmically spaced frequency bands. The wavelet and DCT decomposition models are shown in Figure 3-11.



(a) **(b)**

Figure 3-11: Examples of frequency decomposition models. (a)Wavelet [54**] (b)Block-based DCT [**8**]**

Error normalisation and masking:

The error signal is calculated as the difference between the decomposed channels of the reference and degraded video sequences and is calculated separately for each colour channel. The visibility of the error signal is determined by weighting using visibility thresholds which are calculated based on various masking factors such as contrast masking [51] and spatio-temporal masking mechanisms [52]. Masking is a phenomenon of the human visual system which affects the visibility of certain features in a video scene due to the presence of other features.

Error pooling mechanisms:

Error pooling involves combining error signals from various channels into a single value measure of quality of the degraded video sequence when compared with the reference video sequence. The most commonly used error pooling technique is the Minkowski error metric [55] calculated as shown:

$$Q = \left(\sum_{x,y,l,t,n} \left| I_{reference}(x,y,l,t,n) - I_{\deg raded}(x,y,l,t,n) \right|^{\beta} \right)^{1/\beta} \qquad (15)$$

Where $I_{reference}$ and $I_{\deg raded}$ are the multi-channel decompositions of the reference and degraded video sequences in terms of spatial locations x and y, scale *l*, temporal channel t and frames n. β is the Minkowski exponent with value dependent on the number of dimensions across which the quality measurement is made.

3.5.3 Standardised metrics

In 2000 and 2003, the video quality experts group (VQEG) conducted independent tests to evaluate the performance of several full reference objective video quality metrics in the context of digital television broadcasting. Two large subjective tests were setup to compare the performance of these algorithms. These include the phase I and phase II tests on full reference television (FR-TV) video sequences [56,57]. Based on these studies, the International Telecommunication Union (ITU) has standardised the recommendations ITU-T J.144 [58] and ITU-R BT.1683 [59] for estimating the perceptual video quality in digital television video sequences when the original video sequence is available (full reference models). These include National telecommunication and information administration's video quality metric (NTIA/ITS VQM) [60].

The NTIA/ITS VQM [60] algorithm uses five video quality models to extract different parameters from both the original and compressed video sequences optimised for specific applications based on resolution, frame rate and bit rate information. The five models are: (i) Television model – optimised for television video, (ii) Videoconferencing model – optimised for low bit rate, low resolution multimedia sequences, (iii) General model – optimised for a wide range of bit rates and resolutions, (iv) Developer model – optimised general model with added constraints for fast computation and (v) PSNR-based model – optimised for a wide range of bit rates and resolutions involving the use of a logistic function to estimate quality based on PSNR as given below:

$$VQM = \frac{1}{1+e^{(0.1701*(PSNR-25.6675))}} \text{ , } 10 \leq PSNR \leq 55 \qquad (16)$$

The general block diagram of NTIA/VQM is given in Figure 3-12. The metric divides both the reference and processed video sequences into spatio-temporal regions, i.e. regions of pixels that are spatially and temporally adjacent to each other. Various image based features such as spatial gradients, chrominance, temporal and contrast information are extracted from these spatio-temporal regions before computing visual differences between the two video sequences using comparison functions which model the visual masking of spatio-temporal impairments. The parameters are then integrated into a quality measure of the degraded video sequence.

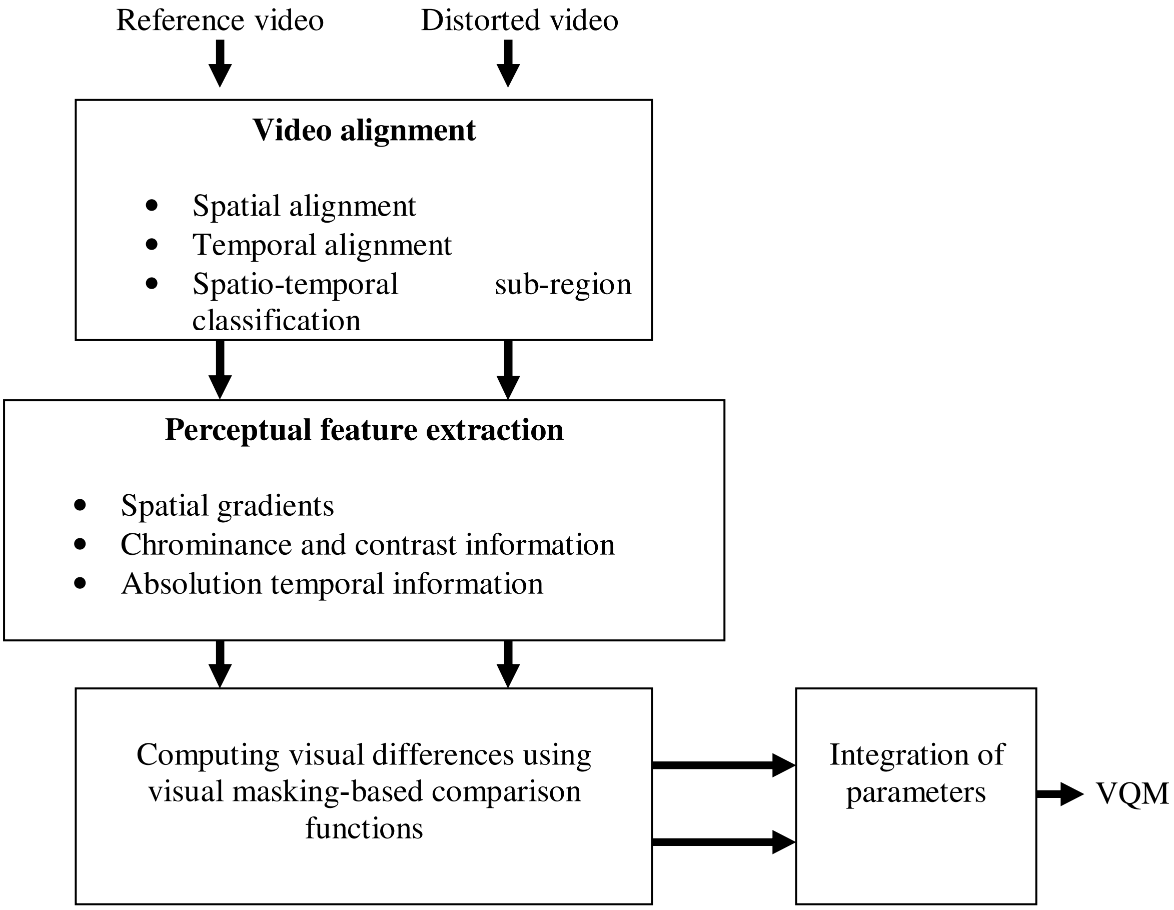

Reference video
Distorted video
Video alignment
Spatial alignment
Temporal alignment
Spatio-temporal sub-region classification
Perceptual feature extraction
Spatial gradients
Chrominance and contrast information
Absolution temporal information
Computing visual differences using visual masking-based comparison functions
Integration of parameters
VQM

Figure 3-12: Block diagram of the NTIA/VQM model

Multimedia video differs from digital TV video in terms of resolution and bandwidth requirements. This resolution typically ranges from Quarter Common Intermediate Format (QCIF) with 176x144 pixels to VGA resolution (640x480 pixels). In 2008, the VQEG conducted a large number of subjective experiments to benchmark the performance of several full reference objective video quality measurement techniques for multimedia scenarios [7]. Based on this VQEG work, four algorithms were standardised in the Recommendation ITU-T J.247 [8]. These algorithms include:

- OPTICOM's Video Quality Measure PEVQ
- NTT's full reference model
- Psytechnics full-reference video quality assessment algorithm
- Yonsei full reference method [61]

The general framework for these four methods is presented in Figure 3-13. The reference and degraded video sequences are aligned spatially and temporally taking into account encoding factors such as frame skip, frame freeze and frame rate. The OPTICOM model also incorporates a pre-processing stage where video frame borders are cropped to take advantage of the fact that distortions at image borders tend to be ignored by viewers. Next, each of the four metrics calculates a different set of visual distortion parameters based on spatial, temporal, luminance, contrast, chrominance and temporal masking properties of the human visual system. Distortions introduced by compression such as blocking, blurring and edge degradation are also taken into account. Finally, these parameters are integrated into a single value measure of estimated subjective quality.

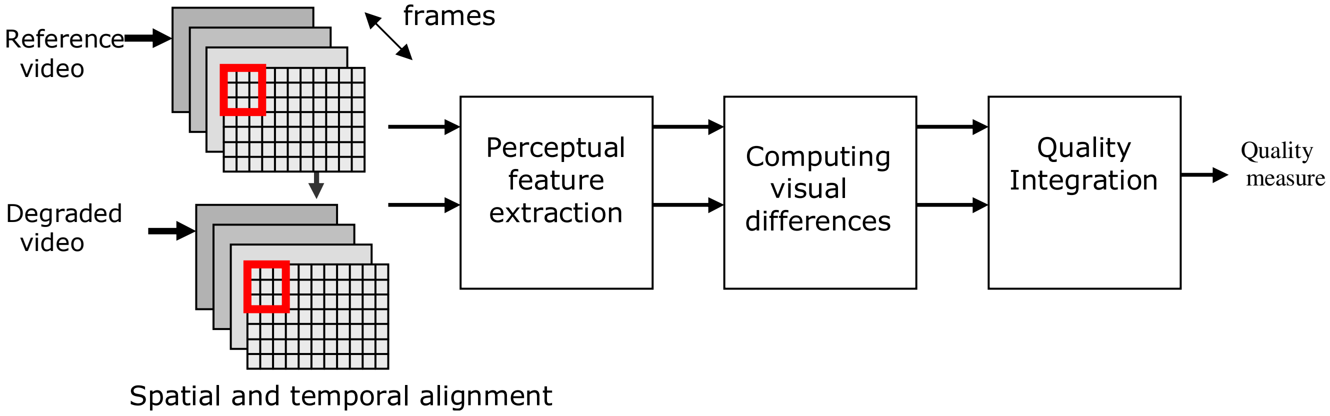

Reference video
frames
Degraded video
Perceptual feature extraction
Computing visual differences
Quality Integration
Quality measure
Spatial and temporal alignment

Figure 3-13: General framework of the video quality models from NTT, OPTICOM, Psytechnics and Yonsei University

The Yonsei University metric [61] measures quality of video based on the degradation in spatial edge areas. The authors found that viewers gave lower quality ratings to video clips with noticeably degraded edge areas despite a relatively low overall mean squared error. Edge detection and thresholding are used to locate edge areas in both the original and degraded video sequences. Degradation in the edge areas is calculated by measuring the PSNR between the edge areas of the original and degraded video clips. Post-adjustments were performed to obtain an estimation of the quality of the degraded video clip.

The correlation between subjective (MOS) and estimated quality for the four metrics as reported in Recommendation J.247 range from 77% to 84% indicating that there is still scope for developing better full Reference objective quality metrics.

3.5.4 Metrics based on masking effects

Masking is an important visual phenomenon which describes why similar artefacts are more visible in certain regions of a video frame while they are hardly noticeable in other regions. Several factors influence the visibility of distortions in video sequences and these include: (a) Spatial texture masking – ability of textured regions to hide more distortions than smoother regions [62,63,64,65,66] (b) Luminance masking – the human visual system is more sensitive to higher luminance contrast than absolute luminance value [50,67,68] (c) Temporal masking – ability of regions undergoing large temporal changes to hide visible distortions [44,49] and (d) Cognition-based factors such as skin colour information – distortions in regions that are important to the viewer (such as human faces) are more visible than similar distortions occurring in other regions [69].

The VSSIM metric [62] gives a measure of similarity between the reference and processed video sequences based on luminance, contrast and spatial texture masking characteristics of the human visual system. The luminance, contrast and structural components between the two sequences are measured and subjected to comparison functions at block-, frame- and sequence- levels before being pooled into an overall similarity measure. The metric has been demonstrated in [62] to perform better than the metrics reported in the VQEG phase I test on full reference television (FR-TV) video [56].

In [48], a perceptual quality metric for estimating perceived quality of compressed multimedia sequences taking into account compression induced artefacts such as blocking effects and HVS-based characteristics such as contrast, spatial texture, colour and temporal masking effects is presented. The metric measures quality using three frame-based parameters: distortion invisibility measure (D), block fidelity measure ($^{\mathrm{F}}\mathrm{BF}$) and content richness fidelity measure (F_{RF}) as shown below:

$$\lambda_{motion} = D * F_{BF} * F_{RF} \qquad (17)$$

The distortion invisibility measure (D) is based on spatial texture, colour and temporal masking effects of the HVS. The block fidelity measure F_{BF} estimates the distortion at block boundaries and is used to identify blocking artefacts. The content richness fidelity measure F_{BF} calculates the colourfulness and contrast of the video scene to exploit the sensitivity of the HVS to brighter colour tones and increased contrast. The three parameters are measured at frame level and integrated across the video sequence to produce a single value measure of perceived quality. Results presented in [48] have shown that this metric produces 91.6% correlation with subjective test results. However, the computational cost for the quality measurement is not discussed and deriving the various parameters based on various masking and image-based factors make the metric impractical for real-time applications.

In [70] a perceptual video quality metric based on three frame-based parameters: visual masking error, blurring distortion and contrast distortion is presented. The visual masking error is measured based on luminance and spatial texture masking. The blurring distortion parameter measures the amount of blurring in the degraded video frame and the contrast distortion measure attempts to measure the amount of structural distortion. The three parameters are measured at frame level and integrated using a simple linear combination method. The performance of the metric has been demonstrated in [56], on the VQEG test video sequences with digital TV resolution compressed at full TV frame rate and high bit rates.

A perceptual sensitivity weighting scheme has been proposed in [34] for bit allocation in a rate control algorithm for videophone applications. The method extracts perceptual features using spatial masking factors such as luminance adaptation and texture masking, and cognitive-based factors such as skin colour information. These features are used to develop a perceptual sensitivity weight map for each video frame to indicate regions that are sensitive to visual distortions. The perceptual weights are used to determine the quantisation parameter of a rate control algorithm in order to achieve improvements in perceived quality for a videophone application. Although the authors claim that the technique produced perceptual gains, the technique is not a generalised perceptual weighting scheme as it has been built for a specific application (i.e., rate control for videophone applications) with human faces.

3.5.5 Increasing the prediction accuracy of objective measures: PSNRplus [71]

Mean squared error (MSE) and peak signal to noise ratio (PSNR) is a popular full reference objective measure used in modern block-based video compression algorithms such as H264/AVC. It is employed by the Rate-Distortion Optimised (RDO) mode selection process as a quality measure for choosing the best compression option that gives an optimal trade-off between picture quality and data rate [32].

Whilst a common approach is to use MSE to choose the best coding option, MSE is a mathematical error measure which does not consider the human visual system and has been found to be an inaccurate measure of perceived quality [9,47]. It may be possible to improve the subjective quality performance of a rate-constrained video codec by replacing MSE with a distortion metric that correlates more closely with subjective quality in the mode selection process. Previous work has found that although the overall correlation between MSE and MOS is poor [4], there is a higher correlation between these parameters for a single sequence coded at several bit rates with the same codec [71]. This correlation decreases with increasing number of different video sequences added to the test data set.

Based on this hypothesis, the authors of [71] have developed a method (PSNRplus) for increasing the correlation between subjective and estimated video quality by estimating the parameters of the linear regression line for each video sequence as:

$$PSNRplus = (PSNR - o) / s \qquad (18)$$

The regression parameters: slope (s) and offset (o), are determined using two additional instances of the original video: PSNR at high quality ($PSNR_{high\_quality}$) and PSNR at low quality ($PSNR_{low\_quality}$). Figure 3-14 gives a visual representation of the method.

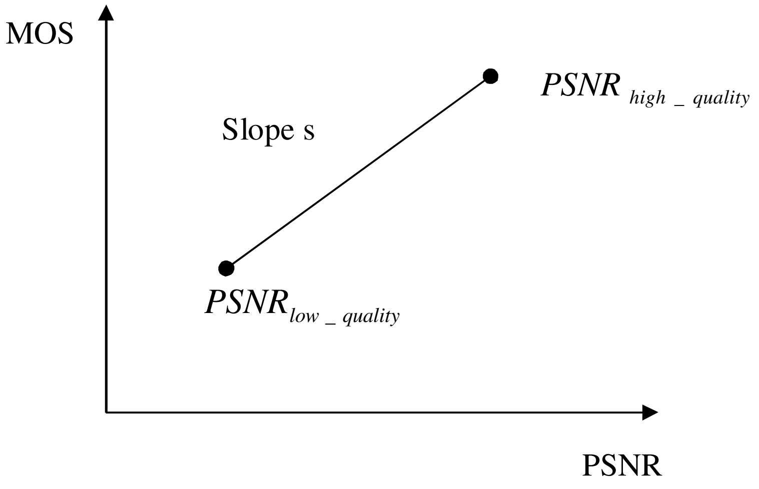


Figure 3-14: Increasing prediction accuracy of PSNR [71]

Although this method produces improved results compared to previous methods in the literature, it requires every sequence to be coded three times in order to obtain the two additional instances hence making this technique unsuitable for real time applications.

3.6 Metric standardisation

The video quality experts group is composed of experts in video quality assessment from Industry, Universities and other International organisations. The group was formed in 1997 to evaluate performance and develop recommendation for objective quality measurement systems using reliable subjective test results for a well-defined set of test material. The main responsibilities of the VQEG are to:

- select and solicit objective models to be included in the evaluation.
- select test material
- develop objective test plans for running selected objective models on the test video data
- develop subjective test plans for conducting subjective tests in accordance to the ITU-R BT.500 recommendations [5]
- conduct objective tests for evaluating the proposed models
- conduct subjective tests for acquiring subjective data
- analyse objective and subjective results using standard comparison metrics such as correlation.

- present findings of the evaluations to the International Telecommunications Union (ITU) for standardisation

In 2000 and 2003, the video quality experts group (VQEG) conducted independent tests to evaluate the performance of several full reference objective video quality metrics in the context of digital television broadcasting. Two large subjective tests were setup to compare the performance of these algorithms. These include the phase I and phase II tests on full reference television (FR-TV) video sequences [56,57]. Based on these studies, four models were recommended to the International Telecommunication Union (ITU). These include models from British Telecom (UK), Yonsei University (Korea), CPqD (Brazil) and NTIA/ITS (USA). On the basis of the VQEG evaluations, the International Telecommunication Union (ITU) has standardised the recommendations ITU-T J.144 [58] and ITU-R BT.1683 [59] for estimating the perceptual video quality in digital television video sequences when the original video sequence is available (full reference models).

In 2008, the VQEG conducted a large number of subjective experiments to benchmark the performance of several full reference objective video quality measurement techniques for multimedia scenarios [7]. Based on this VQEG work, four algorithms were standardised in the Recommendation ITU-T J.247 [8]. These algorithms include models from OPTICOM, NTT, Psytechnics and Yonsei University. The VQEG is currently conducting [72]: (a) Phase II test on full reference video quality metrics for multimedia application and (b) Evaluation of reduced-reference and no-reference metrics for digital television video sequences.

3.7 Metric performance evaluation

Subjective ratings (MOS) acquired from a panel of human observers is the benchmark for evaluating the performance of an objective video quality metric depending on how well it correlates with MOS. There are several methods to compare the performance of quality metrics [4]. Three of these performance evaluation methods have been adopted by the video quality experts group (VQEG) [7,56,57] for evaluation and benchmarking of objective quality metric. These include: Pearson correlation to measure prediction accuracy, Spearman correlation to measure predict monotonicity and outliers ratio to measure prediction consistency.

Prediction accuracy is the ability of an objective quality metric to predict subjective ratings with minimum average error [4]. It is determined using the Pearson's correlation coefficient between predicted results and subjective results. For a set of N data pairs (x_i, y_i), Pearson's correlation (r_P) is defined using means $\overline{x}$ and $\overline{y}$ as follows:

$$r_p = \frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)}\sqrt{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)}} \qquad (19)$$

This method makes a relative comparison between the two data sets assuming a linear relation between them. The Pearson correlation value (r_p) ranges between [0,1] where 1 indicates perfect match between predicted measures and the subjective ratings and 0 indicates no correlation.

Prediction monotonicity determines how well the estimated result reflects an increase or decrease in the actual subjective result regardless of the magnitude of increase or decrease [4]. Spearman rank-order correlation coefficient (r_s) is generally used to measure prediction monotonicity. For a set of N data pairs with ranks, the Spearman correlation (r_s) is defined using mid-ranks $\overline{X}_i$ and $\overline{Y}_i$ as follows:

$$r_s = \frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}} \qquad (20)$$

Spearman correlation value (r_s) also ranges between [0,1] where 1 indicates perfect match between predicted measures and the subjective ratings and 0 indicates no correlation. An advantage of the Spearman rank-order correlation is that it is non-parametric; hence it makes no assumptions about the shape of the relationship between the predicted data and the subjective ratings [4].

The outliers ratio (OR) is a measure of prediction consistency. A data point is considered to be an outlier if the difference between the predicted value and the actual subjective value exceeds ± 2 times the standard deviation of the subjective results. Outliers ratio is the ratio of the number of outlier (D_O) to the total number of data points (N) as shown below:

$$OR = {D_O}/{N} \tag{21}$$

A lower outliers ratio indicates better prediction consistency.

3.8 Discussion

Video quality measurement is imperative for comparing, evaluating and benchmarking video communication systems. Subjective assessment remains the most accurate method of measuring perceived quality of compressed video sequences. However, it is expensive in terms of time and resources and cannot be easily embedded into real-time applications. Hence several objective assessment methods have been developed to predict the subjective results based on video content and the characteristics of the human visual system. The video quality expert group (VQEG) have performed several evaluation tests to benchmark the performances of these quality metrics which have resulted in the standardisation of a few in the ITU-T Recommendations. These metrics have varying degrees of success in predicting subjective (human) test scores, with reported correlations of between 70% and 84% between each objective metric and measured subjective quality scores indicating that there is still scope for developing better approaches to estimate subjective quality. Although several techniques have been proposed in the literature as alternatives to pixel-based approaches, MSE and PSNR still retain their popularity in video processing algorithms due to implementation simplicity and computation speed. Previous research has shown that the correlation between subjective results (MOS) and pixel-based methods such as MSE and PSNR, is high for a single sequence coded to various bit rates. By exploiting this correlation between subjective results and pixel-based measures, it may be possible to accurately predict the subjective results from pixel-based metrics such as MSE and PSNR.

Part 2: Experiment Work

4 Experimental Methodology

4.1 Introduction

This chapter outlines the experimental methods used in this research project. Sections 4.2 and 4.3 give a description of the test material, test equipment and software used for conducting experiments. Subjective video quality evaluation methodology for obtaining mean opinion scores and comparing visual quality of different algorithms is given in section 4.4. Data analysis techniques used for data modelling purposes are mentioned in section 4.5 and performance testing procedures for evaluating the performance of developed algorithms is given in section 4.6.

4.2 Test Material

Video quality assessment depends on key factors such as the application and video content. Test video sequences used in the research work have been selected from video material which is widely used by the video coding community. The choice of test material used in this research work is based on the application, video format, resolution and video content.

This project focuses on developing a new video quality measurement technique and improving the perceptual quality of compressed multimedia video sequences. Multimedia video applications include video conferencing, internet video streaming, mobile video messaging and surveillance. A commonly used video format in these applications is the 4:2:0 format which requires only half the resolution of the chrominance samples when compared to the luminance samples. Multimedia video typically includes the Common Intermediate Format (CIF) with 352x288 pixels and has been popularly used in video quality evaluation tests for benchmarking performance of video quality metrics in multimedia scenarios [7]. Hence, multimedia sequences of 4:2:0 CIF format have been used in this research work. The test dataset includes (a) 16 popularly used multimedia video sequences in the video coding research community obtained from the Xiph.org test media website [73] and (b) The VQEG multimedia data set [72] which was used by the video quality experts group to benchmark multimedia video quality metrics [7].

The selected test sequences may be broadly classified as:

A: Video conferencing video

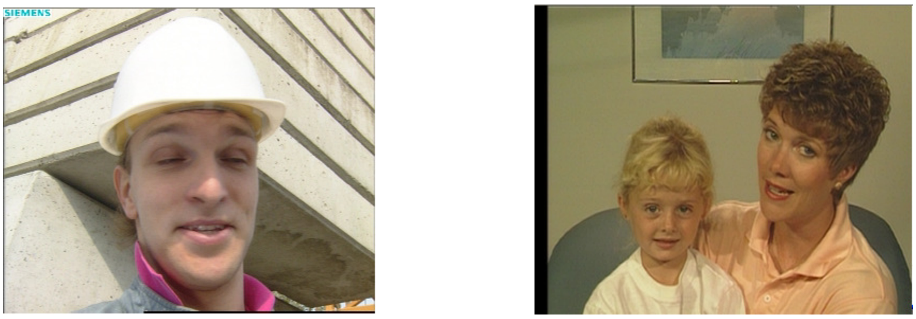
B: Broadcasting type video

C: Sign language video

D: Natural scenes

E: High speed vehicle tracking

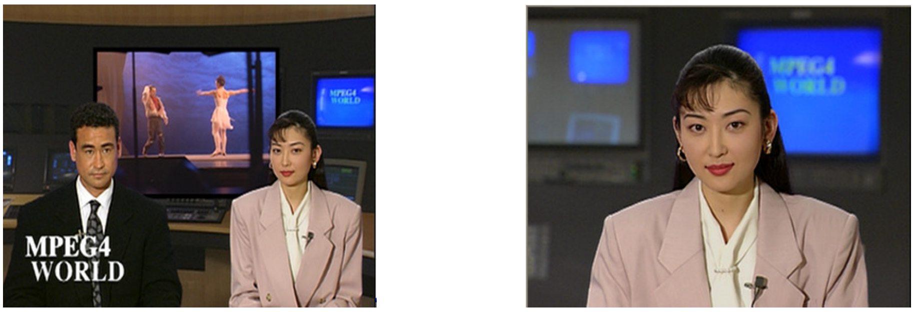
Video conferencing videos typically include 'head and shoulder' shots of person(s) speaking to the camera. The camera is usually static with fixed or changing backgrounds. The video usually contains medium level of spatial detail and motion with prominent facial and/or hand movements. Video conferencing type videos used in this project include: Foreman, Carphone, Mother and Daughter, Salesman and Deadline. Sample frames from Foreman and Mother and Daughter are given in Figure 4-1.



Foreman Mother and Daughter

Figure 4-1: Head and shoulder shots

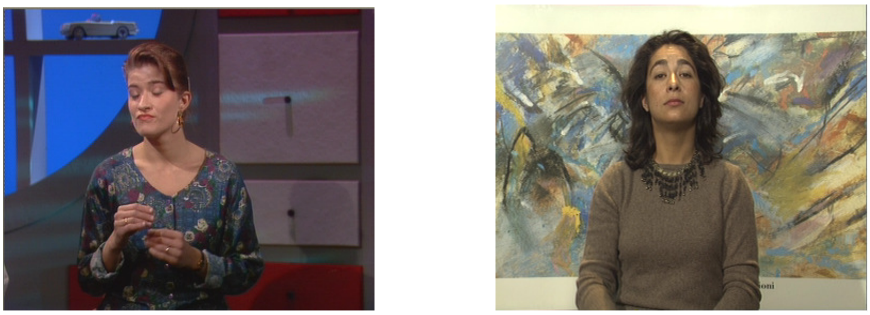
Examples of broadcasting type videos are news programs with one or two presenters reading the news. These video have frequent scene changes and the camera is usually static with fixed or changing background. There is medium detail with medium to high motion. Akiyo and News sequences shown in Figure 4-2 fall into this category.



News Akiyo

Figure 4-2: Broadcasting type video sequences

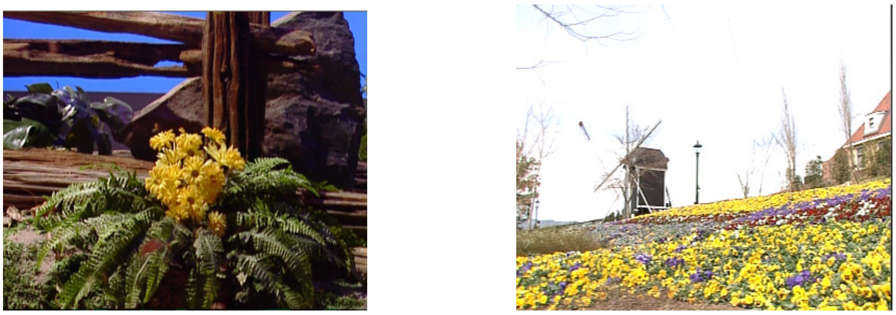
Sign language video clips such as Sign Irene and Silent (Figure 4-3) contain person(s) signing to the camera. The camera is usually stationary. Video clips falling into this category have prominent hand and facial movements.



Sign Irene Silent

Figure 4-3: Sign language video

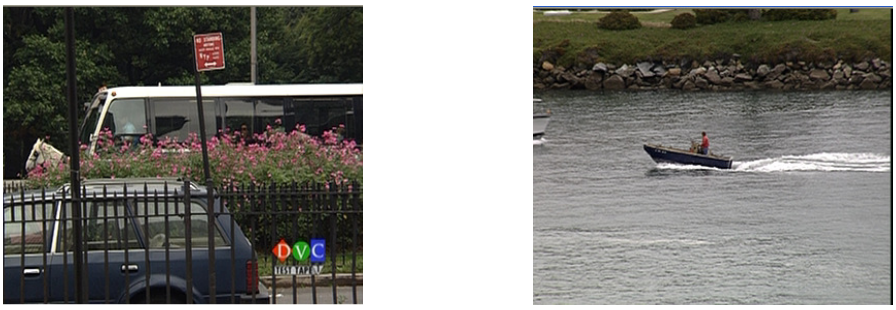
Natural video sequences contain outdoor scenes from nature such as flora, waterfall, trees, etc. These video clips are usually filmed using hand held cameras with changing or moving backgrounds. There is high detail and motion due to swaying of trees, leaves, grasses, etc. Natural sequences in the test material include Tempete and Flowers as shown in Figure 4-4.



Tempete Flowers

Figure 4-4: Natural scenes

Videos of fast travelling vehicles being tracked by the camera usually contain camera panning, zooming and translation. There is very high motion and spatial detail information with constantly changing foregrounds and backgrounds. Examples for high speed vehicle tracking include Container, Mobile, Coastguard and Bus sequences as shown in Figure 4-5.



Bus Coastguard

Figure 4-5: High speed vehicle tracking scenes

The duration of the video sequences used in the test material is around 10 seconds each and the playback frame rate was 25 fps.

4.3 Test Equipment and Software

Techniques and algorithms developed in the research work have been implemented and tested by software simulation using:

A: Testing platform

B: A software video codec called the JM software running on the personal computer

C: Programming software platforms

4.3.1 Testing platform

A computer with the following specifications is used as the test platform for developing and testing the algorithms, running the software video codec and conducting video quality evaluation tests:

Processor:	Intel Pentium M Processor 730, 1.6 GHz
Memory:	512 MB
Operating System:	Microsoft Windows XP Professional
Display Screen:	14.1″ XGA TFT LCD

4.3.2 Software video codec

A software video codec has been used for implementing and testing the algorithms developed and for producing the compressed video sequences. The H.264/AVC Reference software called the JM software (version 12.1) [74] is used as the reference video codec. The JM software is widely used in the video coding community for testing and implementing new algorithms. This software enables new algorithms to be compared and benchmarked with algorithms developed by other researchers. The software is free to download from [75]. The revised manual for the H.264/AVC reference software [76] gives a description of the usage of the reference software including software installation and compilation information.

The JM software contains the source code for the encoder and the decoder along with their configuration files. Figure 4-6 shows the input and output files for the JM encoder. These include the input video sequence, the configuration file, H.264 bit stream, the reconstructed video sequence, the output log file and the trace file (optional).

JM software supports popular formats of raw YCbCr video including the 4:2:0 format. The output file will have the same format as the input file. The configuration files for the encoder and decoder provide the input parameters to the encoder and decoder respectively. The encoder configuration file parameters include: input/output video sequence parameters such as file name, file size and number of frames to be encoded, and encoder control parameters such as profile type (baseline/main/extended), quantisation parameters for I, P and B slices, frame skip, number of reference frames and intra and inter prediction search options.

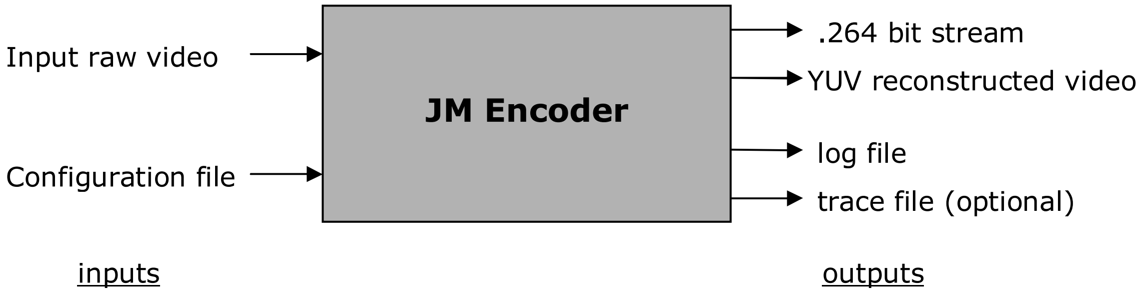


Figure 4-6: Input/output files of the JM encoder

The .264 bit stream is the encoded bit stream which is used for storage and transmission. The YUV reconstructed file is the decoded video sequence which has the same file size as the raw input video but with lower quality as it has been reconstructed after compression. The log file contains encoding statistics such as peak signal to noise ratio (PSNR) of luminance and chrominance components, encoding time and bitrate. The trace file contains the syntax elements used in the encoding process, their values (in decimal format) and number of bits used. The trace file is used for identifying and eliminating errors in the JM encoder and is often used during algorithm implementation for debugging purposes.

4.3.3 Programming software platforms

Software packages used in this research work include MATLAB (version 7.0) and Microsoft Visual C++ professional (version 6.0). MATLAB was used for off-line development, implementation and testing of new algorithms. Microsoft Visual C++ was used for reading, editing and compiling the JM codec. It was also used for modifying the JM software in order to incorporate the algorithms developed for testing and benchmarking purposes.

4.4 Subjective Video Quality Evaluation

Subjective quality measurement involves assessing the picture quality of video using a number of observers who rate the quality using a grading scale. The result of subjective video quality test is the mean opinion score (MOS) which is the average rating of each video sequence compressed to a certain level (fixed QP or bitrate). Subjective evaluations were conducted in this research work to: (a) determine the correlation between subjective and objective video quality measurement techniques, (b) compare the visual quality of compressed video sequences obtained using the reference codec and the algorithm under test.

4.4.1 Test methodology

The first step in subjective video quality evaluation is to design the test process. This involves choosing the appropriate:

A: subjective test method
B: grading scale for video quality rating
C: presentation of test sequences
D: environmental setup
E: test subjects

Choice of subjective test method:

The choice of test method depends on the application area and the quality level of the video sequences under test. The three main categories of subjective evaluation methods are double stimulus, single stimulus and pair comparison. A detailed description of these methods is given in Chapter 3, section 3.4.1. Double stimulus methods use an explicit reference, are thought to be less sensitive to contextual effects and are preferred when high quality video sequences are being evaluated [7,56,57]. In single stimulus methods, only the distorted video sequences are displayed. This method is appropriate if video sequences at comparably low bitrates are being evaluated because in showing the high quality reference, the distorted video sequences may be perceived as poor quality and no distinction between different levels of low quality may be made by the observers. The video sequences used in this research are multimedia sequences compressed to a wide range of bitrates from high quality (QP=6) to very low quality (QP=45). Hence, the single stimulus method has been used in the subjective video quality experiment for the estimation of mean opinion score of compressed sequences.

Presentation of video sequences:

In single stimulus subjective evaluation method, contextual effects occur when the subjective rating of a video sequence is influenced by the order of presentation and the nature of other video sequences in the same test session [77]. This effect is created when there are variations in the subjective rating of sequences based on the impairment present in the preceding video sequences. For example, a video sequence with moderate impairment that follows a set of sequences with weak impairment may be judged lower in quality than if it followed sequences with strong impairment. A common method used to try and counterbalance the contextual effect is the randomization of the test trial presentation order across the different viewers [78].

The presentation order of the video sequences in the single stimulus experiment was randomized between participants such that each participant viewed the sequences in a different presentation order that is, either with increasing magnitude of distortion or with decreasing magnitude of distortion.

Choice of grading scale:

The grading scale used to rate the quality of the video sequences should preferably be detailed enough to allow discrimination between small quality differences and be simple enough to be used in a meaningful way. The ITU-T Recommendation P.910 [6] specifies that a five-point, nine-point or eleven-point grading scale may be used depending on the required discriminative power. In the single stimulus subjective video quality experiment, a discrete five-grade scale is used as shown in figure 4-7. The corresponding numerical values for the opinion scores are: Excellent = 1.0, Good = 0.75, Fair = 0.5, Poor = 0.25 and Bad = 0.

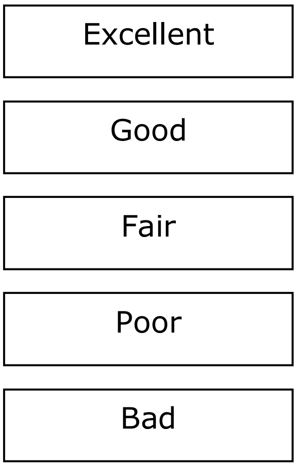


Figure 4-7: Discrete five-grade rating scale.

