NTNU
Innovation and Creativity

Subjective quality evaluation of the effect of packet loss in High-Definition Video

Sander Sunde Vorren

Master of Science in Electronics
Submission date: June 2006
Supervisor: Andrew Perkis, IET
Co-supervisor: Odd Inge Hillestad, IET

Norwegian University of Science and Technology
Department of Electronics and Telecommunications

Problem Description

Video streamed over packet-switched networks such as the Internet are vulnerable to packet loss, which result in a degradation. This degradation can be measured subjectively or by objective measures.
The Internet is a best-effort media-unaware environment where all packets receive equal quality of service (QoS), disregarding the fact that some packets are more essential in streaming multimedia applications. Using the differentiated services (DiffServ) model, unequal degrees of QoS are offered, resulting in essential packets being prioritized through the network.

This shall study the performance of such networks considering both objective video quality models and subjective assessments, and how well the objective results correlate with the subjective assessments.

The video sequences shall be encoded by using the H.264/AVC video compression standard, and further transmitted in RTP packet as described in [1]. Packet loss shall be introduced by using a DiffServ simulator, where decoded distorted sequences are assessed as recommended in [2].

[1] S. Wenger. H.264/AVC over IP. IEEE Trans. Circuits and Systems for Video Technology, vol. 13(no. 7):645–656, July 2003.

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

Assignment given: 16. January 2006
Supervisor: Andrew Perkis, IET

Summary

Video streamed over packet-switched networks such as the Internet are vulnerable to packet loss, which result in a degradation of quality. This degradation can be measured subjectively or by objective measures. The Internet is a best-effort media-unaware environment where all packets receive equal quality of service (QoS), disregarding the fact that some packets are more essential in streaming multimedia applications. Using the differentiated services (DiffServ) model, unequal degrees of QoS are offered, resulting in essential packets being prioritized through the network.

The main objective of this thesis is to conduct an informal subjective evaluation experiment, where the test material used consists of high-definition video distorted by various packet loss rates, using both the best effort Internet and DiffServ as underlying channel models. The results from the subjective evaluation experiment are compared to those of the objective video quality estimation to see how well the objective models perform.

The video sequences are encoded by using the H.264/AVC video compression standard, and further transmitted in RTP packets. Packet loss is introduced by using a DiffServ simulator, where decoded distorted sequences are assessed.

Results show that the NTIA and SSIM were the video quality models with respectively the highest and the lowest performance regarding PLCC, SRCC and RMSE. The NTIA model had statistically significant higher performance than SSIM using PLCC and SRCC with a 95% confidence interval.

When comparing packet loss rate versus objective measures, the performance of best effort degrades more rapidly than the performance of DiffServ. However, the results from the subjective evaluations did not show any statistically significant differences between the two channel models using a 90% confidence interval.

The DMOS values were categorized into low, medium and high packet loss rates. Studying the high (5-10%) packet loss rate category, the DiffServ model achieved a higher mean DMOS value compared to the Best Effort model. For low (0-2.5%) and medium (2.5-4%) packet loss rates categories the Best Effort model achieved a higher mean DMOS value compared to the DiffServ model.

Preface

This diploma thesis, written between January and June 2006 at the Norwegian University of Science and Technology (NTNU), serves as a culmination of my higher academic education. When looking back on the years spent as a student, I find myself satisfied with the knowledge I have acquired, the people I have met, and experiences I have had.

I would like to extend my deepest gratitude to professor Andrew Perkis and his doctoral student Odd Inge Hillestad. They have both contributed greatly in terms of motivating and guiding me through the proceedings of this thesis. I regard them highly, and hope they are pleased with the effort I have put into this thesis. Furthermore, I would like to thank my friends and my family for supporting me through all these years.

Contents

List of Figures

List of Tables

List of Abbreviations

ACR-HRR	Absolute Category Rating with Hidden Reference Removal
CI	Confidence Interval
CODEC	COder-DECoder
DiffServ	Differentiated Services
DMOS	Difference Mean Opinion Score
DSCP	DiffServ Code Point
DT	Drop Tail
FMO	Flexible Macroblock Ordering
GOP	Group of Pictures
HDTV	High-Definition Television
HRC	Hypothetical Reference Circuit
IDR	Istant Decoding Refresh
IntServ	Integrated Services
ITU	International Telecommunication Union
JVT	Joint Video Team
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MTU	Maximum Transfer Unit
PLCC	Pearson Linear Correlation Coefficient
PSNR	Peak Signal-to-Noise Ratio
PVS	Processed Video Sequence (PVS = HRC $*$ SRC)
QoS	Quality of Service
PLR	Packet Loss Rate
RTP	Real-Time Transfer Protocol
SSIM	Structural SIMilarity video quality model
SRC	Source Video Sequence
SRCC	Spearman Rank Order Correlation Coefficient
TCP	Transmission Control Protocol
TOS	Type Of Service
UDP	User Datagram Protocol
VQM	Video Quality Metric or Video Quality Model
VQR	Video Quality Rating
VQEG	Video Quality Experts Group

Chapter 1

Introduction

During the past decade, technological advancements in video coding standards have created new areas of application. The ever-growing Internet with its increasing bandwidth and advanced access technologies is one of the major distribution channels for delivery of multimedia content. The Internet is a challenging environment for real-time streaming applications since the provided best effort service may result in delays and packet loss. The impact of packet loss, when streaming video, highly depends on the contents of the lost packet. A differentiated services (DiffServ) architecture can be employed to avoid losing the most important packets by differentiating the quality of service packets receive.

When a video stream is subject to packet loss, the perceived quality may suffer. The perceived quality of a video can be measured by using subjective metrics where viewers rate video quality. This method of quality assessment is expensive and time consuming. Therefore, objective metrics are employed as a reasonable replacement where they estimate the perceived quality of a video.

This thesis conducts an informal subjective evaluation experiment, where the test material consists of high-definition video distorted by various packet loss rates, using both the best effort Internet and DiffServ as transmission channels while introducing packet loss. Objective metrics are also employed to estimate perceived quality of video for the test material. The results from the subjective evaluation experiment are compared with results from objective video quality estimates to see how well the objective models perform. The main objective in this thesis is to determine if employing DiffServ results in higher subjective ratings compared to the best effort Internet ratings.

This thesis is structured as follows: Chapter 2 presents theory relevant to video coding and streaming over both best effort Internet and DiffServ networks and a description of video quality models. Chapter 3 contains a contains the system desciption for conducting the experiments. Chapter 4 gives the results and a discussion. Chapter 5 contains the conclusion and suggestions for further work.

Chapter 2

Theory

This chapter provides background theory for topics used to produce and evaluate test material for the subjective evaluation procedures.

2.1 The H.264/AVC video compression standard

The H.264/AVC (Advanced Video Coding) is the newest video coding standard from the Joint Video Team (JVT), which is a collaboration between the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group (MPEG). The standard has been approved by ITU-T as Recommendation H.264[1] and by ISO/IEC as International Standard 14496-10 (MPEG-4 part 10) Advanced Video Coding (AVC). The H.264/AVC is a block-based motion compensated hybrid video coding scheme, similar to prior ITU-T and MPEG video coding standards, and it is known for its high coding efficiency over a wide variety of application scenarios, suitable for both low and high bit rates, as well as low and high resolution video.

In order to develop a flexible and customizable video coding scheme, JVT divided the codec into a layered structure, separating the compression and coding of video layer from the network adaptation layer. These two layers, depicted in figure 2.1, are denoted as the video coding layer (VCL) and network abstraction layer (NAL).

The following sections introduce these two layers. For a more in-depth description, see Wiegand et al. [2].

Video Coding Layer

The video coding layer's main task is to provide an efficient representation of the content of video data. This done by employing the classical hybrid coding structure combined with motion compensation. When encoding, a video sequence is divided into pictures. Each picture is then

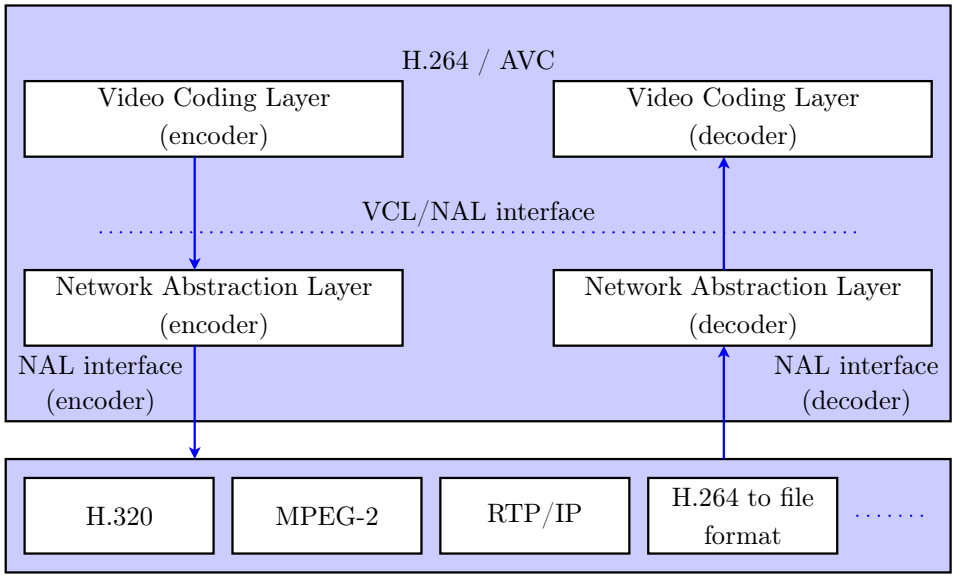


Figure 2.1: The layered structure and the transport environment of H.264/AVC.

sub sampled to the 4:2:0 sampling format in the YCbCr color space. The sub-sampled picture is further divided into 16x16 sample blocks, also referred to as macroblocks. Each macroblock is processed as depicted in the coding structure shown in figure 2.2. The macroblock is motion compensated, transformed, quantized and entropy coded before it is sent to the NAL.

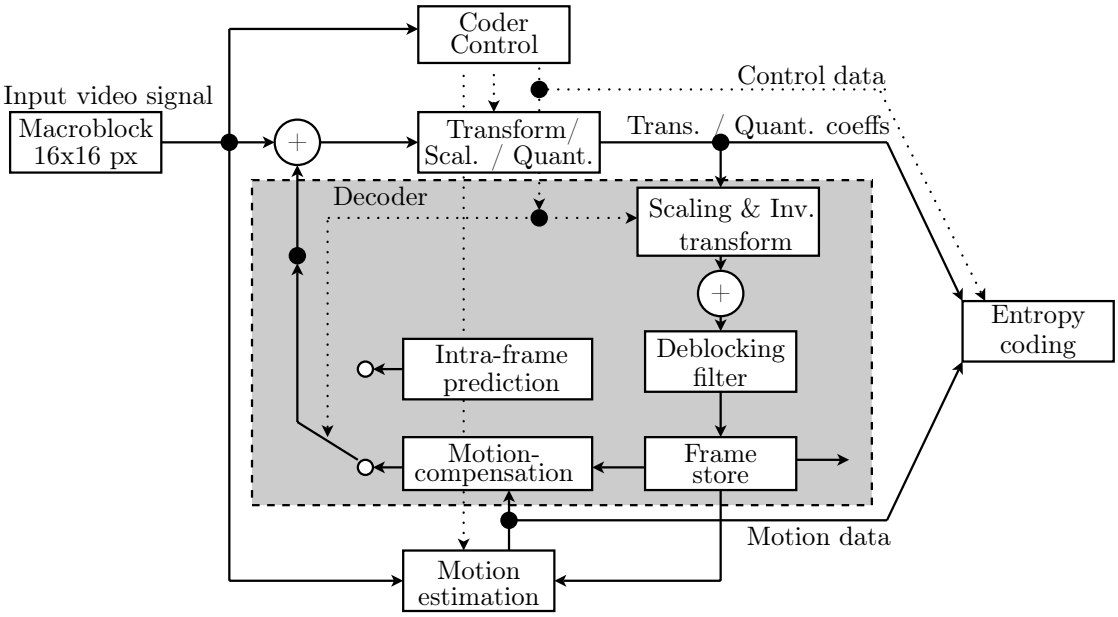
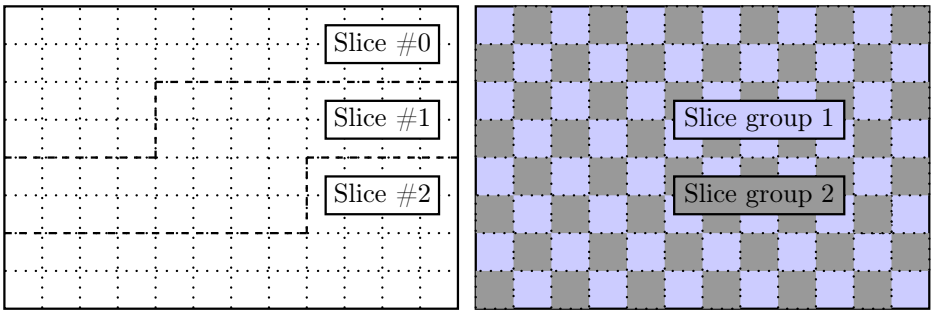


Figure 2.2: Basic coding structure for a macroblock as depicted in Wiegand et al.[2].

Macroblocks are grouped into slices, which are the smallest self-contained video coding units, in a raster scan order. A frame will always consist of one or more slices, which are all coded independently, meaning that they can be decoded without data from other slices. The baseline and extended profile in H.264/AVC also offer flexible macroblock ordering (FMO), where macroblocks

can be grouped in slice groups in a non-raster scan order defined as a slice group map. These maps are the interleaved slices, dispersed macroblock organization, foreground with left-over etc. Within a slice group, macroblocks can be further grouped into slices, where a raster scan order within the slice group map is used. The use of several slices or the use of slice groups will add some degree of error-resilience in preventing errors to propagate across slice boundaries. However, this is done at the expense of coding efficiency, where the spatial redundancy between slices is not removed.



(a) A frame divided into three slices. (b) Dispersed FMO with two slice groups.

Figure 2.3: Slicing by raster scan order and dispersed FMO.

Regardless using of FMO or not, slices can be coded by the following types:

- I slice:
  All macroblocks within a slice are coded using intraprediction. These slices are used as a reference picture for prediction of P and B slices.

- P slice:
  Macroblocks are coded by using interprediction with only one vector per block. Other P slices and I slices are used for prediction. The slice can contain macroblocks coded either by inter- or intraprediction.

- B slice:
  Macroblocks can be coded using interprediction with two motion vectors per block, also known as bi-predictive coding. Either I, P or B slices are used as prediction reference.

As an addition to these slices, a switching P slice and a switching I slice is defined in the standard. These provide functionalities for bit stream switching, random access, error resilience and error recovery.

A typical temporal dependency between slices is shown below in figure 2.4. In order to prevent either prediction or error propagation across GOP boundaries instantaneous decoding refresh (IDR) pictures can be employed. The IDR picture consists only of I slices, where after the decoding an IDR picture, all following coded pictures can be decoded without inter prediction from reference pictures prior to the IDR picture.