Environmental set up:

Calibrated equipment and well-defined test environment deliver more accurate and reproducible subjective test results. Standard environmental set up parameters have been defined in ITU-T Recommendation P.910 [6] for multimedia applications. These include the video display device specifications, testing room conditions and viewing conditions.

All subjective experiments in this research were conducted using a computer with 14.1″ LCD computer display set at a native resolution of 1280x1024 pixels. The choice of an LCD monitor was motivated by the fact that it is considered representative of target end-terminals (e.g. computer monitors and mobile devices). It is noted that screen size has an effect on the visibility of distortion. Distortion in smaller screen sizes such as mobile phones may look different on bigger screens such as computer monitors. The computer was setup in the Centre for Video Communications (CVC) research lab in the Robert Gordon University. The video files were stored locally on this test computer and presented to viewers using an in-house YUV video player called the 'Imagicity Viewer' software developed for playing uncompressed CIF and QCIF video sequences. The viewing distance between the observer and the monitor is specified based on the image resolution. Although minimum recommended viewing distances have been specified in the ITU-T Recommendation J.247 [8], a free viewing distance was used in the subjective tests reported in this thesis. In other words, a fixed viewing distance was not enforced and the viewers were allowed to adjust to their most comfortable viewing distance in order to maintain real world scenarios of watching multimedia video sequences. The viewing device was adjusted to a preferred viewing condition. However, the viewers were instructed to adjust the chair according to their normal computer viewing distance and keep their back in contact with the chair as much as possible to avoid extreme variation of viewing distance during the experiment.

Test subjects:

In order to reach statistical significance, the recommended number of participants for a subjective video quality test ranges between 4 and 40 [5]. In this research work, for each subjective test, 30 non-expert participants were recruited from the Robert Gordon University. They were either members of staff or students. None of them had previously participated in a subjective evaluation and all of them reported to have normal vision.

4.4.2 Experimental procedure

Each subjective evaluation experiment was broken down into three phases: the explanation phase, the training phase and the actual subjective test. In the explanation phase, an oral description of the test procedure was given to the test subjects. Details of the oral description given to the participants are in Appendix B. The training phase was aimed to make the test subjects familiar with the test procedure. The video sequences used for training were representative of the range of quality and the types of degradations included in the actual test. In this experiment, the 'Silent' CIF sequence was used for training purposes. The results of the training phase have been excluded in the data analysis.

Once the training phase was complete, the actual subjective test was conducted on the test video sequences mentioned in section 4.2 which have been compressed at wide range of bitrates. In the case of the single stimulus method, each video sequence was presented one at a time and rated individually. After each video presentation, the viewers were asked to judge the overall picture quality. Voting period was not time-limited. The presentation order of the video data was randomised between viewers.

4.4.3 MOS measurement and validity

Subjective test results have been analysed according to the Recommendation ITU-R BT.500-11 [5]. The first step in the analysis of the subjective tests is the calculation of the mean opinion score. The mean opinion score is the mean value of the ratings from a number of observers for a test sequence with a certain test condition (such as a sequence encoded at a certain quantisation parameter or at a certain bitrate):

$$MOS = \frac{1}{N}\sum_{n=1}^{N} R(n) \tag{22}$$

where N is the number of observers and R is the rating from each observer which is based on a five-grade rating scale between 0 and 1 (0=Bad, 0.25=Poor, 0.5=Fair, 0.75=Good and 1=Excellent). MOS is calculated for each test condition and each test sequence.

Since a limited number of observers are used to represent the entire population, the reliability of the subjective test must be calculated. This is performed using:

A: Standard Deviation

B: 95% Confidence Interval

Standard deviation of a data set is a measure of variability of data from the mean value (i.e., MOS in case of subjective tests). It is calculated for N data points using the data point x_i and the mean value $\bar{x}$ as:

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{x} - x_i\right)^2} \tag{23}$$

A low standard deviation indicates that data points are close to the mean value, whereas high standard deviation indicates that data points are spread out over a large range of values. In subjective evaluation, low standard deviation of the subjective ratings is preferred as it indicates high reliability of the subjective scores.

Confidence interval (CI) is used to indicate the reliability of an estimate. The Recommendation ITU-R BT 500-11 proposes the use of 95% confidence interval for calculating the reliability of MOS scores derived from the standard deviation (SD) of the subjective ratings, the mean value $\bar{x}$ and number of observers (N) as:

$$95\%CI = \left[\bar{x} - 1.96\frac{SD}{\sqrt{N}}, \bar{x} + 1.96\frac{SD}{\sqrt{N}}\right] \tag{24}$$

With a probability of 95%, the absolute value of the difference between the experimental (or estimated) mean score and the "true" mean score for a large number of observers is smaller than the 95% confidence interval, on the condition that the distribution of the individual scores is a normal distribution. The confidence interval is dependent on the number of observers. For the mean opinion score of [0,1] to be reliable, the 95%CI should ideally be below 0.05. It has been observed [5] that increasing the number of participants in a subjective test decreases the 95% CI.

4.5 Data modelling techniques

Data obtained from subjective evaluations and encoding tests are analysed for modelling and performance comparison purposes. Data modelling techniques used in the research work include exponential curve fitting, convex hull fit and linear regression analysis.

4.5.1 Exponential curve fitting

In this research work, exponential curve fitting has been used in:

- the development of the perceptual video quality metric for automatic estimation of metric parameters from video sequences characteristics.
- Calculating the Lagrange multiplier (λ) as a function of the quantisation parameter (QP) in the mode selection algorithm.

Exponential curve fitting involves constructing an exponential curve or exponential function that has the best fit to a set of data points.

Exponential functions have the general form:

$$f(x) = a^{bx} \tag{25}$$

where x is the data point, the amplitude of the exponential curve depends on 'a' and the shape of the exponential curve depends on whether b>0, b=0 or b<0 as shown using an example in Figure 4-8. The exponential curve fitting were performed using the 'cftool' feature in MATLAB programming package.

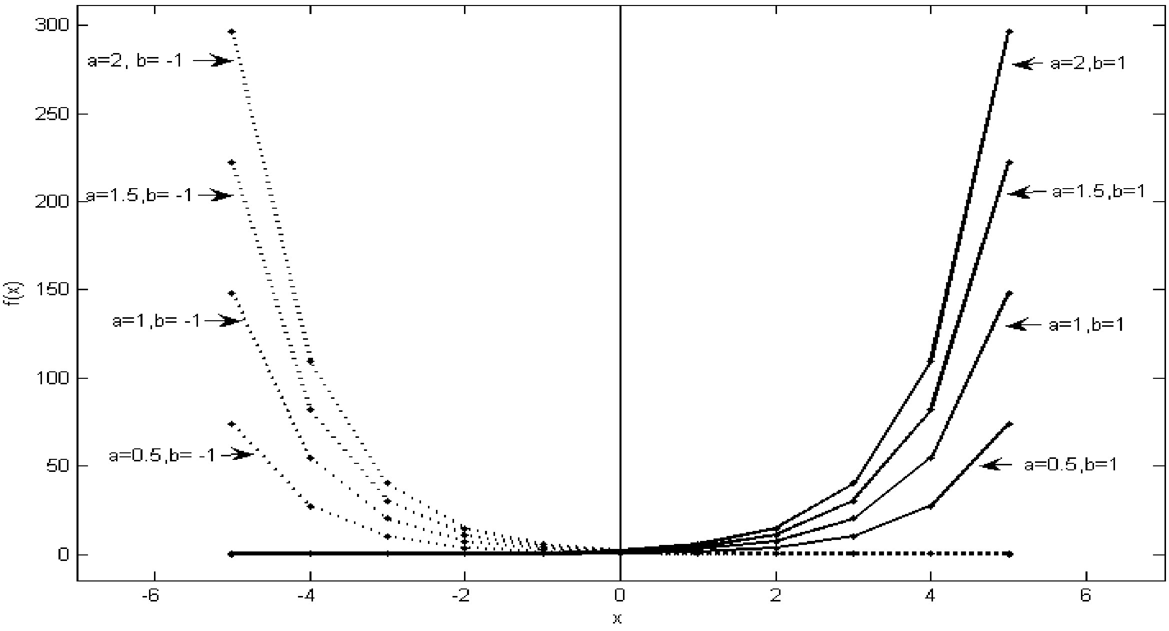


Figure 4-8: Exponential curves

4.5.2 Convex hull fitting

Convex hull of a set of data points is the minimal convex set containing the data points. Convex hull fitting is used to determine the parameters for the perceptually optimised mode selection algorithm. It is used to obtain the best achievable rate-distortion points for a given source as shown below:

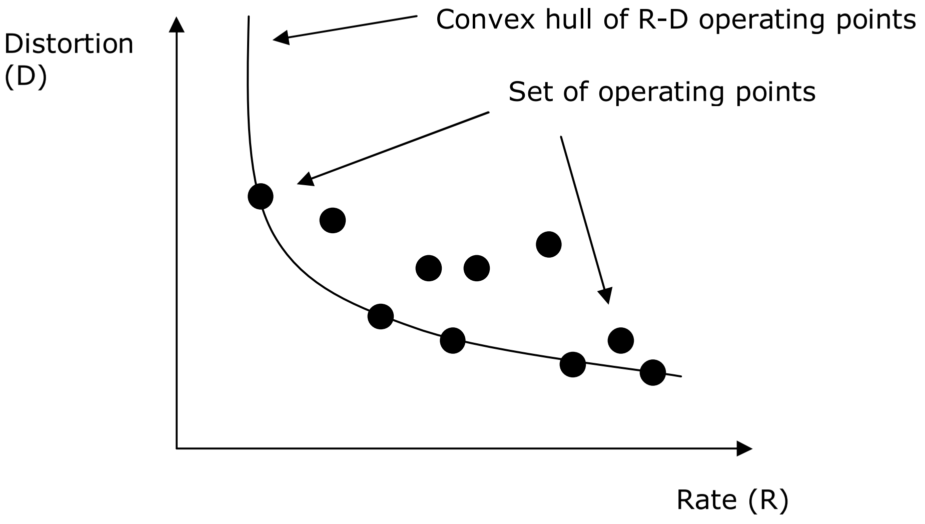


Figure 4-9: Convex hull fitting to obtain best operating R-D points for a given coding condition

The convex hull fitting were performed using the 'cftool' feature in MATLAB programming package.

4.5.3 Linear regression modelling

Linear regression modelling has been used in this research work to investigate the relationship between subjective and objective video quality measures such as MOS and MSE. Linear regression model attempts to explain the functional relationship between two variables (for example: x and y) using a straight line. This relationship may be expressed as:

$$y = \beta_0 - \beta_1 x \tag{26}$$

where β_0 and β_1 are the regression coefficients. It is noted that the regression line in equation (26) is an estimated relationship between the predictor values (x) and the observed values (y). The 'true' regression line is usually never known.

The coefficients β_0 and β_1 are estimated from the observed data set and can be calculated using the least square estimates method as:

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \frac{\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_i\right)}{N}}{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2} \tag{27}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{28}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the regression coefficients β_0 and β_1. x_i and y_i are the predictor and observed values, and N is the number of observations used to fit the model. $\overline{x}$ and $\overline{y}$ are the mean values of x and y respectively.

Once the estimated coefficients are known, the estimated (or fitted) regression line can be written as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{29}$$

An example plot of fitted regression line for a set of data points is shown in Figure 4-10.

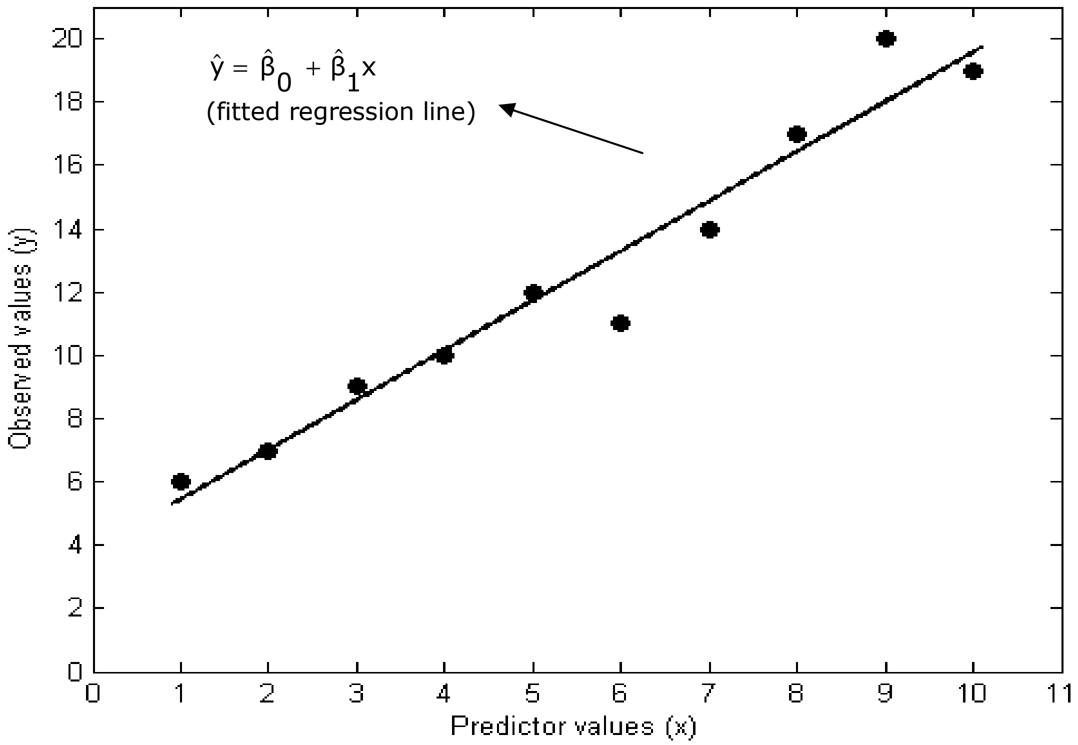


Figure 4-10: Sample data points fitted with a linear regression line

The difference between the actual observed value y_i and the estimated observed value $\hat{y}_i$ obtained from equation (8) is called the residual ε_i which is calculated as:

$$\varepsilon_i = y_i - \hat{y}_i \tag{30}$$

The goodness of the estimated (or fitted) regression line is assessed using two parameters:

A: Coefficient of determination (R-squared or R^2)

B: Sum of square error (SSE).

Coefficient of determination (R^2) is a key output of regression analysis. It is indicates the extent to which the estimated variable (i.e., $\hat{y}$) can be correctly estimated from the predictor variable (i.e., x). It is calculated as:

$$R^2 = \left[\frac{1}{N} \frac{\sum_{i=1}^{N} (\bar{x} - x_i)(\bar{y} - y_i)}{\sigma_x \sigma_y} \right] \tag{31}$$

In equation (31), N is the number of observations, $\bar{x}$ and $\bar{y}$ are the mean values of the predictor and estimated values x_i and y_i. σ_x and σ_y are the standard deviations of x and y.

The coefficient of determination ranges from 0 to 1 with 0 indicating y cannot be estimated from x and 1 indicating that y is estimated from x without any errors. For example, R^2 of 0.25 means that 25% of the variance in y can be estimated using x. R^2 of 0.90 means that 90% of the variance in y can be estimated using x. Higher R^2 values indicate better the estimation result.

Sum of square errors (SSE) of a regression line is a measure of the average amount by which the regression equation over- or under predicts. It is calculated from the residual ε_i as:

$$SSE = \sum_{i=1}^{N} \varepsilon_i{}^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{32}$$

The higher the coefficient of determination, the lower the sum of square errors and the more accurate the estimates are likely to be.

4.6 Algorithm performance testing

Algorithm performance testing involves comparing the performance of developed algorithms with existing techniques for benchmarking purposes. Section 4.6.1 outlines the performance parameters used to evaluate the performance of the new perceptual video quality metric developed in this project with existing video quality metrics. Section 4.6.2 describes the various video coding parameters used to compare the performance of the perceptually optimised mode selection algorithm which was developed with the reference video codec.

4.6.1 Video quality metric performance parameters

The performance of objective quality metrics depends on how well it correlates with subjective ratings. Hence mean opinion score obtained from subjective tests is used to benchmark the performance of the developed quality metric. There are several methods to compare the performance of quality metrics [72]. Three of these performance evaluation methods have been adopted by the video quality experts group (VQEG) for evaluation and benchmarking of objective quality metric. These include: Pearson correlation to measure prediction accuracy, Spearman correlation to measure predict monotonicity and outliers ratio to measure prediction consistency. Brief descriptions of these parameters are given in Chapter 3, section 3.7.

4.6.2 Video coding performance parameters

Coding parameters used to compare the performance between the reference JM encoder and the modified JM encoder are peak signal to noise ratio (PSNR), bitrate and encoding time.

Peak signal to noise ratio (PSNR)

PSNR is the most widely used objective video quality measure. It is a pixel-based measure calculated from mean squared error between the reference video and the video under test, and the square of the highest possible value in the video image (i.e., 255 for an 8-bit image) [13]. PSNR is calculated for an n-bit image as:

$$PSNR = 10\log_{10}\frac{\left(2^{n}-1\right)^{2}}{MSE} \qquad (33)$$

PSNR measures how closely the video under test resembles the original video sequence. The higher the PSNR better the quality of video under test.

Bitrate

Bitrate is the average bits per second and is measured from the total number of bits, number of encoded frames and frame rate as:

$$\text{bit rate} = \frac{\text{total number of bits x frame rate}}{\text{number of frames}} \qquad (34)$$

Bitrate information is obtained from the log file produced by the JM encoder. Comparison of bitrate information between the reference JM encoder and the algorithm under test gives an idea of the bit consumption caused by the algorithm.

Bitrate savings between the modified and reference JM codec is measured as the proportion of bits saved with respect to the reference JM codec bitrate and is calculated as:

$$\text{Bit rate savings(\%)} = \frac{\text{reference codec bitrate - modified codec bitrate}}{\text{reference codec bitrate}} \text{ x } 100 \qquad (35)$$

Encoding time

Encoding time is the time taken by the encoder to perform a specific task. This task could be encoding a specific number of frames or the execution of a certain process such as motion estimation or mode selection. Encoding time is usually used as a measure of computational complexity. A higher encoding time indicates higher computation complexity. In this research work encoding time for the reference JM encoder and the modified JM encoder is measured as the average time taken by each encoder to encode a specific number of frames. The average time is calculated using five repetitions of the encoder running under the same test conditions.

The computational complexity (CC) of the developed algorithm is measured using the encoding times of the modified and reference JM codec as the proportion of increase (or decrease) in encoding time with respect to the reference JM codec and is calculated as:

$$CC(\%) = \frac{\text{reference codec time - modified codec time}}{\text{reference codec time}} \times 100 \qquad (36)$$

4.6.3 Algorithm testing scenario

Figure 4-11 shows the algorithm testing scenario. The original JM encoder is the reference encoder and the algorithm under test is incorporated into the JM encoder to form the modified JM encoder. Each raw test sequence is processed using both the reference and modified encoders.

Coding performance parameters such as PSNR, bitrate and computation time are measured from each encoder and compared. The encoded video is decoded to perform subjective evaluation to compare visual quality of the video sequences from the reference and modified JM encoders at similar bitrate.

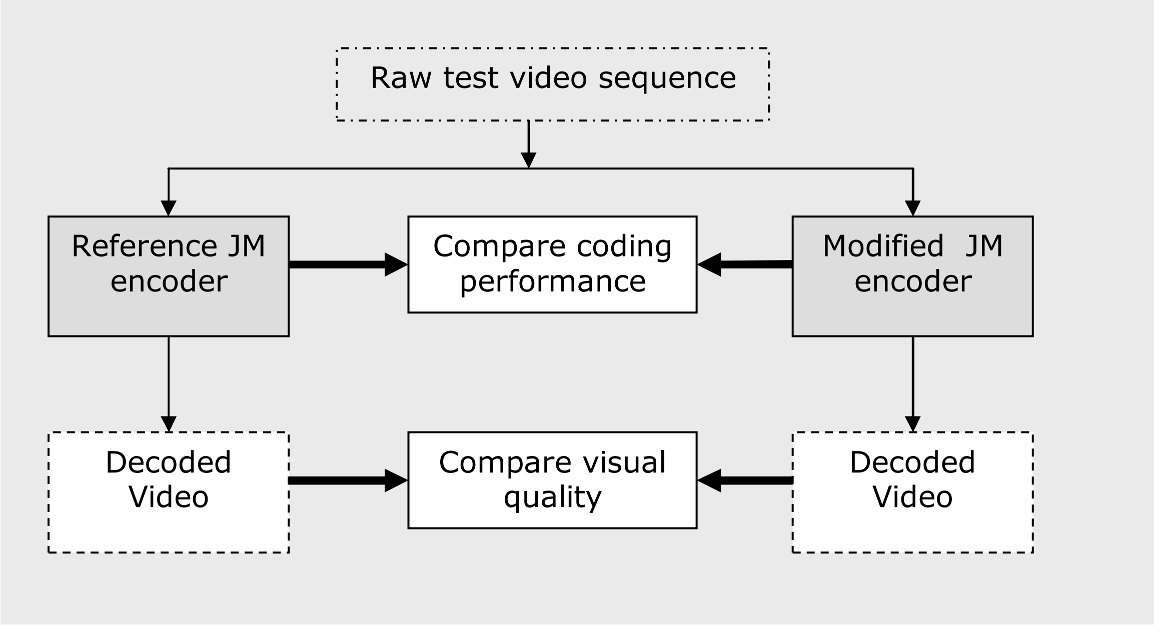


Figure 4-11: Algorithm testing scenario

4.6.4 Algorithm performance analysis

The video compression performance of the algorithms under test is assessed using codec output parameters such as video quality, bitrate and encoding time. The rate distortion plots are obtained by plotting video quality (PSNR or MOS) against bitrate. These plots are used to determine if there is a quality gain and bitrate savings as shown in figure 4-12. Quality gain is achieved if the algorithm under test produces better quality than the reference algorithm at same bitrate. Bitrate saving is obtained if the algorithm which has been developed achieves a particular video quality (Q) with a smaller bitrate compared to the reference algorithm.

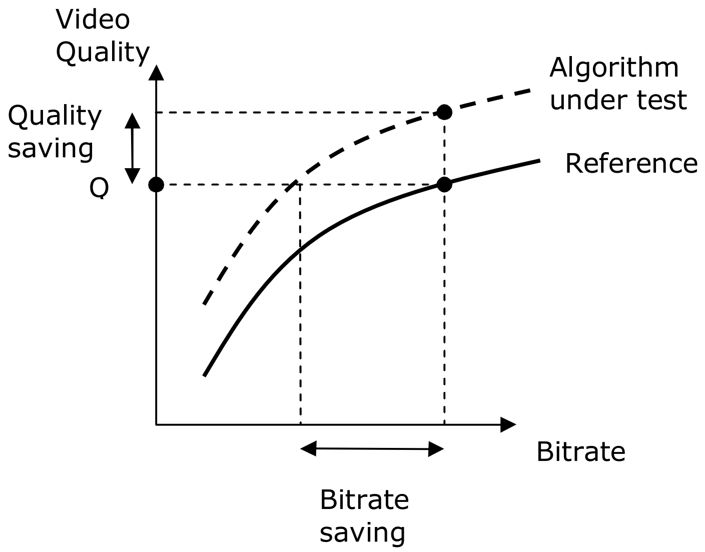


Figure 4-12: rate distortion plots for performance analysis

4.7 Summary

This chapter gives a description of the experimental methods used in the research project. The algorithms developed during the project have been tested using software simulation. Software packages such as MATLAB and Microsoft Visual C++ have been used in developing and testing of algorithms. The H.264/AVC reference software codec is used as the reference video codec. Developed algorithms have been incorporated into the reference codec for performance comparison. Raw standard test sequences of YCbCr 4:2:0 common intermediate format (CIF) have been used for testing the performance of developed algorithms and conducting video quality tests. Data modelling techniques such as linear regression analysis have been used to analyse test data. Finally, the performance of developed algorithms is evaluated using as subjective and objective video quality tests and video compression output parameters such PSNR, bitrate and encoding time.

5 MOSp: A New Perceptual Video Quality Metric For Compressed Video

5.1 Introduction – need for a new quality metric

Video quality measurement is necessary for comparing, evaluating and benchmarking video compression systems. As it is typically the viewer who judges video quality, the subjective measurement of mean opinion score (MOS) [4] is considered to be an accurate way to determine the perceived video quality of compressed video [5,6]. However, evaluating MOS is expensive in terms of time and resources and cannot be calculated automatically within real-time video applications. Hence several objective assessment methods have been developed to automatically predict the subjective results based on video content and the characteristics of the human visual system. The video quality experts group (VQEG) have performed evaluation tests to benchmark the performance of a number of quality metrics in the context of multimedia sequences. This has resulted in the standardisation of a selection of video quality metrics in the ITU-T Recommendations. These metrics have varying degrees of success in predicting the subjective test scores, with reported correlations of 70% to 84% [7] between each objective metric and the measured subjective quality scores, indicating that there is a need for better approaches to estimate subjective quality. Although several techniques have been proposed in the literature as alternatives to pixel-based approaches, mean squared error (MSE) and peak signal to noise ratio (PSNR) still retain their popularity in video compression algorithms due to implementation simplicity and computation speed. Previous research has shown that the correlation between subjective results (MOS) and pixel-based methods such as MSE and PSNR is high for a single sequence coded to various bitrates [71]. By exploiting this high correlation between subjective results and pixel-based measures, it may be possible to accurately predict the subjective results from simple pixel-based metrics such as MSE and PSNR.

This chapter presents a new full reference perceptual video quality metric called the MOSp metric. The metric predicts the mean opinion score of multimedia sequences compressed using block-based video coding schemes. The metric is based on: (i) the high correlation between MSE and MOS for a single sequence compressed at several bitrates and (ii) the visual masking of distortion in a video scene which makes it possible for distortions to be visible in some areas of the video scene while they are unnoticed in other areas.

The chapter is organised as follows: section 5.2 investigates the correlation between MOS and pixel-based measures such as MSE and PSNR. Based on the observations made from experiments in section 5.2, a new video quality metric called the MOSp metric is presented in section 5.3. The hypothesis behind the new metric is also described in this section. The various parameters used in the new metric are described in detail in section 5.4 along with possible methods to automatic calculations of these parameters. Finally, section 5.5 presents a summary of the chapter including highlights of the experimental findings which have led to the development of a new perceptual quality metric.

5.2 Experiment: Quality measurement of compressed video

The following video quality experiment has been conducted to investigate the correlation between subjective results (MOS) and pixel-based metrics such as mean squared error (MSE) and peak signal to ratio (PSNR) for a variety of multimedia video sequences compressed at a wide range of bitrates ranging from high bitrates to very low bitrates using the H.264/AVC video coding standard.

5.2.1 Source Material

The variation of MOS with MSE across various video data was determined using a training data set of eight different video sequences. The sequences were Carphone, Foreman, Deadline, Tempete, News, Bus, Paris and Akiyo. These sequences were chosen to represent a wide variety of video content and they are popularly used in the research community. The test sequences range from low motion and low detail 'head and shoulder' scenes such as Akiyo and News to high motion and high detail scenes such as the Bus sequence. The sequences were in c*ommon intermediate format (CIF)* resolution, 4:2:0 YCBCR format, 10 seconds in duration and were coded using the H264/AVC compression standard. Each test sequence was compressed at a wide range of bitrates using a fixed set of quantisation parameter (QP) values, QP = {6, 26, 34, 36, 38, 40, 42, 45}. There is a closer spacing between the chosen QP values in the range of QP=34 to QP=42, which translate into a 'useful' medium- to low-bitrate range for multimedia sequences.

5.2.2 Coding parameters

This research is focused on video quality of compressed multimedia sequences. Hence, the choice of coding parameters was made in the context of multimedia applications. The test video sequences used in this experiment were compressed using the H.264/AVC JM reference software (version 12.1) available at http://iphome.hhi.de/suehring/tml/, with the following parameters:

- Profile used is Main profile to allow performance enhancing features such as CABAC to be used.
- Level setting is set 4.0 to accommodate higher resolution, frame rate (2,048×1,024@30.0) and bit rate (up to 80,000 kbps)
- Frame Skip: no frames were skipped
- Number of reference frames for Inter motion search is set to 5 after taking into account the increase in computation with increase in reference frames.
- Number of B-pictures used: None used because the extra coding delay introduced by B-frames may not be suitable in certain real time multimedia applications.
- Entropy coding method is set to CABAC to achieve better efficiency when compared to CAVLC.
- RD-Optimisation: High complexity mode which uses exhaustive mode selection for improved compression performance.
- Rate Control: DISABLED to allow the use of fixed QP.
- Slice QP: QPISlice and QPPSlice parameters used and both set to the same value as the sequence QP. This is to keep the variation of parameters in the video codec to a minimum during encoding.

H264/AVC supports up to 16 reference frames for inter motion search. More reference frames can increase compression quality; however it is also computationally intensive during encoding, and requires more memory during decoding. The choice of the number of reference frames also depends on the level settings and the input frame size [79]. Since level 4.0 is chosen and the input file format is CIF, the number of reference frames in this experiment is set to 5.

5.2.3 Methodology and Experimental procedure

The subjective tests involved 30 non-expert evaluators and followed the guidelines in ITU-T Recommendation P.910 [6]. Each evaluator took less than 20 minutes to complete the test. The subjective test method used in this experiment is the single stimulus impairment scale (SSIS) evaluation method because a wide range of bitrates have been used in the test dataset ranging from very high bitrate (QP=6) to very low bitrate (QP=45). This method is appropriate when video sequences at comparably low bitrates are used because the reference video is not displayed. In showing the high quality reference video, the distorted video sequences may be perceived as poor quality and no distinction between different levels of low quality may be made by the observers. To counterbalance the influence of contextual effects of the SSIS method, the presentation order of the video sequences were randomized between participants such that each participant viewed the sequences in a different presentation order that is either with increasing or decreasing magnitude of distortion.

A 5-grade discrete scale ranging from 0 to 1 was used to rate the quality of the test video sequences where 0=bad, 0.25=poor, 0.5=fair, 0.75=good and 1=excellent. Reliability of subjective test scores was tested using the 95% confidence interval measure. The average mean 95% confidence interval for the subjective ratings for all the test sequences was 0.0415 for the MOS scale of [0, 1] where 0=bad picture quality and 1=excellent picture quality.

The MOS for a sequence was calculated as the average of all scores obtained for the sequence compressed at a certain QP. The sequence MSE was calculated as the mean of the sum of squared differences (SSD) between the luminance pixels of the original and the constructed compressed video sequence as given below:

$$SSD_{MB} = \sum_{i=1}^{16}\sum_{j=1}^{16}\left[Y(i,j) - Y_C(i,j)\right]^2 \tag{37}$$

$$MSE_{MB} = \frac{SSD_{MB}}{(16*16)} \tag{38}$$

$$MSE_{Frame} = \frac{1}{N}\sum_{i=1}^{N} MSE_{MB}(i) \tag{39}$$

$$MSE_{sequence} = \frac{1}{T}\sum_{t=1}^{T} MSE_{Frame}(t) \tag{40}$$

SSD_{MB} is the sum of squared difference of each macroblock. Y and Y_C are the luminance values of the original and reconstructed compressed frames. N is the number of macroblocks in each frame and T is the total number of frames in each sequence. Since CIF sequences of 4:2:0 YCbCr format are used in this work, every macroblock contains 16x16 luminance samples, 8x8 Cb samples and 8x8 Cr samples.

5.2.3 Data Analysis

The graphs of MSE versus MOS for all the eight test sequences are shown in Figures 5-1 and 5-2. It can be observed from the graphs that there is high correlation between MOS and MSE values within each sequence compressed at various bitrates. The graphs show characteristic 'hockey stick' shaped curves. The curves are approximately linear from MOS = 1.0 down to MOS = 0.1 with a tail-off below MOS = 0.1. This tail-off occurs because at very low bit-rates (below MOS=0.1), the picture quality is very poor and the users tend to rate the video as ¨Bad¨ quality after a certain error threshold with little discrimination in picture quality. Hence a cut-off may be introduced at MOS=0.1 and data points above this cut-off may be used for modelling purposes.

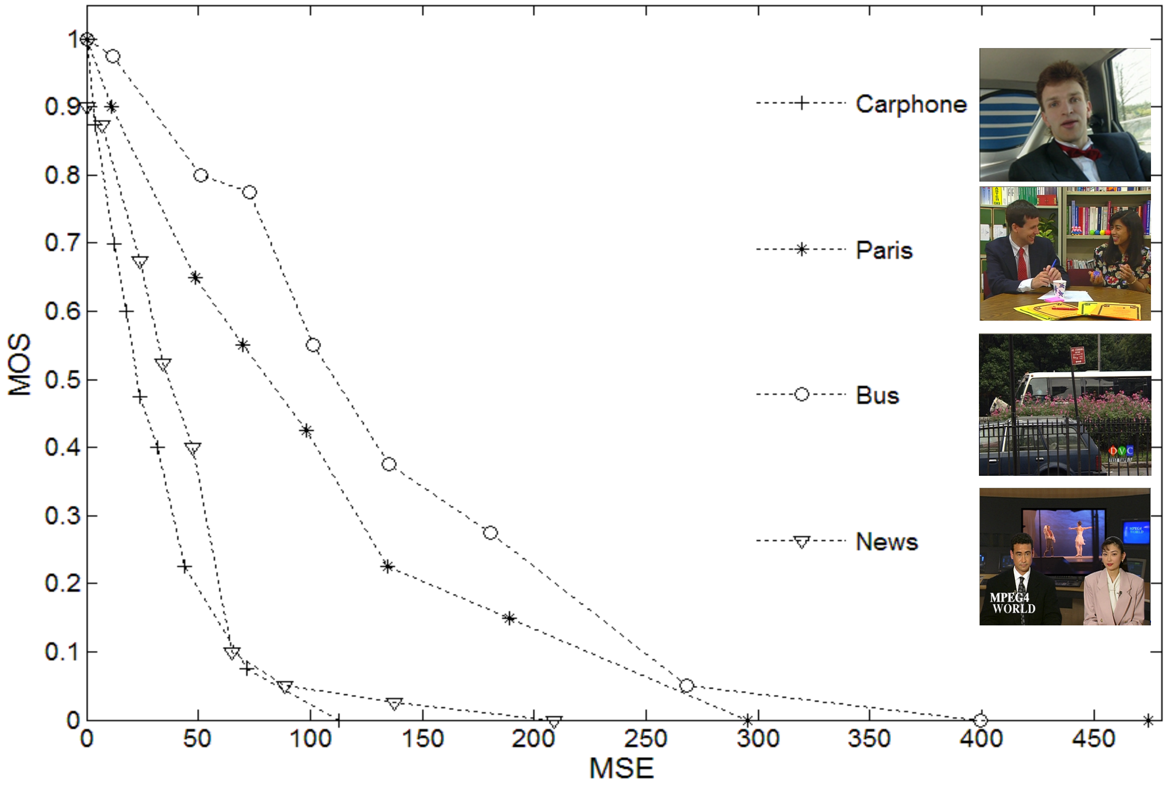


Figure 5-1. Graph of MSE versus MOS for four training sequences: Carphone, Paris, Bus and News

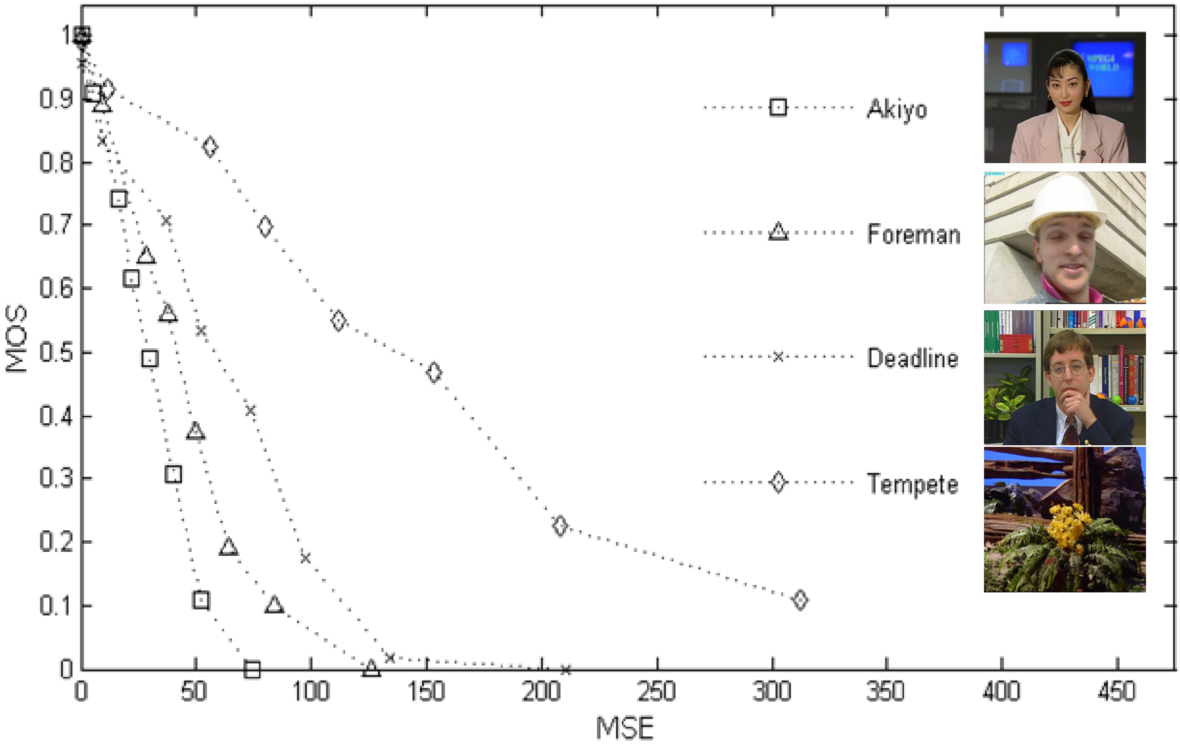


Figure 5-2. Graph of MSE versus MOS for four training sequences: Akiyo, Foreman, Deadline and Tempete.