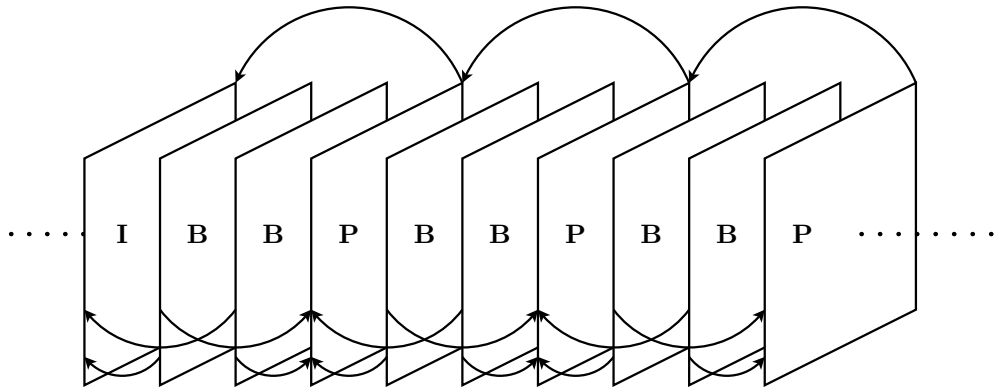


Figure 2.4: The temporal dependencies in a typical GOP.

Network Abstraction Layer

The network abstraction layer receives bit strings from the VCL, and adapts these to various networks and multiplex environments. To simplify, this thesis focuses on packetization of these bit strings[1] The bit strings the NAL receives, are placed in NAL units, which is a byte string containing coded slices, data partition slices, sequence or picture parameter sets. NAL units also consist of a header, which describes the content and importance of unit. A full description of the NAL unit and its header structure can be found in section 2.3.

2.2 The Internet and DiffServ

Currently, the Internet is a packet-switched network that does not provide a guaranteed service for the transmission of data. A sender splits the data into packets, and sends them to a receiver, without knowing what route packets travel, or whether the packet has reached the recipient, and if received, in what order and at what delay. For these reasons, the Internet is well described as a *best effort* network, where the Internet makes its best effort to deliver packets without making any guarantees.

Protocols

Protocols can be divided into a connectionless unreliable service and a connection-oriented reliable service, where the latter guarantees the delivery of packets eventually. An example of a connection-oriented service is the transmission control protocol (TCP[3]). This protocol includes the well-known three-way handshake, as well as retransmission and congestion control. These functions are vital in many applications such as file-transfer, but introduce some adverse effects regarding real-time multimedia streaming. Examples of these are the three-way handshake set-up that introduces delay, and the constant feedback from acknowledgements that generate traffic. The

[1] Annex B in the H.264/AVC standard specifies the bit stream format of encoded H.264/AVC video.

possibility of retransmitting lost packets is often considered unnecessary due to the imposed low-delay requirements of inelastic real-time multimedia systems.

Real-time applications, such as multimedia streaming, are more loss-tolerant than applications that use e.g. TCP, but often, require a minimum rate and low delay. For such applications, the connectionless UDP (User Datagram Protocol[4]) datagrams are suitable as transport protocol. UDP gives no guarantee that a datagram will reach the receiver or in which order datagrams arrive. Furthermore, UDP has no form for congestion control, meaning that UDP can send packets at desired rates, with less overhead than TCP, regardless of the amount of cross traffic in the network.

A vast number of applications use the Real-Time Transport Protocol (RTP[5]) for streaming both real-time and stored multimedia content over the Internet. Streaming applications commonly use RTP over UDP to benefit from RTP's services. The protocol neither assumes that the underlying transport is reliable, nor does it assume that packets arrive in correct sequence.

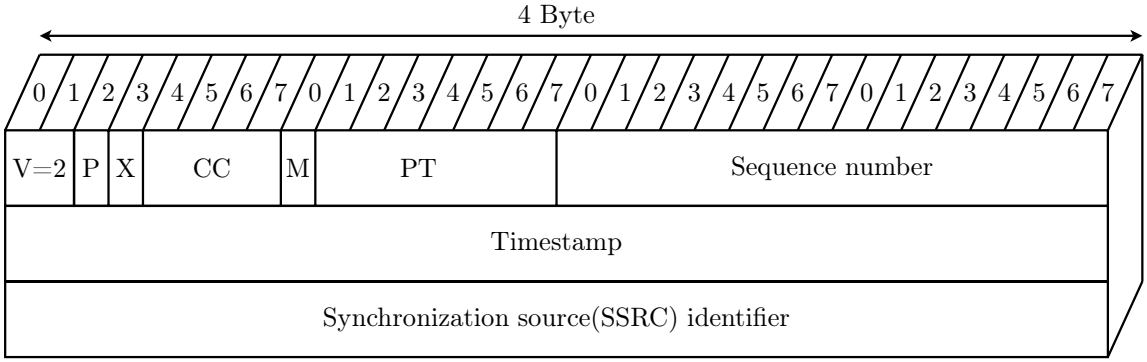


Figure 2.5: RTP header syntax. These twelve octets are present in every RTP packet.

As shown in figure 2.5, each packet is marked with a sequence number, a payload type and a time stamp(For a full description of each field, see [5, Section 5.1]):

- The payload type(PT) is a 7-bit value which identifies the format of the RTP payload to determine how an application should interpret data.
- The sequence number is a 16-bit value that is incremented for each RTP packet sent. This field can be used to detect packet loss and to restore the correct packet sequence.
- The time stamp field is 32 bits long and reflects the sampling instant of the first byte in the RTP packet payload. This field is used for synchronization and jitter calculations by using a clock that linearly increases in time. The clock frequency is dependent on the payload type specified in the RTP header.

As an addition to RTP, systems can use the RTP Control Protocol (RTCP) to monitor and control the transmission of data. RTCP packets do not include any multimedia content, but sender and/or receiver reports containing statistics regarding the quality of service being delivered. These statistics include the number of packets sent, packets lost and the inter-arrival jitter, and is periodically sent to all participants in the streaming session. RFC 3550[5] do not dictate how these statistics should be used, but one example is by increasing/decreasing transmission rate (a form of congestion control).

Congestion and queue management

The Internet is a *best effort* network where senders hope a packet is received. This is a result of all packets being treated equally once they have entered the public Internet. A common scenario in the Internet is congestion. To make the most of the resources available, transport protocols increase the rate at which they is transmitting until congestion symptoms appear. When congestion occurs, routers discard packets because of full queues or buffers. The router's queue management policy determines which and how the router drops the packets. Several different queue management policies exists, but traditionally Internet routers employ the *drop tail*(DT) policy with the *First In First Out* (FIFO) forwarding scheme. If the queue or buffer is full when a packet arrives, FIFO drops that packet. Internet applications often send packets in a bursty manner, leading to large bursty packet losses when using DT.

An alternative to this method is Random Early Detection (RED). RED is an active queue management (AQM) policy that actively tries to prevent congestion. Figure 2.6 shows how the RED determines when to drop a packet. Another variant of the RED policy is discussed later in this chapter.

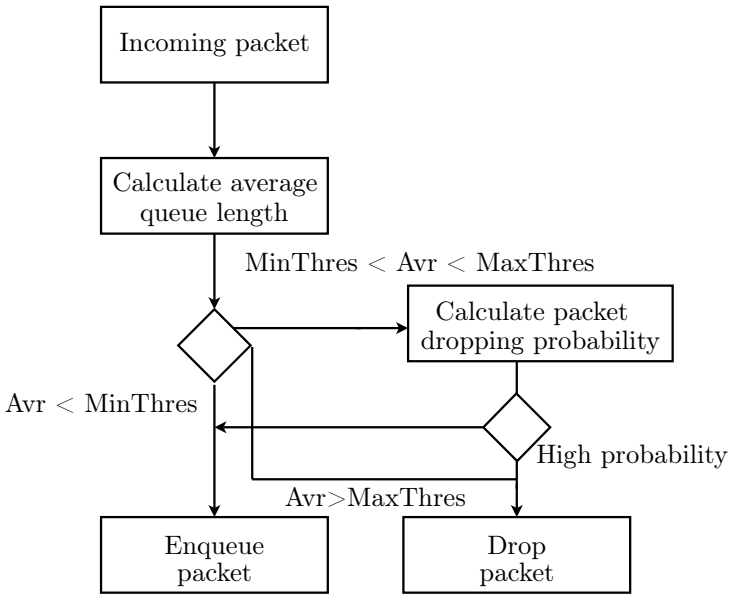


Figure 2.6: Random Early Detection block diagram.

Best Effort and DiffServ

The Internet, with its *best effort* performance, makes no promise regarding the quality of service an application can expect to experience. Applications will receive the level of performance that the network is able to provide at the given moment, without the possibility of requesting e.g. low end-to-end packet delays and low packet loss rates. In order to accommodate such requests, the Internet Engineering Task Force (IETF) has developed two architectures, *integrated services* (IntServ)[6] and *differentiated services* (DiffServ)[7], which provide quality of service in packet-switched networks. This section will briefly explain the differences between the two architectures, followed by a more detailed description of the latter.

The first architecture IETF proposed was the Integrated Services or IntServ. This approach uses the Resource Reservation Protocol (RSVP)[8] to reserve resources through the Internet, providing either a guaranteed quality of service or a controlled-load network service on a per-flow basis. The reservation of resources through the Internet results in the need to maintain state for each flow through a router, which restricts the architecture's level of scalability. Another difficulty associated with this approach is the lack of flexible service models.

A more scalable and flexible approach is the DiffServ architecture which has the ability to handle different classes of traffic in different ways within the Internet. Instead of maintaining a per-flow state in each router as IntServ does, DiffServ classifies and marks each packet according to a service level agreement (SLA). The edge routers, which also shape or drop packets that are out-of-profile (not in accordance with the SLA) do the marking and classification. The metering function, as shown in figure 2.7, compares an incoming packet flow with a negotiated traffic profile and determines if a packet is within the negotiated traffic profile.

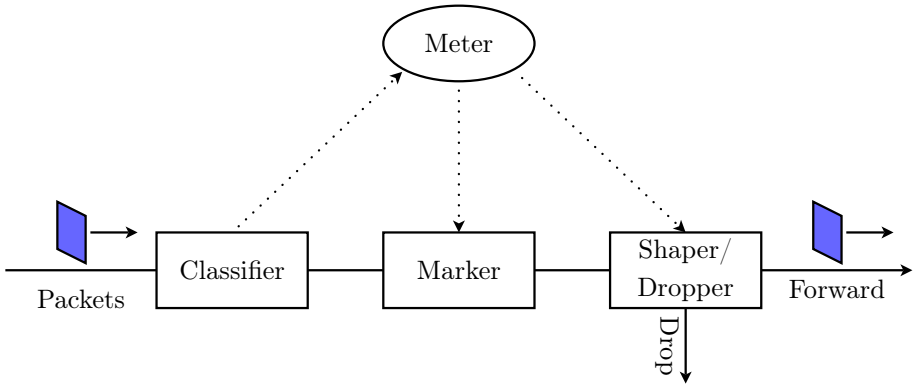


Figure 2.7: Packet classification and traffic conditioning at the edge router.

When a packet has been classified, marked and shaped it gets forwarded from the edge router into a DiffServ domain (DS), which is a set of routers operating with a common set of service provisioning policies and *per-hop behaviour* (PHB) definitions. Within the DiffServ domain, marked packets receive the service associated with their marks. As opposed to IntServ, the DiffServ core routers do not need to maintain a per-flow or a per-packet state, resulting in a lower accumulated workload and signalling as compared to IntServ routers workload. DiffServ is a more scalable model for guaranteeing quality of service in both small and large networks, because only edge routers classify, mark and shape packets. IETF specify two different PHBs to classify and mark packets:

- Expedited forwarding (EF)
  The expedited forwarding[9] behaviour is defined as the highest service class in the DiffServ architecture. DiffServ routers guarantee that this class will receive enough bandwidth such that the output rate equals or exceeds a minimum configured rate. The EF traffic class is guaranteed this regardless of other traffic classes. Even if the router links and queue resources are depleted by traffic from other traffic classes, resources will be freed to accomodate EF traffic.

- Assured forwarding (AF)
  The assured forwarding[10] class is a more complex structure, where the class again is divided into four subclasses, where each of them is guaranteed a minimum amount of bandwidth and buffering. Each of the four classes can further be divided into three drop precedence categories, which are often referred to as the gold, silver and bronze classes of service.

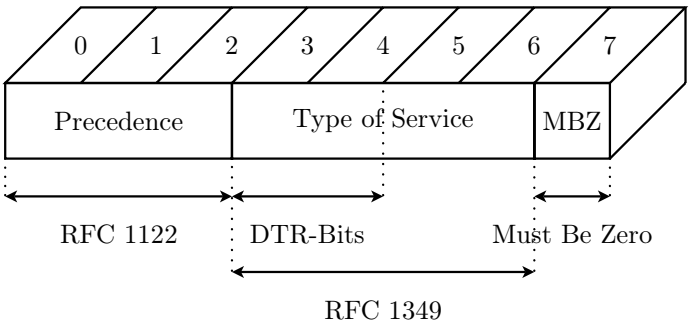


Figure 2.8: The IPv4 TOS Byte.

Packets are marked by employing the IPv4 type of service field (see figure 2.8) or the IPv6 traffic class field. The field is redefined as a DiffServ code point (DSCP) field, as shown in figure 2.9. Only the first six bits are used to identify the traffic class the packet is marked with. Table 2.1 show the relationship between bit pattern and traffic class.

PHB	DSCP
EF	1 0 1 1 1 0
AF11	0 0 1 0 1 0
AF12	0 0 1 1 0 0
AF13	0 0 1 1 1 0
AF21	0 1 0 1 0 0
AF22	0 1 0 1 1 0
AF23	0 1 1 0 0 0

PHB	DSCP
AF31	0 1 1 0 1 0
AF32	0 1 1 1 0 0
AF33	0 1 1 1 1 0
AF41	1 0 0 0 1 0
AF42	1 0 0 1 0 0
AF43	1 0 0 1 1 0
BE	0 0 0 0 0 0

Table 2.1: Relationship between traffic classes and bit patterns in the DSCP field. AFx1, AFx2 and AFx3 respectively equals the gold, silver and bronze drop precedence categories, where x denotes one of the four assured forwarding traffic classes.

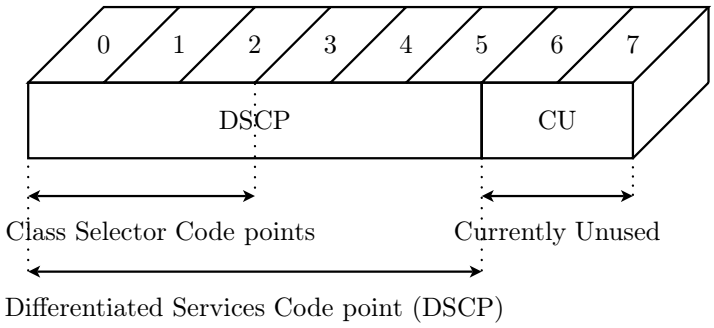


Figure 2.9: DiffServ Code point Field.

According to [11], most implementations of assured forwarding utilize a RED or a similar AQM policy. The Multi-Level RED (MRED) is an extended variant of the RED queue management where multiple sets of RED parameters are configured. Figure 2.10 shows how the three different drop precedence categories within an assured forwarding class, are given individual thresholds and drop probabilities. The figure illustrates a partially overlapped parameter setting for MRED within one queue. Each packet arriving the router is identified by its DSCP field and subsequently

placed in its designated queue. Depending on the available resources in the specific queue, the packet is processed as described in figure 2.6.