5.2.4 Experimental observation

The following observations can be made from the video quality experiment:

- Subjective rating (MOS) decreases with increase in MSE. The rate of decrease varies between sequences as noted from the MSE versus MOS curves in Figures 5-1 and 5-2.

- There is a high correlation between MSE and MOS for a single sequence compressed at various bitrates produced using the same video codec.

- This high correlation can be approximated to a linear relationship when data points at very low bitrates (MOS<0.1) are not considered.

- The high overall correlation between MSE and MOS decreases with increase in the number of different sequences added to the data set.

- Finally, one of the key observations in the experiment is that for the same MSE value, the subjective quality varies significantly for different sequences depending on the video content. This is discussed in detail in section 5.3.

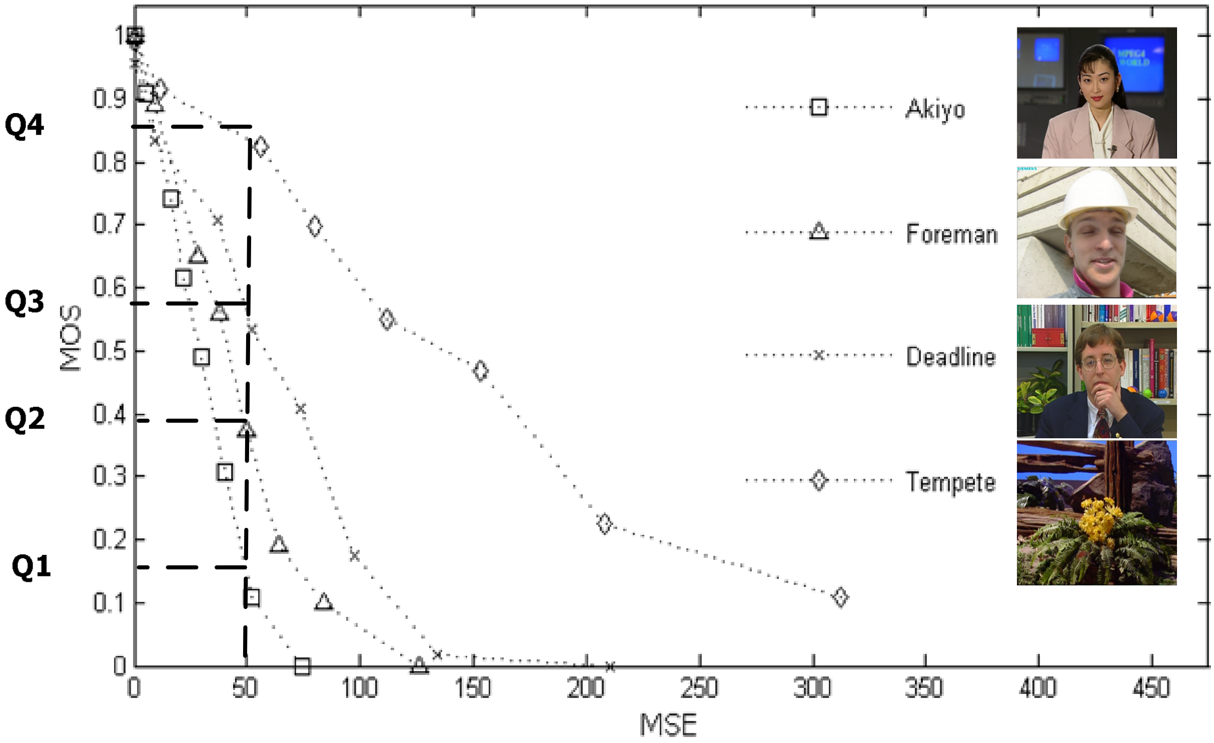
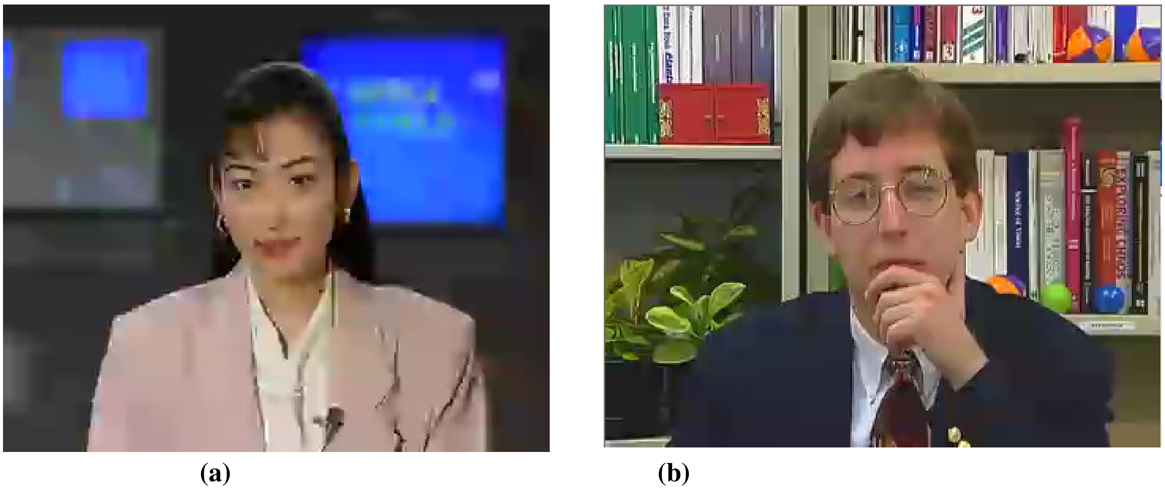


Figure 5-3. Graph of MSE versus MOS for four training sequences: Akiyo, Foreman, Deadline and Tempete. For a fixed MSE (=50), the corresponding subjective ratings for the four sequences are Q1, Q2, Q3 and Q4.

5.3 The concept of predicting MOS from MSE

Experimental results in Figures 5-1 and 5-2 show that for the same MSE value, the subjective quality varies for different sequences depending on the video content. This is illustrated in Figure 5-3. For the same MSE (=50), the four test sequences have different subjective quality. This indicates that sequence content may be one of the contributing factors to the visibility of distortion. Sequence content could be image features, objects in the video scene that relate to task in hand and viewer interest. Therefore, perceived quality of two video images with very similar MSE may be very different. This is illustrated using video frames from two test sequences used in this experiment: (a) Akiyo, with average frame MSE = 74.46 and (b) Deadline with average frame MSE = 74.61. It can be seen that although the average MSE of both the video frames are similar, the visual quality of video frame (b) may be ranked as better than video frame (a). The background and facial features in frame (b) is much clearer than frame (a).



(a) (b)

Figure 5-4: Video frames of two test sequences (a) Akiyo, average frame MSE = 74.46 and (b) Deadline, average frame MSE = 74.61

Based on the experimental observations, it is evident that by exploiting the linear relationship between MSE and MOS, the subjective quality of compressed video may be predicted using:

(i) the MSE between original and compressed video sequences

(ii) the slope of the regression line between MSE and MOS.

This concept of predicting MOS using MSE is illustrated in Figure 5-5 for a video sequence with a known slope 'K'. Based on this hypothesis, a new perceptual video quality metric called the MOSp metric is proposed in section 5.4.

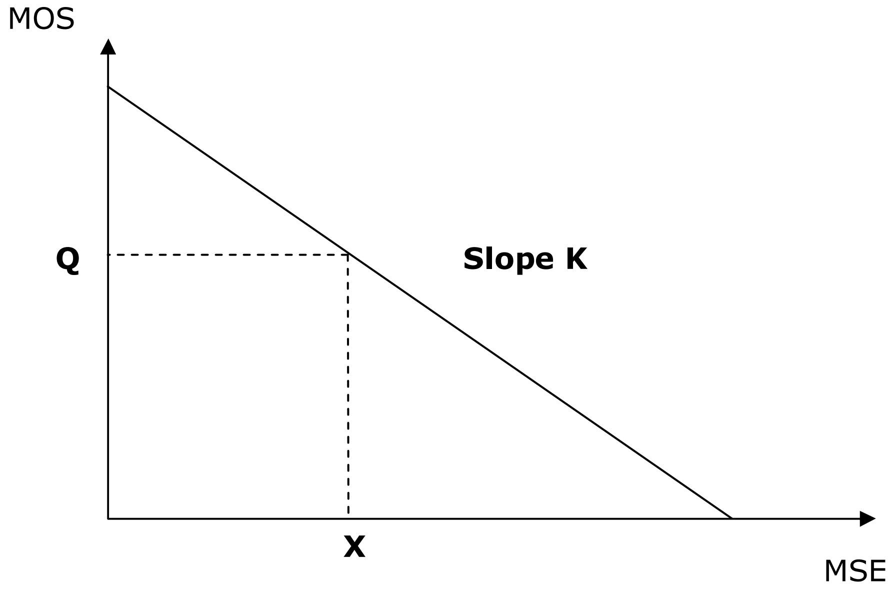


Figure 5-5. The concept of predicting MOS from MSE.

5.4 MOSp: A new perceptual video quality metric

The aim of the proposed MOSp metric is to:

(a) Predict perceived video quality automatically,

(b) Be in close agreement with MOS,

(c) Maintain computational simplicity, with a view of incorporating the metric into real-time block-based video coding algorithms.

Based on the experimental observations obtained from investigating the relationship between MOS and MSE in section 5.2, as shown in Figures 5-1 to 5-5, the new perceptual metric is proposed as:

$$MOS_P = 1 - k_s MSE \qquad (41)$$

The MOSp metric measures the predicted mean opinion score of a compressed sequence using the mean squared error (MSE) between the original and compressed video sequences, and the slope of the regression line (k_s) which may be calculated from the original sequence content. The range of the MOSp metric is similar to that of MOS and it ranges between [0,1] where MOSp = 1 indicates highest quality video and MOSp = 0 indicates lowest quality video. Figures 5-6 illustrates the proposed model which represents the linear relationship between MOS and MSE where the maximum perceived quality (MOS = 1) is observed when there are no pixel errors (MSE = 0).

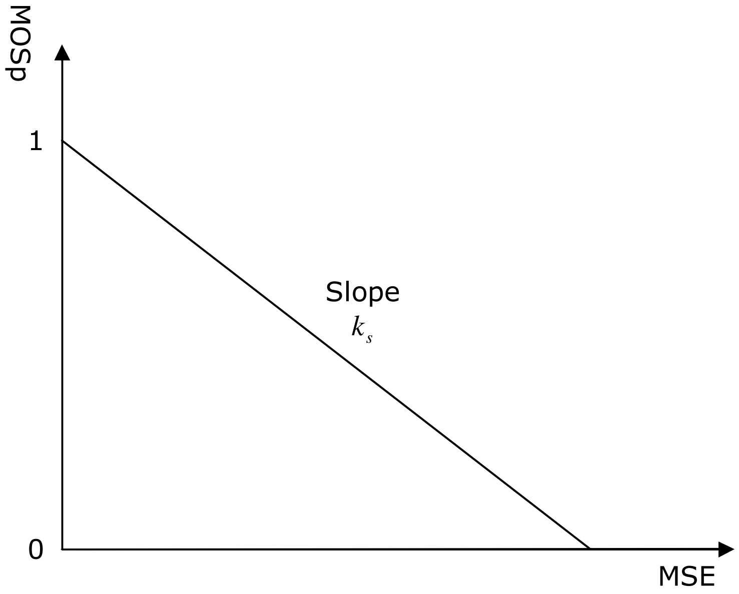


Figure 5-6: Graphical representation of the MOSp metric.

Figures 5-7 shows the proposed model (bold lines) fit to four test sequences used in the experiment. The data points used to obtain the straight line fit for each sequence were considered using the following two conditions:

<u>Condition 1:</u> Include data point MOS = 1 for MSE = 0 for all sequences when calculating the straight line fit. This is done to make the line intercept the MOS axis at 1. The hypothesis behind this condition is that when there are no pixel errors (MSE = 0), the compressed video quality is mathematically identical to the reference video. Hence, no compression induced distortion exists in the processed video clip. In context to this research, which involves measuring perceptual quality of compressed video using a full reference metric, it means that the visual quality of the processed video sequence is identical to the reference video sequence. Hence, at MSE = 0, the subjective quality rating MOS is set to 1 for modelling purposes. It must be noted that in reality, obtaining mean opinion score of 1 is extremely rare as it would mean that every single viewer in the experiment has judged the video quality as 'Excellent'. It can be observed from the subjective test

results presented in figures 5-1 and 5-2 that the MSE versus MOS curves intercept the MOS axis at different points (usually > 0.96).

Condition 2: Exclude data points below MOS < 0.1 during modelling of the straight line fit. This condition is imposed to ignore the 'tail-off' that occurs at very low MOS region, as noticed in Figure 5-7. Perceptual quality at very low bit rates is usually very poor and the viewers tend to rate the video quality as 'Bad' (MOS=0) after a certain error threshold with little discrimination in picture quality. From experimental results, it was observed that the standard deviation between test scores in the region between MOS = 0.1 and MOS = 0 was very high indicating that data points in this region may be unreliable.

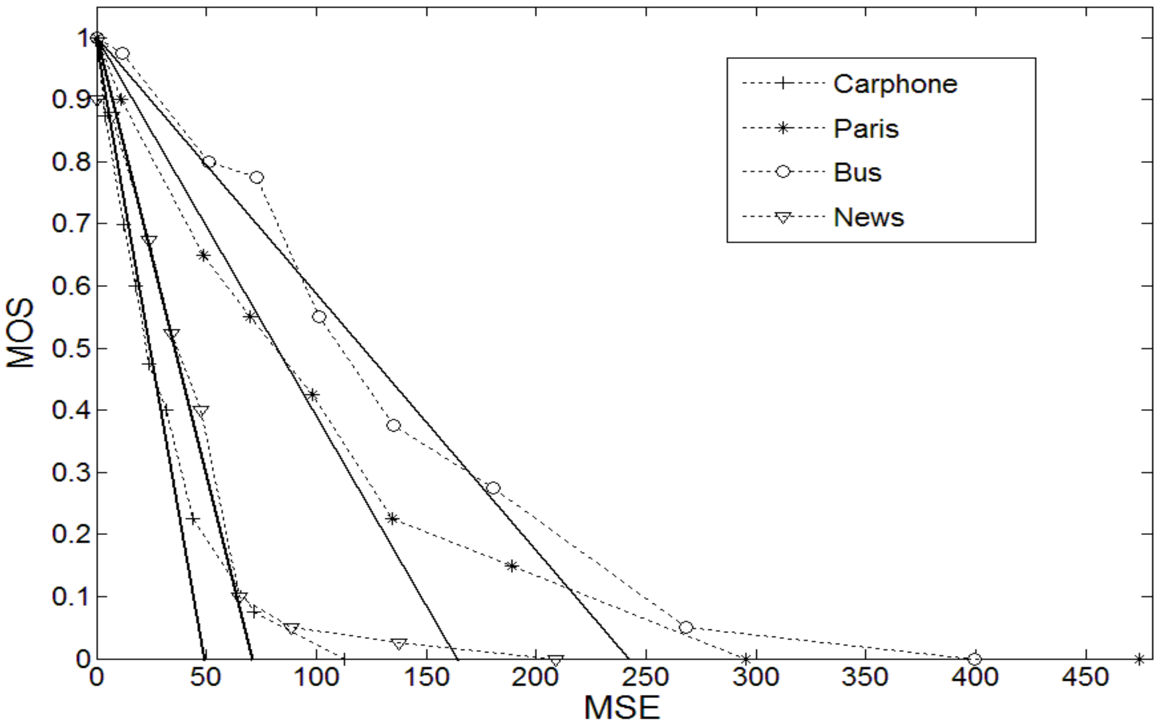


Figure 5-7. Proposed (bold lines) curves for Carphone, Paris, Bus and News..

5.5 MOSp metric parameters

The MOSp metric, as described using equation (41), requires three parameters to measure quality of a compressed video: the mean squared error, the slope (Ks) of the regression line between MSE and MOS, and the y-axis intercept of the regression line (which is set to 1). These parameters are described in detail below.

5.5.1 Metric parameter – Slope Ks

The slope (ks) of the regression line is a key element in the MOSp metric and it acts as a weighting factor for MSE. The amount of weighting, as observed in figures 5-7, is dependent on the video content and varies between sequences. The slope parameter of the MOSp metric also determines the variation of MOS with MSE. A large slope will produce a steeper regression line on the MSE versus MOS graph when compared to a smaller slope. Sequences such as Carphone, Akiyo and Foreman, in Figure 5-7, have steeper regression lines compared to sequences such as Bus and Tempete. This indicates that in sequences with steeper regression lines, a small change in MSE leads to a large change in MOS when compared to sequences such as Bus for the same amount of change in MSE. This is illustrated in figure 5-8.

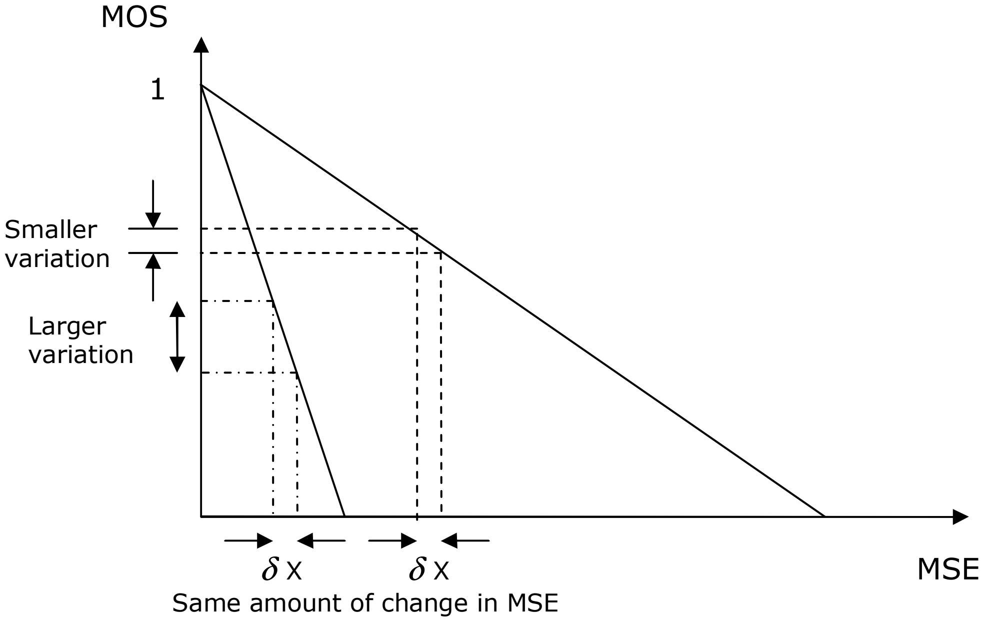


Figure 5-8. Graph illustrating the impact of slope on MSE versus MOS variation.

As noted in section 5.2, video content may have an influence on the slope of the regression line between MSE and MOS. Video content could ‘hide’ or ‘enhance’ the visibility of distortion resulting in a shallower or a steeper slope. This could be objects in the video such as presence of humans, which attract viewer attention. Video sequence containing humans may have a steeper regression line when compared to a sequence containing random objects. On the other hand, image features such as texture and temporal change may also have influence on the slope of the regression line.

Therefore, it may be possible to calculate the slope parameter of the MOSp metric using the video content information. This relationship between the slope parameter and video content is investigated in Chapters 6 and 7 of this thesis.

5.5.2 Metric parameter – MSE

The MSE parameter in the MOSp metric is measured as the mean squared error between the luminance frames of the original and compressed video sequences as explained in equations (37) to (40). The reference video sequence is required in the measurement of MSE thus making the MOSp metric is a full reference video quality metric.

5.5.3 Metric parameter – y-axis intercept = 1

The y-axis intercept of the regression line between MOS and MSE is set to 1 as explained in section 5.4. Therefore the predicted mean opinion score (MOSp) is equal to 1 when MSE = 0. This condition applied to all types of sequences with varying content.

5.6 Summary

This chapter investigated the correlation between subjective quality (MOS) and objective quality (MSE) for compressed multimedia video sequences. It was found that there is high correlation between MSE and MOS for a single sequence compressed to several bit rates using the same coding scheme. Based on this observation, a new metric called the MOSp was proposed to predict MOS from MSE. The MOSp metric is designed to predict perceptual quality of compressed video with compression induced distortions. It was also observed that the regression line between MSE and MOS varies between sequences and may be dependent on video content. Video content may include image features (such as texture, colour and motion) and objects that attract viewer attention based on viewer interest and task in hand. Calculating the parameters of the metric (i.e. slope of the regression line) from video content could make the metric fully automatic. Therefore, the next two chapters investigate the relationship between video content and the parameters of the MSE versus MOS regression line.

6 MOSp Metric Based On MSE And Video Content

6.1 Introduction

Experiments in the previous chapter, conducted to investigate the correlation between subjective and objective measures, showed that there is high correlation between MSE and MOS for a sequence coded at several bit rates using the same coding algorithm. Based on this approximately linear relationship, a new video quality metric called the MOSp metric was proposed to predict mean opinion score from mean squared error and the slope of the regression line between MSE and MOS. It was also noted from the experiments that the slope of the regression line varies between sequences and may be dependent on video content.

Video content may be contributing to the 'hiding' or 'enhancing' of visibility of distortions which in turn may produce a steeper or shallower slope on the MSE versus MOS graph. Video content could include image features such as spatial texture and temporal change. This chapter investigates the relationship between video content and the slope parameter of the regression line between MOS and MSE with a view to automatically estimate the slope parameter from video content and hence the MOSp metric itself.

The chapter is organised as follows: section 6.2 gives a detailed description of the various features that may be used to quantify video content such as spatial texture and temporal change. Section 6.3 describes the experiment conducted to investigate the relationship between video content and the slope of the regression line. Based on this investigation, methods to automatically estimate the slope parameter from video content are given in section 6.4. Video quality measurement at macroblock, frame and sequence level using the MOSp metric is presented in section 6.5. Performance of the MOSp metric is evaluated in section 6.6 and finally, the performance results are discussed in section 6.7.

6.2 Quantifying video content

The slope of the regression line between MSE and MOS varies with video content. Features in a video sequences may have an effect on the visibility of compression induced distortions which could in turn have effect on the steepness or shallowness of the regression line that relates MSE and MOS. Masking effect is an important visual phenomenon which describes why similar levels of distortion may be more visible in certain regions of a video frame while they are hardly noticeable in other regions [63]. As mentioned in Chapter 3 (section 3.2.3), factors contributing to visual masking include spatial texture masking and temporal masking. Features in the video scene such as texture and motion may contribute to the 'hiding' or 'enhancing' of visible distortions. Hence spatial texture masking and temporal masking information are used to quantify sequence content in order to estimate the slope of the regression line (k_s) for each video sequence.

6.2.1 Spatial texture information

Spatial texture masking occurs because regions in a video frame that are rich in texture can mask distortions more effectively than other regions [63]. Spatial edges give a good estimate of the amount of detail in a region and are related to object boundaries, surface crease and other important visual events. Considering this, the spatial edge strength can be used as a measure of spatial texture information. Sobel edge detecting filters [80] are popularly used to obtain edge information due to computational simplicity and robustness to noise.

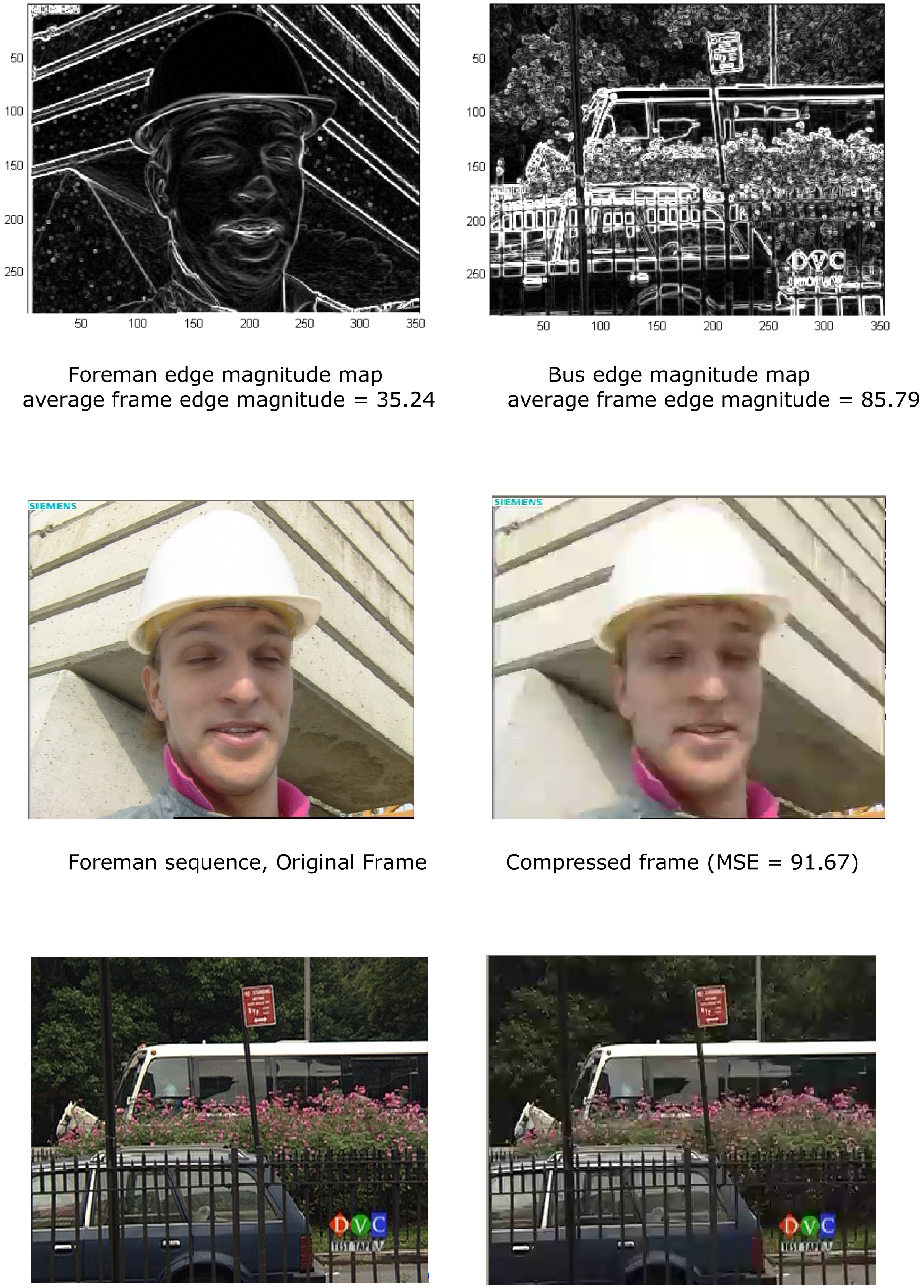
Hence, in this work the Sobel filters have been used to obtain edge information from the luminance component of the original uncompressed video frame. The horizontal edge image ($G_{horizontal}$) and the vertical edge image ($G_{Vertical}$) are separately computed using the Sobel filters, and the edge magnitude image is computed as follows:

$$G(x,y) = \left|G_{horizontal}(x,y)\right| + \left|G_{vertical}(x,y)\right| \tag{42}$$

where, G is the edge magnitude image and (x,y) is the pixel location. Spatial edge strength is measured using local regions, hence the edge magnitude image is divided into 16x16 non-overlapping blocks or macroblocks[1], and the spatial-texture information of each macroblock (STI_{MB}) is computed as the average edge strength of all the pixels in that macroblock. The average edge strength of a macroblock is used as a measure of its spatial texture because a highly textured macroblock will tend to have larger average edge strength due to the presence of strong edges and conversely, a smooth or low-textured macroblock will tend to have smaller average edge strength.

The effect of spatial texture masking on the visibility of distortion can be explained using Figure 6-1. Consider two sequences: Foreman and Bus. The edge magnitude maps of the first frame from both sequences are shown. Figure 6-1 also shows the original frames and the corresponding reconstructed compressed frames. The mean squared error between the original and the compressed frames for both the sequences is similar (approximately = 92). However, the visual quality of the compressed Foreman frame shows more visible loss of detail, especially in the facial area and the boundary between the cap and the face, when compared to the original Foreman frame. For similar MSE, the compressed Bus frame has comparatively less visible loss of detail. From the edge magnitude maps, it can be noted that the Bus frame has large amount of texture in the video scene when compared with the Foreman sequence. Hence, due to the spatial texture masking effect, the distortion in the Bus frame is less visible than the distortion in the Foreman frame.

[1] Video content and MSE are calculated at macroblock level in order to facilitate incorporating the metric into block-based video coding algorithms.



Foreman edge magnitude map
average frame edge magnitude = 35.24

Bus edge magnitude map
average frame edge magnitude = 85.79

Foreman sequence, Original Frame

Compressed frame (MSE = 91.67)

Bus sequence, Original Frame

Compressed frame (MSE = 92.29)

Figure 6-1: Spatial texture masking and visibility of distortion.

6.2.2 Temporal change

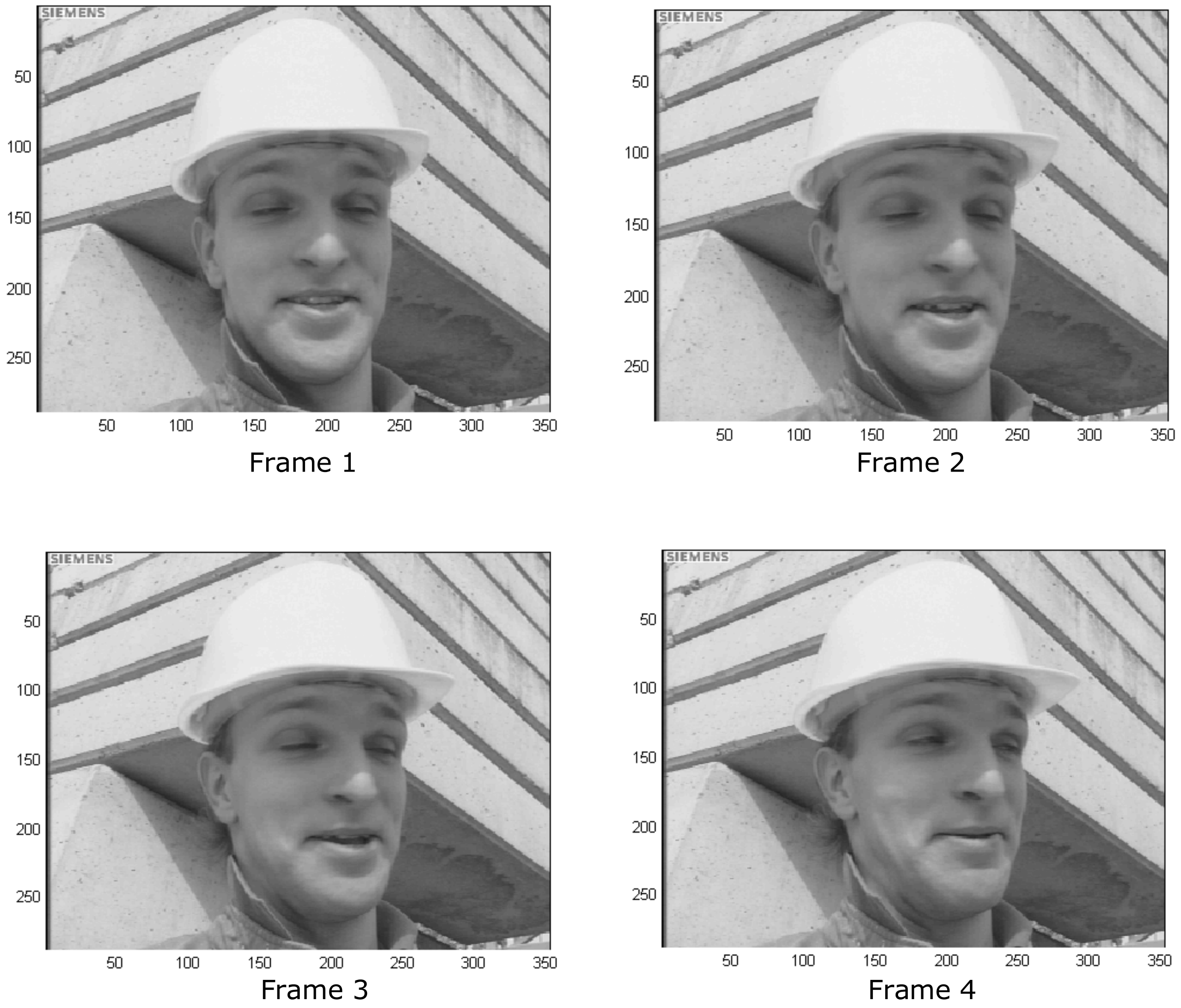
Temporal masking occurs because regions that undergo large temporal changes can mask or 'hide' distortions more effectively than other regions due to the limited cognitive and temporal response of the human viewer [81]. There are several approaches in the literature to measure temporal change in video [14]. These include image differencing and calculation of motion vectors. Although image differencing is computationally simple when compared to motion vector estimation, it is known to enhance image noise. In this research, the temporal change is calculated as the gradient magnitude of the absolute difference between the current luminance frame (Y_n) and the previous luminance frame (Y_{n-1}):

$$Y_{diff} = abs(Y_n - Y_{n-1}) \tag{43}$$

$$TI_n = \left|GT_{horizontal}(x,y)\right| + \left|GT_{vertical}(x,y)\right| \tag{44}$$

where TI_n is the temporal gradient magnitude image of the current frame, $GT_{horizontal}(x,y)$ and $GT_{vertical}(x,y)$ are the horizontal and vertical Sobel gradient images of Y_{diff} image.

Equation (44) is used as a measure of temporal information because it is more robust to noise than simple image differencing [14]. Since equation (44) is also a measure of gradient magnitude, it may be comparable with the spatial edge magnitude measure described in equation (42). And most importantly, it gives an accurate measure of temporal change because a large temporal change between current and previous frame pixels will result in a large absolute difference value and hence a large gradient magnitude. This is explained using Figures 6-2 and 6-3. Figure 6-2 shows four original frames from the Foreman sequence. Figure 6-3 shows the frame difference images between frames 1 and 2, frames 1 and 3 and frames 1 and 4. This combination of frame difference has been chosen to demonstrate the ability of the temporal edge magnitude measure to effectively reflect temporal change. The corresponding temporal gradient magnitude images are also given in figure 6-3. The average temporal edge magnitude of the difference image between frames 1 and 2 is 40.93. Similarly, the average temporal changes between frames 1 and 3 and frames 1 and 4 are 58.76 and 63.91 respectively. This indicates that with the increase in temporal difference, there is a corresponding increase in the average temporal edge magnitude. Hence this measure is an effective indicator of temporal change. The temporal information of each macroblock (TI_{mb}) is computed as the average temporal gradient strength of all the pixels in the macroblock. The temporal gradient strengths of all the macroblocks in a frame are averaged to obtain the measure for frame temporal change.