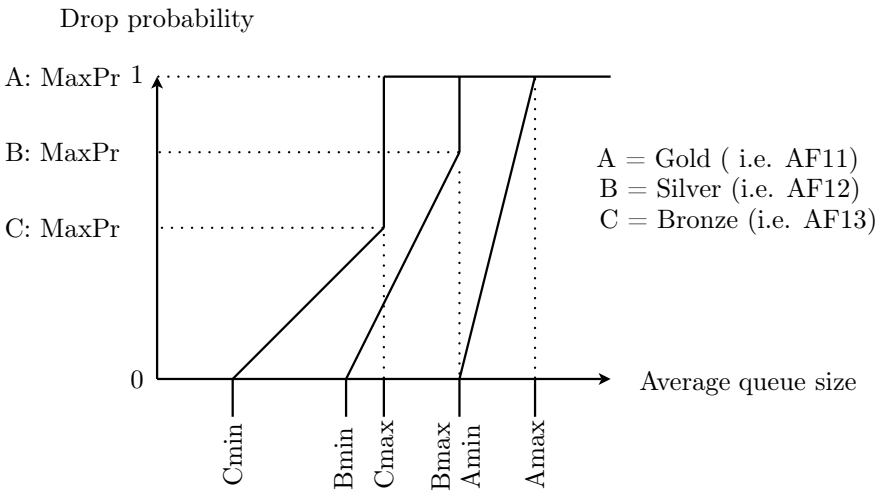


Figure 2.10: Multilevel Random Early Detection. The thresholds are partially overlapped.

2.3 Streaming H.264 video over the Internet

When packetizing H.264 video, the NAL interface generates NAL units which are placed in RTP packets. The NAL unit is byte string containing a coded slice, data partitions or sequence and parameter sets. The unit contains a 8 bit header, which can be divided into three fields. Figure 2.11 depicts the unit header structure:

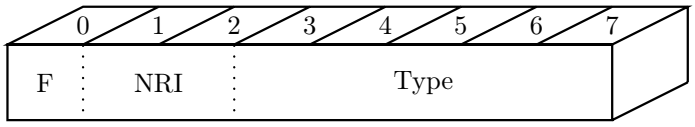


Figure 2.11: NAL unit header as described in [12].

The first field, denoted as F, is a 1 bit field defined by the H.264 specification. This bit must be zero or else it will interpreted as a syntax violation.

The NRI field describe the importance of the NAL unit. A value of zero indicates that the unit is not used for prediction. Values higher than zero indicate that the packet is used for prediction, and a loss of this unit will lead to drifting effects. The higher the value, the more important the unit is for reconstruction of the video signal. Table 2.2 presents an overview of the meaning of the NRI value.

NRI value:	The NAL unit contains:
3 (1 1)	An IDR picture with I slices
2 (1 0)	A non-IDR coded slice (I or P) A coded slice data partiton A
1 (0 1)	A coded slice data patition B A coded slice data partiton C
0 (0 0)	A coded B slice

Table 2.2: Examples of NRI values and their meaning.

The last field, which is a 5-bit field, indicates the type of NAL unit. According to [12], there are 32 different types, where the types 1-12 are defined by the H.264/AVC standard. Types 24-31 are reserved for use outside of H.264/AVC, employed by the RTP payload specification for signal aggregation and fragmentation packets. The H.264/AVC standard reserve all other values for future used. Table 2.3 is an excerpt from table 7-1 in [1]:

NAL unit type	The NAL unit contains
1	Coded slice of a non-IDR picture
2	Coded slice data partition A
3	Coded slice data partition B
4	Coded slice data partition C
5	Coded slice of an IDR picture
6	Supplemental enhancement information (SEI)
...	...

Table 2.3: A few NAL unit type codes.

The payload can be divided into the following payload structures:

Type	Packet	Type name
0	undefined	
1-23	NAL unit	Single NAL unit packet per H.264
24	STAP-A	Single-time aggregation packet
25	STAP-B	Single-time aggregation packet
26	MTAP16	Multi-time aggregation packet
27	MTAP24	Multi-time aggregation packet
28	FU-A	Fragmentation unit
29	FU-B	Fragmentation unit
30-31	undefined	

Table 2.4: Summary of NAL unit types and their payload structures[13]. These values are used inNAL unit header presented in figure 2.11.

Single NAL unit packet

This packet structure is the simplest and contains only one NAL unit and its NAL unit header. The transmission order of the NAL units, determined by the RTP sequence number, is identical

to the NAL unit decoding order. The time stamp increase only when a new frame or field is transmitted.

Aggregation unit packet

Some NAL units, such as supplemental enhancement information units or parameter set units, are often small. In order to minimize overhead by packetizing such packets individually, aggregation of several NAL units can be done. Two basic types of aggregation packets have been defined, where the single-time aggregation packets (STAP) contain NAL units with identical timestamps, and the multi-time aggregation packets (MTAP) can contain NAL units with different timestamps. The former of them are often used in low-delay environments, while the latter is more suitable for high delay environments.

Fragmentation unit packet

RTP packets are placed in the payload of a UDP datagram, which is encapsulated in an IP packet. The IPv4 packet is limited to maximally be 64 Kbytes in size, but the underlying network can further limit the allowed packet size. This boundary is denoted as the *maximum transfer unit* (MTU) and is commonly set to 1500 byte, which is the largest allowed packet size transferred on an Ethernet. If an IPv4 packet exceeds the MTU of the underlying network, fragmentation and reassembly will be performed in the network layer of the OSI model. The network layer fragments packets by dividing the payload into several packets, each with the original packet header. Despite these limitations, NAL units larger than 64 Kbytes can be transferred by employing fragmentation on the application layer.

2.4 Video Quality Assessment

Two assessment classes are used to assess video quality. The first method is subjective assessment where human viewers rate the perceived quality of the video clips, while the second one is the objective approach where quality is measured with mathematical models.

Subjective Video Quality Metrics

Subjective evalution methods can be divided into two main categories which are the singe-stimulus method and the double-stimulus method, which indicate the method the test material is presented to the viewers. These methods are described and standardized in the ITU-R BT.500-11 recommendation[14] and in the ITU-T P.910[15].

Single Stimulus Method

When using the single stimulus method, subjects or viewers are presented with one test sequence at a time, where assessment is done independtly of other presented test sequences. This fashion of assessing test material replicate the home viewing conditions, i.e. regular TV broadcasting, where viewers do not have the possibility of comparing the viewed video sequence with a reference video.

Single stimulus methods can further be divided into single stimulus (SS) and single stimulus with multiple repetition (SSMR). The SS method present the test sequences or pictures only once in the test session, while the SSMR present the test material three times, randomly scattered throughout the test session. This is done to stabilize the observer's opionion of the presented material, where the first rating is discarded from further analysis.

Furthermore, single stimulus methods can be divided by the manner assessment is performed. Rating of the test material can be done either in a post-presentation fashion or in by contiously rating the material. The Single stimulus continous quality evaluation (SSCQE) method, as indicated by its name, provide continuous rating of the presented test material. This method addresses the problems regarding selection of representative test material, which is often limited to a duration of 10 seconds. By employing continous quality evalution, longer sequences can be presented that are more representative of realistic video content and error statistics. An example of the post-presentation assessment is the single stimulus absolute category rating scale with hidden reference removal (ACR-HRR) method, where the unprocessed reference sequence is included in the session, without the knowledge of the viewers. The viewer's opinion of the reference sequence is subtracted from the viewer's opinion of each impaired sequence, resulting in a difference mean opinion score. This post-processing step is known as hidden reference removal.

Different rating scales also exist. Some examples of these are:

- The absolute category rating scale (ACR), which is a adjectival category rating scale. Viewer's rate the quality of the test material with the following five-grade category scale:

Adjectival category:	Value:
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

- The degradation category scale (DCR) is similar to the ACR, but the degree of perceived impairment is rated rather than the perceived quality :

The numerical value paired with each grade on the above scales indicate the mapping between category rating and mean opinion scores (MOS). These values however are not presented to the assessor. In addition to these rating scales several other scales exist. Two examples of these are

Adjectival category:	Value:
Imperceptible	5
Perceptible, but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

the numerical categorical rating scales, where both category and numerical value is presented, and the non-categorical judgement scales, where only a numerical value is used for assessment.

Double Stimulus Method

The double stimulus method presents two test sequences simultaneously. Viewers are asked to rate the sequences compared to each other. Similar to the single stimulus method, this can be done either by post-presentation assessment or by continous assessment, and sequences can either be presented once or mulitple times. The rating scales earlier mentioned are also used for double stimulous video quality assessment.

Some examples of these methods are the double stimulous impairment scale (DSIS) method, the double stimulous continous quality scale (DSCQS) scale and the simultaneous double stimulus for continous evaluation (SDSCE) method. The latter method is an extension of the SSCQE, which was described earlier.

To adress typical assessment problems, the following table can be used for selection of test methodology:

Assessment problem	Method used
Measure the quality of systems relative to a reference	DSCQS
Measure the robustness of systems	DSIS
Quantify the quality of systems	Ratio-scaling method or categorical scaling (i.e. ACR)
Compare the quality of alternative systems	Method of direct comparison, ratio-scaling method or categorical scaling (i.e. ACR
...	...
...	...
Measure the fidelity between two impaired video sequences	SDSCE
Compare different error resilience tools	SDSCE

Table 2.5: Selection of test methods for assessments. The table is an excerpt from BT.500-11 [14].

Objective Video Quality Metrics

Objective metrics can be grouped into three categories, depending on the availability of the original video sequence:

- Full-reference Methods (FR)
  The complete original video sequence is available for comparison with the received or distorted video sequence.

- Reduced-reference Methods (RR)
  Some part of the original video sequence or a set of extracted parameters are available as side information.

- No-reference Methods (NR)
  The video quality is predicted without any knowledge of the original video sequence.

PSNR

The peak signal-to-noise ratio(PSNR) is a full reference metric which measures the mean square error(MSE) between two images of equal size.

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

$$PSNR = 10 * log_{10}(\frac{255^2}{MSE}) = 20 * log_{10}(\frac{255}{\sqrt{MSE}}) \tag{2.2}$$

Extending this formula to a sequence of images, the average of the MSE for each picture is used when calculating PSNR.

$$PSNR = 10 * log_{10}(\frac{255^2}{\sum_{k=0}^{k-1} MSE}) \tag{2.3}$$

SSIM

The structural similarity (SSIM) is an full-reference objective video quality metric which is based on the assumption that the human visual perception is highly adapted for extracting structures in a scene, as apposed to earlier developed methods which modify the MSE estimation. The metric compare local patterns of pixels intensities that have been normalized for luminance and contrast. When measuring the SSIM index, a value of 1 indicate that the two compared pictures or video sequences are identical. Lower values indicate structural dissimilarity. Figure 2.12 depicts the SSIM block diagram. An indepth description of the video quality model can be found in Wang et al.[16]

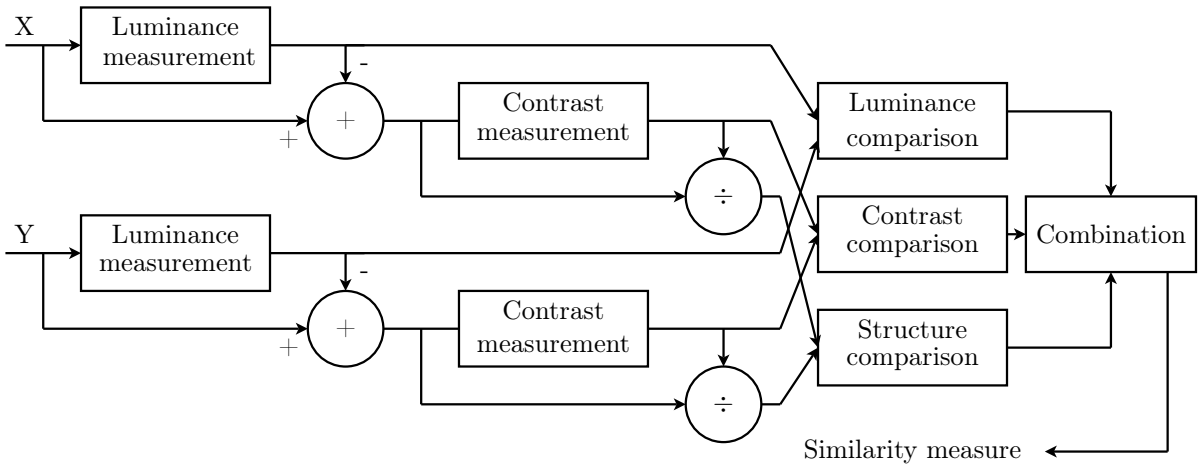


Figure 2.12: The structural similarity measurement system[16].

NTIA VQM

The National Telecommunications and Information Administration (NTIA) has developed a new video quality model known as the NTIA General Model. The model uses objective parameters to measure perceptual effects such as blurring, block distortion, error blocks, noise and unnatural motion. The model's output values range from zero to one, which respectively denote no perceptual impairment and maximum perceived impairment. [17] gives a detailed account of how the General Model is calculated.

MSU VQM

This VQM is a modification of the the digital video quality (DVQ)[18] metric, which is based on the discrete cosine transfom (DCT), and is a full-reference video quality model. The model divides frames into blocks by employing DCT, followed by a temporal filtering of the resulting DCT coefficients. The differences between the temporally filtered coefficients of the reference video sequence and the processed video sequence calculated, followed by a pooling of the calculated difference. The structure of this model is depicted in figure 2.13. A pooled value of zero indicates no estimated impariment, while increasing values indicate an increasing degree of impairment.

The modified model, hereafter denoted as MSU Video Quality Model[2], is described in [19].

Objective Video Quality Model Performance Attributes

This section describes a set of supporting metrics that calculates how well the objective video quality models (VQM) act as an estimator of video quality. The supporting metrics presented

[2]This is due to the fact that it has been implemented in MSU Video Quality Measurement Tool (http://www.compression.ru/video/

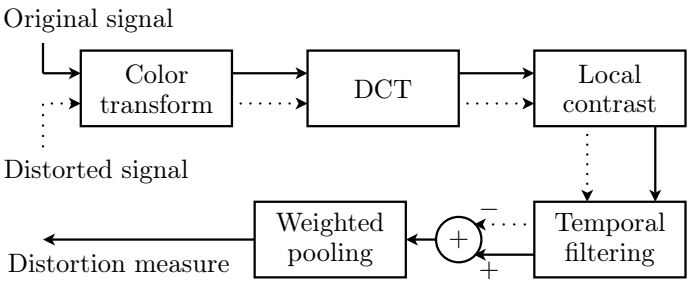


Figure 2.13: Block diagram for the MSU VQM.

are recommended and described in [20],[21],[22] and [23]. The following attributes can be used to characterize the performance of the objective VQMs:

- Prediction accuracy
- Prediction monotonicity
- Prediction consistency

To determine these attributes, three different metrics are used. Each metric measures the figure of merit based on the relationship between *differential mean opinion score* (DMOS) and *predicted differential mean score* (DMOSp) for each and for subsets within each VQM. The figures presented in this section are idealized examples to illustrate each attribute and are reconstructions of figures found in [24].

Prior to evaluating any performance metric, a mapping between the video quality ratings (VQR) for each VQM and the DMOS rating scale is performed. Because subjective ratings often are compressed at the ends of the scales, a non-linear mapping between VQR and DMOS is performed. According to VQEG[22], the following non-linear mapping has performed empirically well:

$$DMOS_p = \frac{b_1}{1 + \exp\left(-b_2 * (VQR - b_3)\right)} \tag{2.4}$$

If the logistic rescaling results in a high mean-square error, the following cubic polynomial monotonic regression is applied:

$$DMOS_p = VQR^3 * b_1 + VQR^2 * b_2 + VQR * b_3 + b_4 \tag{2.5}$$

Using a 5-grade scale like the ACR-HRR does, DMOS is defined as

$$DMOS = MOS(PVS) - MOS(SRC) + 5 \tag{2.6}$$

where $MOS(PVS)$ is the MOS of the *processed video sequence* (PVS), and MOS(SRC) is the MOS of the *source reference circuit* (SRC).

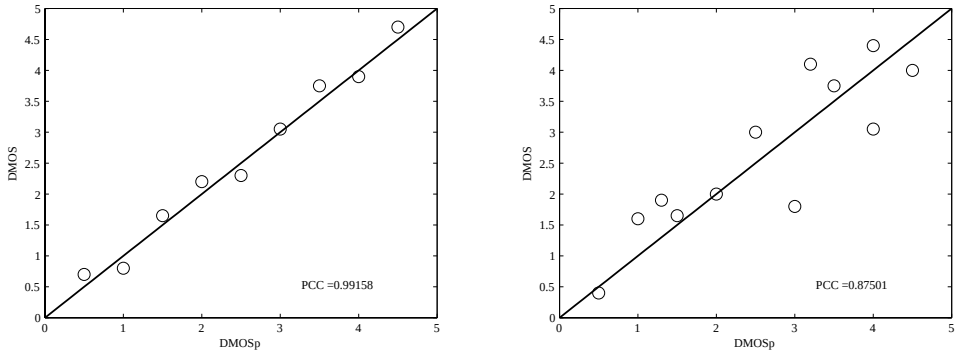
Prediction accuracy

This attribute measures linear relationship between the predicted DMOSp and the subjective ratings. Figure2.14(a) illustrates a high linear correlation between DMOS and DMOSp, while figure2.14(b) illustrates lower prediction accuracy. To calculate the correlation, the *Pearson linear correlation coefficient* (PLCC) is used (as defined in formulas 2.7 and 2.8).

$$r_p = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^{n} x_i - \overline{x})^2}\sqrt{(\sum_{i=1}^{n} y_i - \overline{y})^2}} \tag{2.7}$$

Formula 2.7 can be simplified, as shown in 2.8, where calculation of the mean values is no longer necessary.

$$r_p = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i^2)}{n})}} \tag{2.8}$$



(a) Model with greater accuracy.

(b) Model with less accuracy.

Figure 2.14: Prediction accuracy[24].

The confidence interval for the PLCC is calculated by using a Fisher-z transform, which is defined as follows:

$$z = 0.5 * \ln\frac{1+R}{1-R} \quad \text{with the standard deviation} \quad \sigma_z = \sqrt{\frac{1}{N-3}} \tag{2.9}$$

As an addition to the Pearson linear correlation coefficient, the Root Mean Squared Error(RMSE) is calculated. The absolute prediction error is defined as the difference between the DMOS and DMOSp:

$$Perror(i) = DMOS(i) - DMOS_p(i) \tag{2.10}$$

where the index i denotes the PVS. The RMSE of the absolute prediction error is calculated as follows:

$$rmse = \sqrt{\frac{1}{N-d} * \sum_{i=1} nPerror(i)^2} \tag{2.11}$$

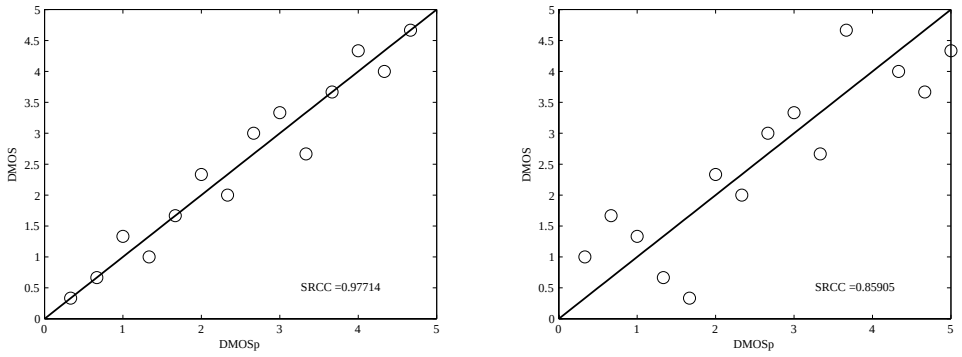
where N denotes the number of samples and d the number of degrees of freedom of the mapping function(formula 2.5 or 2.4).

Prediction monotonicity

Monotonicity measures if an increase in one variable is associated with an increase in another variable, regardless of the magnitude of increase. The relationship between the variables is insignificant, as long as the variables increase or decrease identically. The degree of monotonicity is quantified by the non-parametric *Spearman rank-order correlation coefficient* (SRCC), which is defined as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{N}(x_i - y_i)^2}{N(N^2-1)} \tag{2.12}$$

Figure 2.15(a) illustrates a model with higher monotonicity than the model illustrated in 2.15(b).



(a) Model with more monotonicity.

(b) Model with less monotonicity.

Figure 2.15: Prediction monotonicity[24].

Prediction consistency

This attribute describes how consistently a VQM predicts the DMOS. To quantify the degree of consistency, the total number of outliers is measured. An outlier is defined as a prediction with

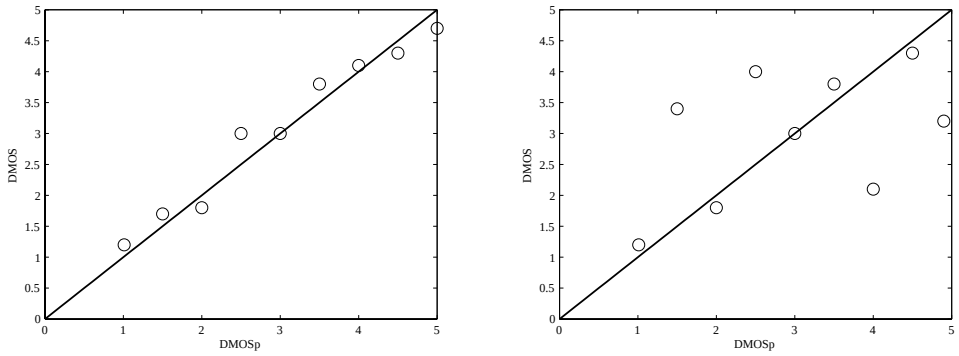
a prediction error greater than a given threshold. This threshold is usually defined as twice the standard deviation σ_{y_i} of the subjective rating differences for the given point:

$$|x_i - y_i| > 2\sigma_{y_i} \tag{2.13}$$

The total number of outliers can be used to calculate the outlier ratio, where lower values indicate prediction consistency. The ratio is defined as

$$r_O = \frac{N_o}{N} \tag{2.14}$$

where N_o is the total number of outlier points and N is total number of points in the dataset.



(a) Model with consistent errors.

(b) Model with large outlying errors.

Figure 2.16: Prediction consistency[24].

Chapter 3

System Description

3.1 Introduction

This chapter describes the video sequences used in the evaluation, how they are encoded, and how transmission errors typical for best effort and DiffServ IP networks are simulated.

3.2 Video sequences

Three different video sequences were used as source video sequences (SRC) in the informal subjective evaluation described later in this chapter. Prior to encoding with the H.264 video coding standard, sequences were downsampled to conform to the 720p High-Definition (HD) video format. Table 3.0(a) below indicates the sequences' original video formats and properties, while table 3.0(b) indicates the adjusted resolutions and properties for the video sequences coded with the H.264/AVC video coding standard.

Sequence	Frame/Field rate	Resolution	Frames		Video length/duration	Encoding
StEM	24	4096x1714	16605	≈	11 minutes 53 seconds	Progressive
Raven	60	1280x720	600	=	10 seconds	Progressive
Tandberg	60	1440x1080	800	≈	13 seconds	Interlaced

Sequence	Frame/Field rate	Resolution	Frames		Video length/duration	Encoding
StEM	24	1280x720	274	≈	11.42 seconds	Progressive
Raven	30	1280x720	300	=	10 seconds	Progressive
Tandberg	30	1280x720	338	≈	11.3 seconds	Progressive

Table 3.1: The original and the encoded video sequences.

The sequence referred to as StEM is the "Standardized Evaluation Material" sequence produced by Digital Cinema Initiatives (DCI)[25]. The excerpt, which is used as SRC in the subjective

evaluation procedures, consists of 274 frames ranging from frame 3680 up to and including frame 3954 from the original sequence. The clip contains two scene shifts with camera panning, a large number of moving objects and a high level of texture detail. Figure 3.1 shows the first frame in the sequence excerpt.

Figure 3.1: First frame in the StEM video sequence excerpt.

The Raven sequence contains only one scene with camera panning and little rapid movement. The clip contains a moving object centered in the foreground, while the background is out of focus containing a low degree of motion. During the sequence, a shift in focus is performed, rendering the foreground out of focus for a short interval. The first frame of the sequence is presented in figure 3.2.

Figure 3.2: First frame in the Raven video sequence.

The last sequence is the Tandberg clip, which is produced by Q2S[26] for Tandberg[1]. Originally, this clip had a spatial resolution of 1440 by 1080 pixels encoded interlaced with a temporal resolution of 60 fields per second. In order to convert this interlaced signal to the progressive 720p format, the top field of the original signal was filtered by a interpolation filter followed by cropping as shown in the equation 3.1 below. The filter coefficients and source code for this process can be

[1] http://www.tandberg.no

found in appendix C.

$$1440x1080i \overset{\text{remove bottom field}}{\Longrightarrow} 1440x540p \overset{\text{filter upsampling}}{\Longrightarrow} 1440x720p \overset{\text{crop}}{\Longrightarrow} 1280x720p \tag{3.1}$$

The Tandberg clip includes a static background and moving objects in the foreground. This video clip targets video-conferencing scenarios, where neither scene shifts nor camera panning is present. Figure 3.3 depicts the first frame from the sequence.

Figure 3.3: First frame in the Tanberg video sequence.

The H.264 reference software, described in section 3.5, encodes the three SRCs with a set of mutual parameters. The key encoder parameters are presented below, while table 3.2 presents the quantization parameters, period of I frames and resulting rate.

- GOP structure fit to frame rate, resulting in one I slice/frame every second.
- Only I and P slices are used. Every I slice is coded as an IDR picture.
- A maximum number of five reference frames can be used for motion compensation.
- Data partitioning is not employed.
- UVLC entropy coding.
- Dispersed FMO with two slice groups, with maximum 1450 bytes in a slice.

SRC	I period	QP I	QP P	Rate (kb/s)
StEM	24	24	24	5934.40
Raven	30	25	25	5246.74
Tandberg	30	23	22	5995.21

Table 3.2: Individual coding parameters and rate for the video sequences.

After encoding the video sequences, the following objective VQM results, per SRC, were calculated.

SRC	PSNR(Y)	PSNR(U)	PSNR(V)	NTIA	MSU	SSIM
StEM	40.05	43.06	45.87	0.112	0.949	0.956
Raven	41.22	45.44	44.34	0.138	0.628	0.964
Tandberg	41.23	47.09	46.55	0.104	0.738	0.971

Table 3.3: Objective video quality after encoding.

3.3 Test bed

The test bed can be considered as a system where an input signal is influenced by a chain of manipulation elements, resulting in a distorted or altered output signal as figure 3.4 illustrates. A streaming media test bed is a realization of this test bed structure, where media content is encoded and compressed, thereafter transmitted over various channels and decoded by a media client. The test bed realization discussed and employed in this report consists of the streaming H.264 video content over an IP-based network using the RTP transport protocol, and employs two network models. These are the well-known best effort model and the DiffServ model, where the flow of packets is distorted by provoking packet loss using a network simulator. This test bed is a realization of the system presented by Hillestad et al.[27]

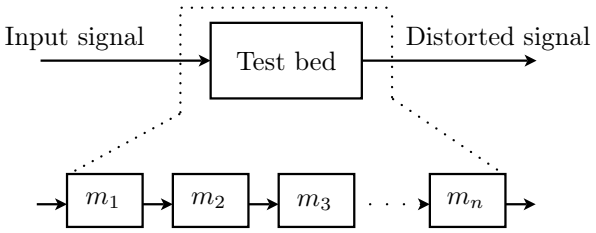


Figure 3.4: Principal mode of operation for a test bed.

The test bed consists of an application part and a network part, where the latter generally introduces distortion. Henceforth, only the network part is considered to introduce distortion and the signal generated by the H.264/AVC JM encoder is considered the original (distortion less) signal. The network part can consist of any number of different physical routers and switches as well as software routers and network simulators. Figure 3.5 depicts the employed streaming media test bed.

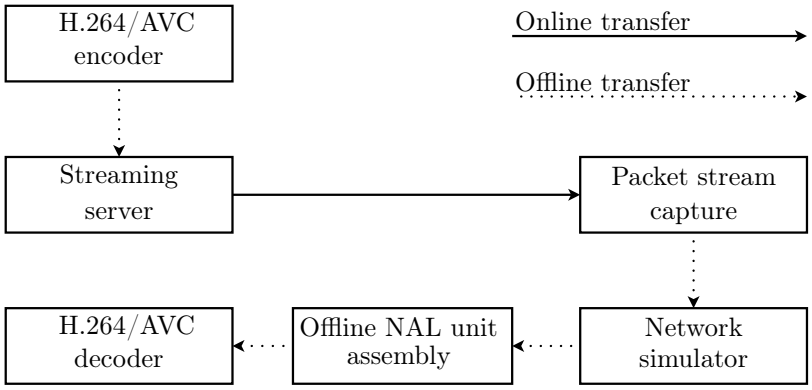


Figure 3.5: Block diagram describing the test bed.

The original video sequence is first encoded and subsequently placed into an mp4 container and hinted with a MTU of 1450 bytes. The server broadcasts the mp4 file over the physical test bed network, and by using a network interface monitoring card[28] the packet flow is captured and stored in a *trace file*, which contains the entire packet flow and time stamps accurately indicating when each packet was captured. The trace file is further converted into a *pcap file*, from which the contents of each packet is analysed to determine the values stored in the NAL units header (As earlier described, these are stored in the payload of the RTP packet). Information extracted from the pcap file is then used to classify and mark the packets with desired DSCP values, differentiating the service the packets experience in the network simulator. Finally, the output from the simulator is reconstructed to a H.264/AVC encoded file conforming with the bit stream syntax described in Annex B in the video coding standard[1], with the help of the *pcap2avc* utility. The pcap2avc utility and the network simulator are respectively described in section 3.5 and 3.4.

3.4 The network simulator

The simulator, which was mentioned earlier while describing the test bed, is an implementation of a DiffServ router and is developed using the object-oriented *Discrete Event Modelling on Simula* (DEMOS)[29] and is developed at Q2S[26] by [30].

The IP network can be simulated by using a single router and connect one or more sources to generate cross traffic. The router offers six queues, where all except the EF queue can use one of three implemented queue management policies. The previously discussed queue management policies MRED and DT (see section 2.2) are implemented in the simulator. The third policy is the *Priority Drop* (PD), a queue policy that drops an enqueued packet marked with high drop precedence if a packet with lower drop precedence arrives at the full queue. The assured forwarding and best effort queues depicted in figure 3.6 are forwarded using a *deficit round robin* (DRR) scheduler, while packets marked with the expedited forwarding DSCP are always forwarded directly without the use of a scheduler.

Based on parameters defined in a configuration file, the simulator exposes the input studied source to cross-traffic, where the packet flows in the aggregated traffic compete for the router's resources. *Pcap2st* (See section 3.5) produces a representation of the source traffic under study, which contains the header information, appurtenant time stamps and packet sizes extracted from the pcap file. Based on the DSCP field in a packet, the classifier places it in its designated queue. The different queue management policies determines if a packet is forwarded or dropped, where the forwarded packets from the studied source are stored in an out-file that can be used with the *st2pcap*(See section 3.5) application to create a new pcap file.