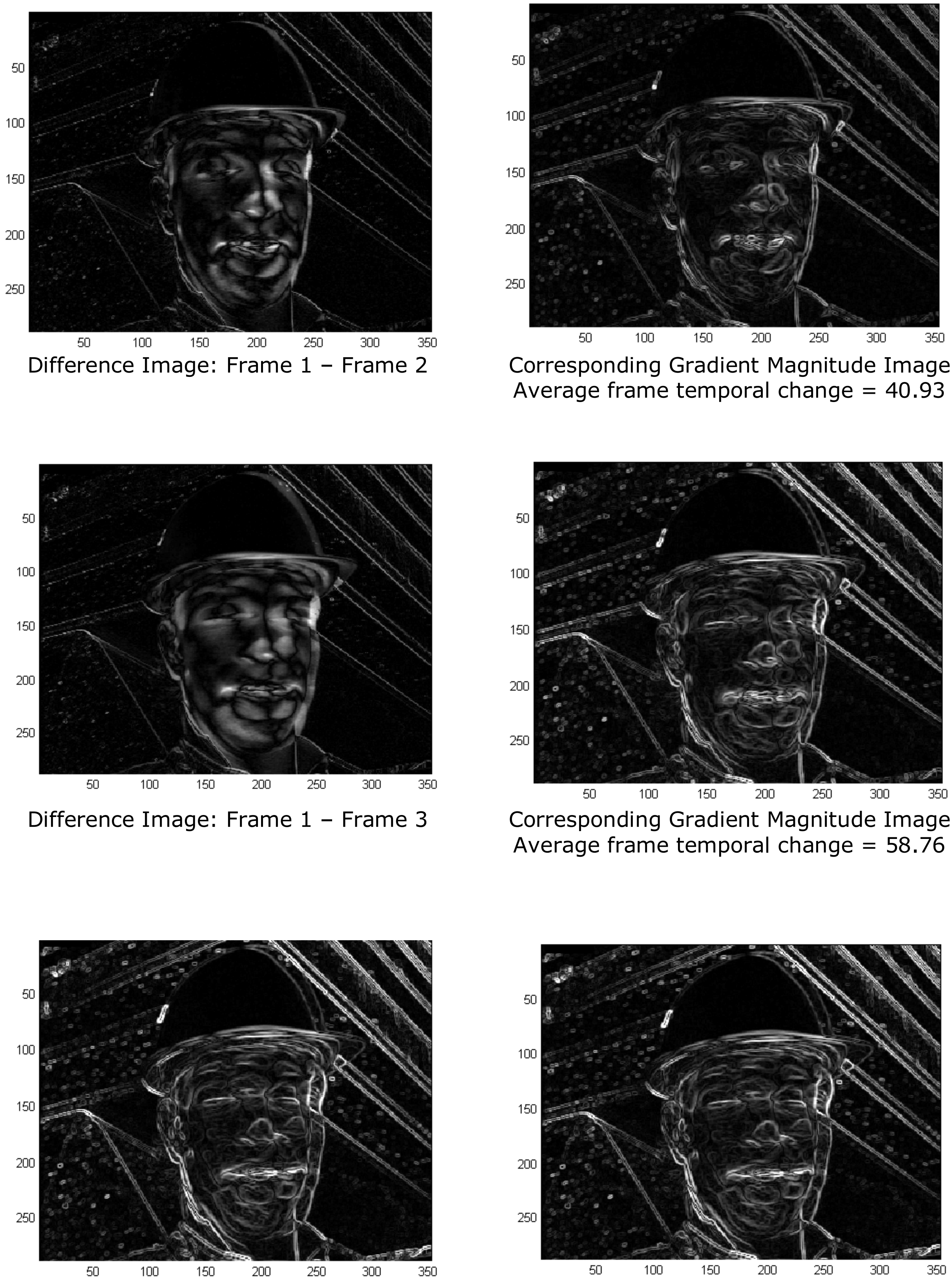
Frame 1

Frame 2

Frame 3

Frame 4

Figure 6-2: Original frames from Foreman sequence



Difference Image: Frame 1 – Frame 2

Corresponding Gradient Magnitude Image
Average frame temporal change = 40.93

Difference Image: Frame 1 – Frame 3

Corresponding Gradient Magnitude Image
Average frame temporal change = 58.76

Difference Image: Frame 1 – Frame 4

Corresponding Gradient Magnitude Image
Average frame temporal change = 63.91

Figure 6-3: Difference images and the corresponding gradient magnitude image for the Foreman sequence

6.3 Sequence Activity measure

Mean opinion score is a subjective score for the entire sequence. Therefore, video content of the whole sequence must be considered during slope parameter estimation. In this research work, the term 'sequence activity' is used to represent a measure of sensitivity of the sequence to visual distortion and is derived from sequence content based on visual masking information.

The sequence activity measure will be used to estimate the slope (k_s) of the regression line in the MOSp metric for each sequence. The sequence activity measure is a value between [0,1] and indicates the sensitiveness of the sequence to visible distortions. The activity measure is calculated at macroblock level first. The activity of a frame is calculated as the average of activities of all the macroblocks in the frame. The sequence activity A_{seq} is the average value of activities of all the frames and is given as:

$$A_{seq} = \frac{1}{T} * \sum_{j=1}^{T} \left(\frac{\sum_{i=1}^{P} A_{i,j}}{P} \right) \tag{45}$$

Where 'i' is the macroblock number and P is the total number of macroblocks in a video frame. 'j' is the frame number and T is the total number of frames in the video sequence.

6.4 Slope estimation from video content

From the video quality experiments in section 5.2, it has been noted that one of the main factors contributing to subjective quality is the sensitivity of video content to visible distortions. In section 6.3, methods of quantifying sequence content from spatial texture information and temporal change information were described. The visibility of distortions in video may depend on the individual contribution or a combined contribution of the above mentioned features. This section investigates that relationship. Therefore, this section presents the slope estimation for MOSp metric using the following video content measures:

1.) Spatial texture information

2.) Combination of spatial texture and temporal change information

These are described in detail in this section.

6.4.1 Slope estimation using spatial texture

The sensitivity of sequence content to the visibility of distortions may be measured using spatial texture masking. As explained in section 6.2.1, the spatial-texture information of each macroblock (STI_{MB}) is computed as the average edge strength of all the pixels in that macroblock. The average edge strength is used as a measure of sequence activity because it is hypothesised that highly textured regions with larger average edge strengths are more tolerant to visual distortions than smoother regions with lower average edge strengths due to the spatial texture masking effect. The sequence activity is calculated from the average edge strength of all the macroblocks in the sequence using equation (42).

The relation between slope and the sequence activity is acquired using the eight training sequences mentioned in section 5.2. The 'data points' in Figure 6-4 are the slopes of the MSE versus MOS curves of these eight training sequences. The relation between slope and sequence activity is derived using the exponential fit as:

$$k_s = 0.03585 * \exp(-0.02439 * A_{seq}) \qquad (46)$$

k_s is the estimated slope and A_{seq} is activity of the sequence derived from its spatial texture information using equation (42).

Equation (46) is the curve fit plotted as the dotted line in Figure 6-4. It is clear from the graph that (46) is a good prediction of slope k_s. The goodness of this fit was measured using R-squared value as 96.2%.

From Figure 6-4 it can be observed that low textured sequences such as the Carphone sequence with sequence activity of 30.72 produce steeper regression lines in the MSE versus MOS graph. Highly textured sequences such as the Mobile sequence with sequence activity of 111.49 have shallower regression lines. This indicates that in low-activity sequences, a small change in MSE leads to a larger change in MOS when compared to high-activity sequences for the same amount of change in MSE.

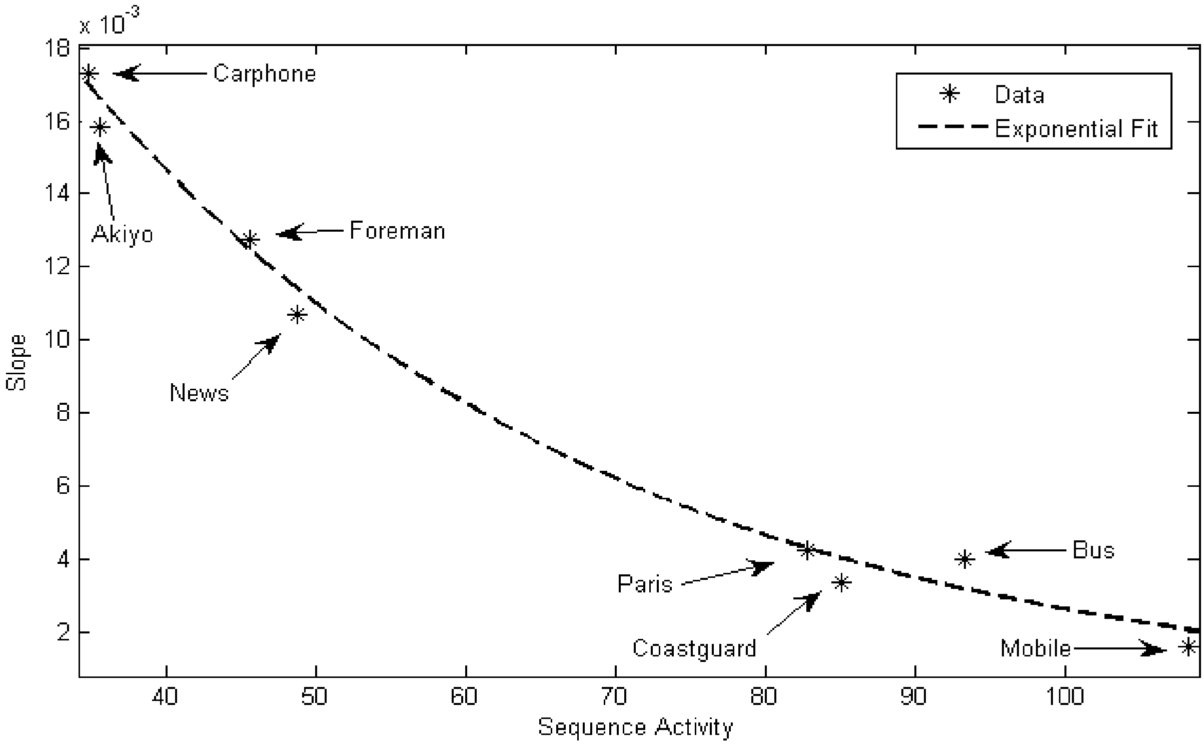


Figure 6-4: Slope estimation from spatial texture information

6.4.2 Slope estimation using spatial texture and temporal change

The activity of a macroblock (A_{mb}) may be obtained from the spatial-texture information and the temporal information of the macroblock as:

$$A_{mb} = \max(STI_{MB}, TI_{MB}) \quad (47)$$

where STI_{MB} is the spatial texture information derived from equation (1), TI_{MB} is the temporal change information derived from equation (44).

Equation (47) is used as a measure of macroblock activity because both spatial texture and temporal change contribute to the masking of distortions. Due to the complex nature of the human visual system, there is very little evidence in literature of the combined effect of spatial texture and temporal information on human perception. Also, it was observed from experiments that combining spatial-texture measure (STI_{mb}) and temporal change measure (TI_{mb}) into a single value may obscure the contribution of each element to masking the distortion. Hence the maximum of the two measures is considered as the activity of a macroblock. The relation between slope and the sequence activity is acquired using the eight training sequences mentioned in section 5.2. The 'data points' in Figure 6-5 are the slopes of the MSE versus MOS curves of these eight training sequences. We derive the relation between slope and sequence activity using the exponential fit as:

$$k_s = 0.03697 * \exp(-0.02236 * A_{seq}) \qquad (48)$$

k_s is the estimated slope and A_{seq} is activity of the sequence derived from its spatial and temporal masking information using equation (48). Figure 6-5 shows the exponential curve fitted to the data points. R-squared is 91.27%. From Figure 6-5 it can be observed that low activity sequences which have low to medium amount of detail and motion such as the Carphone sequence with sequence activity of 34.93 produce steeper regression lines in the MSE versus MOS graph. High activity sequences such as the Bus sequence with sequence activity of 123.92 have shallower regression lines.

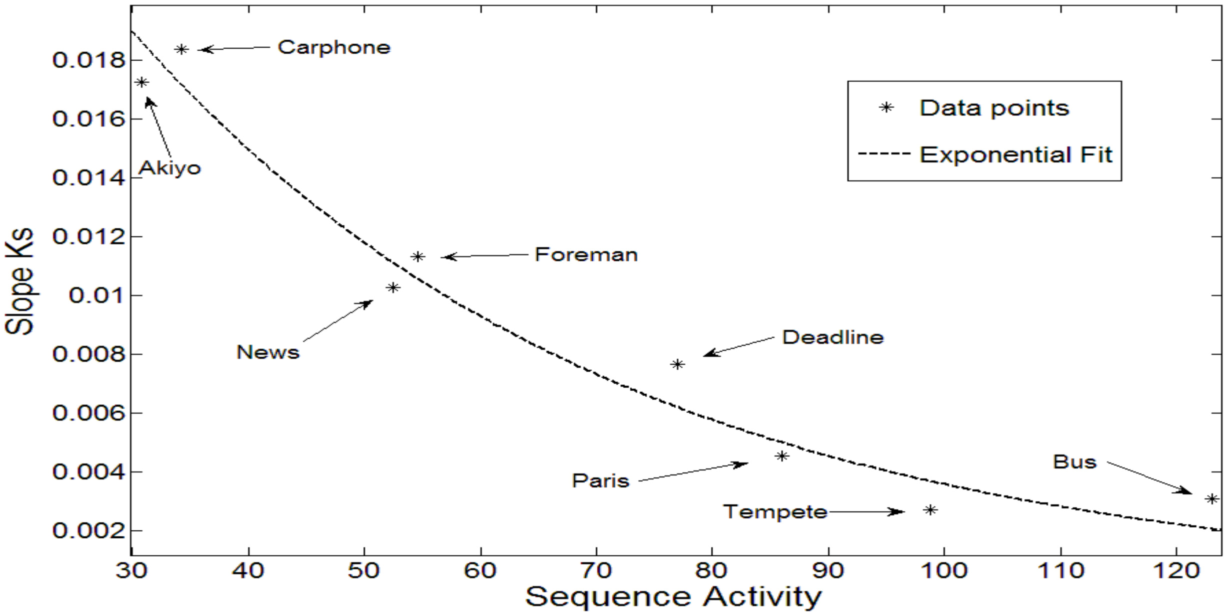


Figure 6-5. Graph showing relation between slope and sequence activity derived from spatial texture and temporal change information

6.5 Macroblock, frame and sequence level quality estimation using MOSp metric

During subjective evaluation of video quality by human observers, a judgement is made based on the overall quality of the sequence under test. Video sequences compressed at low bit-rates may have good picture quality in some parts of the sequence while other parts may have poor picture quality. The overall sequence quality rating reflects the quality of the entire video sequence, at least for short sequences [43]. Hence the use of combined average of MOSp is proposed for all the macroblocks in a frame as the frame-level quality measure and the average of MOSp of all the frames in the sequence as the overall quality measure of the video sequence.

Quality is first evaluated at macroblock-level, then combined into frame-level quality and finally averaged into a single valued sequence-level quality measure. The proposed metric computes the predicted subjective quality (MOSp) of each macroblock in the processed frame. The activity of every macroblock is calculated using methods described in section 6.2 in order to determine the slope k_{mb} using one of the approaches in section 6.4. The MSE between macroblocks of the original and reconstructed compressed macroblocks is computed using equations (37-40) in chapter 5. The MOSp for each macroblock is computed as:

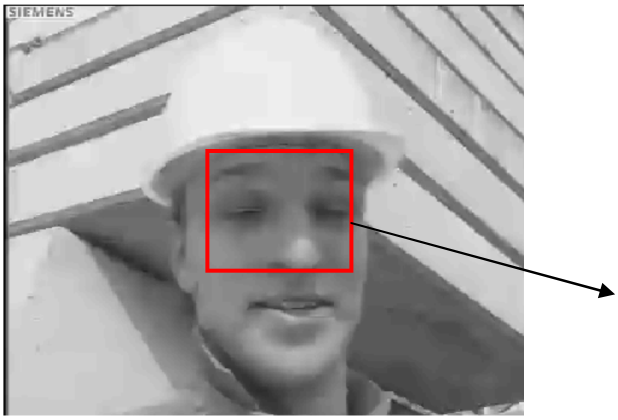
$$MOSp_{mb} = 1 - k_{mb}(MSE_{mb}) \qquad (49)$$

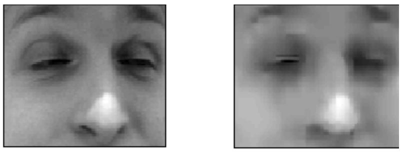
Figures 6-6 and 6-7 give an illustration of the hypothesis behind: (i) the prediction of perceived quality using MOSp at macroblock-level and (ii) the weighting of MSE based on the visual sensitivity of the video content. Figures 6-6 and 6-7 are video frames from the Foreman and Bus sequences compressed at QP = 36. A region in each frame has been selected for analysis purposes. It is a group of 5x5 macroblocks indicated by the red box in Figures 6-6 and 6-7. The corresponding MSE and MOSp values of these macroblocks are given. To compare the performance of the MOSp metric calculated using slope estimation methods described in sections 6.4.1 and 6.4.2, the MOSp values obtained using both the methods for the two 5x5 regions are presented. From the figures, the following observations can be made:

i. The MSE values of the 5x5 region in Bus sequence (Figure 6-7) are higher than those of the Foreman sequence (Figure 6-6). However, the visual quality of the 5x5 compressed region in the Foreman sequence is worse than that of the bus sequence. There is significant loss of detail and contrast in the facial region of the Foreman frame compared to the 5x5 uncompressed region of the Foreman frame.

ii. The average MOSp value of the 5x5 macroblock region obtained using both the slope estimation methods are presented . The average MOSp values of the Foreman region obtained using both the methods are lower than the Bus region although the average MSE of the Bus region is higher. This is because the MOSp metric is designed to identify regions with low texture and motion as being more sensitive to visible distortions when compared to regions with high texture and motion. Therefore, the distortions in low texture and low motion areas, such as the facial regions in the Foreman frame, produce lower MOSp score to indicate lower perceived quality. High texture/high motion regions, such as Figure 6-7, are more tolerable to visible distortions. Hence the MOSp values are higher in the macroblock of Figure 6-7 although the MSE values are comparatively higher than Figure 6-6.

This example demonstrates that the MOSp metric is a more effective predictor of subjective quality than MSE, for these selected 5x5 macroblock regions from the Foreman and Bus sequences.





22.29	17.06	4.33	8.07	13.9
24.58	24.4	9.8	15.74	33.06
29.29	18.42	9.31	17.25	23.98
12.02	6.71	16.4	25.43	10.59
4.32	9.95	26.68	23.85	21.77

Macroblock MSE values, average MSE = 17.18

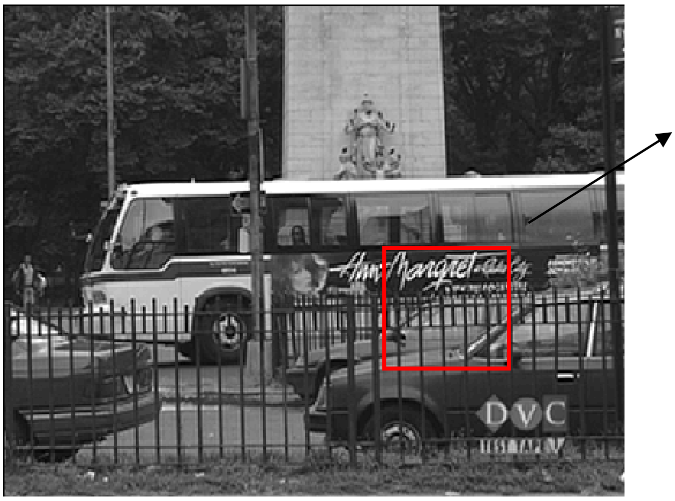
0.6	0.54	0.65	0.68	0.64
0.38	0.5	0.72	0.65	0.38
0.49	0.66	0.65	0.64	0.58
0.57	0.58	0.59	0.58	0.61
0.66	0.59	0.49	0.61	0.42

Macroblock MOSp values using spatial texture only, average MOSp = 0.5784

0.61	0.5	0.74	0.76	0.73
0.42	0.56	0.81	0.62	0.44
0.53	0.69	0.69	0.67	0.58
0.67	0.6	0.71	0.59	0.64
0.71	0.63	0.58	0.66	0.53

Macroblock MOSp values using spatial texture and temporal change, average MOSp = 0.6268

Figure 6-6: Video frame from Foreman sequence compressed at QP = 36. Note: MOSp = [0,1] where 0 = bad and 1 = excellent picture quality.



Original region

Compressed region

111.4	128.2	195.3	147.9	150.9
115.3	128.7	162.3	173.4	190.5
163.1	195.1	116.9	116.8	125.4
75.3	69.1	103	105.1	120.3
65.9	45.9	42.8	45	52.3

Macroblock MSE values, average MSE = 117.84

0.896	0.898	0.895	0.887	0.846
0.854	0.926	0.987	0.993	0.974
0.898	0.914	0.899	0.893	0.956
0.882	0.837	0.893	0.897	0.893
0.898	0.834	0.836	0.883	0.797

Macroblock MOSp values using spatial texture only, average MOSp = 0.895

0.910	0.913	0.916	0.894	0.872
0.878	0.937	0.992	0.997	0.979
0.902	0.893	0.907	0.896	0.961
0.893	0.839	0.908	0.903	0.925
0.902	0.846	0.851	0.895	0.821

Macroblock MOSp values using spatial texture and temporal change, average MOSp = 0.904

Figure 6-7: Video frames from Bus sequences compressed at QP = 36. Note: MOSp = [0,1] where 0 = bad and 1 = excellent picture quality.

6.6 Metric performance evaluation

Performance of an objective video quality metric depends on whether the metric is in close agreement with subjective results (MOS), whether it can be calculated automatically in real time and whether it has computational simplicity. In this section, the performance of the MOSp metric is evaluated based on its correlation with subjective test results and comparison with existing full reference objective video quality metrics. Section 6.6.1 describes a subjective experiment conducted to obtain subjective scores (MOS) for 32 multimedia video sequences of varying video content. The MOS scores are used for benchmarking the performance of the objective video quality metrics. Section 6.6.2 provides details of the subjective evaluation process conducted to obtain the subjective scores (MOS) for the test sequences. Following the performance evaluation methods adopted by the video quality experts group (VQEG), three evaluation metrics have been used to benchmark performance of the MOSp metric: Pearson Correlation, Spearman's rank correlation and the outliers ratio. Details of the software implementations of these evaluation metrics and the correlation results of the MOSp and the five other metrics with respect to subjective results (MOS) are presented in section 6.6.3. Visual representation of the correlation between subjective quality and the predicted quality may be presented using 'scatter plots'. Hence, scatter plots of MOSp and the other metrics are given in section 6.6.4. The processing time of video quality metrics is important in real time video applications. Hence, the processing times for the MOSp metric and other metrics are evaluated in section 6.6.5. Section 6.6.6 gives performance comparison of the two methods for calculating the MOSp metric based on sequence content parameters described in section 6.2. This comparison is performed to investigate the advantages and limitations of the MOSp metric.

6.6.1 Test Material

For evaluating the performance of the metric, 32 multimedia video sequences of CIF resolution format were used in order to include a wide variety of video content. These include:

- 8 training sequences used in modelling the MOSp metric. These sequences include: Carphone, Foreman, News, Bus, Coastguard, Deadline, Paris and Tempete.
- 8 video sequences popularly used in the video compression research community. These include: Salesman, Mother and Daughter, Container, Grasses, Mobile, Husky, Akiyo and Sign Irene.
- 16 video sequences from the VQEG data set of multimedia sequences [73].

These sequences were chosen to represent a wide variety of video content. The VQEG test data set is very widely used in the video quality measurement research community. The test sequences range from low motion and low detail 'head and shoulder' scenes such as Mother and Daughter to high motion and high detail scenes such as the Husky sequence. The sequences were in c*ommon intermediate format (CIF)* resolution to represent multimedia sequences.

The sequences were in 4:2:0 YCbCr format, 10 seconds in duration and were coded using the H264/AVC compression standard. Each test sequence was compressed at a wide range of bitrates using a fixed set of quantisation parameter (QP) values, QP = {6, 26, 34, 36, 38, 40, 42, 45}. There is a closer spacing between the chosen QP values in the 'useful' medium- to low-bitrate range for multimedia sequences (QP=34 to QP =42).

6.6.2 Coding parameters

The test video sequences used in this experiment were compressed using the H.264/AVC JM reference software (version 12.1) available at http://iphome.hhi.de/suehring/tml/, with the following parameters:

- Profile used is Main profile.
- Level IDC setting is set 4.0
- Frame Skip: no frames were skipped
- Number of reference frames for Inter motion search is set to 5.
- Number of B-pictures used = 0
- Entropy coding method is set to CABAC.
- RD-Optimisation: High complexity mode
- Rate Control: DISABLED to allow the use of fixed QP.
- Slice QP: QPISlice and QPPSlice parameters used and both set to the same value as the sequence QP.

Note that the coding parameters used in the MOSp metric evaluation is identical to those used to produce the training sequences in section 5.2.

6.6.3 Subjective evaluation

The subjective evaluation process involved 30 non-expert evaluators and followed the guidelines in ITU-T P.910 Recommendation [6]. The single stimulus impairment scale (SSIS) evaluation method was used. A 5-grade scale from 0 to 1 was used to rate the quality of the test video sequences where 0=bad, 0.25=poor, 0.5=fair, 0.75=good and 1=excellent. The experimental procedure was carried out as described in section 5.2. The MOS for a sequence was calculated as the average of all scores obtained for the sequence compressed at a certain QP. The mean 95% confidence interval for the whole data set was 0.0473. The sequence MSE was calculated as the mean of the sum of squared differences (SSD) between the luminance pixels of the original and the constructed compressed video sequence.

6.6.4 Performance comparison of MOSp with existing metrics

Following the performance evaluation methods adopted by the video quality experts group (VQEG), three evaluation metrics have been used to benchmark performance of the MOSp metric: Pearson Correlation, Spearman's rank correlation and the outliers ratio. The Pearson's correlation coefficient is used to measure the prediction accuracy of the MOSp metric. The Spearman's correlation is used to measure the prediction consistency of the MOSp metric. Both Pearson and Spearman correlation value range between [0,1] where 1 indicates very high correlation between predicted measures and the subjective ratings and 0 indicates no correlation. The correlation coefficients and outliers ratio between subjective score (MOS) and the corresponding objective measure were calculated as described in Chapter 3 (section 3.7). Experimental results are illustrated in Table 6-1, Table 6-2 and Table 6-3. Tables 6-1 and 6-2 give the Pearson's correlation between the estimated and actual perceptual quality for MOSp and the five popular metrics using the test sequences. It is noted that the non-training sequences have not been used in developing the proposed metric. For each sequence in the table, the highest correlation coefficient is highlighted in bold font. From the table, the following observations can be made:

- The MOSp metric produces high correlation (>90%) with subjective ratings for a variety of video sequences ranging from low activity such as Akiyo and News, to high activity sequences such as Bus, Mobile and Coastguard. The metric also produces good results with sequences which are a combination of both low-activity and high-activity scenes such as Foreman and Tempete sequences.
- The PSNRplus and NTIAVQM metrics also produce high correlations (>90%) with MOS for most sequences. However, PSNRplus has a very large computational cost due to the need for encoding each video sequence three times in order to make a quality estimation. This makes PSNRplus unsuitable for real time multimedia applications.
- The MOSp metric has higher correlation with subjective ratings when compared to NTIAVQM for 28 out of the 32 test sequences.
- The two methods of calculating MOSp from various video content are also presented in Table 6-1. It can be noted that the MOSp based on spatial texture produces higher correlation with MOS compared to the MOSp metric based on spatial texture and temporal information for 21 out of 32 sequences. This could be because the slope estimation model based on

spatial texture is more accurate than the model which uses a combination of spatial texture and temporal change.

Table 6-1. Pearson Correlation between popular metrics and MOS for 16 test sequences

Sequence	PSNR	VSSIM	PSNRplus	NTIAVQM	Yonsei	MOSp based on spatial texture	MOSp based on spatial texture and temporal
Training							
Foreman [73]	0.775	0.794	0.958	0.965	0.872	**0.988**	0.997
Carphone[73]	0.672	0.849	0.957	0.933	0.868	0.943	**0.968**
Bus [73]	0.719	0.838	0.856	0.962	0.896	0.988	**0.989**
Deadline [73]	0.773	0.834	0.927	**0.985**	0.847	0.924	**0.939**
News [73]	0.849	0.771	0.931	0.916	0.863	0.915	0.944
Paris [73]	0.809	0.797	0.948	0.968	0.870	0.942	0.964
Tempete [73]	0.712	0.785	0.890	**0.975**	0.917	**0.97**	0.964
Akiyo [73]	0.752	0.811	0.932	0.908	0.939	0.901	**0.986**
Non-training							
Husky [73]	0.702	0.775	0.901	**0.961**	0.841	**0.925**	0.921
Salesman[73]	0.801	0.818	0.919	0.948	0.941	0.874	0.927
Container[73]	0.795	0.825	0.953	0.976	0.876	0.975	**0.989**
Grasses [73]	0.774	0.727	0.879	0.963	0.916	0.962	**0.994**
Mobile [73]	0.697	0.713	0.919	0.949	0.884	0.947	**0.984**
Sign Irene [73]	0.763	0.746	0.955	0.949	0.945	0.892	**0.957**
Mother & daughter[73]	0.725	0.764	0.924	0.935	0.869	0.93	**0.952**
Coastguard [73]	0.762	0.724	0.883	0.969	0.866	0.991	**0.995**

Table 6-2. Pearson Correlation between popular metrics and MOS for 16 CIF VQEG Multimedia test dataset

Sequence	**PSNR**	**VSSIM**	**PSNRplus**	**NTIAVQM**	**Yonsei**	**MOSp based on spatial texture**	**MOSp based on spatial texture and temporal**
ITU_SRC_MobileCalendar_cif [72]	0.697	0.713	0.919	0.949	0.884	**0.973**	0.901
ITU_SRC_Football_cif [72]	769	0.772	0.887	0.942	0.914	**0.962**	0.945
ITU_SRC_FlowerGarden_cif [72]	0.72	0.867	0.892	0.945	0.942	**0.96**	0.912
ITU_SRC_Stephen_cif [72]	0.774	0.821	0.957	0.959	0.89	**0.913**	0.875
ANSI_SRC_Crew_cif [72]	0.703	0.729	0.856	0.975	0.857	**0.93**	0.891
ANSI_SRC_MissAmerica_cif [72]	0.754	0.766	0.927	0.916	0.923	**0.981**	0.915
CBC_SRC_BetesPasBetes_cif [72]	0.81	0.768	0.931	**0.952**	0.941	**0.916**	0.904
ANSI_SRC_washdc_cif [72]	0.683	0.722	0.948	0.949	0.927	**0.955**	0.899
ANSI_SRC_vtc2mp_cif [72]	0.762	0.819	0.89	0.918	0.872	**0.917**	0.868
ANSI_SRC_vtc1nw_cif [72]	0.775	0.827	0.861	0.905	0.861	**0.932**	0.895
ANSI_SRC_5row1_cif [72]	0.751	0.824	0.952	0.941	0.896	**0.949**	0.939
ITU_SRC_Cheerleaders_cif [72]	0.724	0.816	0.947	0.952	0.847	**0.957**	0.944
ANSI_SRC_boblec_cif [72]	0.699	0.794	0.902	0.964	0.87	**0.984**	0.964
CRC_SRC_Redflower_cif [72]	0.736	0.818	0.873	0.952	0.872	**0.969**	0.947
ANSI_SRC_vtc2zm_cif [72]	0.71	0.829	0.925	0.928	0.907	0.895	**0.919**
CBC_SRC_BetesPasBetes_cif [72]	0.752	0.801	0.937	0.931	0.893	**0.951**	0.906

Table 6-3 illustrates the overall performance of MOSp and the popular quality metrics when compared with the actual subjective results (MOS). The table consists of the Pearson correlation, Spearman correlation and outliers ratio for each metric when all the 32 test video sequences mentioned in Tables 6-1 and 6-2 are included.

Table 6-3. Overall Comparison of MOSp with popular metrics

Metric	Pearson Correlation	Spearman Correlation	Outliers Ratio
PSNR	0.696	0.711	0.857
VSSIM	0.723	0.779	0.797
PSNRplus	0.886	**0.959**	0.596
NTIA VQM	0.901	0.913	0.516
Yonsei University	0.863	0.878	0.628
MOSp based on spatial texture only	**0.942**	0.948	**0.410**
MOSp based on spatial texture and temporal change	0.926	0.938	0.48

The overall performance of objective quality metrics is important because it reflects the accuracy of prediction of perceived quality for a variety of sequences ranging from low activity sequences such as Akiyo to very high activity sequences such as Husky.

Pearson correlation measures the ability of an objective quality metric to predict subjective ratings with minimum average error assuming a linear correlation between the two quality metrics. The overall Pearson correlation for MOSp based on spatial texture is 0.942, which are the highest amongst the metrics compared in Table 6-3. The closest metric to this performance is the MOSp metric based on spatial texture and temporal change with Pearson correlation of 0.926.

Spearman correlation determines how well the estimated result reflects an increase or decrease in the actual subjective result regardless of the magnitude of increase or decrease. It also makes no assumptions about the shape of the relationship between the predicted data and the subjective ratings. From table 6-3, it can be observed that PSNRplus produces the highest Spearman correlation with MOS compared to other metrics. The closest to this performance is the MOSp metric based on spatial texture. Although PSNRplus produces improved results compared to MOSp, it requires every sequence to be coded three times in order to obtain the two additional instances for making quality estimation. Therefore, PSNRplus may have limited applications in real-time video quality estimation as explained further in section 6.6.5 when computation times of various metrics are compared.

The outliers ratio (OR) measures prediction consistency and is a ratio of “false” scores to the total number of scores [82]. The "false" scores are the scores that lie outside the interval [MOS-2σ, MOS+2σ]. A lower outliers ratio indicates better metric performance. The MOSp metric based on spatial texture has the lowest outliers ratio compared to other metrics in Table 6-3.

MOSp metric based on spatial texture performs better than the MOSp metric based on spatial texture and temporal changes. This may be due the following two reasons:

1. Spatial texture masking has more effect on the visibility of distortion when compared to temporal masking.
2. The slope estimation model based on spatial texture only is more accurate compared to the model based on spatial texture and temporal change information.

The overall performance of the MOSp metric demonstrates that it produces high correlation with MOS (>90%) for a variety of video content compressed to a wide range of bitrates.

6.6.4 Scatter plots

Figures 6-8 and 6-9 show the scatter plots of subjective ratings (MOS) versus the proposed metric (MOSp) and other five popular metrics. The scatter plots contain all the 32 test sequences.

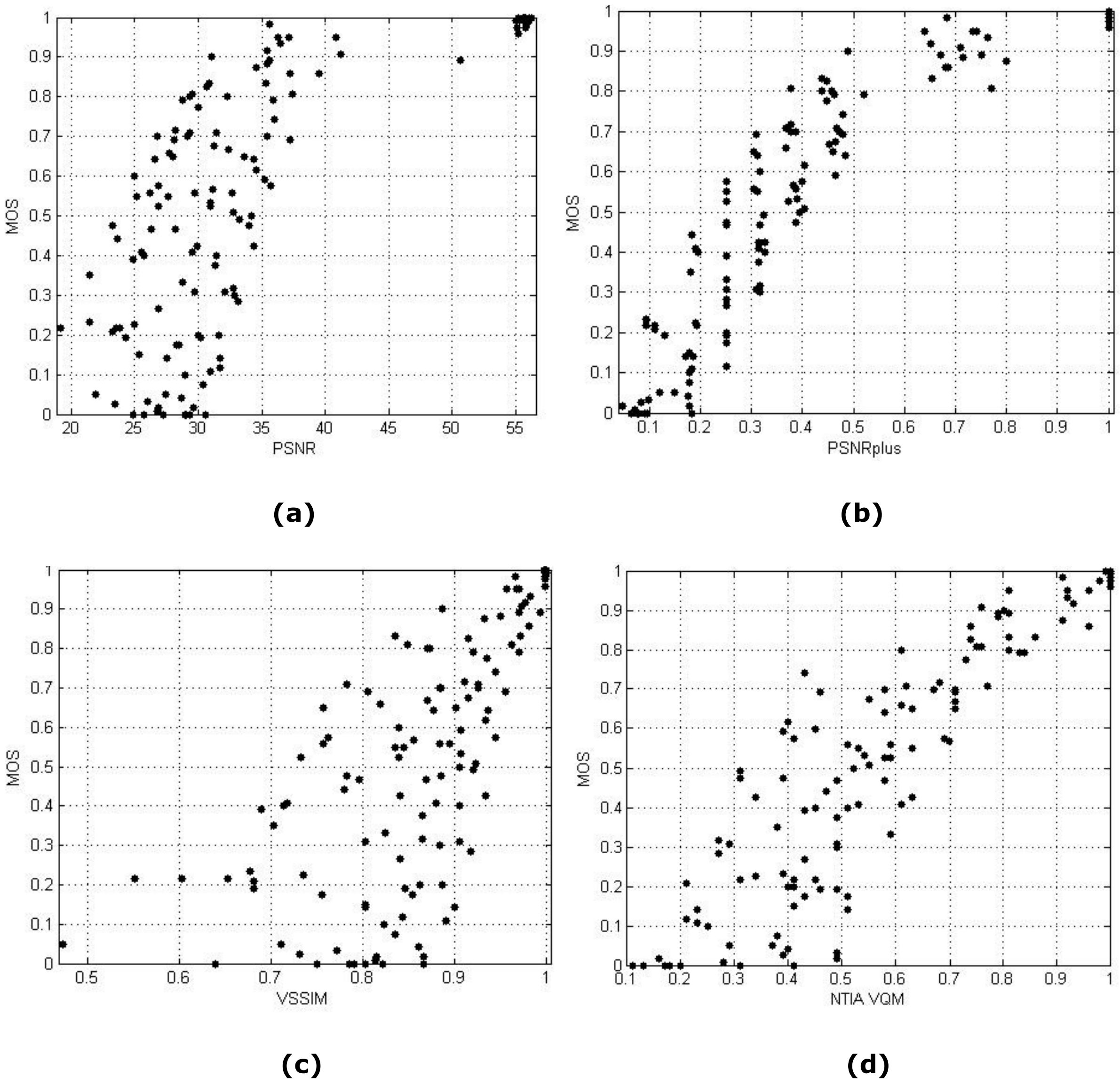
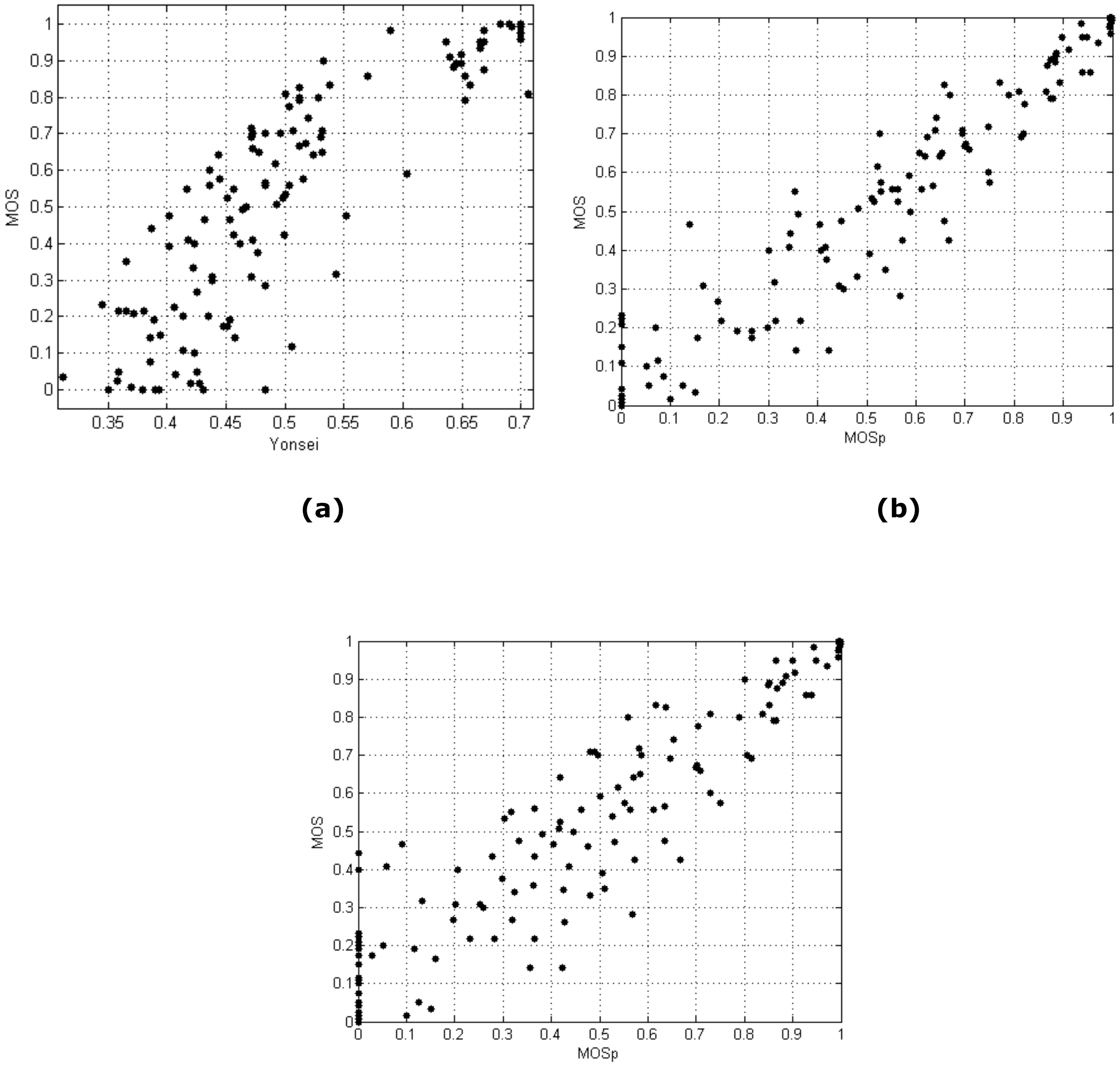


Figure 6-8: Scatter of subjective ratings (MOS) versus (a)PSNR, (b)PSNRplus, (c)VSSIM and (d)NTIA VQM for all the 32 CIF test sequences compressed using H264/AVC.



(a) **(b)**

(c)

Figure 6-9: Scatter plot of subjective ratings (MOS) versus (a)Yonsei University metric and (b)MOSp based on spatial information only, (c)MOSp based on spatial texture and temporal change information, for all the 32 CIF test sequences compressed using H264/AVC.

6.6.5 Processing time

Processing time for quality measurement systems is important in real-time video applications. This section investigates the computational complexity of the various video quality metrics discussed in this chapter. Table 6-4 shows the percentage increase in coding time when compared to the coding time of the reference H.264/AVC software codec called the JM software [83]. All the quality metrics were implemented into the software codec for performance evaluation. The coding time is taken for JM software to encode the 'Paris' sequence of CIF resolution with 150 frames. The implementations were carried out on a 1.5 GHz, 512 MB RAM desktop PC. The NTIA/ITS VQM algorithm is not included in the comparison of Table 6-3 because the software implementation of this metric [60] requires user input during the measurement process, whereas the software codec implementations of the metrics listed in the speed comparison Table 6-4 are fully automatic. It can be observed from the table that the PSNRplus metric consumes the most processing time due to the requirement of encoding each video sequence three times to make a quality estimation. The second most computationally expensive metric is the VSSIM metric which requires each video frame to be spatio-temporally aligned and perceptual features to be extracted from a large set of parameters in order to quantify video quality. The metric requiring least processing time is the PSNR metric but it produces poor correlation with MOS as noted from the evaluation results in Tabled 6-1 and 6-2. The second most efficient quality metric in terms of processing time is the Yonsei University metric which produces around 86% correlation with subjective results.

The processing times for the MOSp metric derived from sequence content using the two different methods are presented in Table 6-3. The MOSp metric based on spatial texture information requires the least processing time (4.9%) which is less than the MOSp metric based on both spatial texture and temporal change. From Table 6-4 it is evident that MOSp is faster than VSSIM and PSNRplus but requires more processing time compared to PSNR and the Yonsei University metric. However, the MOSp metric gives improved correlation with subjective quality compared to other metrics evaluated in this work.

Table 6-4. Running speed of MOSp and popular metrics

Metric	Increase in coding time (%)
PSNR	Negligible
VSSIM	32.7%
PSNRplus	200.18%
Yonsei University	4.2%
MOSp based on spatial texture	4.9%
MOSp based on spatial texture and temporal change information	7.7%

The processing times for MOSp metric presented in Table 6-4 includes all the evaluation processes including spatial edge strength calculation and temporal change calculations performed on each frame. For the purposes of this research, the algorithms for all the metrics presented in Table 6-4 have been implemented in the C programming language. The coding time is the time taken by H264 compression algorithm to process a video sequence. The percentage increase in coding time presented in Table 6-4 is calculated from the average coding times recorded using five repetitions.

The MOSp metric predicts perceptual quality using MSE which is a widely used quality measure in video compression algorithms. The only additional requirement for MOSp calculation is the slope estimation using one of the two methods using spatial texture and temporal change information. From Table 6-4, it is can be observed that quality estimation using MOSp metric increases computation time by 4.9% to 7.7% depending on the method of slope estimation. Hence the choice of slope estimation for calculating MOSp metric will depend on the application, the required prediction accuracy and the available computational resources.

6.7 Discussion

The MOSp metric estimates the mean opinion score (MOS) of compressed multimedia sequences using: (i) MSE between the original and compressed video sequences and (ii) video content of the original video sequence which is measured using features such as spatial texture and temporal change. Two methods of estimating the slope of the regression in the MOSp metric have been proposed using spatial texture and temporal change. The performance evaluation results indicate that the MOSp metric produces high correlation with MOS (>90%) with an increase in coding time between 4.9% to 7.7%. The metric correlates with subjective results better than popular metrics PSNR, PSNRplus, VSSIM, Yonsei and NTIAVQM metric.

Video content which influences the visibility of distortions may also include objects in the video which attract human attention depending on viewer interest and task in hand. These aspects of video content may also influence the relationship between MSE and MOS. From experiments in sections 5.2, it can be observed that sequences containing human figures tend to have steeper slope when compared to sequences without humans. Hence investigations are carried out in the following chapter to see if the MOSp metric can be improved by incorporating cognition-based factors such as the presence of humans in video sequences.

7 MOSp Metric Based On MSE, Video Content And Cognition Factors

7.1 Introduction

The MOSp metric described in Chapters 5 and 6 is designed to predict the MOS of compressed video using MSE and video content. It exploits the linear relationship between MSE and MOS for a sequence coded at several bitrates using the same coding algorithm. The slope of the regression line between MSE and MOS varies with video content. Masking effects mean that video content has an influence in 'enhancing' or 'hiding' video compression artefacts. Based on this phenomenon, the slope parameter of the MOSp metric was derived from spatial texture and temporal change information in Chapter 6. Performance results show that the MOSp metric produces very high correlation (>90%) with MOS and performs better than other popular metrics for a test dataset containing a wide variety of multimedia sequences compressed to a large range of bit rates using the H264/AVC encoder.

Video content such as objects in the video scene which attract human attention may also have effect on the visibility of distortions. These objects could relate to viewer interest, task in hand and prior knowledge [84]. Therefore, by considering these factors, it may be possible to extend the MOSp metric based on spatial texture and temporal change to incorporate cognition based factors which attract viewer attention. This chapter investigates the relationship between video content and the slope parameter of the regression line between MOS and MSE with a view to incorporate cognition factors into the existing MOSp metric.

The chapter is organised as follows: section 7.2 gives a description of cognition factors which may have an effect on the visibility of distortion including the presence of humans in video. Section 7.3 describes the experiment conducted to investigate the relationship between video content and the slope of the regression line. Based on this investigation, methods to automatically estimate the slope parameter from video content are given in section 7.4. Video quality measurement at macroblock, frame and sequence level using the MOSp metric is presented in section 7.5. Performance of the MOSp metric is evaluated in section 7.6 and finally, the performance results are discussed in section 7.7.

7.2 Cognition based factors affecting visual quality of video

Cognition based factors that attract human attention while watching video may be used to classify video content into foreground and background regions. These factors include objects or patterns in the video scene that are 'recognised' by the viewer based on viewer interest, prior knowledge or task-in-hand. Research has shown that presence of humans and particularly human faces in a scene attract visual attention [85]. In certain applications such as sign language, hand movements are equally important. Hence in general, skin colour can be used as a cognitive-driven factor, as it is an indicator of the presence of humans and human faces. Previous studies on the effects of artefacts on perceived quality [69] have found that distortions in foreground areas such as human faces caused lower subjective ratings while similar artefacts in the background areas went unnoticed. Therefore, objects in the video scene which attract viewer attention may contribute to enhancing or masking of visible distortions in compressed video and have effect on the slope of the regression between MSE and MOS.

Skin colour is a popular cognition-driven perceptual cue and has been proven to be an effective feature in many applications such as face detection and hand tracking [86,87]. Skin colour detection involves choosing an appropriate colour space and identifying a cluster associated with skin colour in this space. Pixels are then classified as skin if they belong to the skin cluster. YCbCr is a commonly used colour space for skin detection due to its ability to separate luminance and chrominance information and popularity in the image and video compression algorithms. Skin clusters are more compact in the YCbCr colour space compared to other colour spaces [88].

In this research, Hsu's nonlinear transform [89] of chroma in YCbCr colour space is used to classify pixels as skin. The transform exploits the nonlinear dependency of skin colour on luminance and hence overcomes the difficulty of detecting skin in changing lighting conditions. The transform converts the chroma components (Cb and Cr) of each pixel into functions of the luminance component (Y) as: $C_b^{'}(y)$ and

$C_r^{'}(y)$. The skin cluster in the transformed $C_b^{'}C_r^{'}$ colour space is modelled using an ellipse and is described as:

$$\frac{(x-e_{cx})^2}{a^2}+\frac{(y-e_{cy})^2}{b^2}=1 \tag{50}$$

$$\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} C_b^{'}-cx \\ C_r^{'}-cy \end{bmatrix} \tag{51}$$

where cx=109.38, cy=152.02, θ=2.53 radians, e_{cx}=1.60, e_{cy}=2.41, a=21.39 and b=14.03. A pixel is classified as skin if the $C_b^{'}$ and $C_r^{'}$values of the pixel lie on or with in the ellipse described above. The algorithm produces a rough segmentation of skin regions in a video frame.

Although skin detection using colour is popularly used for its computational simplicity and efficiency, it is known to produce false positives which include regions in the image which have similar colour to skin tone. To overcome this problem, a general approach to reducing false positives is to perform morphological filtering operations such as erosion and dilation to the detected regions. These filtering methods are described in detail in [89].

Figure 7-1 shows skin maps of four test sequences: Akiyo, Foreman, Sign Irene and Bus sequences. The maps represent macroblocks in the video frame which contain skin pixels. These frames have been produced after performing skin detection, erosion and dilation operations. Frames from Akiyo, Foreman and Sign Irene have approximately detected skin regions with no skin region in the Bus sequence.

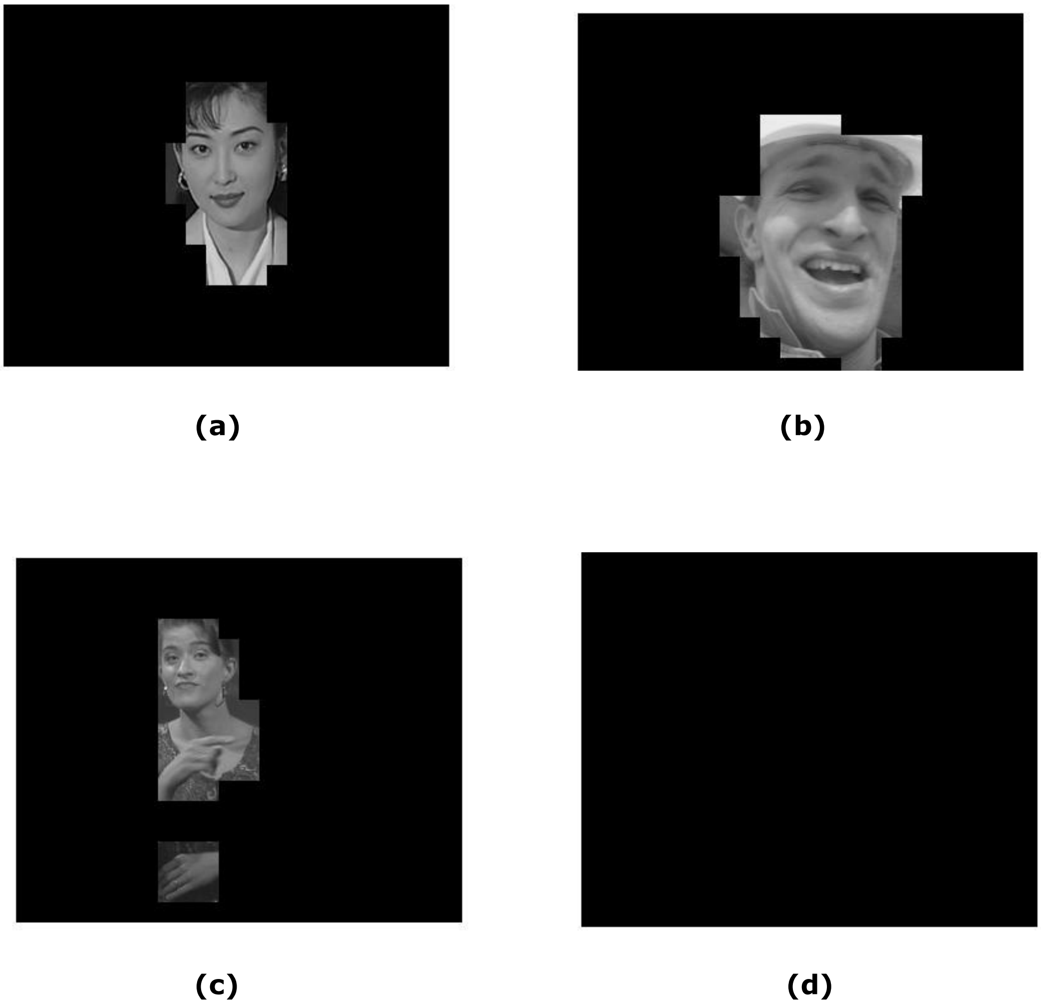


Figure 7-1: Skin maps of four test sequences. (a) Akiyo, (b) Foreman, (c) Sign Irene and (d) Bus

7.3 Slope estimation from video content and cognition factors

Video content can be quantified using spatial texture information, temporal change information and skin information. The visibility of distortions in video may depend on the individual contribution or a combined contribution of the above mentioned features. Hence, the sequence activity measure is calculated from the following feature combinations with a view to automatically estimate the slope parameter of the regression line from video content:

- spatial texture masking and skin information
- Spatial texture masking, temporal masking and skin information

7.3.1 Slope estimation using spatial texture and skin information

Sequence activity is calculated as the combined average of all the macroblock activities in the sequence. Macroblock activity obtained from the spatial-texture information and the skin colour information is calculated as:

$$A_{mb} = \begin{cases} 1 & SkinMacroblock \\ 1 - STI_{MB} & Otherwise \end{cases} \qquad (52)$$

Macroblock activity ranges between [0, 1] where 0 indicates low activity, such that the macroblock has high tolerance to visual distortion and 1 indicates high activity, such that distortions in the macroblock may be visible. According to (52), a macroblock containing skin pixels will have the maximum activity value of 1 indicating that it is most intolerable to distortion. For macroblocks which do not contain skin pixels, the activity is equal to 1-STI_{MB} where STI_{MB} is the average spatial edge strength of a macroblock with values ranging between 0 and 1 and is derived using equation (42) in chapter 6. A highly textured macroblock has a large STI_{MB} and a small macroblock activity value (1-STI_{MB}) due to its high tolerance to visual distortions. Conversely, a low textured macroblock will have a smaller STI_{MB} and a large activity value due to its high sensitivity to visual distortions.

In order to automatically calculate slope (k_s) from the spatial texture and skin colour information of the video sequence, a relation between slope k_s and sequence activity must be found. This relation is acquired using the seven training sequences: Foreman, Akiyo, News, Deadline, Husky, Bus and Coastguard. The 'data points' in Figure 7-2 are the slopes of the MSE versus MOS curves of seven training sequences. The relation between slope and sequence activity is derived using the exponential fit as:

$$k_s = 0.0001108 * \exp\left(5.705 * A_{seq}\right) \qquad (53)$$

k_s is the estimated slope and A_{seq} is activity of the sequence derived from its spatial texture masking and skin information. Equation (53) is the curve fit plotted as the dotted line in Figure 7-2. The goodness of the curve fit was measured using R-squared value as 98.74% indicating that equation (53) is a good estimation of the data points. From Figure 7-2, it can be observed that high-activity sequences such as the Akiyo sequence with sequence activity of 0.942 produce steeper regression lines in the MSE versus MOS graph. Low activity sequences such as Bus and Husky have shallower regression lines. This indicates that in high-activity sequences, a small change in MSE leads to a larger change in MOS when compared to low-activity sequences for the same amount of change in MSE.

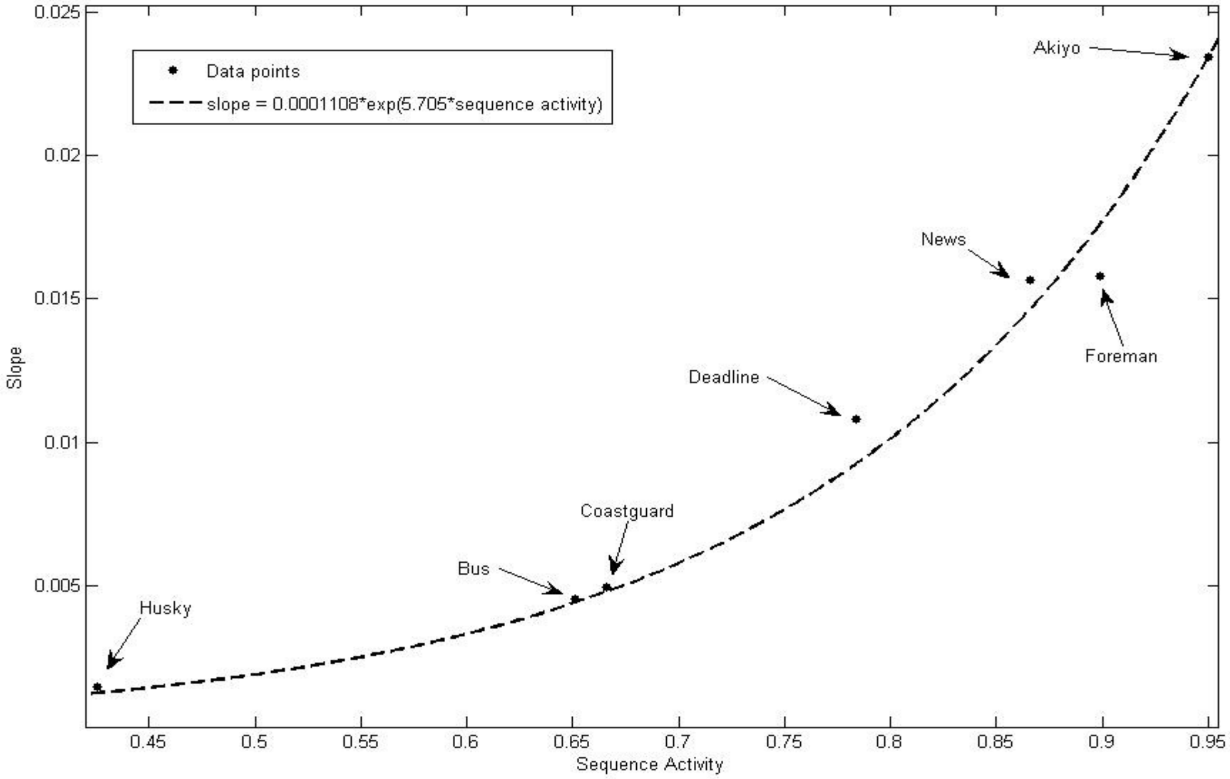


Figure 7-2. Graph showing relation between slope and sequence activity derived from spatial texture and skin information

7.3.2 Slope estimation using spatial texture, temporal change and skin information.

Macroblock activity may also be obtained from the spatial-texture, temporal change and the skin colour information of the macroblock as:

$$A_{mb} = \begin{cases} 1 & SkinMacrob\ lock \\ 1 - \max\left(STI_{MB}, TI_{MB}\right) & Otherwise \end{cases} \qquad (54)$$

Macroblock activity ranges between [0, 1] where 0 indicates the macroblock has high tolerance to visual distortion and 1 indicates that distortions in the macroblock may be visible. According to equation (54) a macroblock containing skin pixels will have the maximum activity value of 1 indicating that it is most intolerable to distortion. For macroblocks which do not contain skin pixels, the activity is equal to 1-$\max(STI_{mb}, TI_{mb})$ where STI_{mb} is the average spatial edge strength of a macroblock with values ranging between 0 and 1 derived from equation (42) in chapter 6, TI_{mb} is the temporal change information and is derived from equation (44) in chapter 6. A macroblock which is highly textured or undergoes larges temporal change has a large $\max(STI_{mb}, TI_{mb})$ value due to its high tolerance to visual distortions. Conversely, a macroblock with low detail or motion will have a smaller $\max(STI_{mb}, TI_{mb})$ value due to its high sensitivity to visual distortions.

This relation is acquired using the eight training sequences mentioned in section 6.4.1. The relation between slope and sequence activity is derived using the exponential fit as:

$$k_s = 0.0411 * \exp\left(-0.01583 * A_{seq}\right) \qquad (55)$$

k_s is the estimated slope and A_{seq} is activity of the sequence derived from its spatial texture masking and skin information.

7.4 Macroblock, frame and sequence level quality estimation using the MOSp metric

As explained in section 6.5, MOS is a subjective score made while considering the overall quality of the sequence under test, at least for short sequences [43]. Hence the use of combined average of MOSp is proposed for all the macroblocks in a frame as the frame-level quality measure and the average of MOSp of all the frames in the sequence as the overall quality measure of the video sequence. Quality is first evaluated at macroblock-level, then combined into frame-level quality and finally averaged into a single valued sequence-level quality measure. MOSp for each macroblock is computed as:

$$MOSp_{mb} = 1 - k_{mb}(MSE_{mb}) \tag{56}$$

Figures 7-4 and 7-5 present a macroblock-level analysis of quality estimation using the MOSp metric. The figures show video frames from Foreman and Bus sequences compressed at QP = 36. A region in each frame has been selected for analysis purposes. It is a group of 5x5 macroblocks indicated by the red box in Figures 7-4 and 7-5. The corresponding MSE and MOSp values of these macroblocks are given. To compare the performance of the MOSp metric calculated using the four different slope estimation methods based on spatial texture, temporal change and skin information, the MOSp values obtained using the four methods for the two 5x5 regions are presented. From the figures, the following observations can be made:

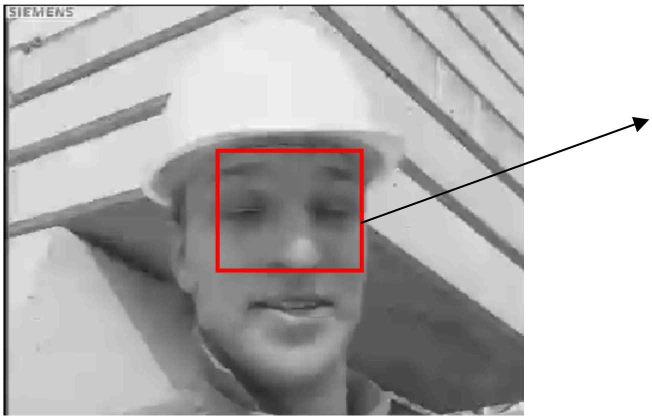
i. The MSE values of the 5x5 region in Bus sequence (Figure 7-5) are higher than those of the Foreman sequence (Figure 7-4). However, the visual quality of the 5x5 compressed region in the Foreman sequence is worse than that of the Bus sequence. There is significant loss of detail and contrast in the facial region of the Foreman frame compared to the 5x5 uncompressed region of the Foreman frame.

ii. The average MOSp value of the 5x5 macroblock region obtained using the four different methods of slope estimation are presented . The average MOSp values of the Foreman region obtained using the four methods are lower than the Bus region although the average MSE of Bus region is higher. This is because the MOSp metric identifies regions with skin, low texture and low motion as being more sensitive to visible distortion when compared to regions with no skin, high texture and high motion. Therefore, the distortion in the facial regions in the Foreman frame, produce lower MOSp score to indicate lower perceived quality. High texture/high motion regions, such as Figure 7-5, are more tolerable to visible distortions. Hence the MOSp values are higher in the macroblock of Figure 7-5 although the MSE values are comparatively higher than Figure 7-4.

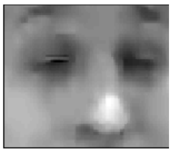
iii. In Figure 7-4, there is a noticeable difference in the average MOSp values obtained from the four methods for the 5x5 region of the Foreman sequence. The average MOSp values calculated using skin information are significantly lower than the average MOSp values calculated without using skin information. This is because the MOSp metric which incorporates skin information classifies macroblocks in the skin regions as being sensitive to visible distortions compared to the MOSp metric which does not incorporate skin information. Therefore, for the same value of MSE, the MOSp metric based on skin information produces lower MOSp score when compared to the MOSp metric which is not based on skin information.

iv. In Figure 7-5, the 5x5 region of the Bus sequence does not belong to skin region and the average MOSp values obtained using the four slope estimations methods are similar.

This example demonstrates that: (a) the MOSp metric is a more effective predictor of subjective quality than MSE, for these selected 5x5 macroblock regions from the Foreman and Bus sequences. (b) In regions where skin pixels are present, the MOSp metric based on skin information produces significantly different results compared to the MOSp metric without skin information for the same value of MSE.



Original region



Compressed region

22.29	17.06	4.33	8.07	13.9
24.58	24.4	9.8	15.74	33.06
29.29	18.42	9.31	17.25	23.98
12.02	6.71	16.4	25.43	10.59
4.32	9.95	26.68	23.85	21.77

Macroblock MSE values, average MSE = 17.18

0.6	0.54	0.65	0.68	0.64
0.38	0.5	0.72	0.65	0.38
0.49	0.66	0.65	0.64	0.58
0.57	0.58	0.59	0.58	0.61
0.66	0.59	0.49	0.61	0.42

Macroblock MOSp values using spatial texture only, average MOSp = 0.5784

0.61	0.5	0.74	0.76	0.73
0.42	0.56	0.81	0.62	0.44
0.53	0.69	0.69	0.67	0.58
0.67	0.6	0.71	0.59	0.64
0.71	0.63	0.58	0.66	0.53

Macroblock MOSp values using spatial texture and temporal change, average MOSp = 0.6268

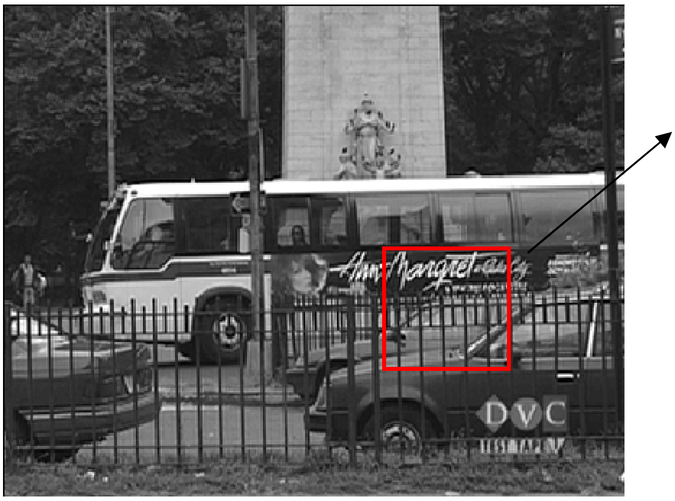
0.26	0.43	0.86	0.73	0.54
0.18	0.19	0.67	0.48	0
0.03	0.39	0.69	0.43	0.20
0.59	0.78	0.45	0.15	0.65
0.86	0.67	0.11	0.21	0.28

Macroblock MOSp values using spatial texture and skin colour, average MOSp=0.433

0.08	0.3	0.82	0.67	0.43
0	0	0.6	0.35	0
0	0.24	0.62	0.29	0.01
0.51	0.72	0.33	0	0.57
0.82	0.59	0	0.02	0.11

Macroblock MOSp values using spatial texture, temporal change and skin colour, average MOSp=0.323

Figure 7-4: Video frame from Foreman sequence compressed at QP = 36. Note: MOSp = [0,1] where 0 = bad and 1 = excellent picture quality.



Original region

Compressed region

111.4	128.2	195.3	147.9	150.9
115.3	128.7	162.3	173.4	190.5
163.1	195.1	116.9	116.8	125.4
75.3	69.1	103	105.1	120.3
65.9	45.9	42.8	45	52.3

Macroblock MSE values, average MSE = 117.84

0.896	0.898	0.895	0.887	0.846
0.854	0.926	0.987	0.993	0.974
0.898	0.914	0.899	0.893	0.956
0.882	0.837	0.893	0.897	0.893
0.898	0.834	0.836	0.883	0.797

Macroblock MOSp values using spatial texture only, average MOSp = 0.895

0.910	0.913	0.916	0.894	0.872
0.878	0.937	0.992	0.997	0.979
0.902	0.893	0.907	0.896	0.961
0.893	0.839	0.908	0.903	0.925
0.902	0.846	0.851	0.895	0.821

Macroblock MOSp values using spatial texture and temporal change, average MOSp = 0.904

0.886	0.875	0.883	0.876	0.832
0.853	0.906	0.978	0.959	0.961
0.887	0.881	0.893	0.884	0.879
0.861	0.834	0.863	0.881	0.862
0.896	0.828	0.837	0.864	0.781

Macroblock MOSp values using Spatial texture and skin colour, average MOSp = 0.8776

0.893	0.885	0.889	0.877	0.851
0.853	0.906	0.978	0.986	0.973
0.891	0.896	0.903	0.9	0.879
0.892	0.84	0.895	0.906	0.901
0.899	0.840	0.848	0.891	0.802

Macroblock MOSp values using Spatial texture, temporal change and skin colour, average MOSp = 0.8910

Figure 7-5: Video frames from Bus sequences compressed at QP = 36. Note: MOSp = [0,1] where 0 = bad and 1 = excellent picture quality.

7.5 Metric performance evaluation

Performance results of the MOSp metric based on spatial texture, temporal change and skin information are present in this section. Performance results include Pearson, Spearman correlation and Outlier's Ratio between MOSp and MOS to investigate prediction accuracy and consistency, scatter plots for visual representation of the correlation between MOSp and MOS, and processing times. The aim of this evaluation is to investigate whether the integration of cognition factors such as skin information in to the existing MOSp metric which is based on spatial texture and temporal changes produces better MOS prediction results.

7.5.1 Correlation coefficients and Outliers ratio

Following the evaluation procedure presented in section 6.6 of Chapter 6 to evaluate the performance of MOSp metric based on spatial texture and temporal changes, experimental results are illustrated in Table 7-1, Table 7-2 and Table 7-3. Tables 7-1 and 7-2 give the Pearson's correlation between the estimated and actual perceptual quality for MOSp and the five popular metrics using the test sequences. For each sequence in the table, the highest correlation coefficient is highlighted in bold font. From the table, the following observations can be made:

- The MOSp metric produces high correlation (>90%) with subjective ratings for a variety of video sequences ranging from low activity such as Akiyo and News, to high activity sequences such as Bus, Mobile and Coastguard. The metric also produces good results with sequences which are a combination of both low-activity and high-activity scenes such as Foreman and Tempete sequences.
- The PSNRplus and NTIAVQM metrics also produce high correlations (>90%) with MOS. However, PSNRplus is computationally expensive to implement in real time multimedia applications.
- The MOSp metric has higher correlation with subjective ratings when compared to NTIAVQM for 28 out of the 32 test sequences.
- The four methods of calculating MOSp from various video content are also presented in Table 7-1. It can be noted that the MOSp based on spatial texture and skin information produces higher correlation with MOS compared to the other three methods of calculating MOSp. This high correlation may be due to higher correlation of skin information and spatial

texture with the visibility of distortion resulting in a more accurate slope estimation model as shown in section 7.3.1.

- It can also be noted that the MOSp based on spatial texture and skin information has the highest correlation with MOS in sequences containing people. This indicates that the MOSp metric is a good predictor of perceived quality in sequences where humans are present.
- In sequences which do not have the presence of people, the MOSp metric based on spatial texture has higher correlation with MOS.

Table 7-1. Pearson Correlation between popular metrics and MOS for 16 test sequences

Sequence	**PSNR**	**VSSIM**	**PSNRplus**	**NTIAVQM**	**Yonsei**	**MOSp based on texture**	**MOSp (texture & temporal)**	**MOSp (texture & skin)**	**MOSp (texture temporal & skin)**
Training									
Foreman	0.775	0.794	0.958	0.965	0.872	0.988	0.997	**0.998**	0.991
Carphone	0.672	0.849	0.957	0.933	0.868	0.943	0.968	**0.980**	0.972
Bus	0.719	0.838	0.856	0.962	0.896	0.988	**0.989**	0.981	0.983
Deadline	0.773	0.834	0.927	**0.985**	0.847	0.924	0.939	0.953	0.944
News	0.849	0.771	0.931	0.916	0.863	0.915	0.944	**0.989**	0.962
Paris	0.809	0.797	0.948	0.968	0.870	0.942	0.964	**0.973**	0.957
Tempete	0.712	0.785	0.890	**0.975**	0.917	0.97	0.964	0.963	0.961
Akiyo	0.752	0.811	0.932	0.908	0.939	0.901	0.986	**0.994**	0.990
Non-training									
Husky	0.702	0.775	0.901	**0.961**	0.841	0.925	0.921	0.912	0.917
Salesman	0.801	0.818	0.919	0.948	0.941	0.874	0.927	**0.981**	0.935
Container	0.795	0.825	0.953	0.976	0.876	0.975	**0.989**	0.969	0.972
Grasses	0.774	0.727	0.879	0.963	0.916	0.962	**0.994**	0.955	0.919
Mobile	0.697	0.713	0.919	0.949	0.884	0.947	**0.984**	0.976	0.947
Sign Irene	0.763	0.746	0.955	0.949	0.945	0.892	0.957	**0.985**	0.963
Mother & daughter	0.725	0.764	0.924	0.935	0.869	0.93	0.952	**0.979**	0.955
Coastguard	0.762	0.724	0.883	0.969	0.866	0.991	**0.995**	0.982	0.968

Table 7-2. Pearson Correlation between popular metrics and MOS for 16 CIF VQEG Multimedia test dataset

Sequence	PSNR	VSSIM	PSNRplus	NTIAVQM	Yonsei	MOSp based on texture	MOSp (texture & temporal)	MOSp (texture & skin)	MOSp (texture temporal & skin)
ITU_SRC_MobileCalendar_cif	0.697	0.713	0.919	0.949	0.884	0.973	0.901	**0.976**	0.917
ITU_SRC_Football_cif	769	0.772	0.887	0.942	0.914	**0.962**	0.945	0.959	0.948
ITU_SRC_FlowerGarden_cif	0.72	0.867	0.892	0.945	0.942	0.96	0.912	**0.972**	0.953
ITU_SRC_Stephen_cif	0.774	0.821	0.957	0.959	0.89	0.913	0.875	**0.98**	0.916
ANSI_SRC_Crew_cif	0.703	0.729	0.856	0.975	0.857	0.93	0.891	**0.979**	0.912
ANSI_SRC_MissAmerica_cif	0.754	0.766	0.927	0.916	0.923	**0.981**	0.915	0.973	0.894
CBC_SRC_BetesPasBetes_cif	0.81	0.768	0.931	**0.952**	0.941	0.916	0.904	0.891	0.907
ANSI_SRC_washdc_cif	0.683	0.722	0.948	0.949	0.927	**0.955**	0.899	0.943	0.990
ANSI_SRC_vtc2mp_cif	0.762	0.819	0.89	0.918	0.872	0.917	0.868	**0.989**	0.932
ANSI_SRC_vtc1nw_cif	0.775	0.827	0.861	0.905	0.861	0.932	0.895	**0.981**	0.919
ANSI_SRC_5row1_cif	0.751	0.824	0.952	0.941	0.896	0.949	0.939	**0.973**	0.942
ITU_SRC_Cheerleaders_cif	0.724	0.816	0.947	0.952	0.847	0.957	0.944	**0.969**	0.928
ANSI_SRC_boblec_cif	0.699	0.794	0.902	0.964	0.87	**0.984**	0.964	0.918	0.964
CRC_SRC_Redflower_cif	0.736	0.818	0.873	0.952	0.872	**0.969**	0.947	0.894	0.951
ANSI_SRC_vtc2zm_cif	0.71	0.829	0.925	0.928	0.907	0.895	0.919	**0.928**	0.925
CBC_SRC_BetesPasBetes_cif	0.752	0.801	0.937	0.931	0.893	**0.951**	0.906	0.895	0.898

Table 7-3 illustrates the overall performance of MOSp and the popular quality metrics when compared with the actual subjective results (MOS). The table consists of the Pearson correlation, Spearman correlation and outliers ratio of MOSp and five popular quality metrics when all the 32 test video sequences mentioned in Table 7-1 and 7-2 are included.