Simulator setup

A set of simulation parameters are used to produce simulations with desired packet loss rates. The simulations consists of two subsets, where the first use the MRED policy for simulating a DiffServ router's behaviour while the latter use DT for the best effort model. When simulating the DiffServ subset, all packets in the aggregated traffic are classified with one of the three drop precedences belonging to the AF1 queue. The MRED policy use the thresholds and probabilities presented in table 3.4. Figure 2.10 in section 2.2 depicts the partially overlapped MRED.

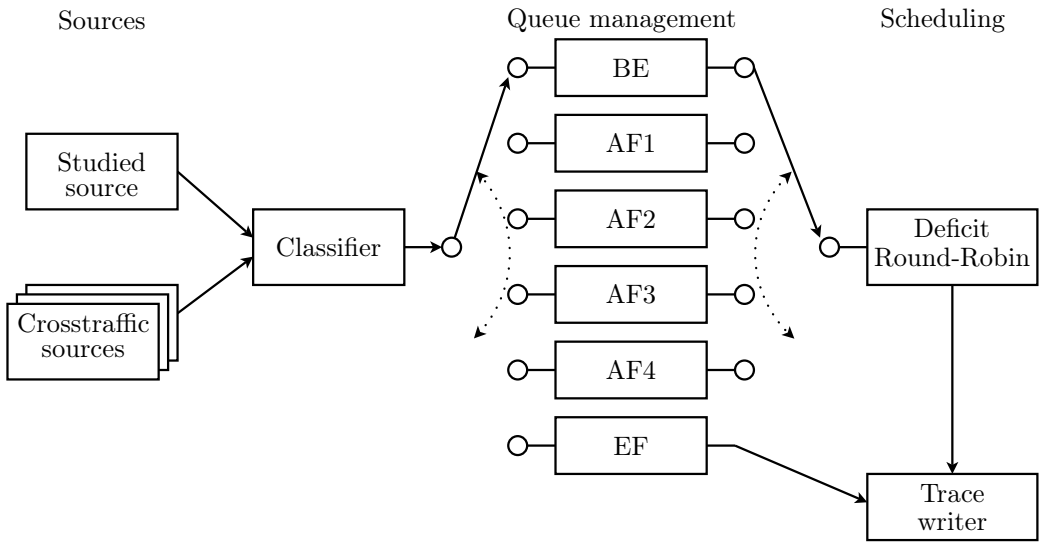


Figure 3.6: The DiffServ simulator as depicted in the documentation of the software.

Drop precendence	Min. Threshold	Max. Threshold	Max. Probability
AF11 (gold)	80%	100%	100%
AF12 (silver)	50%	80%	75%
AF13 (bronze)	20%	60%	50%

Table 3.4: Thresholds used in the MRED policy when simulating DiffServ.

Based on the extraction of NAL unit headers from the pcap file, the in file generated by pcap2st is modified prior to simulation. Table 3.5 show the novel mapping between DSCP and NRI values (Table 2.2 shows typical NRI values). Packets containing IDR pictures are assigned the lowest drop precedence, while other packets containing either I or P slices are assigned a medium drop precedence.

NRI	DCSP	Drop precedence category	Packet contains
3	0 0 1 0 1 0	AF11 (gold)	IDR pictures with I slices
2	0 0 1 1 0 0	AF12 (silver)	I or P slices

Table 3.5: Mapping between NRI and DSCP values.

When simulating the best effort model, all packets in both the studied source and the cross-traffic are marked with the same drop precedence. The aggregated packet flow uses only one queue that employs the DT policy.

Packet Loss Rates

In order to achieve the desired packet loss rates (PLR), parameters such as delay until inserting source traffic, outgoing link capacity, simulation seed and cross traffic were varied. Ten different packet loss rates were desired, where the first packet loss in the studied sequence could not occur before a given period of time had passed. This time restriction was included to prevent the occurrence of distortions in the decoded video signal prior to length of one second of displaying at the individual frame rates. The following packet loss rates were desired:

0.00 %	0.10 %	1.00 %	2.00 %	2.50 %
3.00 %	4.00 %	5.00 %	7.50 %	10.00 %

Cross-traffic

Different sets of cross-traffic, which differed in both duration and rate, were used for the simulations. The purpose of using different sets was to avoid creating similar or identical simulations. Figure 3.7 shows an example of cross-traffic used in the simulations. The plot on the figure represents an aggregated packet flow consisting of several streamed video sequences, all marked with the novel-marking scheme presented earlier. The MergeTrace tool, which is part of the simulator software pack, merges several packet streams in order to create cross-traffic. The different streams are introduced in the cross-traffic file with a uniform delay between 1 and 4 seconds. Appendix D includes a table describing the encoded video sequences used in the cross-traffic.

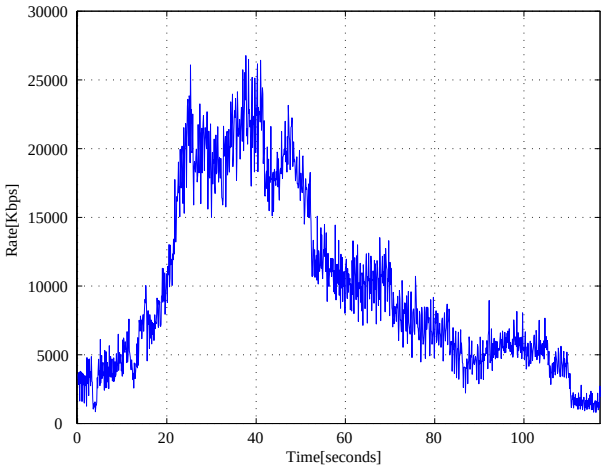


Figure 3.7: Example of cross-traffic used in the DiffServ router.

3.5 Other software and tools used

JM - H.264/AVC reference software

JM[31] is the H.264/AVC reference software copyrighted by ITU. As this software package is under ongoing development, version 10.2 is used for encoding and decoding video sequences in this thesis. To use the software with the earlier described parameters, some modifications had to be done. In order to encode 1280 by 720 frames with a slice size fixed to 1450 bytes, the number of allowed slices had to be increased to 100 slices per picture. This change was done in both the encoder and decoder. The usage of the encoder/decoder is described in the reference software manual[32].

The decoder tries to conceal errors caused by e.g. packet loss. If a packet contains intrapredicted slice data, the decoder locates the corrupted macroblocks and uses pixel interpolation functions to correct them one block at a time. Scanning is done vertically and each corrupted column is

corrected bi-directionally. If a lost packet contains an interpredicted slice, two methods of error concealment is performed. If the average motion vector of the correctly received macroblocks is less than a given threshold, concealment is done by using the motion vectors of a reliable neighbour macroblock. If this average motion vector is above the threshold, then a copy of the macroblock from the reference picture at the same location is used.

Figure 3.8 shows a frame from the Tandberg sequence that has lost packets while being streamed over the Internet. By studying the hand railing located on the left side of the picture, we can see the use of interpolation caused by the loss of packets containing I slice data. The macroblock copy method is also presented in the figure, where a macroblock containing a part of the orange is copied from a reference P or I slice.

Figure 3.8: Frame 252 from the Tandberg sequence with 5% PLR.

If an entire frame is lost, error concealment can be done either using a copy of the previously decoded frame or by creating a new frame based on a motion copy algorithm.

MPEG4IP

MPEG4IP[33] is an open-source software package that provides an end-to-end system for streaming of multimedia. The package includes a broadcaster for several video coding formats such as MPEG-4, H.261, MPEG-2 and H.263. The package also includes a MP4 file creator and hinter, which is used in this thesis. The software creates an MP4 file container that encapsulates the encoded video sequence and hints an audio/video track adapted to MTU.

Pcap2avc

Pcap2avc[34] is an application for reconstruction of streamed H.264/AVC video using RTP. Pcap2avc parse each packet in a pcap file, extracting the H.264/AVC content from the RTP payload, and

storing them correctly in an encoded H.264 file, conforming to bitstream format described in Annex B of the H.264/AVC standard. The reconstruction of the sequence is based on the information given in the NAL unit header and the RTP sequence number.

SimTraceTools

SimTraceTools defines text based file format describing a packet stream, denoted st. Each line in the file represents a packet with properties such as time stamp, size, destination etc. The st file is used as input and output in the simulator, and can be created by an application in the SimTraceTools package. To create st files, information from a pcap file is extracted. After simulations has been conducted, the SimTraceTools package can recreate the pcap file, where the pcap file is modified with the packet loss, delays and reordering described in the st file. The SimTraceTools package is developed by [35].

3.6 Subjective test procedure

To evaluate the perceived quality of the video samples, the absolute category rating scale (ACR) is used. The ACR is a single-stimulus method where processed video sequences are presented individually, without being paired with a reference sequence. The reference sequence is included in the session, without the knowledge of the viewers. The viewer's opinion of the reference sequence is subtracted from the viewer's opinion of each impaired sequence, resulting in a difference mean opinion score. This post-processing step is known as "hidden reference removal". The absolute category scale with hidden reference removal (ACR-HRR) is a standardized method of subjective video quality assessment and is recommended by VQEG [36].

General Description

The Absolute Category Rating is a method where video sequences are presented one at a time and are rated independently on a category scale. After each presentation, the subjects evaluate the quality of the video sequence that was presented. The following scale is used for quality evaluation:

5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

According to the VQEG Multimedia Test Plan 1.11 [22], the video sequences should be 8 seconds of length, after which a grey screen is presented until vote is given. The ITU-R BT.500-11 [14] recommends a mid-grey adaptation field, a stimulus and a mid-grey post-exposure field of 3, 10 and 10seconds respectively. Based on these two similar recommendations, we choose to use a mid-grey adaption field, a stimulus and a mid-grey post-exposure field of 3, 10-14 and 10 seconds respectively. A timeline of the stimulus presentation is shown in figure 3.9.

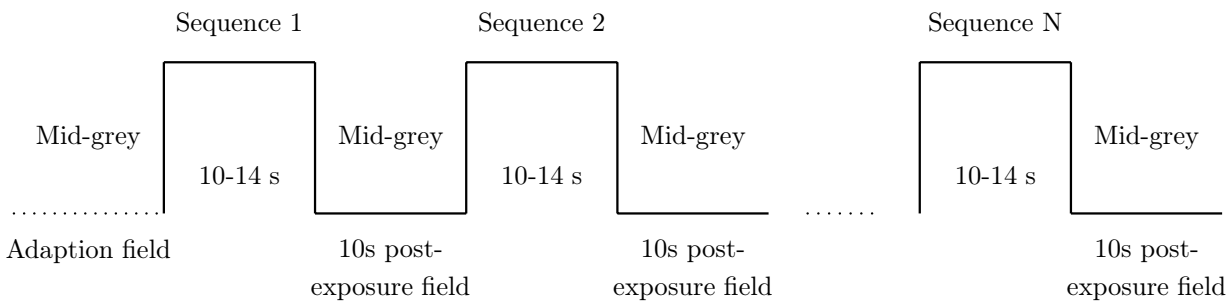


Figure 3.9: ACR-HRR stimulus timeline

Video formats and display

The 720p HDTV format is the only video format used in this test plan. Figure 3.10 is presented in ITU-R BT.500-11 [14, p. 4], and is valid for both SDTV and HDTV video formats. The figure shows the ratio between the viewing distances and the picture height. The test subjects are not expected to maintain the same viewing distance throughout the tests, but they are encouraged to stay near the proposed Preferred Viewing Distandce(PVD). The tests asses video-quality only, imposing no demands regarding acoustics.

Display Specification and Setup

The subjective tests will use LCD displays, and in accordance to VQEG MM Test plan conforms to the specifications given in table 3.6

Test Method

All subjective tests were done on the same computer and LCD monitor using the same software package throughout all tests. The subjects' personalia were registered in a form, where information such as age, gender, name and subject number is stored. All voting was done by a form when the sequences were assessed.

The subjects

ITU-R P.910 [15, p. 10] states that 40 viewers are enough to complete the subjective evaluations. Post-experiment screening was performed to validate each viewers voting. This validation or screening of viewers discarded voting done randomly. Additional viewers were not required if the post-experiment screening resulted in less than 40 valid viewers. Annex VI in VQEG MM Test Plan v1.11 proposes two methods for post-experiment screening of results. Using the first method, the rejection criteria verified the level of consistency of the raw scores between one viewer according to the average raw scores of all the viewers.

Screen diagonal (in)		Screen height (H)	PVD
4/3 ratio	16/9 ratio	(m)	(H)
12	15	0.18	9
15	18	0.23	8
20	24	0.30	7
29	36	0.45	6
60	73	0.91	5
> 100	> 120	> 1.53	3-4

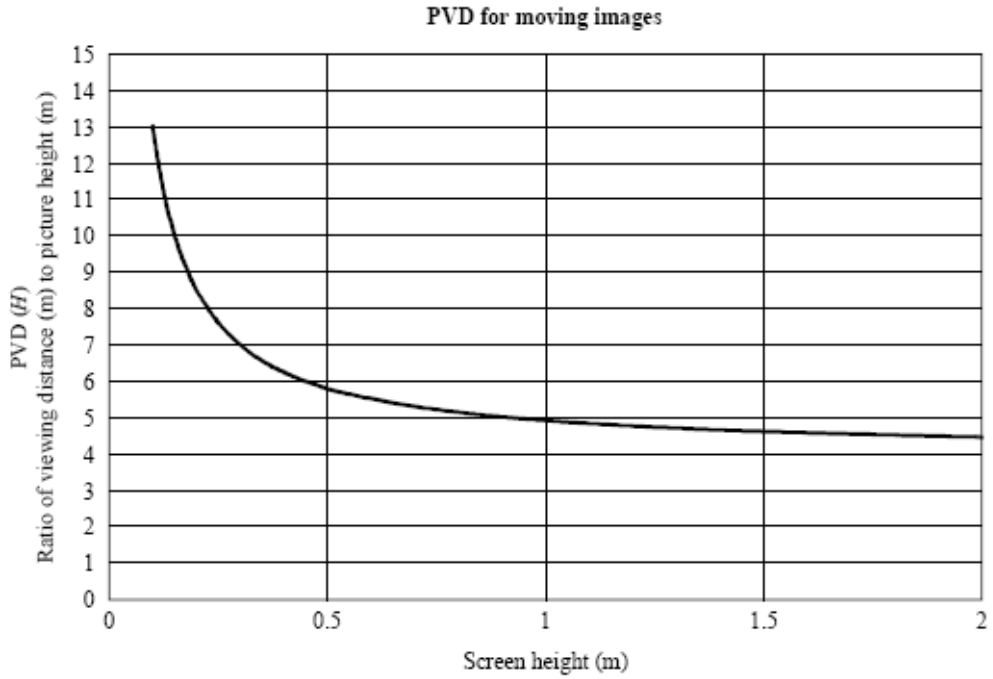


Figure 3.10: Preferred Viewing Distandce(PVD)

Each viewer was presented with a unique order of video sequences. Each viewer was only allowed to participate once, disregarding a discarded previous participation. Only non-expert viewers are allowed to participate. A viewer was considered as a non-expert if the viewer was not an experienced assessor or work with video picture quality.

Viewing conditions

The viewer performed assessment of video individually. The viewer was seated directly in front of the video display at the given viewing distance.