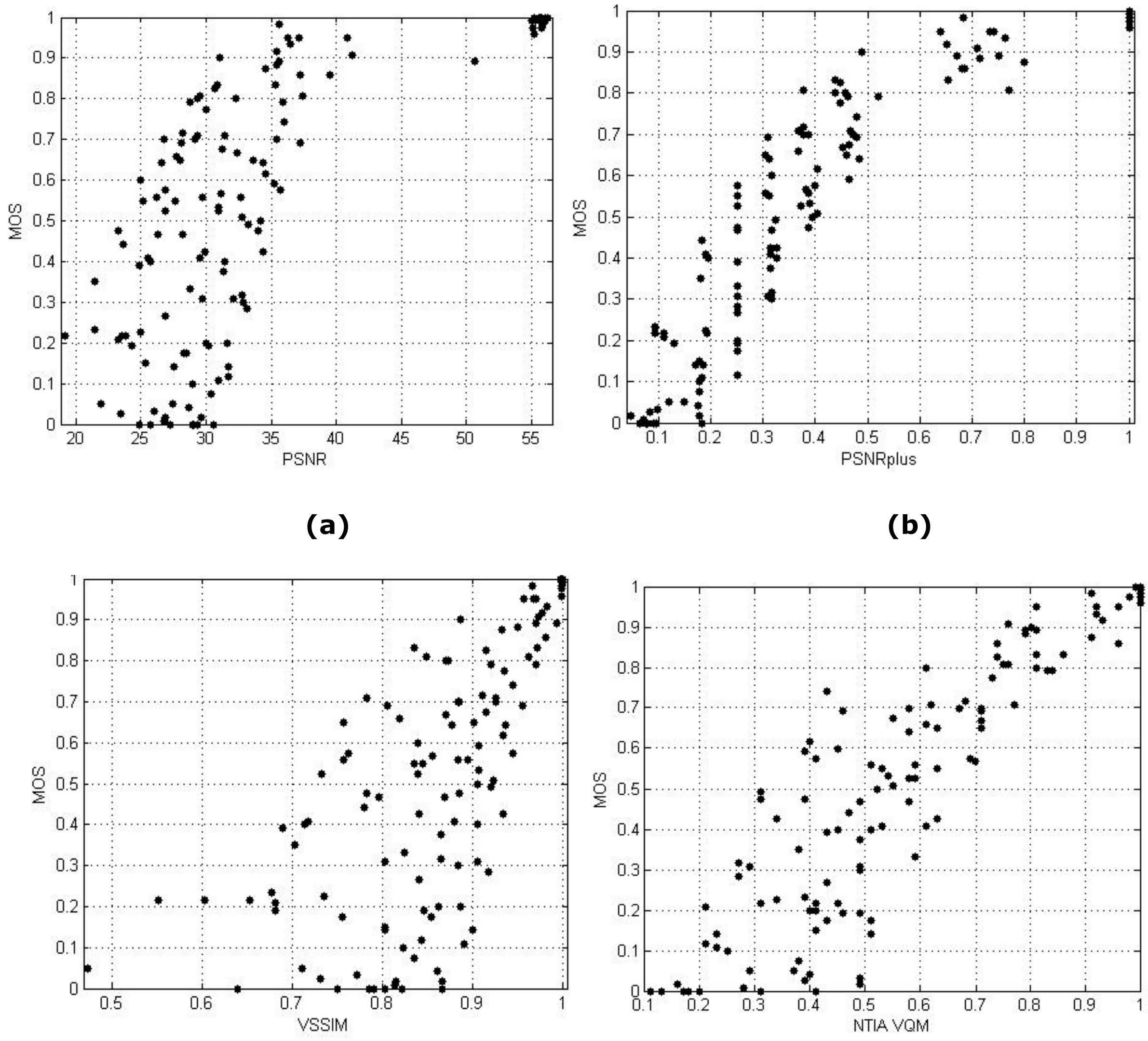
Table 7-3. Comparison of MOSp with popular metrics including all training and non-training sequences

Metric	Pearson Correlation	Spearman Correlation	Outliers Ratio
PSNR	0.696	0.711	0.857
VSSIM	0.723	0.779	0.797
PSNRplus	0.886	0.959	0.596
NTIA VQM	0.901	0.913	0.516
Yonsei University	0.863	0.878	0.628
MOSp based on spatial texture only	0.942	0.948	0.410
MOSp based on spatial texture and temporal change	0.926	0.938	0.48
MOSp based on spatial texture and skin colour	**0.954**	**0.961**	**0.402**
MOSp based on spatial texture, temporal change and skin information	0.946	0.949	0.415

The overall performance of objective quality metrics represents prediction accuracy the quality metric across a variety of video content. The overall Pearson and Spearman correlation values for MOSp based on spatial texture and skin colour are 0.954 and 0.961 respectively, which are the highest amongst the metrics compared in Table 7-3. The outliers ratio for this metric is 0.402, which is the lowest in all the metrics. The closest metric to this performance is the MOSp metric based on spatial texture, temporal change and skin information with Pearson and Spearman correlation values of 0.946 and 0.953 and outliers ratio of 0.415. The high correlation of the MOSp metric based on spatial texture and skin information may be due to a better prediction results for video sequences containing people as indicated in Tables 7-1 and 7-2. The overall performance of the MOSp metric is that it produces better correlation with MOS compared to the five objective quality measures: PSNR, VSSIM, PSNRplus, NTIA/ITS VQM and Yonsei University metric for a variety of video scenes compressed to a wide range of bitrates.

7.6.2 Scatter plots

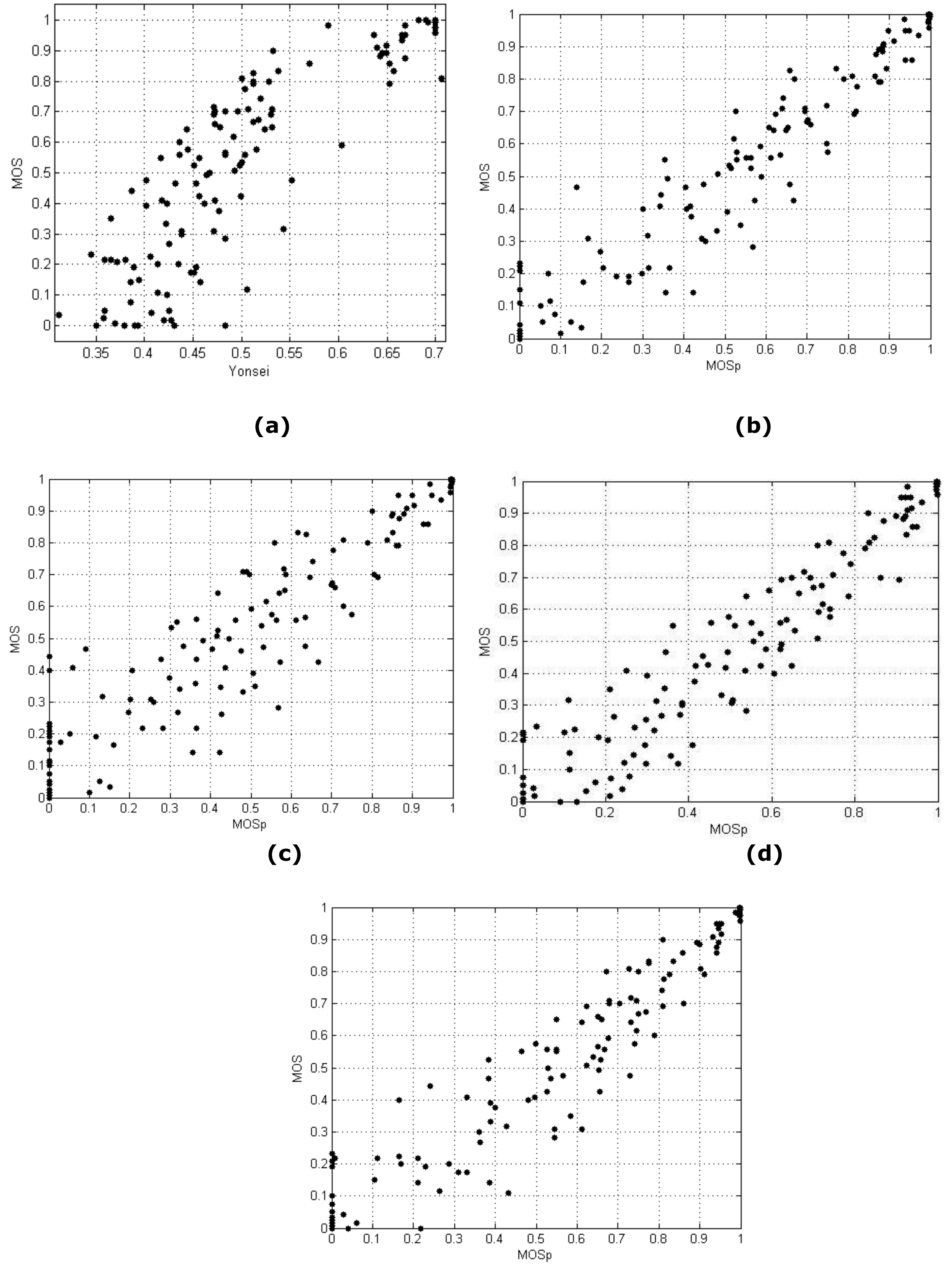
Figures 7-6 and 7-7 show the scatter plots of subjective ratings (MOS) versus the proposed metric (MOSp) and other five popular metrics. The scatter plots contain all the 32 test sequences.



(a) **(b)**

(c) **(d)**

Figure 7-6: Scatter of subjective ratings (MOS) versus (a)PSNR, (b)PSNRplus, (c)VSSIM and (d)NTIA VQM for all the 32 CIF test sequences compressed using H264/AVC.



(a)

(b)

(c)

(d)

(e)

Figure 7-7: Scatter plot of subjective ratings (MOS) versus (a)Yonsei University metric, (b)MOSp based on texture only, (c)MOSp based on texture and temporal change, (d) MOSp based on texture and skin (e) MOSp based on texture, temporal change and skin for all the 32 CIF test sequences compressed using H264/AVC.

From the scatter plots it can be observed that the data points of the PSNRplus metric (Figure 7-6(b)) and MOSp metric based on spatial texture and skin information (Figure 7-7(d)) are most concentrated when compared to the other scatter plots. This is reflected in the high Spearman rank correlation for the two metrics with respect to MOS in Table 7-3 (PSNRplus=0.959 and MOSp based on spatial texture and skin=0.961).

7.5.3 Processing time

Following the procedure used in section 6.6.5 in chapter 6, Table 7-4 shows the percentage increase in coding time when compared to the coding time of the reference H.264/AVC software codec called the JM software [75]. The coding time is taken for JM software to encode the 'Paris' sequence of CIF resolution with 150 frames. The processing times for the MOSp metric derived from sequence content using the four different methods are presented in Table 7-4. Compared to the four methods, the MOSp metric based on spatial texture information requires the least processing time (4.9%). The MOSp metric based on both spatial texture and skin information increases coding time to process a CIF sequence with 150 frames by 8.2%. This increase is nearly double the time required by MOSp based on texture only because of the skin detection algorithm. From Table 7-4 it is evident that MOSp is faster than VSSIM and PSNRplus but requires more processing time compared to PSNR and the Yonsei University metric. However, MOSp based on both spatial texture and skin information gives improved correlation with subjective quality compared to other metrics evaluated in this work.

Table 7-4. Running speed of MOSp and popular metrics

Metric	Increase in coding time (%)
PSNR	Negligible
VSSIM	32.7%
PSNRplus	200.18%
Yonsei University	4.2%
MOSp based on spatial texture	4.9%
MOSp based on spatial texture and temporal change information	7.7%
MOSp based on spatial texture and skin colour	8.2%
MOSp based on spatial texture, temporal change and skin information	11.6%

The processing times for MOSp metric presented in Table 7-4 includes all the evaluation processes including spatial edge strength calculation, temporal change estimation and skin detection performed on each frame. The coding time is the time taken by the H264 compression algorithm to process a video sequence. The percentage increase in coding time presented in Table 7-4 is calculated from the average coding times recorded using five repetitions.

The MOSp metric predicts perceptual quality automatically using MSE which is a widely used quality measure in video compression algorithms. The only additional requirement for MOSp calculation is the slope estimation using one of the four methods using spatial texture, temporal change and skin information. From Table 7-4, it is can be observed that quality estimation using MOSp metric increases computation time by 4.9% to 11.6% depending on the method of slope estimation. Hence the choice of slope estimation for calculating MOSp metric will depend on the application, the required prediction accuracy and the available computational resources.

7.6 Summary

This chapter presented two methods for integrating cognition-based factors such as skin information in to the MOSp metric in order to increase its correlation subjective quality. The performance evaluation results show that the MOSp metric produces high correlation with MOS (>90%) with 4.9% to 11.6% increase in coding time. Results also show that the MOSp metric produces higher correlation with subjective results when compared to popular metrics such as PSNR, PSNRplus, VSSIM, Yonsei and NTIAVQM metric.

Performance comparison between the four methods of calculating MOSp from video content show that the MOSp metric based on spatial texture and skin information produces highest correlation with MOS (95.4%). This high correlation may be due to the following reasons:

i. The combination of spatial texture masking and cognition factors has impact on the visibility of distortion in video.
ii. The slope variation between different video content has better correlation with spatial texture and skin information. This results in a more accurate slope estimation model for calculating the slope parameter of the MOSp

metric and hence a better performing MOSp metric which produces higher prediction accuracy.

The MOSp metric is a full reference objective video quality metric designed to predict MOS of compressed video automatically from MSE and video content within reasonable computation time. Since all the parameters of the metric are calculated at macroblock level, it can be easily incorporated into block-based video coding algorithms for making real time quality estimations. Apart from estimating video quality, the MOSp metric may also be used to replace mathematical error measures which are generally employed by the video encoder to make coding decisions. This application of the MOSp metric to perceptual video coding is investigated in the following chapter.

8 Application Of The MOSp Metric To Perceptual Video Coding

8.1 Introduction

A new full reference video quality metric called the MOSp metric for predicting MOS of compressed video from MSE and video content was presented in earlier chapters. Performance results of the MOSp metric have shown that the metric has very high correlation with MOS compared to other popular metrics. The MOSp metric is designed to predict MOS automatically with reasonable computation time. Since all parameters are calculated at macroblock level, the MOSp metric can be readily incorporated in to block-based video coding algorithms.

Apart from measuring video quality of compressed video, the MOSp metric may also be used in perceptual video coding where coding decisions are made by incorporating a perceptual quality metric into the decision making process. Previous research has shown that perceptual quality based video coding can be achieved by employing quality metrics in motion estimation, mode selection process and rate control processes. Amongst popular objective quality metrics, the structural similarity (SSIM) index [90] has been preferred in perceptual video coding algorithms due to its simplicity and efficiency. It has been incorporated into motion estimation [91], mode selection [92] and rate control [93] processes in hybrid video coding algorithms. In [91] and [92], a SSIM-based distortion measure was used in the RD optimised framework. However, a single Lagrange multiplier model was derived without considering input sequence characteristics. In [93], an SSIM motivated rate control scheme was proposed on an approximate RD curve, while the properties of the SSIM index were not fully exploited. In [94], the authors define a reduced reference SSIM-based distortion model and develop a perceptual RDO scheme for mode selection. The results showed bit rate savings for same level of SSIM quality value.

Although several perceptual quality metrics exist in the literature, the application of these metrics to real time perceptual video coding is limited due to computational complexity and speed issues. Hence, this chapter investigates ways of integrating the MOSp metric into the H264/AVC encoder in order to improve perceptual quality of compressed video by making coding decisions based on a MOSp-based distortion measure rather than mathematical distortion measures such as sum of squared difference (SSD) and sum of absolute difference (SAD).

Advanced video coding schemes such as H264/AVC use motion estimation and mode selection processes to find the best coding option for each macroblock. The motion estimation process employs the rate-distortion optimised search method to find the best matching block for the current block. SAD between a search block and current block pixels is used as the distortion measure and it represents pixel differences between the blocks. On the other hand, the mode selection process is used to select the best mode to encode a macroblock. It is also rate-distortion optimised as described in chapter 2. The distortion measure used is SSD between original block and the reconstructed block.

Since the MOSp metric is based on the mean squared error between the original and reconstructed video sequences, it is suitable for integration into mode selection rather than motion estimation. This chapter presents a new MOSp-based mode selection algorithm for H264/AVC encoder which employs the MOSp metric in making mode decisions for each macroblock.

The chapter is organised as follows: section 8.2 presents the hypothesis behind MOSp based video coding. A new MOSp-based mode selection model is described in section 8.3. The parameters used in the model such as MOSp-based distortion measure and the Lagrange multiplier are also derived in this section. The new MOSp-based mode selection algorithm for a H264/AVC encoder is presented in section 8.4. Section 8.5 describes an experiment conducted to investigate if the MOSp-based mode selection algorithm gives better visual quality compared to the reference H264/AVC encoder for similar bitrate. Analysis of the experimental findings is discussed in section 8.6 and finally in section 8.7, the main experimental observations are summarised.

8.2 Hypothesis

Popular objective measures such as sum of squared difference (SSD) and sum of absolute difference (SAD) are used in modern block-based video compression algorithms such as H264/AVC [16]. These measures are employed by the rate–distortion optimised mode selection process as quality measures for choosing the best compression option that gives an optimal trade-off between picture quality and data rate [32]. The RD optimised mode selection process involves minimising the rate–distortion cost $J=D+\lambda R$ where λ is the Lagrange multiplier, R is the rate and D is the SSD between original and reconstructed video data.

While the general approach is to use SSD to choose the best coding option, it is a mathematical error measure which does not consider the human visual system and is therefore not an accurate measure of perceived quality for compressed video sequences. It may be possible to improve the subjective quality performance of a rate-constrained video codec by replacing SSD with a MOSp-based distortion metric that correlates more closely with subjective quality in the mode selection process. Hence this chapter presents a MOSp-based mode selection algorithm where the distortion measure is calculated from the MOSp metric.

8.3 MOSp-based mode selection

The mode selection process in block-based video encoders involves minimising the rate-distortion cost function $J=D+\lambda R$ where λ is the Lagrange multiplier, R is the rate and D is the SSD between original and reconstructed video data. MOSp-based mode selection would involve integrating the MOSp metric into the RD cost function to make the mode selection and choosing the best mode which minimises this cost function. A new MOSp-based mode selection model is presented in this section.

8.3.1 MOSp-based mode selection model

The rate-distortion cost function used in the reference H264/AVC is:

$$J = D + \lambda R \qquad (60)$$

where λ is the Lagrange multiplier, R is the rate and D is the SSD between original and reconstructed video data. Integrating the MOSp metric into equation (44) involves defining a new MOSp-based distortion measure and a new Langrange multiplier. The new MOSp-based rate-distortion cost function model is given as:

$$J = D_{mosp} + \lambda_{mosp} R \qquad (61)$$

where D_{mosp} is MOSp-based distortion measure which replaces the SSD measure and λ_{mosp} is the 'new' Lagrange multiplier which must to be re-modelled. The Lagrange multiplier in the reference H264/AVC is calculated as a function of the Quantisation Parameter (QP) [32, 95] and has been modelled for SSD as the distortion metric. Therefore, changing SSD to D_{mosp} will require re-modelling of the Lagrange multiplier to obtain λ_{mosp}. R is the total bits for coding a macroblock using the mode under test. The parameters for the MOSp-based mode selection model are detailed in sections 8.3.2 to 8.3.4

8.3.2 Model parameter estimation: D_{mosp}

The MOSp metric measures perceived video quality from MSE and video content. It has values between [0,1] with 0 indicating 'very poor' visual quality and 1 indicating 'Excellent' visual quality. A distortion measure derived from the MOSp metric must be inversely related to the MOSp measure. Therefore, the MOSp-based distortion measure, D_{mosp} is given as:

$$D_{mosp} = 1 - MOSp \qquad (62)$$

As presented in Chapter 5 section 5.4, the MOSp metric is calculated from MSE and the slope parameter (Ks) as:

$$MOSp = 1 - k_s MSE \tag{63}$$

Substituting (4) in (3) gives:

$$D_{mosp} = 1 - 1 + k_s MSE \tag{64}$$

$$D_{mosp} = k_s MSE \tag{65}$$

Equation (65) is used as the new MOSp-based distortion measure in the mode selection algorithm and is measured as the product of the slope parameter (Ks) and the MSE. Slope Ks will be derived from macroblock activity and MSE is the mean squared error between the original and the reconstructed macroblocks. Since MOSp values range from [0,1], D_{mosp} will also have values ranging from [0,1] where 0 indicates no visible distortion and 1 indicates maximum visible distortion. From equation (65), D_{mosp} is derived as the product of slope Ks and MSE. The slope Ks is dependent on content and has larger value for content which are sensitive to visible distortions and smaller value for content which can 'mask' visibility of compression-related distortions. Hence, multiplying Ks with MSE will make Ks a 'weighting factor' for MSE and would 'magnify' or 'minimise' MSE based on the content in the macroblock. This weighting will have impact on the resulting rate distortion cost function which strives to keep a balance between distortion and rate.

8.3.3 Model parameter: λ_{mosp}

The Lagrange multiplier in the rate-distortion optimised mode selection acts as a balancing parameter between rate and distortion. The Lagrange multiplier defined in the reference H264/AVC encoder has been derived experimentally using SSD as the distortion metric. Since the distortion metric is changed from SSD to D_{mosp} in the MOSp-based mode selection model, a new Lagrange multiplier λ_{mosp} must be modelled using similar experiments. This is presented in section 8.4.

8.3.4 Model parameter: Rate (R)

The rate parameter in the mode selection model is the total bits required to encode a macroblock using the mode under test. The number of coded bits depends on the type of content in the macroblock. High detail and high motion macroblocks with changing content may require larger number of coded bits. Higher rates generally mean better picture quality. Hence the aim of integrating the MOSp metric into the mode selection process is to allocate modes with higher rates to visually 'important' macroblocks in order to improve the overall visual quality of compressed video.

8.4 Modelling the Lagrange multiplier λ_{mosp}

The Lagrange multiplier for mode selection in the reference H264/AVC has been experimentally modelled as a function of the Quantisation Parameter (QP) [32,111] using SSD as the distortion metric. Following the experiments detailed in [32], λ_{mosp} is modelled as described in this section.

Six multimedia CIF sequences of 4:2:0 format were used for modelling λ_{mosp}. The sequences were 50 frames in duration and have a wide variety of content from 'head and shoulder' shots to high detail and high motion vehicle tracking. These sequences are Foreman, Akiyo, News, Bus, Coastguard and Husky. The test video sequences were compressed using main profile of the reference H.264/AVC JM reference software (version 12.1) to a range of QP values. Each test sequence was encoded several times at a certain QP by incrementing the lambda (λ_{mosp}) value by small amounts and recording the corresponding average bitrate and average D_{mosp} values. The rate-distortion curve for each test sequence was obtained by plotting the average bitrate versus average D_{mosp} recorded for all the QP values and using a convex hull fitting tool to obtain the R-D_{mosp} curve. The QP and lambda (λ_{mosp}) values corresponding to the points on the R-D_{mosp} curve were used to derive the λ_{mosp} - QP relationship for the test sequence under test. This process was repeated for all the six test sequences and the λ_{mosp} - QP plots for these sequences are presented in Figure 8.1. From Figure 8.1, it is observed that λ_{mosp} varies exponentially with respect to QP and this variation is different for different video content.

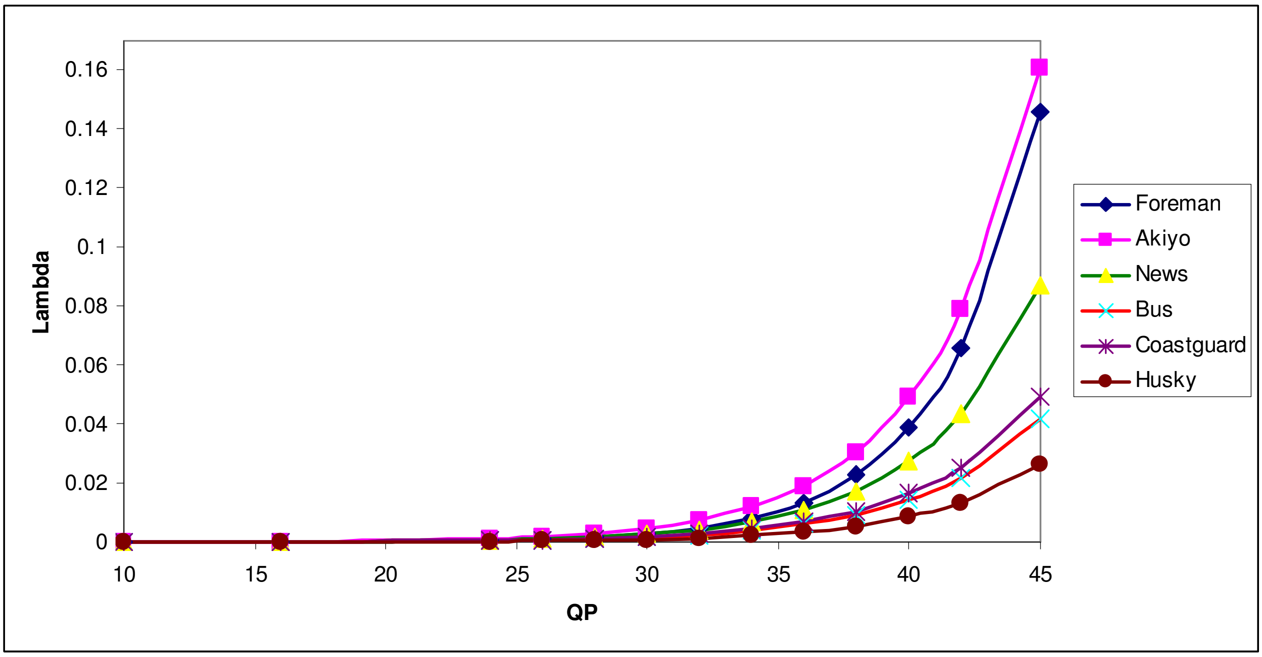


Figure 8-1: Lambda versus QP plots for six test sequences

Based on the QP - λ_{mosp} models for the six test sequences, the generalised model for calculating the Lagrange multiplier λ_{mosp} for the MOSp-based mode selection algorithm QP and video content is given as:

$$\lambda_{mosp} = A * \exp(B * QP) \tag{66}$$

where A and B are parameters of the exponential curve derived from sequence activity as:

$$A = (9.413E\text{-}009*Activity) + 1.152E\text{-}006 \tag{67}$$

$$B = (-0.0003292*Activity) + 0.2685 \tag{68}$$

where 'Activity' is the sequence activity derived from spatial texture and skin information as given in Chapter 7, equation (56). Therefore, the Lagrange multiplier λ_{mosp} value will vary depending on QP and video content and can be automatically calculated using equations (66), (67) and (68).

8.5 MOSp-based mode selection algorithm

The MOSp-based mode selection algorithm is summarised below:

1. For each macroblock, calculate the activity and slope using equations (55) and (56).
2. Calculate the Largrange multiplier λ_{mosp} using equations (66), (67) and (68).
3. Select a macroblock mode
4. Encode the macroblock and calculate D_{mosp} **=** Ks * MSE
5. Compute RD cost function J= $D_{mosp} + \lambda_{mosp}$ R
6. Check if J < Jmin, where J min = minimum RDcost for all modes.
7. If J<Jmin, check if all modes have been evaluated. If NO, then update Jmin = J and go to step 2. If YES, then current mode is the best mode for encoding the macroblock.

8.6 Experiment: Performance evaluation of the MOSp-based mode selection algorithm

The aim of this experiment is to investigate whether MOSp-based mode selection improves visual quality of compressed video when compared to the reference video encoder for similar bit rate. The experiment involves performance comparison between two coding algorithms: the reference H264/AVC encoder and the reference H264/AVC encoder with MOSp-based mode selection algorithm. The following sections describe the experiment in detail, including test material, coding parameters, subjective evaluation conducted to obtain MOS scores for sequences compressed using both the coding algorithms, data analysis of obtained results and discussion based on the experiment results and observations.

8.6.1 Test Material

12 multimedia CIF sequences of 4:2:0 format were used in this experiment, each of 10 seconds in duration. The sequences include:

- Training sequences which were used to obtain the General Lambda model: Foreman, Bus, News, Husky, Akiyo and Coastguard

- Non-training sequences: Carphone, Crew, Football, Miss. America, Stephen and City sequences.

These sequences were chose to represent a wide variety of content and are popularly used in the video compression research community.

8.6.2 Coding Parameters:

The test video sequences used in this experiment were compressed using the H.264/AVC JM reference software (version 12.1) available at http://iphome.hhi.de/suehring/tml/, with the following parameters:

- Profile used is Main profile.
- Level IDC setting is set 4.0
- Frame Skip: no frames were skipped
- Number of reference frames for Inter motion search is set to 5.
- Number of B-pictures used = 0
- Entropy coding method is set to CABAC.
- RD-Optimisation: High complexity mode
- Rate Control: DISABLED to allow the use of fixed QP.
- Slice QP: QPISlice and QPPSlice parameters used and both set to the same value as the sequence QP.

Note that the coding parameters used in the MOSp metric evaluation is identical to those used to produce the training sequences in section 6.2.

- QP values = {24, 26, 28, 30, 32, 34, 36, 38}

Coding algorithms used in the experiment:

Codec A: Reference H264/AVC JM encoder with no changes made to the mode selection process, full mode selection used to include all the available coding modes.

Codec B: H264/AVC JM encoder with MOSp-based mode selection algorithm. Full mode selection is used. Note that the Lagrange multiplier has been calculated as a function of QP and acitivity as explained in section 9.4 using the General model. The mode selection algorithm was implemented as detailed in section 8.5.

8.6.3 Subjective evaluation

The subjective tests involved 30 non-expert evaluators and followed the guidelines in ITU-T P.910 Recommendation [6]. Each evaluator took 19 to 22 minutes to complete the test. The subjective test method used in this experiment is the single stimulus impairment scale (SSIS) evaluation method. Since the visual quality of Codec A is being compared with Codec B, each viewer was shown sequences coded with both the codecs A and B. Considering the time limitations for conducting subjective evaluations [6], each viewer was shown four sets of sequences containing two different video clips compressed using the two codecs A and B. In total, each viewer evaluated 4x8=32 video sequences. The sequences were presented in a randomised presentation order with either increasing or decreasing magnitude of distortion. This was done to counterbalance the influence contextual effects of the SSIS method [78].

A 5-grade discrete scale ranging from 0 to 1 was used to rate the quality of each of the test video sequences where 0=bad, 0.25=poor, 0.5=fair, 0.75=good and 1=excellent. Reliability of subjective test scores was tested using the 95% confidence interval measure. The average mean 95% confidence interval for the subjective ratings for all the test sequences was 0.0447 for the MOS scale of [0, 1] where 0=bad picture quality and 1=excellent picture quality. The MOS for a sequence was calculated as the average of all scores obtained for the sequence compressed at a certain QP.

8.6.4 Experiment Results

To investigate if there is a gain in MOS using MOSp-based mode selection algorithm when compared with the reference H264 encoder, the results are presented as bitrate versus MOS graphs for each of the 12 test sequence. Each graph has two curves, one for each codec. These bitrate versus MOS graphs are presented in:

- Figures 8-2 and 8-3 for training sequences.
- Figures 8-4 and 8-5 for non-training sequences.

Table 8-1 compares the coding performance of the two codecs and includes the following information:

1. Percentage gain (or loss) in visual quality for each sequence which is calculated as:

$$\Delta MOS = MOS_{CodecB} - MOS_{CodecA} \tag{69}$$

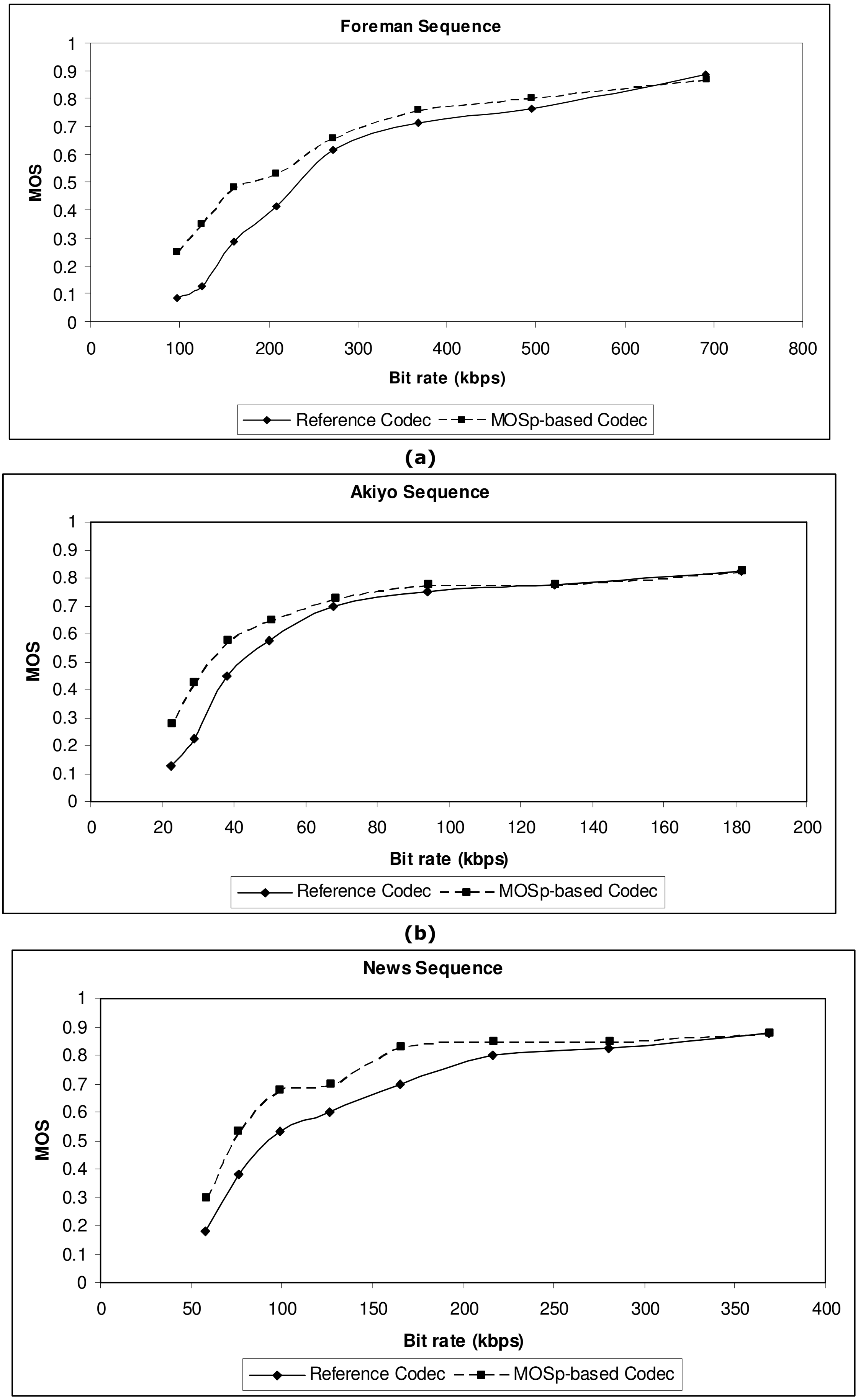
2. Percentage gain (or loss) in bit rate for each sequence calculated as:

$$\Delta Bitrate(\%) = \frac{Bitrate_{CodecB} - Bitrate_{CodecA}}{Bitrate_{CodecA}} X100 \tag{70}$$

3. Percentage gain (or loss) in PSNR for each sequence calculated as:

$$\Delta PSNR(\%) = \frac{PSNR_{CodecB} - PSNR_{CodecA}}{PSNR_{CodecA}} X100 \tag{71}$$

Gain in quality, PSNR and bitrate is represented by a '+' sign and a loss is represented by a '-' sign. The overall range of the above three measures is given in Table 8-1. Table 8-2 lists the maximum improvements in quality and bitrate for each test sequence.

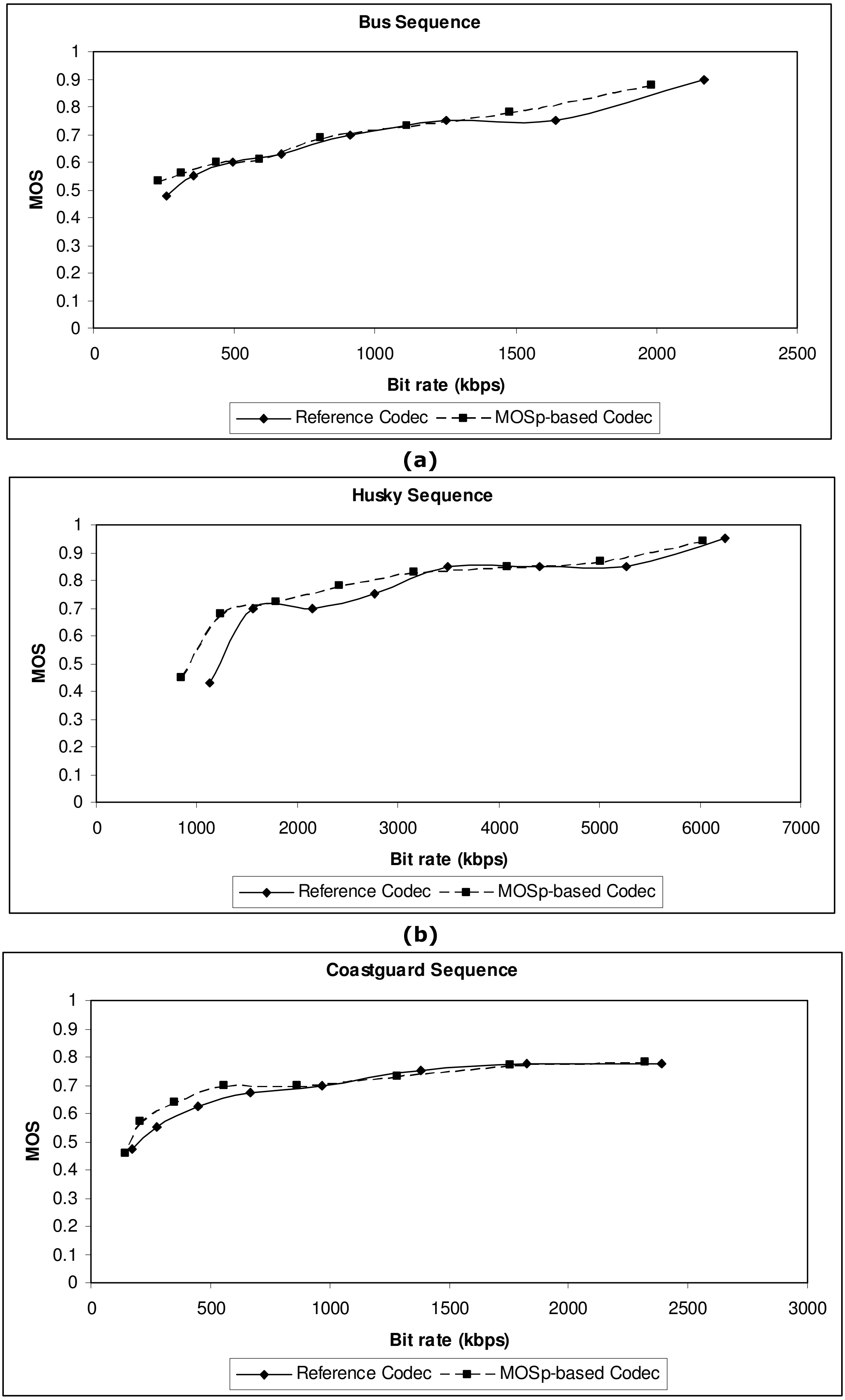


(a)

(b)

(c)

Figure 8-2: Bit rate versus MOS graphs for training sequences

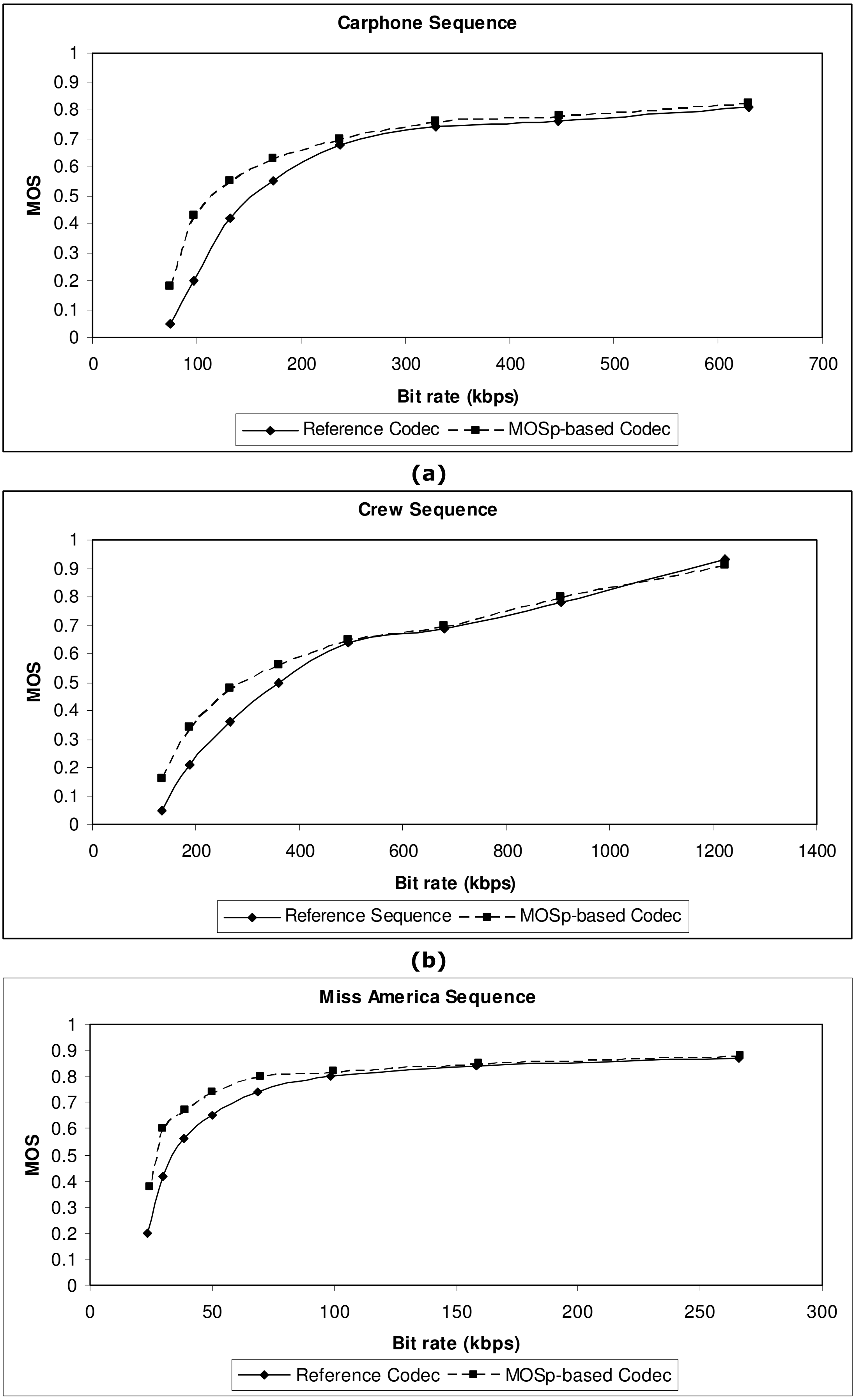


(a)

(b)

(c)

Figure 8-3: Bit rate versus MOS graphs for Training sequences

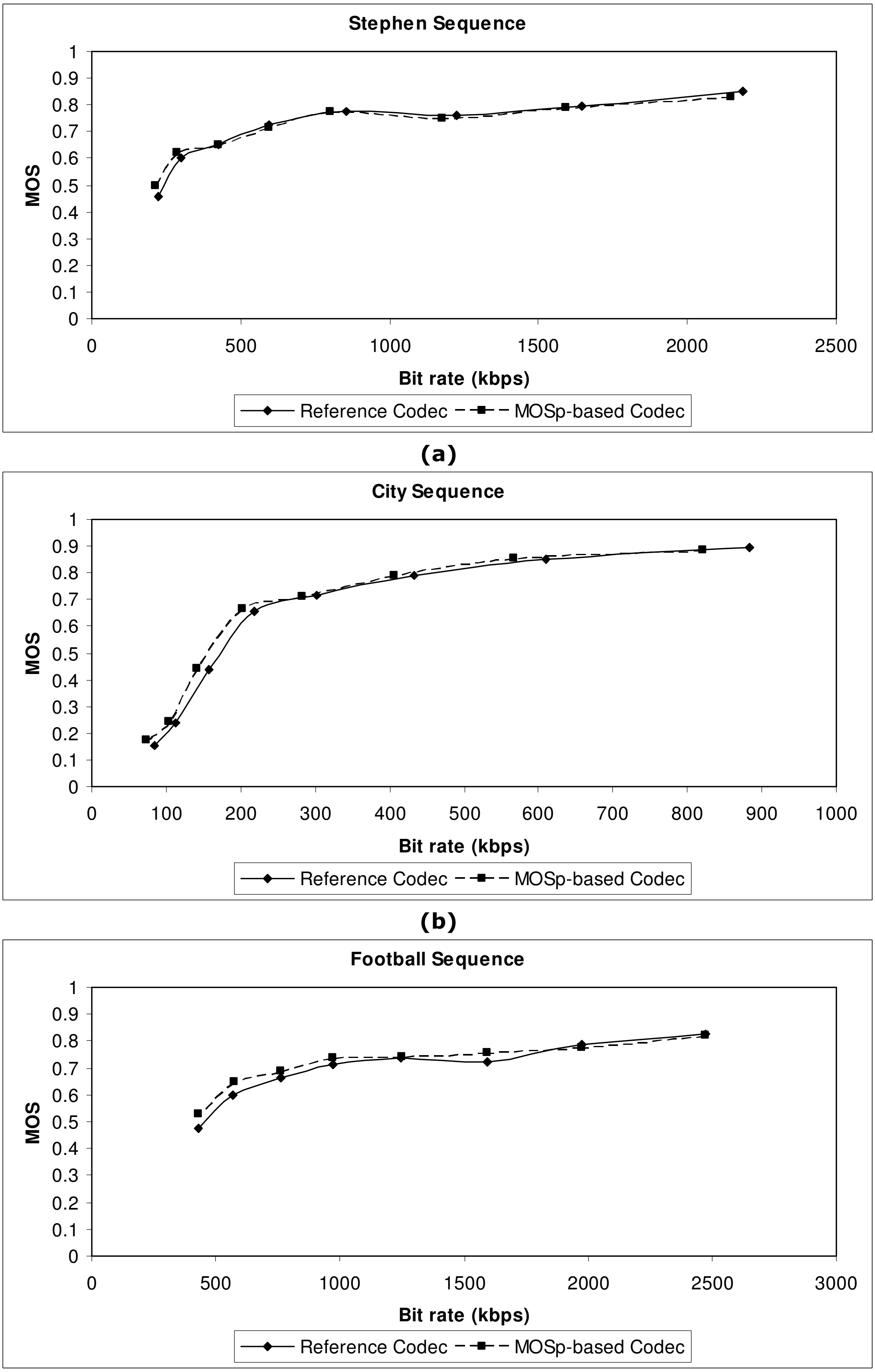


(a)

(b)

(c)

Figure 8-4: Bit rate versus MOS graphs for non-training sequences



(a)

(b)

(c)

Figure 8-5: Bit rate versus MOS graphs for non-training sequences

Table 8-1: Performance comparison of codec with MOSp-based mode selection compared to the Reference H264/AVC encoder

Sequence	**ΔBitrate (%)**	**ΔPSNR (%)**	**ΔMOS**
Foreman	0.06 to 0.83	-0.79 to -0.05	**-0.015 to 0.225**
Akiyo	0.25 to 1.11	-0.28 to -0.094	0 to 0.2
News	0.01 to 1.09	-0.31 to -0.02	0.03 to 0.15
Bus	-8.6 to -11.92	-1.65 to -0.46	-0.01 to 0.04
Husky	**-3.47 to -25.8**	-1.90 to -0.83	-0.04 to 0.02
Coastguard	**-3.09 to -26.22**	-1.94 to -0.72	-0.06 to +0.025
Carphone	0.07 to 0.72	-0.62 to -0.38	**0.02 to 0.23**
Crew	-0.012 to 0.34	–0.56 to -0.14	-0.01 to 0.11
Miss America	0.48 to 1.04	-0.41 to -0.16	0.01 to 0.18
Stephen	-0.01 to -4.45	-0.97 to -0.3	-0.01 to 0.04
City	-6.3 to -12.7	-1.64 to -0.39	-0.008 to 0.01
Football	-3.47 to 0.04	-0.44 to -0.098	-0.005 to 0.055

Table 8-2: Maximum improvements from Codec with MOSp-based mode selection compared to the Reference H264/AVC encoder

Sequence	**Maximum improvements obtained**
Foreman	Δ**MOS = 0.225** for ΔBitrate = 0.83% and ΔPSNR = -0.79%
Akiyo	Δ**MOS = 0.2** for ΔBitrate = 1.11% and ΔPSNR = -0.28%
News	Δ**MOS = 0.15** for ΔBitrate = 1.09% and ΔPSNR = -0.31%
Bus	Δ**Bitrate = -11.92%** for ΔMOS =-0.01 and ΔPSNR=-1.65%
Husky	Δ**Bitrate = -25.8%** for ΔMOS =-0.04 and ΔPSNR=-1.90%
Coastguard	Δ**Bitrate = -26.22%** for ΔMOS =-0.06 and ΔPSNR=-1.94%
Carphone	Δ**MOS = 0.23** for ΔBitrate **=** 0.72% and ΔPSNR = -0.62%
Crew	Δ**MOS = 0.11** for ΔBitrate **=** 0.34% and ΔPSNR = –0.56%
Miss America	Δ**MOS = 0.18** for ΔBitrate **=** 1.04% and ΔPSNR=–0.56%
Stephen	Δ**Bitrate = -4.45%** for ΔMOS **=** -0.01 and ΔPSNR=-0.97%
City	Δ**Bitrate = -12.7%** for ΔMOS **=**-0.008 and ΔPSNR=-1.72 %
Football	Δ**Bitrate = -3.47%** for ΔMOS **=**-0.005 and ΔPSNR = -0.44%

The following observations can be made from these results:

From Figures 8-2, 8-3, 8-4 and 8-5, it can be noted that at high bitrates, the gap between the two curves is very small indicating that the gain in visual quality at high bitrates is negligible. The gap between the two curves (reference and MOSP-based) increases with decrease in bitrate. This gap is more prominent in some sequences when compared to other sequences. This variation in MOS gain with bitrates may be negligible because at high bitrate, the variation in MOSp for different macroblock modes may be small. Hence switching between different macroblock modes using the MOSp-based mode selection may not result in a significant gain in overall visual quality of the sequence. On the other hand, at lower bitrates, the variation of MOSp for different macroblock modes may be larger and hence switching between modes based on the MOSp-based mode selection may result in a significant overall gain in visual quality of the compressed sequence. A macroblock-level analysis is required to further investigate this observation. This analysis is presented in section 8.6.5.

In sequences which contain human faces, such as Foreman, Akiyo, News, Carphone, Crew and Miss. America, the codec with MOS-based mode selection produces higher gain in visual quality at lower bitrates compared to the reference codec. There is a gain in MOS of 0.23 in Carphone sequence with 0.72% increase in bitrate and 0.62% decrease in PSNR when compared to the Reference codec. In Foreman sequence, there is a gain in MOS of 0.225 with 0.83% increase in bitrate and 0.79%. The MOSp metric is designed to identify macroblocks in the video scene which belong to skin and low texture as visually important macroblocks. The MOSp-based distortion measure D_{mosp} is a product of slope (Ks) and MSE. Therefore, the slope (Ks) acts as a 'weighting factor' to MSE. Slope (Ks) is derived from content and has a larger value for skin and low texture content when compared to non-skin, high-texture content. A macroblock classified as 'skin macroblock' will have a large slope and hence a larger D_{mosp} when compared to a 'non-skin macroblock' with same MSE value. This magnification of MSE based on the slope parameter will have impact on the rate-distortion cost function ($J=D+\lambda R$) because higher distortion would mean higher RD cost function resulting in higher quality modes being selected for encoding the macroblock. Hence, sequences where human faces are present have better visual quality when coded with the MOSp-based mode selection algorithm when compared to the Reference codec for similar bitrates.

In Sequences which don't contain humans, such as Bus, Husky, Coastguard and City, the gain in MOS for codec using the MOSp-based mode selection is very low (around 0.05) even at low bit rates. This indicates that the quality of video produced using both the codecs for sequences without humans very similar. However, it is noted from Table 8-1 that there is a gain in bitrate for similar quality in these sequences at low bit rate. Coastguard and Husky sequences have approximately 26% gain in bitrate with nearly 2% drop in PSNR but the difference in visual quality between the two codecs is insignificant (around 0.02%). This insignificant gain in visual quality and significant gain in bitrate may be because the MOSp metric is designed to classify high texture, non-skin macroblocks as 'visually unimportant' macroblocks with high resistance to visible distortion. Therefore the slope (Ks) would have a smaller value compared to skin/low texture macroblocks resulting in a smaller D_{mosp} compared to a skin/low texture macroblock with same MSE. This scaling of D_{mosp} will have impact on the RD cost function ((J=D+λR) because lower distortion would mean lower RD cost function resulting in lower quality/bitrate modes being selected for encoding the macroblock. Hence, in sequences where humans are absent, the MOSp-based mode selection algorithm gives a gain in bitrate for similar visual quality when compared with the reference encoder.

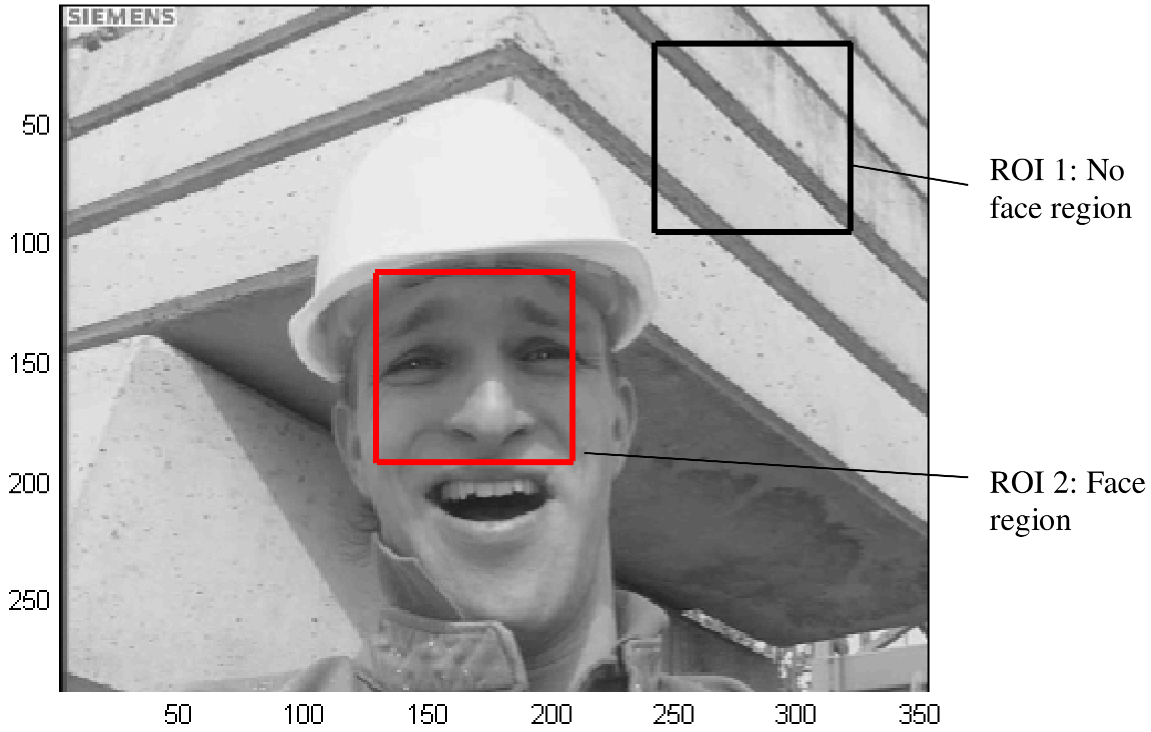
It has also been noted that in sequences such as Football and Stephen, although humans are present, gain in visual quality and bitrate is very small when compared to the Reference codec. Stephen and football sequences categorise as sports video and the video content include very high motion, camera panning and high detail. Although humans are present, the focus of viewer attention in sport video may not be limited to human faces and the attention may be more focused on other things such as tracking the football or tennis ball, looking out for goals, etc. Since the MOSp metric incorporates spatial texture and skin information, it is limited to identifying human faces and low textured objects in the video scene as being visually important.

8.6.5 Macroblock level Analysis

Performance results presented in section 8.6.4 showed that the MOSp-based mode selection algorithm produces a gain in visual quality for sequences where humans are present. In sequences where humans are absent, there is a gain in bitrate for very similar visual quality compared to the reference codec. It was also observed that the gain in visual quality is very low at high bitrates and it increases with decrease in bit rate. This section gives a macroblock level analysis in support of these findings. Figures 8-6 and 8-7 are video frames from Foreman and Coastguard sequences compressed at QP = 36. Two regions in each frame have been selected for analysis purposes. They are a group of 4x4 macroblocks indicated by the red box in Figures 8-6 and 8-7. The corresponding MOSp values obtained from using the MOSp-based codec and the Reference codec are given. From the figures, the following observations can be made:

i. The MOSp values of the 4x4 regions in both Foreman and coastguard sequence show that MOSp-based mode selection algorithm produces higher average MOSp compared to the reference codec. This indicates that making mode decisions based on the MOSp metric can improve visual quality at macroblock level.

ii. In the foreman sequence, the gain in average MOSp is higher in the face region (nearly double) when compared to the non-face region indicating that MOSp-based mode selection produces higher gain in visually sensitive regions such as the human face when compared to other regions in the video scene.



ROI 1: Non-face region:

0.43	0.61	0.42	0.58		0.46	0.59	0.45	0.61
0.47	0.53	0.61	0.56		0.53	0.58	0.63	0.54
0.52	0.60	0.58	0.47		0.57	0.66	0.63	0.45
0.63	0.59	0.61	0.49		0.69	0.64	0.69	0.52

MOSp values (Reference Codec)
Average MOSp = 0.54

MOSp values (MOSp-based Codec)
Average MOSp =0.58

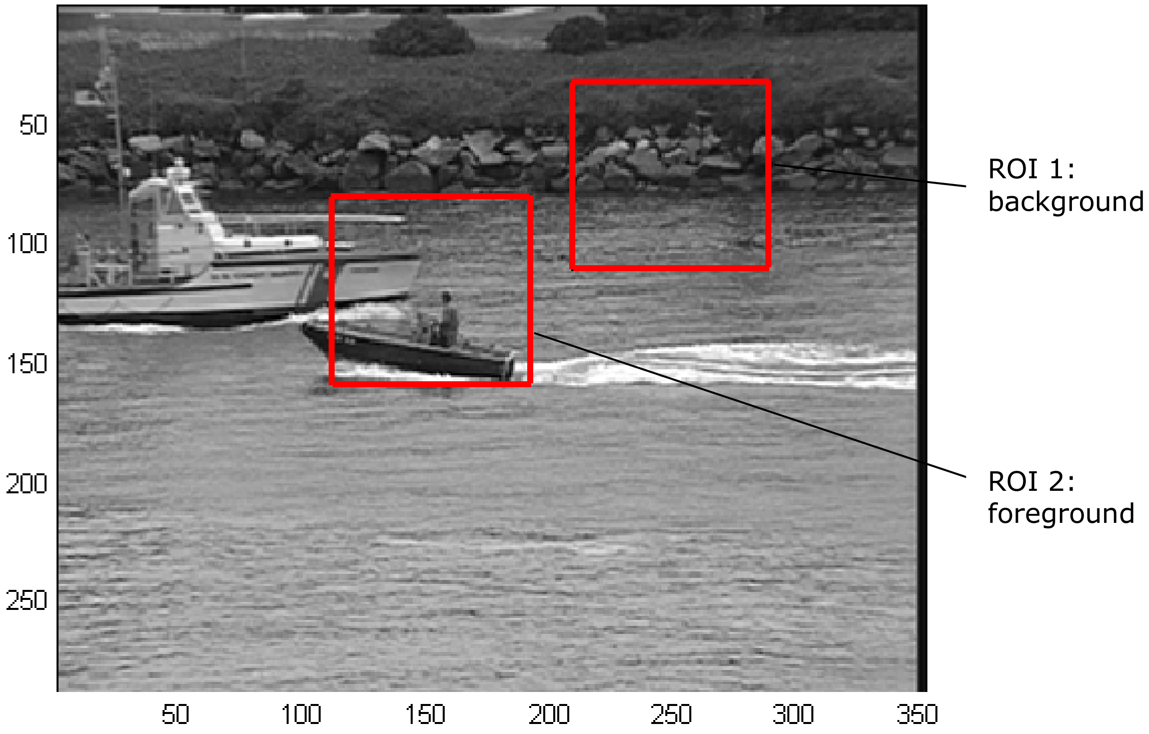
ROI 2: Face region:

0.26	0.33	0.17	0.42		0.41	0.45	0.39	0.54
0.09	0.11	0.06	0.02		0.38	0.3	0.28	0.31
0.21	0.39	0.19	0.23		0.47	0.53	0.45	0.49
0.25	0.28	0.04	0.15		0.44	0.41	0.29	0.35

MOSp values (Reference Codec)
Average MOSp = 0.2

MOSp values (MOSp-based Codec)
Average MOSp =0.41

Figure 8-6: MB-level analysis for Foreman CIF sequence, QP=36, frame 37



ROI 1: Backgound:

0.67	0.63	0.64	0.61
0.68	0.69	0.61	0.69
0.62	0.64	0.62	0.61
0.57	0.54	0.52	0.52

MOSp values (Reference Codec)
Average MOSp = 0.616

0.68	0.65	0.64	0.6
0.7	0.68	0.64	0.71
0.63	0.64	0.61	0.63
0.56	0.56	0.5	0.55

MOSp values (MOSp-based Codec)
Average MOSp =0.623

ROI 2: foreground:

0.58	0.56	0.62	0.61
0.62	0.59	0.6	0.64
0.63	0.61	0.59	0.62
0.62	0.6	0.59	0.61

MOSp values (Reference Codec)
Average MOSp = 0.605

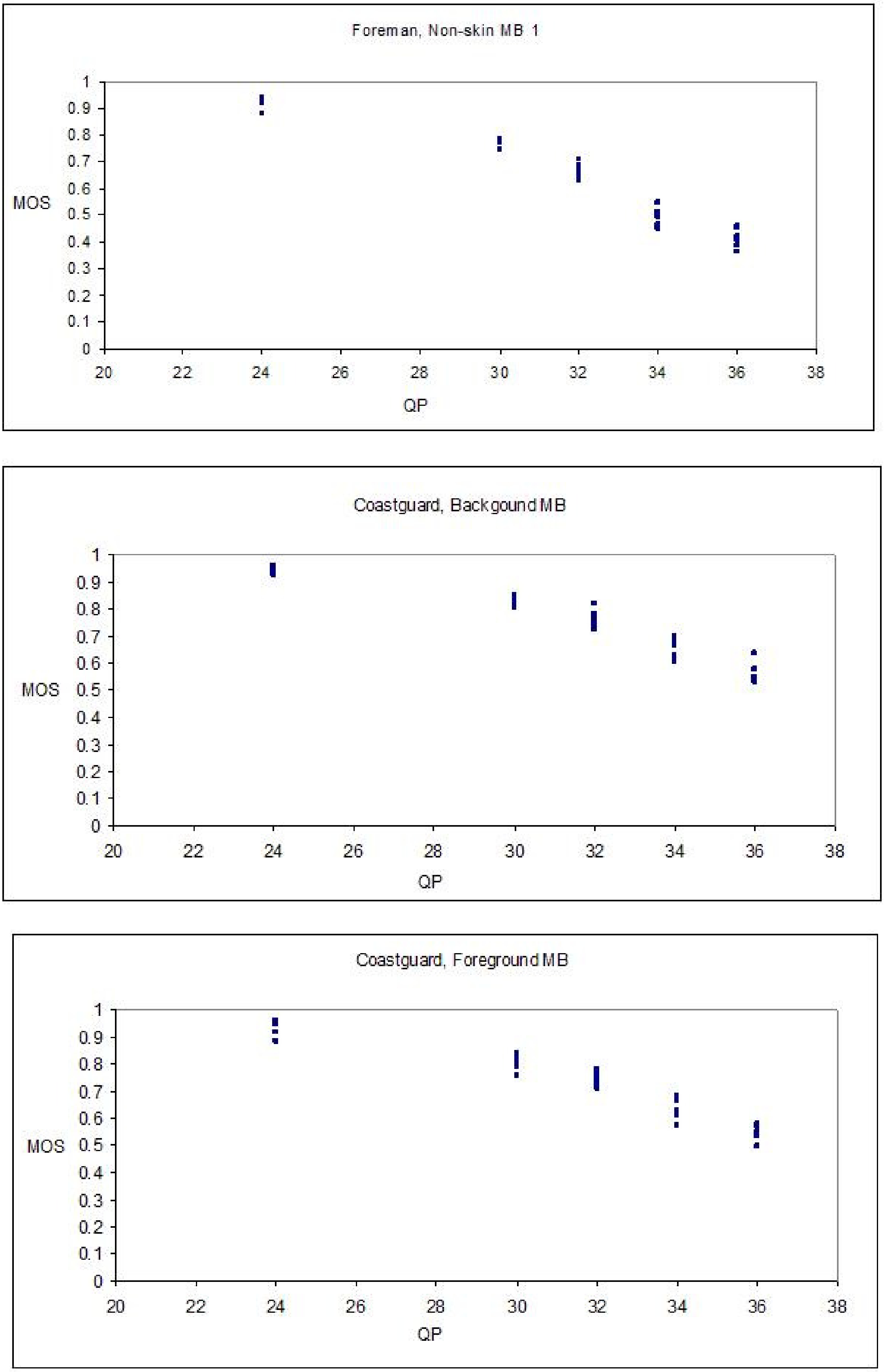
0.62	0.56	0.63	0.65
0.64	0.61	0.62	0.67
0.63	0.63	0.64	0.66
0.65	0.62	0.61	0.63

MOSp values (MOSp-based Codec)
Average MOSp =0.628

Figure 8-7: MB-level analysis for Coastguard CIF sequence, QP=36, frame 29

Figures 8.8 and 8.9 present QP versus MOSp plots for fives sample macroblocks taken from the Foreman and Coastguard regions shown in figure 8.6 and 8.7. These plots show the variations in MOSp for different modes for each QP value. It can be observed that in non-face macroblocks, the variation of MOSp between modes is less compared to face macroblocks. Therefore, switching between modes in a non-face macroblock may not produce a significant gain in visual quality. However in face macroblocks, since the variation in MOSp between modes for each QP is higher, switching the mode selection may produce a significant change in visual quality of the macroblock.

The amount of variation in MOSp for different modes is dependent on the slope parameter of the MOSp metric. The slope of the regression line can be steeper or shallower depending on content. Face macroblocks are assigned the steepest slope value due to high sensitivity to visible distortions and therefore have large variations in MOSp values when compared to macroblocks with shallower slopes. It can also be observed that the variation in MOSp between modes increases with increase in QP. This is more prominent in face macroblocks. This may explain the low gain in visual quality at high bitrates and increase in visual quality gain with decrease in bitrate.

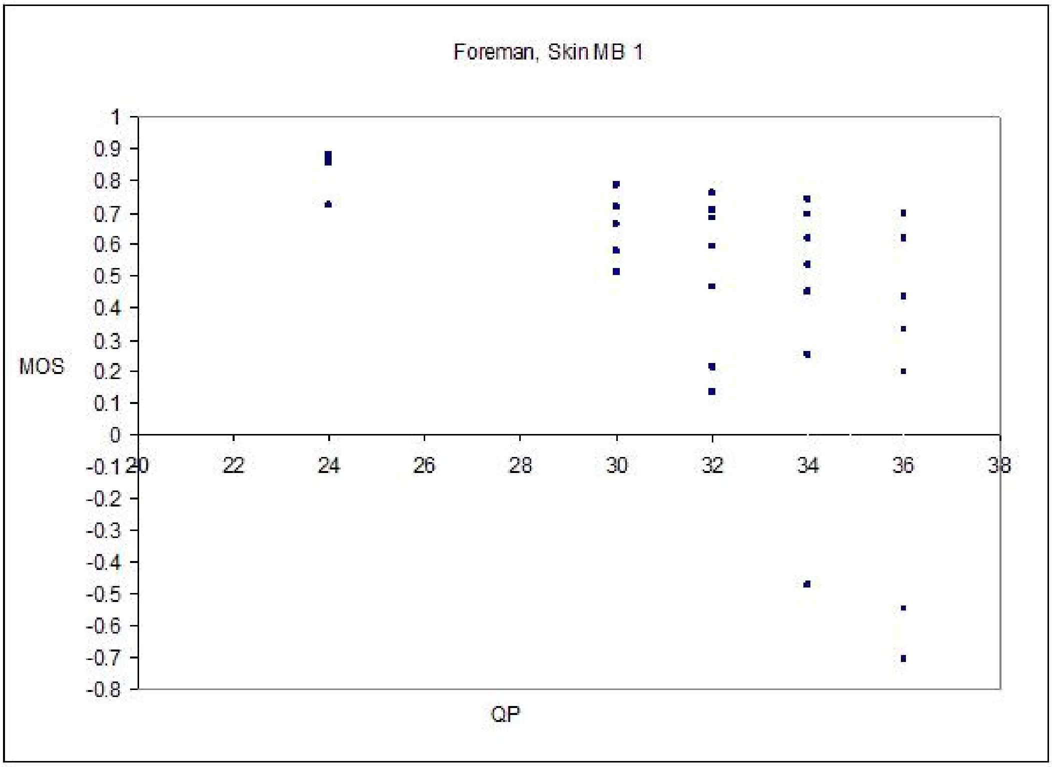


Foreman, Non-skin MB 1

Coastguard, Backgound MB

Coastguard, Foreground MB

Figure 8-8: MB-level analysis using QP versus MOSp plots for various MB modes



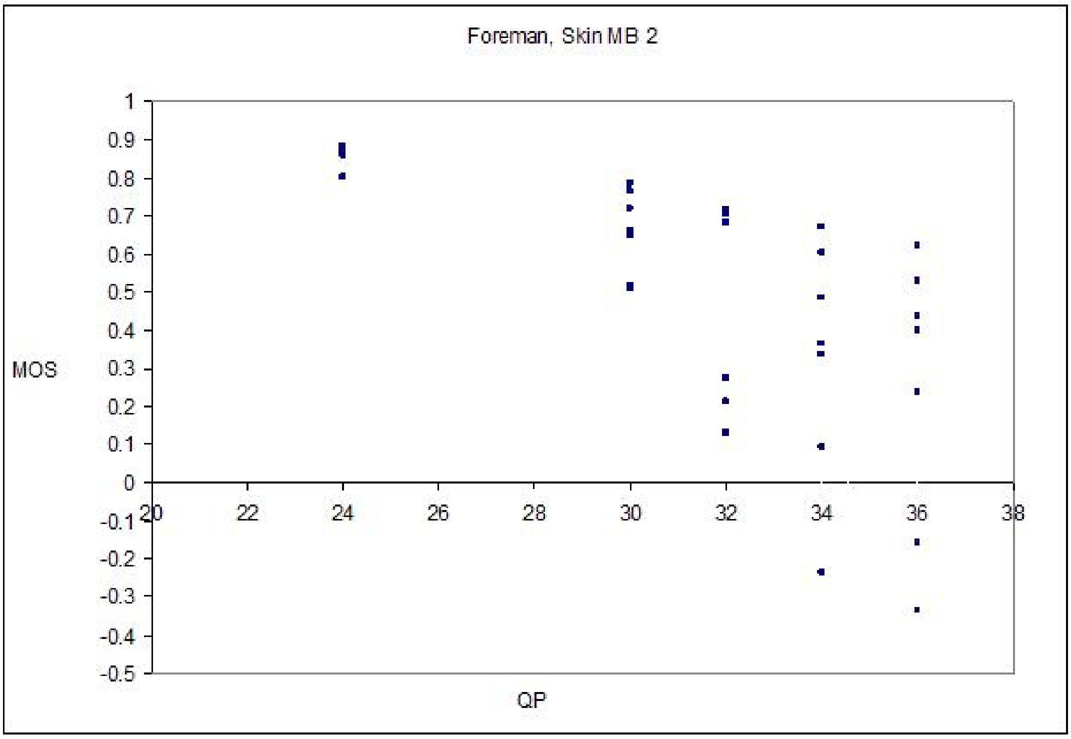


Figure 8-9: MB-level analysis using QP versus MOSp plots for various MB modes

8.7 Summary

This chapter presented a new MOSp-based mode selection algorithm developed to make coding decisions based on visual quality rather than mathematical error measures such as SSD. The MOSp-based rate-distortion model consists of a MOSp-based distortion measure and a new Lagrange multiplier which is derived from QP and activity. The MOSp-based mode selection algorithm was implemented in the H264 JM reference encoder. The performance of MOSp-based mode selection was evaluated using subjective evaluation to investigate whether visual quality gain can be achieved compared to the reference codec for similar bit rate.