Monitor Feature:	Specification:
Diagonal Size:	17-24 inches
Dot pitch:	< 0.30
Gray to Gray Response Time(if specified by manufacturer, otherwise assume response time reposted is white-black)	< 30ms (<10ms if base on white-black)
Colour Temperature:	6500K
Calibration:	Yes
Calibration Method:	Software
Bit Depth:	8 bits/colour
Refresh Rate:	>=60Hz
Standalone/laptop	Standalone
Label:	TCO '03

Table 3.6: Display specification and setup factors

Experiment design

Each viewer was presented with three different movie clips. For each clip, 19 video sequences were presented, including both the processed and the unprocessed reference video. The following procedure was used to perform the assessments:

1. Introduction and instructions to the viewer.

2. A few practise clips are used to train the viewer. These clips will contain the same degrees of impairments (packet loss rates) as used in the experiment. VQEG proposes the use of 6 clips for training.

3. Assessment of the first movie clip. 19 video sequences.

4. A short break.

5. Assessment of the second movie clip. 19 video sequences.

6. A short break.

7. Assessment of the last movie clip. 19 video sequences.

The duration of the test was between 30 and 45 minutes, including the breaks and training.

Post-Experiment Screening

The rejection criteria verified the level of consistency of one viewer corresponding to all viewers. In order to validate results from each viewer analysis on both per PVS and per Hypothetical Reference Circuit (HRC) was performed, because a viewer might have an individual content preference that differs from that of other viewers. Therefore, it was necessary to analyse not only per PVS, but also per HRC.

Pearson linear correlation coefficient per PVS for one viewer vs. all viewers

$$r_{p1} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i^2)}{n})}} \tag{3.2}$$

where
x_i = MOS of all viewers per PVS
y_i = individual score for one viewer for the corresponding PVS
n = number of PVSs
i = PVS index

Pearson linear correlation coefficient per HRC for one viewer vs. all viewers

$$r_{p2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n})(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i^2)}{n})}} \tag{3.3}$$

where
x_i = condition MOS of all viewers per HRC
y_i = individual condtion MOS for one viewer for the corresponding HRC
n = number of HRCs
i = HRC index

If a viewer scored $r_{p1} < 0.75$ or $r_{p2} < 0.80$[2], his or hers assessments were excluded from the evaluation.

[2]These thresholds were initially 0.85 and 0.9. At an official VQEG meeting in Boston April 24-28 2006, the new rejection criterias and rejection method were decided

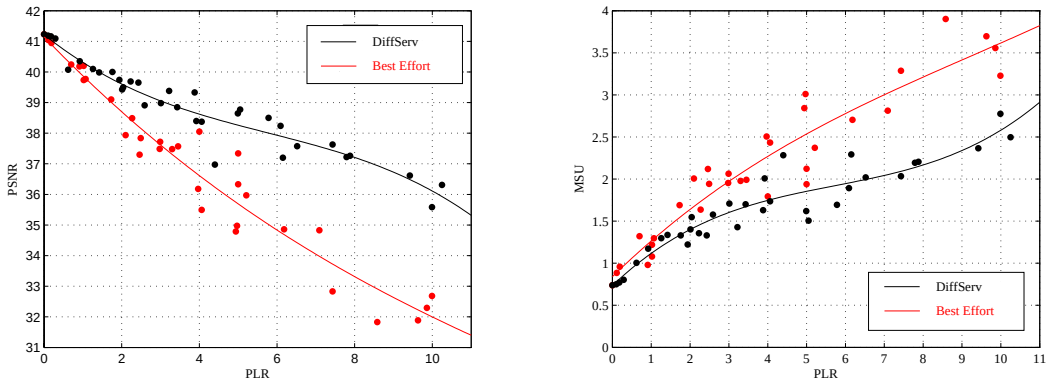
Chapter 4

Results and performance evaluation

In this chapter, we present results from both the objective video quality models and subjective quality assessments. The chapter consists of several sections. The first studies whether the use of a differentiated services network with multi-level random early detection outperforms the drop-tail best-effort model, first objectively then subjectively. Finally, the chapter presents how well the objective video quality models perform according to the subjective assessments.

4.1 Video quality model results

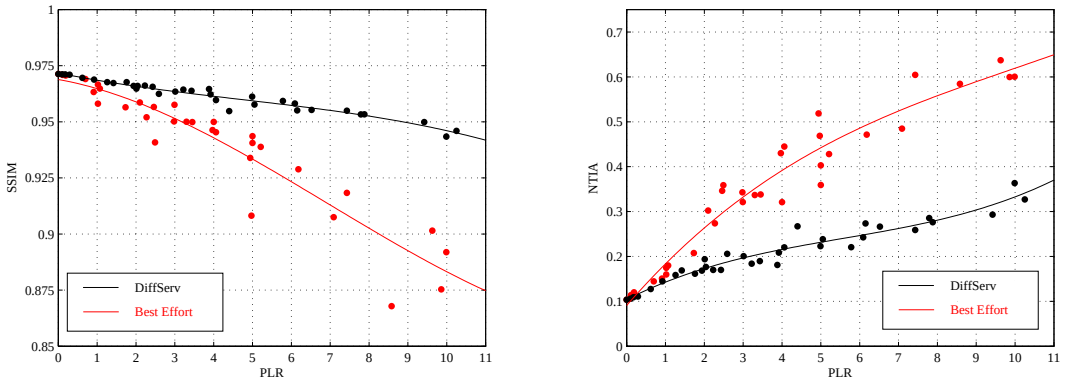
Using the video quality models described in 2.4, objective results were calculated for all processed video sequences. The results for each video quality model using the Tandberg sequence is depicted in figure 4.1 and 4.2, where cubic polynomial regression has been applied to all the plots for the purpose of illustration.



(a) PSNR

(b) MSU

Figure 4.1: PSNR and MSU for the Tandberg sequence.



(a) SSIM (b) NTIA

Figure 4.2: SSIM and NTIA for the Tandberg sequence.

The PSNR and SSIM video quality models denote increasing values as a decreasing degree of distortion, while the opposite applies to MSU and NTIA. The figures presented above show a higher measured level of quality when using the DiffServ model applying to all of the video quality models. At lower packet loss rates such as 1% the two network models yield minor differences for all the video quality models. Increasing the packet loss rate leads to a larger difference between the network models, e.g. at 6% packet loss the DiffServ yields 3 dB higher PSNR and 0.2367 lower NTIA index compared to the best effort network. Considering figure 4.2(a) depicting SSIM indexes, the differences between the two network models are very small (only 0.0332 at 6% PLR of a scale ranging from 0 to 1).

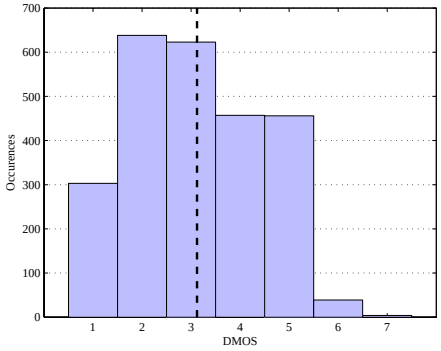
The differences can be explained by considering the novel marking scheme presented in section 3.4. The DiffServ model drops only slices containing IDR frames at high packet rates, while the best effort network drops packets regardless slice type and packet priority. The importance of an IDR frame surpasses the importance of a P slice, since it is used as reference for the all P slices that follow within the GOP. This means a loss of data in an IDR frame will propagate until a new I slice or IDR frame is decoded. Loss of P slice data also results in propagation errors, where the impact depends on placement within the GOP. P slices placed at the end of the GOP will not propagate errors with the same impact as one placed near the start of the GOP.

4.2 Subjective procedure results

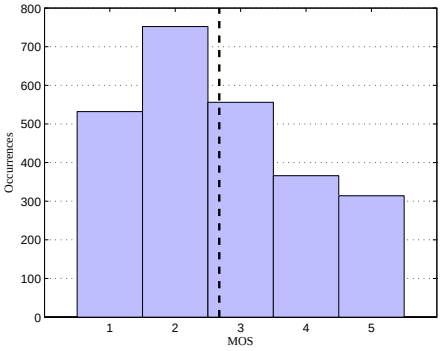
After all 43 participants had assessed the test material; a post-expirement screening process was perfomed to determine the level of consistency for one viewer corresponding to all viewers per PVS and per HRC. All participants were found valid using the thresholds determined by the MM test plan. To validate the participants, equations 3.2 and 3.3 were used.

Figure 4.3 shows the histograms containing MOS and DMOS both for the whole set of test material and for the two different network models. The value 5 in the MOS histograms corresponds to the adjective “Excellent”, 4 to “Good” etc. Using equation 2.6, the MOS is mapped to a set of DMOS values which indicite the differential mean opionion scores for all the subjects. The DMOS histogram in figure 4.3(a) contains some values exceeding the 5-point ACR scale used for

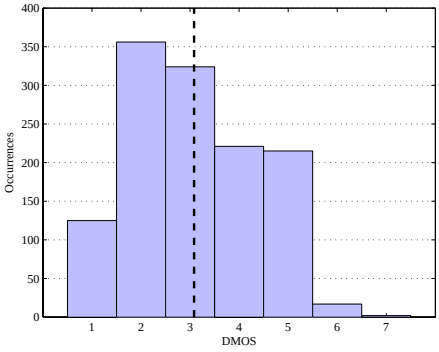
assessments. These values indicate that some viewers rated an impaired PVS as having better quality than the SRC.



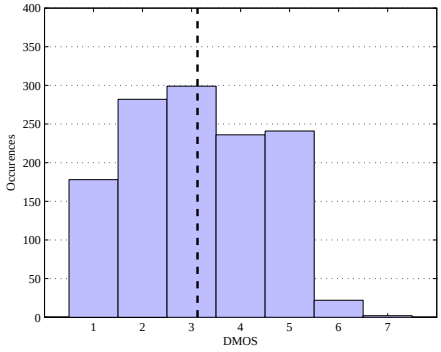
(a) DMOS distribution



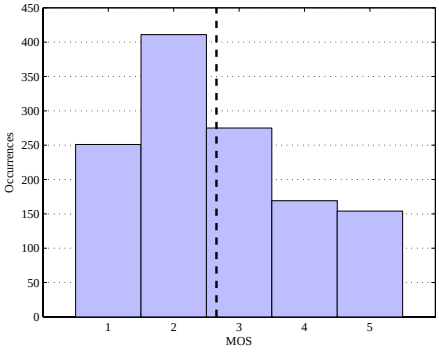
(b) MOS distribution



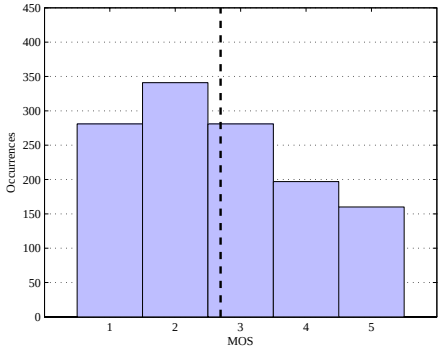
(c) DiffServ DMOS distribution



(d) Best Effort DMOS distribution



(e) DiffServ MOS distribution

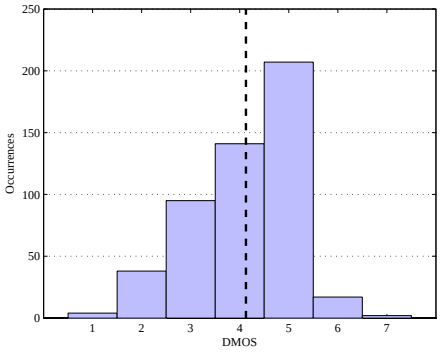


(f) Best Effort MOS distribution

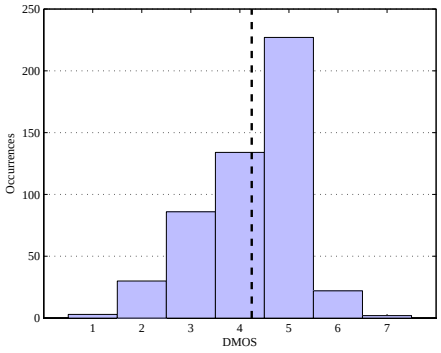
Figure 4.3: MOS and DMOS histograms where the dashed lines indicate the mean values.

When comparing the two network models, the histograms indicate that best effort model has a higher number of occurrences in the upper part of the rating scale. However, the DMOS histograms show an approximately identical mean value. The mean values of the DMOS ratings, per PVS, can be found in table A.1. The ratings was further divided into categories such as low, medium

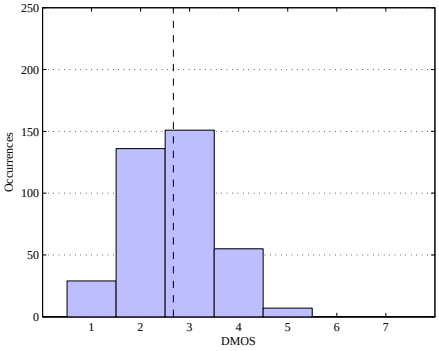
and high packet loss rate that respectively correspond to 0-2%, 2.5-4% and 5-10% packet loss. Figure 4.4 depicts the DMOS values for these categories. The DiffServ model has a lower mean value for the low and medium categories, while for the high packet loss, the mean value is higher than for best effort.



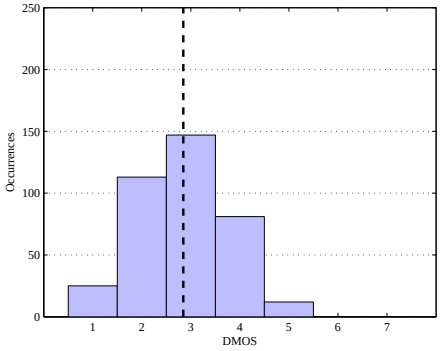
(a) DiffServ DMOS distribution low PLR



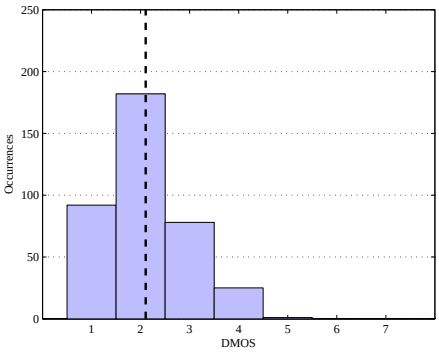
(b) BestEffort DMOS distribution low PLR



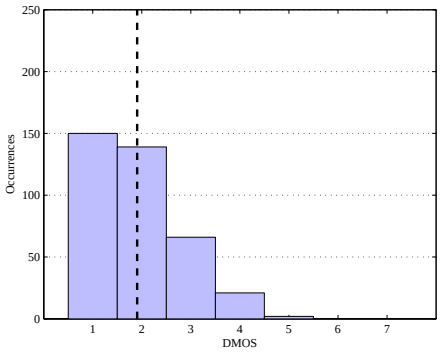
(c) DiffServ DMOS distribution medium PLR



(d) BestEffort DMOS distribution medium PLR



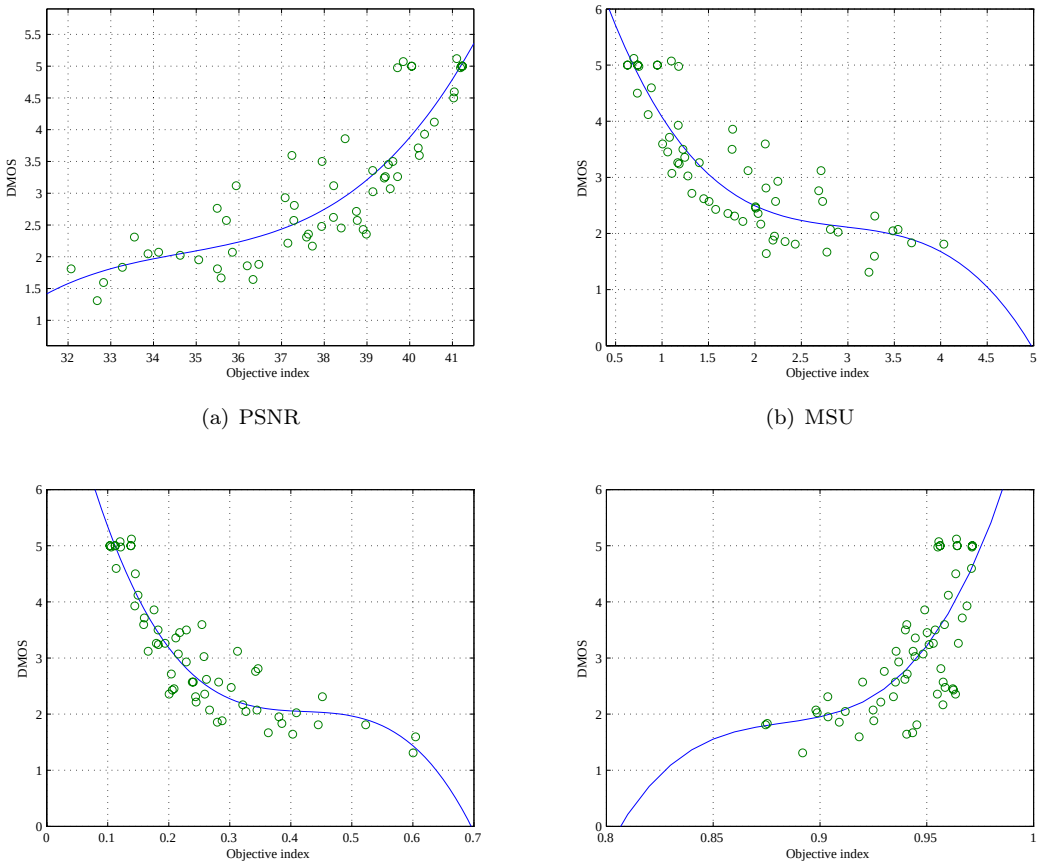
(e) DiffServ DMOS distribution high PLR



(f) BestEffort DMOS distribution high PLR

Figure 4.4: DMOS histogram for low, medium and high packet loss. Mean values are indicated by the dashed lines.