Following is the performance summary of evaluating the MOSp-based mode selection algorithm:

- In sequences where humans are present, the MOSp-based mode selection algorithm produces a gain in visual quality (up to MOS = 0.2 on a scale of [0,1]) when compared to the reference codec for similar bitrates.
- In sequences where humans are not present, the MOSp-based mode selection algorithm produces bitrate savings (of up to 26%) when compared to the reference codec for similar visual quality.
- The gap in bitrate-quality performance between the MOSp-based codec and the reference codec is insignificant at higher bitrates. This gap increases with decrease in bitrate.

Based on these observations, it is evident that by incorporating the MOSp-metric into the reference h264/AVC encoder, coding decisions can be made based on visual quality rather than mathematical measures such as SSD and SAD. The results have shown that visual quality gain can be achieved particularly in regions that are sensitive to visible distortions, such as human faces.

Part 3: Discussion & Conclusion

9 Discussion And Future Work

9.1 Introduction

This chapter presents a detailed summary of the main contributions of this work. The developed algorithms and experimental findings are critically analysed with emphasis to their benefits and limitations. The relevance of the main findings to addressing the research problem is also discussed. Finally, possible directions for further developments and improvements in relation to the contributions of this work are presented.

9.2 Main contributions and critical evaluation of results

The aim of this work has been to develop a new perceptual quality metric for compressed video and incorporate it into block-based video coding algorithms so that visual quality estimations can be made in real time video coding algorithms. The main contributions of this work include:

1. Developing a novel perceptual quality metric called the MOSp metric for measuring subjective quality of compressed video.
2. Developing methods to quantify video content using spatial texture, temporal change.
3. Deriving the MOSp metric from MSE and video content.
4. Extending the MOSp metric based on MSE and video content to incorporate cognition-based factors.
5. As an application of the MOSp metric to perceptual video coding, developing a new MOSp metric based mode selection algorithm for a H264/AVC encoder.

A detailed summary of these contributions along with a critical review of the experimental findings are presented in sections 9.2.1 to 9.2.4.

9.2.1 Developing of a novel perceptual quality metric called the MOSp metric

In chapter 5, an experiment on video quality measurement was conducted to investigate the relationship between subjective and objective measures in context to multimedia video sequences compressed using block-based video coding. The results confirm experimentally that there is high correlation between MOS and MSE for a sequence coded to several bitrates using the same coding algorithm. These experimental findings form the foundation of further research in this work and have resulted in the development of a new perceptual quality metric for compressed video. Based on this linear relationship between MSE and MOS, the MOSp metric has been developed for video with compression-induced distortion. The MOSp metric predicts MOS from MSE and the slope of the regression line between MSE and MOS. It was also noted from the video quality experiment that video content may have an influence on the slope of the regression line between MSE and MOS. Video content may include image features (such as texture, colour and motion) and objects that attract viewer attention based on viewer interest and task in hand. Therefore, calculating the parameters of the metric (i.e. slope of the regression line) from video content would make the metric fully automatic. Therefore, the following stages of this work investigated this relationship between video content and the slope parameter of the MOSp metric.

Advantages:

1. Advantages of predicting subjective quality using the MOSp metric include saving time and resources when compared to conducting subjective evaluations to measure video quality of compressed video.
2. The MOSp metric is based on MSE which is a popular video quality metric employed in block-based video coding algorithms. The only additional requirement for MOSp calculation is the slope estimation from video content. Therefore, the MOSp metric would be very useful for integrating into block-based video encoders for making real-time quality estimation.

Limitations:

1. Impairments in compressed video could include compression-induced distortion and transmission-induced distortion. The sequences used to model the MOSp metric contained only compression-induced distortion in

multimedia sequences. Therefore the MOSp metric is limited to assessing quality of video with compression-induced artefacts.

2. Visibility of distortion in video may also be dependent on factors such as viewing distance, frame resolution and frame rate. Since the MOSp metric has been developed using multimedia video sequences, metric parameters such as the slope, may require remodelling for higher resolution video sequences.

9.2.2 Deriving the MOSp metric from MSE and video content

Video quality experiments in Chapter 5 showed that the slope of the regression line varies between sequences and may be dependent on video content. Video content may be contributing to the 'hiding' or 'enhancing' of visibility of distortions which in turn may produce a steeper or shallower slope on the MSE versus MOS graph. This part of the research work, presented in Chapter 6, investigated the relationship between video content and the slope of the regression line between MSE and MOS with a view to automatically estimating the slope parameter for each video sequence.

Video content such as spatial texture and temporal change may contribute to the visibility of distortions. Spatial texture and temporal change may be quantified using spatial edge strength and temporal edge strength measures. Hence, the relationship between these measures and the slope parameter of the regression line between MSE and MOS was investigated in Chapter 6 and two methods for slope estimation have been proposed.

The performance evaluation of the MOSp metric based on these two methods indicate that the MOSp metric produces high correlation with MOS (>90%) with an increase in coding time between 4.9% to 7.7%. The metric also produces higher correlation with subjective results compared to popular metrics such as PSNR, PSNRplus, VSSIM, Yonsei and the NTIAVQM metric.

Other factors influencing the visibility of distortion may include objects in the video scene which attract viewer attention. Hence investigations were carried out in the next stage of this work to see if the MOSp metric could incorporate cognition-based factors such as presence of human in video with a view to further increase its correlation with MOS.

9.2.3 Extending the MOSp metric based on MSE and video content to incorporate cognition-based factors.

Cognition based factors that attract human attention while watching video may be used to classify video content into foreground and background regions. These factors include objects or patterns in the video scene that are 'recognised' by the viewer based on viewer interest, prior knowledge or task-in-hand. Previous research has shown that presence of humans, particularly human faces, in a scene attract visual attention and distortions in these areas caused lower subjective ratings while similar distortion in other areas went unnoticed. Therefore, objects in the video scene which attract viewer attention may contribute to enhancing or masking of visible distortions in compressed video and have effect on the slope of the regression between MSE and MOS. This phenomenon was noticed in the MSE versus MOS graphs in chapter 5. Sequences with human faces, such as Foreman, Akiyo and News have steeper slopes compared to sequences without human faces, such as Bus and Coastguard. Skin colour is a popular cognition-driven perceptual cue and has been proven to be an effective feature in many applications such as face detection and hand tracking.

Hence, in Chapter 7, two methods for integrating skin information in to the MOSp metric were proposed in order to increase its correlation subjective quality:

1. Spatial texture and skin information
2. Spatial texture, temporal change and skin information.

The performance evaluation results show that the MOSp metric produces high correlation with MOS (>90%) with 4.9% to 11.6% increase in coding time and it has higher prediction accuracy compared to popular metrics such as PSNR, PSNRplus, VSSIM, Yonsei and NTIAVQM metric.

Performance comparison between the four methods of calculating MOSp from video content and cognition factors show that the MOSp metric based on spatial texture and skin information produces highest correlation with MOS (95.4%). This high correlation may be due to the following reasons:

(i) The combination of Spatial texture masking and cognition factors has higher influence on the visibility of distortion in video.

(ii) The slope variation between different video content has better correlation with spatial texture and skin information. This results in a more accurate slope estimation model for calculating the slope parameter of the MOSp metric and hence a better performing MOSp metric which produces higher prediction accuracy.

Performance results have shown that the MOSp metric has high prediction accuracy with MOS for a variety of video content. Hence, this proves that the initial hypothesis of predicting MOS by exploiting the linear relationship between MSE and MOS holds for a variety of video content compressed using block-based coding scheme.

Advantages:

1. The MOSp metric has high prediction accuracy with subjective quality.
2. Unlike existing perceptual quality metric which are based on complex models of the human visual system, the MOSp metric is simple to implement and requires reasonable computing time for video quality estimation.
3. Since all the parameters of the MOSp metric are calculated at macroblock level, it can be easily integrated into block-based video coding algorithms for real-time quality estimation.

Limitations:

1. The accuracy of MOSp measurement depends on the slope parameter estimation which is estimated using features in the video that have influence on the visibility of distortion. Therefore, the choice of features used to quantify video content for slope estimation is important and has impact on the prediction accuracy of the MOSp metric.
2. The MOSp metric uses spatial texture and temporal change to quantify video content; it is limited to identifying these regions as being visually important. Other image features such as colour, brightness and contrast may also have

influence on the visibility of distortion. Therefore, incorporating these factors into the MOSp metric may make the metric more robust for different types of video content.

3. The skin detection algorithm used in this work to identify skin regions is based on skin colour detection. Although this method is popularly used for its simplicity and efficiency, it is known to produce false positives. This may have impact on the overall accuracy of the MOSp metric because falsely detected regions in the video scene will have inaccurate MOSp values.
4. In sequences such as sports video, the viewer attention may not be limited to humans in the video scene and the attention may be more focused on other things such as tracking the football or tennis ball, looking out for goals, etc. Incorporating these factors into the MOSp metric may make the metric more robust for different types of video content.

9.2.4 As an application of the MOSp metric to perceptual video coding, developing a new MOSp metric based mode selection algorithm for a H264/AVC encoder.

A new MOSp-based mode selection algorithm for the H264/AVC encoder which employs the MOSp metric in making macroblock mode decisions is presented in Chapter 8. The MOSp metric based on spatial texture and skin information is used for this application due to high prediction accuracy compared to the other three methods of MOSp estimation. The MOSp-based rate-distortion model consists of a MOSp-based distortion measure and a new Lagrange multiplier which is derived from QP and video content. The MOSp-based mode selection algorithm was implemented in the H264 JM reference encoder. Performance of MOSp-based mode selection was evaluated using subjective evaluation to investigate if visual quality gain can be achieved compared to the reference codec for similar bit rate.

Advantages:

1. Performance results show that by integrating the MOSp metric into the mode selection algorithm, coding decisions can be made based on visual quality rather than mathematical measures such as SSD and SAD.

2. In sequences where humans are present, the MOSp-based mode selection algorithm produces a gain in visual quality (up to MOS = 0.2) when compared to the reference codec for similar bitrate. This indicates that incorporating the MOSp metric into the mode selection process produces visual quality gain in content that are identified as visually important by the MOSp metric. Therefore, the MOSp-based mode selection algorithm may be useful in video conferencing and broadcasting applications where improvement in visual quality is more important than bitrate savings.
3. In sequences where humans are absent, the MOSp-based mode selection algorithm produces bitrate savings (of up to 26%) when compared to the reference codec for similar visual quality. The MOSp metric is designed to identify high texture and non skin regions as visually unimportant and hence the mode selection process allocates modes that produce lower bits resulting in overall bitrate savings. Hence, this algorithm may be suitable in applications where bitrate savings are necessary whilst maintaining a certain level of visual quality.
4. The results also showed that the gap in bitrate-quality performance between the MOSp-based codec and the reference codec is insignificant at higher bitrates and increases with decrease in bitrate. Therefore, the MOSp-based mode selection algorithm is more useful in lower bitrate applications such as video communications on mobile platforms.

Limitations:

The MOSp metric integrated into the mode selection algorithm is based on spatial texture and skin information. Therefore, it is limited to identifying human faces and low textured objects in the video scene as being visually important. In sequences such as sports video where the focus of viewer attention may not be limited to these features, the MOSp-based mode selection algorithm does not produce a significant gain in visual quality or bitrate. This limitation may be overcome by incorporating more features into the MOSp metric for indentifying visually important regions in the video scene.

9.3 Suggestions for Future Work

The algorithms developed in this research work were summarised and critically evaluated in the earlier sections. This section presents some suggestions for further research, mainly focused on addressing the limitations of the above algorithms in order to achieve better performance and flexibility.

1. The MOSp metric was built using multimedia video sequences compressed using H264/AVC video coding algorithm. Therefore, the parameters of the MOSp metric have been modelled for multimedia video. In order to develop a generalised MOSp metric, further experimental work is required to investigate the relationship between MSE and MOS for different resolutions, frame rate and video coding schemes.
2. The slope parameter of the MOSp metric is estimated from spatial texture and temporal changes. Including other image features such as colour, brightness and contrast, to identify visually important regions in the video scene may improve performance and flexibility of the MOSp metric. Further experimental work is required to investigate the relationship between these features and the slope of the regression line between MSE and MOS.
3. The robustness of the skin detection algorithm used in this research may be further improved by including other algorithms such as facial feature detection and/or face tracking.
4. The MOSp metric is limited to identifying skin regions as important regions in the video scene which attract human attention. Therefore, more experiments are required to include other application-dependent object detection methods such as vehicle tracking in surveillance video.
5. Experiments to evaluate the MOSp-based mode selection algorithm have shown that integrating the MOSp-metric into the mode selection process improves visual quality in visually important regions such as human faces. Developing a rate control algorithm based the MOSp metric could be a possibility for further research. The MOSp metric may be used for better bit allocation by classifying macroblocks based on the sensitivity to visible distortion and allocating higher bits to visually important regions in the video.
6. The MOSp metric may also be used in low complexity video coding algorithms in order to produce high perceptual quality at reduced processing resource conditions.

10 Conclusion

Video quality measurement is necessary for developing, evaluating and benchmarking video coding algorithms. The subjective measurement of mean opinion score is an accurate method to determine the perceived video quality of compressed video. However, it is expensive in terms of time and resources and cannot be easily embedded into real-time video applications. Hence several objective assessment methods have been developed to predict the subjective results based on video content and the characteristics of the human visual system. The performance of these measures is often limited due to computational complexity and poor correlation with MOS, indicating that there is still scope for developing better approaches to estimate subjective quality. The objective of this research work was to develop novel algorithms to measure perceived video quality of compressed video with a view of improving perceptual quality of compressed video by making coding decisions based on accurately estimated perceptual quality.

The research project was structured into four stages as presented in the Chapter 1 and each stage has been completed successfully. Below is a brief summary of each stage:

Stage 1: Video quality evaluation of compressed video and development of the MOSp metric.

Stage 1 of the project involved conducting a literature review on existing subjective and objective video quality measurement techniques to gain a strong theoretical background and identify limitations of existing techniques. This review is presented in chapters 2 and 3. This stage also involved evaluating video quality of sequences compressed using block-based coding algorithm and investigating the relationship between subjective and objective video quality measures. The results (presented in Chapter 5) proved experimentally that there is high correlation between MSE and MOS for a sequence coded to several bitrates using the same coding algorithm. Based on this linear relationship between MSE and MOS, a new video quality metric called the MOSp metric was developed to predict MOS from MSE and the slope of the regression line between MOS and MSE.

Stage 2: Deriving the MOSp metric from MSE and video content

Experiments conducted in stage 1 to investigate the relationship between MSE and MOS also showed that the slope of the regression line between MSE and MOS varies for different content. Therefore, stage 2 of the project investigated the relationship between video content and the slope parameter of the MOSp metric. Based on these investigations, two methods for estimating slope from spatial texture and temporal change information have been developed. Performance results of the MOSp metric, presented in chapter 6, show that the MOSp metric produces high correlation with MOS (>90%) with an increase in coding time between 4.9% to 7.7%. The metric also produces higher correlation with subjective results compared to popular objective metrics evaluated in the experiment.

Stage 3: Incorporating cognition based factors into the MOSp metric

Factors affecting visual quality of compressed video may include objects in the video scene that attract viewer attention. Therefore, these factors may have influence on the slope of the regression between MSE and MOS. Stage 3 of the project investigated methods of integrating cognition based features such as skin information into the MOSp metric in order to further improve its prediction performance. Based on these investigations as detailed in chapter 8, two methods for estimating the slope parameter of the MOSp metric from spatial texture, temporal change and skin information were developed. Performance results of the MOSp metric show that the MOSp metric based on spatial texture and skin information produces highest correlation with MOS (95.4%) with 8.2% increase in coding time. Hence this metric was used in stage 4 to investigate whether making coding decision based on the MOSp metric improves visual quality of compressed video.

Stage 4: Development of the MOSp-based mode selection algorithm.

In stage 4 of the project, investigates methods to apply the MOSp metric to perceptual video coding. A new MOSp-based mode selection algorithm for the H264/AVC encoder which employs the MOSp metric in making macroblock mode decisions was developed. The MOSp-based rate-distortion model consists of a MOSp-based distortion measure and a new Lagrange multiplier which is derived from QP and video content. The MOSp-based mode selection algorithm was implemented in the H264 JM reference encoder. Performance of MOSp-based mode selection was evaluated using subjective evaluation to investigate if visual quality gain can be achieved compared to the reference codec for similar bit rate. Performance results, presented in chapter 8, show that by integrating the MOSp metric into the mode selection process, it is possible to make coding decision based on estimated visual quality rather than mathematical error measures such as SSD. In sequences where humans are present, the MOSp-based mode selection algorithm produced a gain in visual quality (up to MOS = 0.2) when compared to the reference codec for similar bitrate. In sequences where humans are absent, the MOSp-based mode selection algorithm produced bitrate savings (of up to 26%) when compared to the reference codec for similar visual quality.

This work achieves the main objective of the research project which is to develop a novel technique of measuring perceived video quality of multimedia sequences compressed using block-based coding algorithm. The application of the developed quality metric to perceptual video coding was also investigated. The main contributions of this work include:

- Development of a novel video quality metric called the MOSp metric to predict perceived quality of video sequences compressed using block-based video coding algorithms.
- Development of techniques of automatically calculating parameters of the MOSp metric from MSE and video content.
- Development of techniques to incorporate cognition based factors into the MOSp metric in order to further improve its prediction accuracy.
- Development of a new MOSp-based mode selection algorithm which employs the MOSp metric in making mode selection in order to achieve better visual quality compared to the reference video encoder for similar bitrate.

In comparison with other published work, the main contributions of this work are based on firm theoretical foundations and experimental proof with minimal use of empirically obtained thresholds. The parameters used in the algorithms are adaptive to changing video content. Unlike other perceptual quality metrics, the MOSp metric is computationally simple to implement and requires reasonable running time. Since all the metric parameters are automatically calculated at macroblock level, it can be very easily integrated in to video coding algorithms.

Novel contributions of this work may be used in applications such as video conferencing, multimedia video communications, surveillance and mobile video communications, for automatic perceptual quality estimation in real time. In conclusion, the novel algorithms developed in this research work are particularly useful for integrating into block based video encoders such as H264/AVC in order to make coding decisions based on estimated visual quality rather than the currently used mathematical error measures.

References

[1] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services," May 2003.

[2] S. Saponara, C. Blanch, K. Denolf, and J. Bormans, "The JVT advanced video coding standard: Complexity and performance analysis on a tool-by-tool basis," presented at IEEE Packet Video 2003, Nantes, France, April 2003.

[3] T. Stockhammer, M. M. Hannuksela, and T. Wiegand, "H.264/AVC in Wireless Environments," *IEEE Trans. Circuits and System. Video Technology*, vol. 13, no. 7, pp. 688-703, July 2003.

[4] S. Winkler, "Digital video quality: vision models and metrics", John Wiley & Sons 2005.

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

[6] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications", ITU-R Std., September 1999.

[7] Video Quality Experts Group, "Final Report from the VQEG on the validation of Objective Models of Multimedia Quality Assessment, Phase I", www.vqeg.org, September 2008

[8] ITU-T Recommendation J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference". August 2008.

[9] A. Bhat, I. E. G. Richardson, S. Kannangara and Y. Zhao, "A new perceptual quality metric for compressed video based on mean squared error", *IEEE Signal Process.: Image Commun.*, vol. 25, no. 8, pp. 588–596, Feb. 2010.

[10] A. Bhat, S. Kannangara, Y. Zhao and I. E. G. Richardson, "A full reference quality metric for compressed video based on mean squared error and video content", *IEEE Trans. Circuits and System. Video Technology*, in publication.

[11] A. Bhat, I. Richardson, and S. Kannangara, "A new perceptual quality metric for compressed video," in *Proc. IEEE ICASSP*, pp. 993–936, Apr. 2009.

[12] A. Bhat, I. Richardson, and S. Kannangara, "A novel perceptual quality metric for video compression," in *Proc. PCS*, pp. 1–4, May 2009.

[13] I.E.G. Richardson, "Video codec design: Developing Image and Video Compression Systems", John Wiley & Sons 2002.

[14] R. C. Gonzalez and R. E. Woods, "Digital Image Processing", Prentice Hall, 2002.

[15] A. K. Jain, "Fundamentals of digital image processing", Prentice Hall, 1989.

[16] I.E.G. Richardson, "H.264 and MPEG-4 Video Compression: Video coding for next generation multimedia", John Wiley & Sons 2003.

[17] K.R.Rao and P.Yip, Discrete Cosine Transform: Algorithms, advantages, applications. San Diego, CA: Academic Press, 1990.

[18] N.Ahmed, T. Natarajan and K.R.Rao, "Discrete Cosine Transform", IEEE Transactions on Computers, pp. 90-93, January 1974.

[19] D.Marpe, H.Schwarz and T.Wiegand, “Context-based Adaptive Binary Arithmetic coding in H.264/AVC video coding standard”, IEEE Trans. on Circuits and System, Video Technology, Vol.13, No.7, pp.620-636, July 2003.

[20] P.List, A.Joch, J.Lainema, G.Bjontegaard and M.Karczewiez, “Adaptive de-blocking filter”, IEEE Trans. On Circuits and system, Video Technology, Vol.13, No.7, pp.614-619, July 2003.

[21] ISO/IEC 11172-2 MPEG-1 Video, “Information Technology coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s – part 2: Video”, 1993.

[22] ISO/IEC 13818-2 MPEG 2 – video, “Information Technology: Generic coding of moving pictures and associated audio information: Video”, 1995.

[23]ITU-T Recommendation H.261, “Video codec for audio visual services at px64 kb/s”, 1993.

[24] ITU-T Recommendation H.263, “Video coding for low bit rate communications”, 1998.

[25] A Luthra, G J Sullivan, and T Weigand, "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits and System. Video Technology, vol. 13, pp. 560-576, July, 2003.

[26] H S Malvar, A Hallapuro, and M Karzewicz, "Low-complexity transform and quantisation in H.264/AVC," *IEEE Trans. Circuits and System. Video Technology*, vol. 13, pp. 598-603, July 2003.

[27] M Wien, "Variable block-size transform for H.264/AVC," *IEEE Trans. Circuits and System. Video Technology*, vol. 13, pp. 604-613, July 2003.

[28] P. List, A. Joch, J. Lainema, G. Bjontegaard and M. Karczewicz, "Adaptive Deblocking Filter," *IEEE Trans. Circuits and System. Video Technology*, vol. 13, pp. 614-619, July 2003.

[29] T. Wiegand, G. J. Sullivan, J. Reichel, H. Schwarz, and M. Wien, “Joint Draft 11 of SVC Amendment”, Joint Video Team, Doc. JVT-X201, Jul. 2007.

[30] I.E.G. Richardson, “The H.264 Advanced Video Compression Standard”, John Wiley & Sons 2010.

[31] The Official website for the International Telecommunications Union (ITU): http://www.itu.int

[32] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," in *IEEE Signal Processing Magazine*, vol. 15, no. 6, November 1998, pp. 23-50.

[33] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," in *IEEE Signal Processing Magazine*, vol. 15, no. 6, November 1998, pp. 74-90.

[34] Ge XianXian, Wang Yu, Hao ChongYang, Yang LiNa. Rate-Distortion Optimized Strategy of H.264/AVC. October 2005, Wireless Communication Techniques Magazine.p14-18.

[35] ITU-T, SG15/WP15/1, Q15-D65, “Video Codec Test Model Number 10 (TMN-10)”, Download via anonymous FTP to standard.pictel.com, Apr 1998.

[36] ITU-T (formerly CCITT), “Video Coding for Low Bitrate Communication!, ITU-T Recommendation H.263; version 1, Nov 1995; version 2, Jan 1998.

[37] T. Wiegand, H. Schwarz, A. Joch, F. Lossentini, and G. J. Sullivan, “Rate-Constrained Coder Control and Comparison of Video Coding Standards”, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 13, No. 7, pp. 688-703, July 2003.

[38] Z. Li et al., "Adaptive Basic Unit Layer Rate Control for JVT," JVT-G012, 7th Meeting: Pattaya, Thailand, March 2003.

[39] Z. Li et al., "Proposed Draft of Adaptive Rate Control," JVT-H017, 8th Meeting: Geneva, May 2003.

[40] G. Sullivan, T. Wiegand and K.P. Lim, "Joint Model Reference Encoding Methods and Decoding Concealment Methods; Section 2.6: Rate Control" JVT-I049, San Diego, September 2003.

[41] B. A. Wandell, Foundations of Vision, Sinauer Associates, Inc., 1995.

[42] H.R. Wu and K.R. Rao, Digital Video Image Quality and Perceptual Coding, CRC Press (ISBN: 0-8247-2777-0), Nov. 2005.

[43] R.Aldridge, J.Davidoff, M.Ghanbari, D.Hands & D.Pearson. “Recency effect in the subjective evaluation of digitally-coded television pictures”, *Proceedings of Fifth IEE International Image Processing Conference*, Edinburgh, pp.336-339, July 1995.

[44] S.Wolf and M.H.Pinson, “Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system”, In proceedings of SPIE conference on multimedia systems and applications II, Vol.3845, pp. 266-277, September 1999.

[45] L.P.Gunawan and M.Ghanbari, “Reduced-reference video quality assessment using discriminative local hormonic strength with motion consideration”, IEEE transactions on Circuits and Systems for video technology, Vol.18, no.1, pp 71-83, January 2008.

[46] R.Venkatesh Babu, A.S.Bopardikar, A.Perkis and O.I.Hillestad, “No-reference metrics for video streaming applications”, In proceedings of International Packet workshop, December 2004.

[47] B. Girod, “What’s wrong with mean-squared error,” in Digital Images and Human Vision, A. B. Watson, ed., pp. 207-220, MIT Press, 1993.

[48] S. Winkler, “A perceptual distortion metric for digital colour video,” *Proc. SPIE*, vol. 3644, pp. 175-184, 1999.

[49] A.M. Eskicioglu, P.S. Fisher, “Image quality measures and their performance”, IEEE Trans. Comm. Vol 43, pp. 2959-2965, December 1995.

[50] Z. Wang, H. R. Sheikh, and A. C. Bovik, “Objective video quality assessment”. The Handbook of Video Databases: Design and Applications (B. Furht and O. Marques, eds.), CRC Press, 2003.

[51] M.Webster, “Human colour perception and its adaptation”, Network computation in Neural systems, vol. 7, pp.587-634, 1996.

[52] A.B.Watson, J.Hu, J.F. McGowan III, “DVQ: a digital video quality metric based on human vision”, Journal of Electronic Imaging Vol.10, no.1, pp.20-29, 2001.

[53] J.Lubin et al., “Method and apparatus for assessing the visibility of differences between two image sequences”, US Patent 5,974,159, 1999.

[54] A.P. Bradley, “A wavelet visible difference predictor”, IEEE Trans. Image Process. Vol. 5, pp 717-730, May 1999.

[55] M.A.Masry and S.S.Hemami, “A metric for continuous quality evaluation of compressed video with severe distortions”, Journal of Signal processing and Image communications, Vol. 19, pp 133-146, 2004.

[56] Video Quality Experts Group, “Final Report from the VQEG on the validation of Objective Models of Video Quality Assessment, Phase I”, *www.vpeg.org,* June 2000.

[57] Video Quality Experts Group, “Final Report from the VQEG on the validation of Objective Models of Video Quality Assessment, Phase II”, www.vpeg.org, August 2003.

[58] ITU-T J.144 “Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference”. February 2004.

[59] Recommendation ITU-R BT.1683, “Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference”, January 2004.

[60] S. Wolf and M. Pinson, “Video quality measurement techniques”, NTIA report 02-392, June 2002.

[61] C.Lee, S.Cho, J.Choe, T.Jeong, W.Ahn and E.Lee, “Objective video quality assessment”, Optical Engineering, Volume 45(1), 2006.

[62] Z. Wang, L. Lu and A.C. Bovik, ¨Video Quality Assessment Based on Structural Distortion Measurement,¨ IEEE Signal Proc. Image Communication, vol. 19, no. 2, pp. 121-132, Feb. 2004.

[63] E.P Ong, W. Lin, Lu Zhongkang, S. Yao, M. H. Loke, "Perceptual Quality Metric for H.264 Low Bit Rate Videos", IEEE ICME, pp.677-680, July 2006.

[64] C.H. Chou and Y.C. Li, “A perceptually tuned subband image coder based on the measure of just-noticeabledistortion profile,” IEEE Trans. Circuits Syst. Video Technol., vol. 5, no. 6, pp. 467–476, 1995.

[65] C.H. Chou and C.W. Chen, “A perceptually optimized 3-D subband image codec for video communication over wireless channels,” IEEE Trans. Circuits Syst. Video Technol., vol. 6, no. 2, pp. 143–156, 1996.

[66] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E.P. Ong and S. Yao, “Rate control for videophone using local perceptual cues”. IEEE trans. On Circuits and systems for video technology. Vol. 15, No. 4, pp. 496-507, April 2005.

[67] J.M.Foley, Human luminance pattern-vision mechanisms: Masking experiments require a new model, Journal of the Optical society of America, vol.11, no. 6, pp.1710-1719, 1994.

[68] E.P Ong, X.Yang, W.Lin, Lu Zhongkang, S. Yao, X.Lin, S.Rahardja and B.C.Seng, "Perceptual quality and objective quality measurements of compressed videos", Journal of Visual Communications and Image Representation, Vol 17, pp.717-737, 2006.

[69] M. Masry and S.S. Hemami, "An analysis of subjective quality in low bit rate video", Proc. ICIP, pp. 465-468, Thesaloniki, Greece, 2001.
[70] S.Yao, W.Lin, Z.Lu, E.P.Ong, M.Etoh, "Objective quality assessment for compressed video", IEEE International symposium on circuits and systems, pp.688-691, May 2003.

[71] T. Oelbaum, K Diepold and W. Zia, ¨A generic method to increase prediction accuracy of visual quality metrics¨, PCS 2007.

[72] *Video Quality Experts Group (VQEG)* official website: www.vqeg.org

[73] *Xiph.org Test Media* [Online]. Available: http://media.xiph.org/video/derf

[74] Joint Video Team (JVT) of ISO/IEC MPEG and ITU T VCEG, "JM12.1 Test Model Codec", http://iphome.hhi.de/suehring/tml/.

[75] JM software download: http://iphome.hhi.de/suehring/tml/download/

[76] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, "JVT-AE010: H.264/14496-10 AVC Reference Software Manual", 31st Meeting, London, UK, 28 June – 3 July 2009

[77] P. Corriveau. Video quality testing. In H.R. Wu and K.R. Rao, editors, *Digital Video Image Quality and Perceptual Coding*, pages 125–153. CRC Press, 2006.

[78] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach. All subjective scales are not created equal: The effects of context on different scales. *Signal Processing*, 77(1):1–9, August 1999.

[79] G. J. Sullivan, P. Topiwala and A. Luthra, "The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions", SPIE Conference on Applications of Digital Image Processing XXVII, August, 2004.

[80] X. Ran and N. Farvardin, ¨A perceptually motivated three-component image model – Part 1: Description of the model¨, IEEE Trans. On Image Proc, Vol. 4(4), pp.401-415, 1995.

[81] Carney, S. A. Klein, Q. Hu, "Visual masking near spatiotemporal edges", Proc. SPIE Human Vision and Electronic imaging, Vol. 2657, 1998, pp-393-402.

[82] Z. Wang, L. Lu and A. C. Bovik, "Video Quality Assessment Based on Structural Distortion Measurement". In Signal Processing: Image Communication, Vol. 19, No 2, pp. 121–132, February 2004

[83] Joint Video Team (JVT) of ISO/IEC MPEG and ITU T VCEG, "JM12.1 Test Model Codec", http://iphome.hhi.de/suehring/tml/.

[84] Y Zhong, I Richardson, A Sahraie and P McGeorge, "Influence of Task and Scene Content on Subjective Video Quality", Lecture Notes in Computer Science, Volume 3211 / 2004, pp. 295.

[85] A.E. Savakis, S.P. Etz, A.C. Loui, "Evaluation of image appeal in consumer photography", in Proc. SPIE, Vol. 3959, pp. 111-120, San Jose, CA, 2000.

[86] M.J. Jones, J.M.Rehg, "Statistical color models with application to skin detection". International Journal of Computer Vision (IJCV), Vol.46, no.1, pp.81-96, 2002.

[87] M.H. Yang, D.J. Kriegman and N. Ahuja, "Detecting faces in images: a survey", IEEE trans. Pattern Analysis and Machine Intelligence, Vol. 24, No. 1, pp.34-58, Jan 2002.

[88] P. Kakumanu, S. Makrogiannis and N. Bourbakis, "A survey of skin-color modelling and detection methods", Pattern Recognition, Vol. 40, No. 3, 2007, pp. 1106-1122.

[89] R.L, Hsu, A.M. Mohammad and A.K. Jain, "Face detection in colour images", IEEE trans. Pattern Analysis and Machine Intelligence, vol 24, no. 5, pp.696-706, May 2002.

[90] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity. IEEE Trans. On Image Processing, Vol. 13, pp. 600-612, April 2004.

[91] Z. Mai, C. Yang, K. Kuang, and L. Po, A novel motion estimation method based on structural similarity for H.264 inter prediction, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 913–916, 2006.

[92] C. Yang, H. Wang, and L. Po, Improved inter prediction based on structural similarity in H.264, *IEEE International Conference on Signal Processing and Communications*, vol. 2, pp. 340–343, 2007.

[93] T. Ou, Y. Huang, and H. Chen, A perceptual-based approach to bit allocation for H.264 encoder, *SPIE Visual Communications and Image Processing*, Jul. 2010.

[94] S. Wang, A. Rehman, Z. Wang, S. Ma and W. Gao, Rate-distortion optimisation for video coding.

[95] K. P. Lim, G. Sullivan, and T. Wiegand, "Text description of joint model reference encoding methods and decoding concealment methods," in *24th Meeting of Joint Video Team (JVT)*, Geneva, Switzerland, Jun. 2007, doc. JVT-X101.

Bibliography

S. Winkler, ”Digital video quality: vision models and metrics”, John Wiley & Sons 2005.

I.E.G. Richardson, “Video codec design: Developing Image and Video Compression Systems”, John Wiley & Sons 2002.

R. C. Gonzalez and R. E. Woods, “Digital Image Processing”, Prentice Hall, 2002.

A. K. Jain, “Fundamentals of digital image processing”, Prentice Hall, 1989.

I.E.G. Richardson, “H.264 and MPEG-4 Video Compression: Video coding for next generation multimedia”, John Wiley & Sons 2003.

K.R.Rao and P.Yip, Discrete Cosine Transform: Algorithms, advantages, applications. San Diego, CA: Academic Press, 1990.

I.E.G. Richardson, “The H.264 Advanced Video Compression Standard”, John Wiley & Sons 2010.

A. Wandell, Foundations of Vision, Sinauer Associates, Inc., 1995.

H.R. Wu and K.R. Rao, Digital Video Image Quality and Perceptual Coding, CRC Press (ISBN: 0-8247-2777-0), Nov. 2005.

Appendix A: List of Publications

Journals:

A. Bhat, I. E. G. Richardson, S. Kannangara and Y. Zhao, "A new perceptual quality metric for compressed video based on mean squared error", *IEEE Signal Process.: Image Commun.*, vol. 25, no. 8, pp. 588–596, Feb. 2010.

A. Bhat, S. Kannangara, Y. Zhao and I. E. G. Richardson, "A full reference quality metric for compressed video based on mean squared error and video content", *IEEE Trans. Circuits and System. Video Technology*, in publication.

Conference papers:

A. Bhat, I. Richardson, and S. Kannangara, "A new perceptual quality metric for compressed video," in *Proc. IEEE ICASSP*, pp. 993–936, Apr. 2009.

A. Bhat, I. Richardson, and S. Kannangara, "A novel perceptual quality metric for video compression," in *Proc. PCS*, pp. 1–4, May 2009.

Appendix B

Training instructions given to viewers during subjective evaluations:

"In this experiment you will be shown short video sequences on the screen one at a time. Each time a sequence is shown, you should judge its picture quality by choosing a five-point scale."

(i)*Excellent*: if the content in the video sequence has no noticeable distortion.

(ii)*Good*: at least one noticeable distortion is detected in the entire sequence.

(iii)*Fair*: several noticeable distortion are detected, spread all over the sequence.

(iv)*Poor*: many noticeable distortion which destroy the scene structure or create new patterns in some parts of the sequence, are detected.

(v)*Bad*: very strong noticeable detected which destroy the scene structure or create new patterns in the major part of the sequence.