4.3 Fitting of video quality models to DMOS



(a) PSNR

(b) MSU

(c) NTIA

(d) SSIM

Figure 4.5: Non-linear regression for mapping objective models to DMOS.

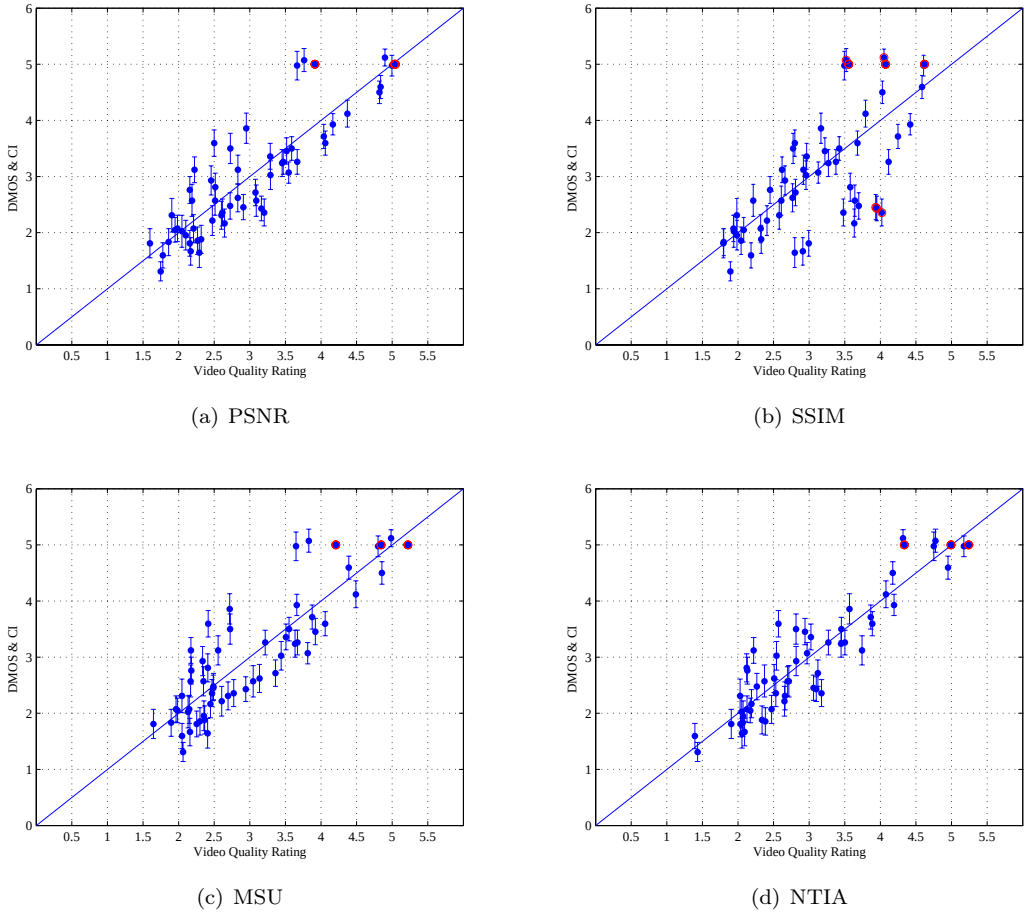
All the objective video quality models have been fitted by cubic polynomial monotic regression. This was done because the mapping function (2.4) recommended by VQEG resulted in a higher sum-of-squares error. Table 4.1 show the coefficients used for each regression and the resulting sum-of-squares error. The table shows that the SSIM model has a higher sum-of-squares than the other models. This can also be seen in figure 4.5(d) where the majority of measurements are densely gathered on the SSIM-axis, but the DMOS ratings for the corresponding PVSs are scattered over the ACR scale, resulting in a high prediction error. The other three models fit well, which can be interpreted from both the figures and the sum-of-square error.

Model	Mapping	Coefficients				SS
PSNR	Cubic	b_1 : 0.007746	b_2 : −0.80568	b_3 : 28.0562	b_4 : −325.0258	0.2569
NTIA	Cubic	b_1 : −94.014	b_2 : 119.11	b_3 : −50.762	b_4 : 9.3215	0.1674
MSU	Cubic	b_1 : −0.20873	b_2 : 1.8516	b_3 : −5.6737	b_4 : 8.1046	0.2751
SSIM	Cubic	b_1 : 3273.7	b_2 : −8665.6	b_3 : 7651.4	b_4 : −2251.7	0.5926

Table 4.1: Regression of objective video quality models.

4.4 Performance of the video quality models

To calculate the performance of the video quality models, the metrics described in section 2.4 are used. Figure 4.6 presents the scatter plots of DMOS versus VQR for the used video quality models. The error bars depicts the 95% confidence interval of assessments done per PVS, while the red points indicate outliers.



(a) PSNR

(b) SSIM

(c) MSU

(d) NTIA

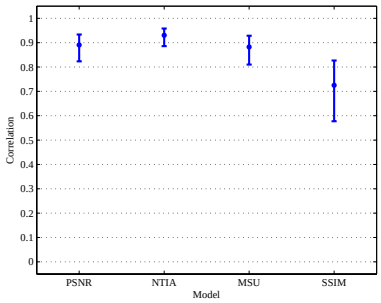
Figure 4.6: DMOS with 95% confidence interval versus predicted models and outliers.

The figures above show how well the predicted DMOS perform using a 95% confidence interval. A high number of confidence intervals crossing the depicted line indicates that the predicted DMOS of the video quality model fits well with the individual ratings. This figure shows how well the video quality models fit in with the subjective ratings per PVS. Based on the illustration of these figures, the NTIA model seem to give the best fit.

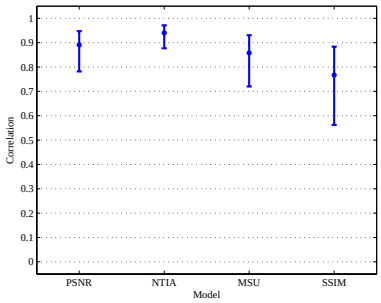
The outlier points shown in the plots can be misleading, because viewers rated the PVS identically using the DMOS (The viewers have rated them differently using the MOS scale, but the mapping to DMOS rendered them equal). This results in $\sigma = 0$, meaning that a DMOS prediction from an objective quality model must be a perfect match to the rated value. To conform to the original test plan and the manner of which an outlier is defined (see section 2.4), all outliers are included for further analysis.

Pearson linear correlation coefficient

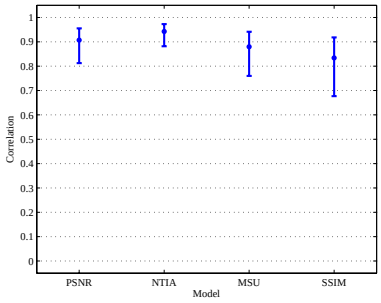
The Pearson linear correlation coefficient is calculated on a variety of subsets within the subjective tests. These subsets and the corresponding results are depicted in figure 4.7 and 4.8. The figures show the correlation coefficients with their 95% confidence intervals calculated by using the Fisher-z transform (equation 2.9).



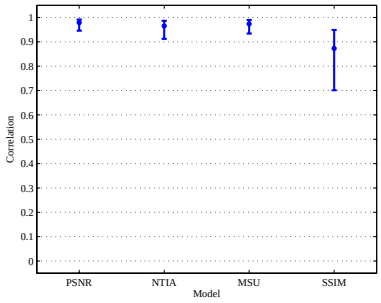
(a) Overall



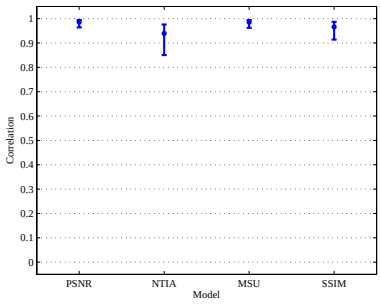
(b) Best Effort Overall



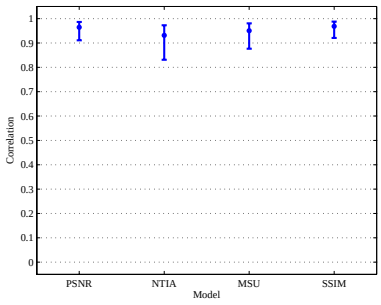
(c) DiffServ Overall



(d) Tandberg Overall

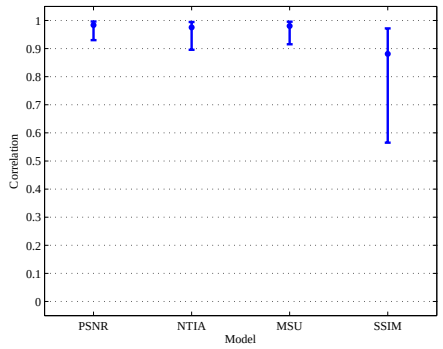


(e) Raven Overall

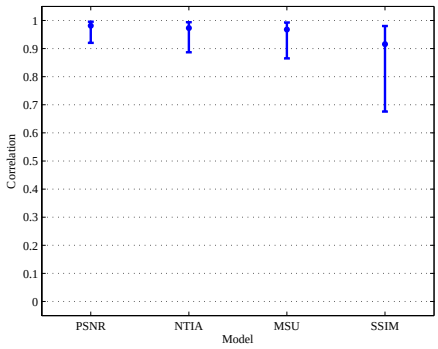


(f) Stem Overall

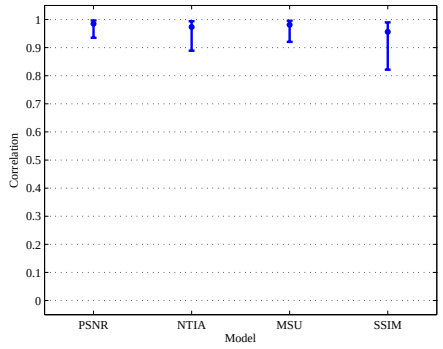
Figure 4.7: PLCC between DMOS and the objective models. The error bars indicate the 95% confidence intervals.



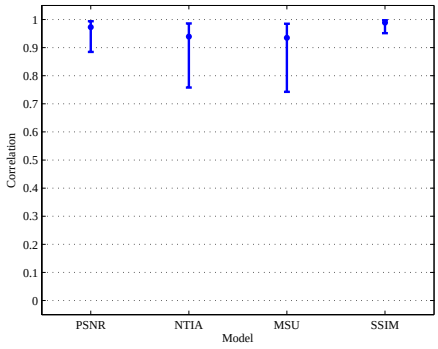
(a) Tandberg and DiffServ



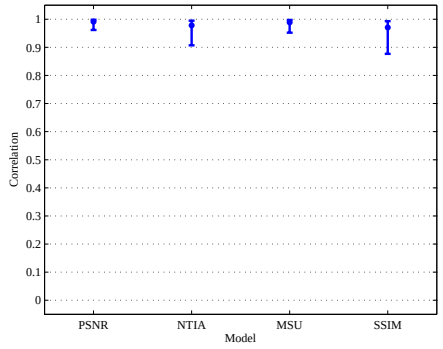
(b) Tandberg and Best Effort



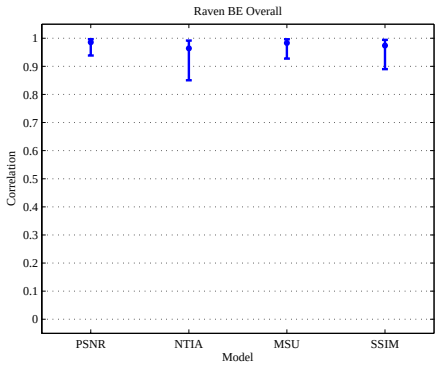
(c) Stem and DiffServ



(d) Stem and Best Effort



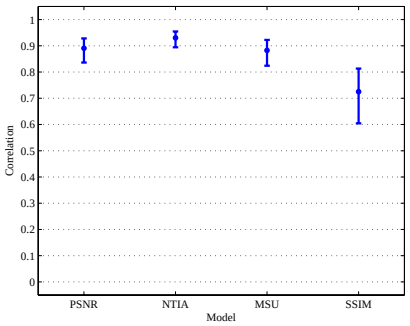
(e) Raven and DiffServ



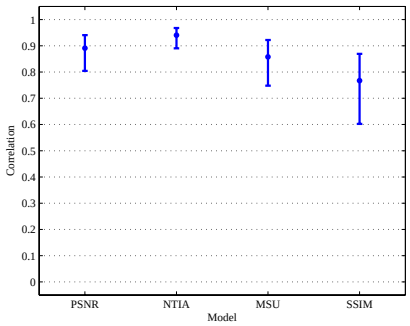
(f) Raven and Best Effort

Figure 4.8: More PLCC between DMOS and the objective models with 95% confidence intervals.

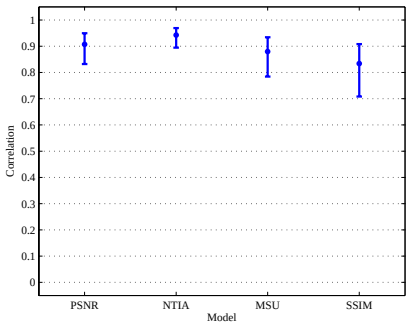
The figures indicate that all models, except SSIM, are highly linearly correlated with all the subjective ratings, both when testing over the whole set and over subsets of assessments. The PLCC of the SSIM model however, varies depending on video sequence used. When using the StEM clip with the best effort network, SSIM outperforms the other clips with the highest PLCC. But, we cannot statistically prove that SSIM is better, due to the overlapping of the confidence intervals of all the video quality models. Regarding figure 4.7(a), where the overall performance of the video quality models is depicted, NTIA significantly outperforms SSIM using a 95% confidence interval. In figure 4.9, the PLCC with 90% confidence interval is depicted, where PSNR, NTIA and MSU are all statistically significant better then SSIM using the entire set of ratings. Using ratings from the best effor PVSs, the NTIA has a statistically significant higher correlation then SSIM.



(a) Overall



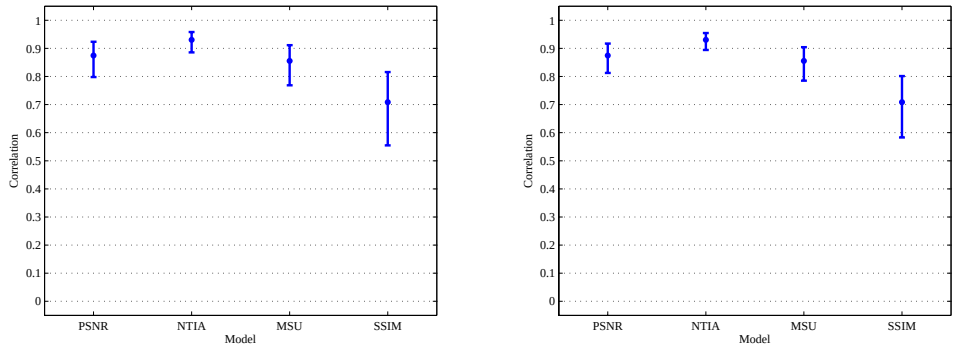
(b) Best Effort Overall



(c) DiffServ Overall

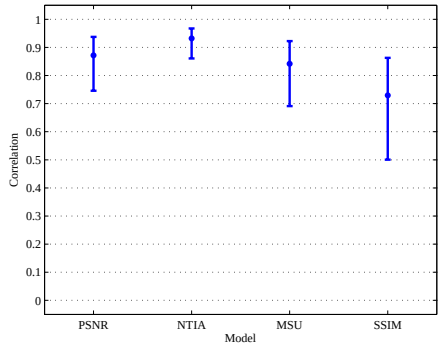
Figure 4.9: More PLCC between DMOS and the objective models with 90% confidence intervals.

Spearman Rank order correlation coefficient

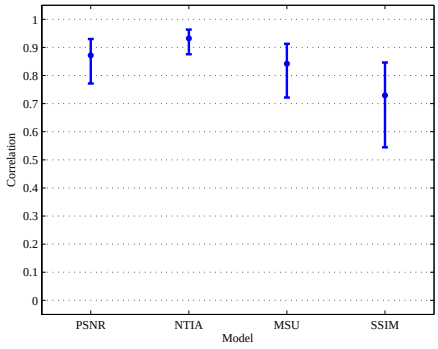


(a) Overall 95% CI

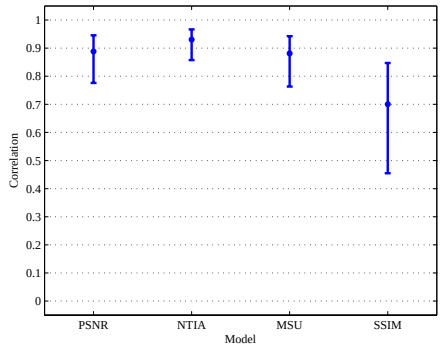
(b) Overall 90% CI



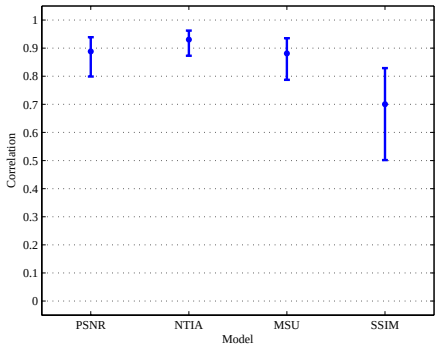
(c) Best Effort Overall 95% CI



(d) Best Effort Overall 90% CI



(e) DiffServ Overall 95% CI



(f) DiffServ Overall 90% CI

Figure 4.10: SRCC between DMOS and the objective models. The error bars indicate the 90% and 95% confidence intervals.

The confidence intervals for the SRCC were calculated using the Fisher-z transform. Figure 4.10(a) shows that the NTIA model is significantly better correlated than the SSIM model, with a confidence interval of 95%. When using 90%, this applies to all the three subsets, and PSNR shows a significantly higher correlation when using the overall set of ratings.

Comparing the results

The following tables show the Pearson linear correlation coefficients and the Spearman rank order correlation coefficient with a 95% confidence interval. All video quality models show no significant difference between the best effort network and DiffServ.

Data Set	PLCC	Low CI	High CI	SRCC	Low CI	High CI
PSNR						
Overall	0.8908	0.8230	0.9335	0.8746	0.7979	0.9235
DiffServ	0.9072	0.8125	0.9553	0.8880	0.7761	0.9457
BestEffort	0.8911	0.7818	0.9473	0.8717	0.7456	0.9376
MSU						
Overall	0.8825	0.8102	0.9284	0.8555	0.7685	0.9114
DiffServ	0.8794	0.7599	0.9414	0.8811	0.7631	0.9423
BestEffort	0.8578	0.7200	0.9305	0.8418	0.6912	0.9224
NTIA						
Overall	0.9303	0.8856	0.9579	0.9303	0.8856	0.9579
DiffServ	0.9401	0.8816	0.9726	0.9303	0.8856	0.9666
BestEffort	0.9425	0.8769	0.9714	0.9320	0.8607	0.9674
SSIM						
Overall	0.7249	0.5773	0.8267	0.7088	0.5547	0.8159
DiffServ	0.8338	0.6768	0.9182	0.7005	0.4550	0.8470
BestEffort	0.7672	0.5624	0.8833	0.7292	0.5004	0.8628

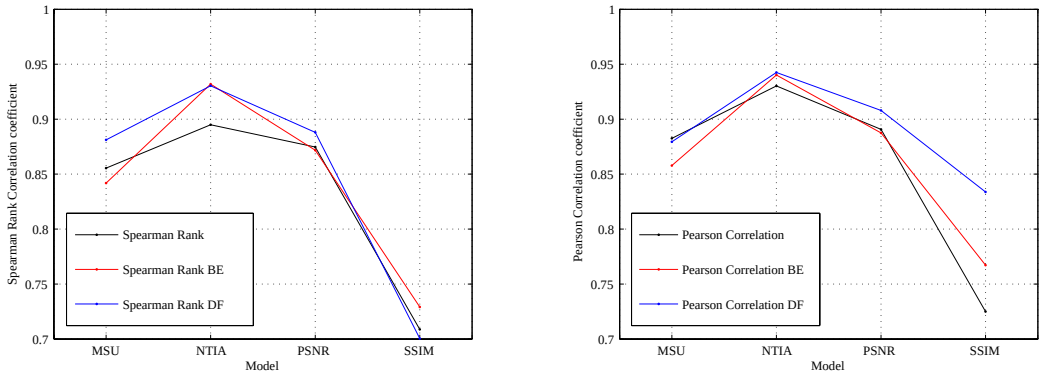
Table 4.2: Video quality models and PLCC and SRCC for 95% CI

Table 4.3 shows the results for all the performance metrics. The NTIA model has the highest performance for the overall, DiffServ and best effort sets in all the performance metrics. This holds true disregarding the outlier ratio, where as described earlier the standard deviation of the ratings given for that PVS is zero.

Model	Network	PLCC	SRCC	Outliers	Outlier Ratio	RMSE
PSNR	Overall	0.89078	0.87463	6	0.100%	0.52464
	DiffServ	0.9072	0.8880	3	0.100%	0.5011
	Best Effort	0.8876	0.8717	3	0.100 %	0.5846
NTIA	Overall	0.93029	0.89492	6	0.100 %	0.42349
	DiffServ	0.9425	0.89492	3	0.100 %	0.4001
	Best Effort	0.9401	0.9320	3	0.100 %	0.4756
MSU	Overall	0.88253	0.85548	6	0.100 %	0.54292
	DiffServ	0.8794	0.8811	3	0.100 %	0.4707
	Best Effort	0.8578	0.8086	3	0.100 %	0.6429
SSIM	Overall	0.72495	0.70876	11	0.183 %	0.79683
	DiffServ	0.8338	0.7005	7	0.2333 %	0.8448
	Best Effort	0.7672	0.7292	4	0.1333 %	0.8086

Table 4.3: Results from performance metrics

Figure 4.11(a) and 4.11(b) show the Spearman rank order and the Pearson linear correlation coefficients for the complete, DiffServ and Best Effort sets of ratings. The figures show that all models perform better for DiffServ clips than the Best Effort clips, except NTIA, which performs equally well in both scenarios.



(a) SRCC (b) SRCC

Figure 4.11: Spearman Rank and Pearson linear correlation coefficients for the different models.

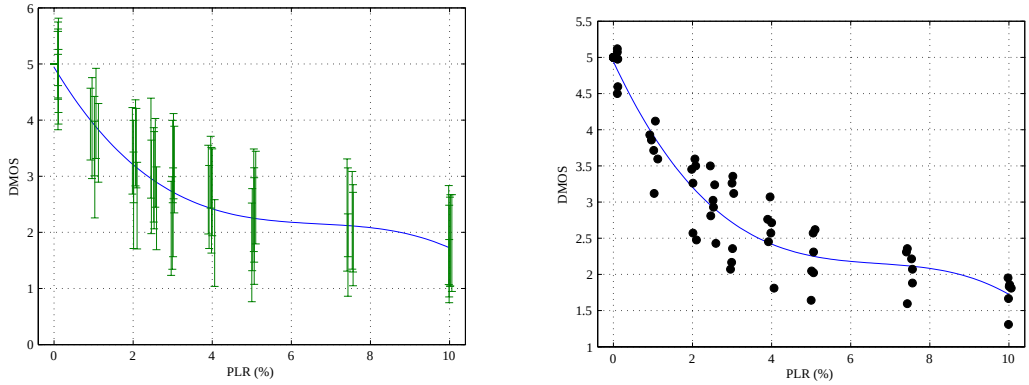


Figure 4.12: DMOS and packet-loss rate. Figure a) shows error bars which indicate the standard deviation for each PVS. Figure b) shows the actual DMOS average for each PVS.

The figures and values presented in this chapter either indicate a tendency or show statistically significant differences. These findings are summarized as follows:

1. Mean Opinion Score and video quality

- The differences in mean value for both MOS and DMOS over all sets of ratings are minor.
- Studying the high (5-10%) packet loss rate category, the DiffServ model receives a higher DMOS mean value compared to the Best Effort model.
- Studying the low (0-2.5%) and medium (2.5-4%) packet loss rate categories, the Best Effort model receives a higher DMOS mean value compared to the DiffServ model.

2. Model prediction based on Pearson linear correlation

- The NTIA model is statistically significant better than SSIM with a 95% confidence interval using the complete set of ratings with PLCC.
- The PSNR, MSU and NTI models are statistically significant better than SSIM with a 90% confidence interval using the complete set of ratings with PLCC.

3. Model prediction based on Spearman rank order correlation

- The NTIA model is statistically significant better than SSIM with a 95% confidence interval using both the complete set of ratings and the DiffServ subset with SPCC.
- The NTIA model is statistically significant better than SSIM with a 90% confidence intervall using the complete set, the best effort subset and the DiffServ subset with SPCC.
- The PSNR model is statistically significant better than SSIM with a 90% confidence intervall using the complete set with SPCC.

4. Video Quality Models scores and video quality

- The use of DiffServ with the described novel marking scheme, result in a gain in quality compared to the best effort model at fixed packet loss rates. This applies to all four video quality models employed.
- The cubic polynomial monotic regression fit well for all models, except for SSIM, where a high sum-of-squares value indicates a bad fit.

The NTIA model has the highest SPCC and PLCC, and lowest RMSE indicating that NTIA has the highest performance of the four models. The SSIM model has the lowest SPCC and PLCC, and highest RMSE indicating that SSIM has the lowest performance of the four models.

Sources of error

All the video quality models used for comparison are full-reference methods, where the original video signal is required for calculating the objective quality measure. This conflicts with the streaming scenario, where receivers do not have access to the undistorted original signal. It is incorrect to imply that these models are fit for video quality measurements in broadcasting or streaming systems, given that they perform well in the subjective evaluation experiment. No-reference or reduced-reference methods should be used for this scenario. The NTIA model however, can be implemented as a reduced-reference method.

To produce the test material, we simulate the behaviour of the two networks and introduce competing cross traffic that depletes available resources in the router. In order to create the desired set of packet loss rates, we vary the link capacity and the moment the studied source is introduced in the simulated router. By studying a plot of the cross-traffic as presented in figure 3.7, the router link and time of entry for the studied source is approximated, assuring that packets containing slice data from the first second of video sequence is not distorted. By the method of trial and error, simulations are done until the set of desired packet loss rates is achieved. This method for generating test material can be viewed as a source of error, when simulations are done until the desired packet loss rates is achieved. Despite this selection of test material, we have no control over which packets are lost and to what degree the loss distorts the video signal.

Another source of error is the encoded video sequences. They all differ in statistical content and require different bit rates to be encoded with similar PSNR values. We disregard this, and encode the three sequences yielding similar bit rates. This results in three different video sequences, with unequal amounts of distortion introduced by the encoder. The degree of distortion will not affect the subjective evaluations, as all analysis is based on DMOS. The objective video quality models however, can suffer from this, causing increased residuals when applying regression functions.

Chapter 5

Conclusion

The main objective of this thesis was to conduct an informal subjective evaluation experiment, where the test material used consists of high-definition video distorted by various packet loss rates, using both the best effort Internet and DiffServ as underlying channel models. We compared the results from the subjective evaluation experiment to results from objective video quality to see how well the objective models perform.

NTIA and SSIM were the video quality models with respectively the highest and the lowest performance regarding PLCC, SRCC and RMSE. The NTIA model had a statistically significant higher performance than SSIM using PLCC and SRCC with a 95% confidence interval.

When comparing packet loss rate versus objective measures, the performance of best effort degrades more rapidly than the performance of DiffServ. However, the results from the subjective evaluations did not show any statistically significant differences between the two channel models using a 90% confidence interval.

The DMOS values were categorized into low, medium and high packet loss rates. Studying the high (5-10%) packet loss rate category, the DiffServ model achieved a higher mean DMOS value compared to the Best Effort model. For low (0-2.5%) and medium (2.5-4%) packet loss rates categories the Best Effort model achieved a higher mean DMOS value compared to the DiffServ model.

Bibliography

[1] ITU-T and ISO/IEC JTC-1. Advanced Video Coding for generic audiovisual services. *ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 part 10) AVC*, March 2005.

[2] T. Wiegand, G.J. Sullivan, G. Bjøntegaard, and A. Luthra. Overview of the H.264/AVC Video Coding Standard. *IEEE Trans. On Circuits And Systems For Video Technology*, vol. 13(no. 7):560–576, July 2003.

[3] J. B. Postel. RFC 793: Transmission Control Protocol. IETF, September 1981.

[4] J. B. Postel. RFC 768: User Datagram Protocol. IETF, August 1980.

[5] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson. RFC 3550: RTP: A Transport Protocol for Real-Time Applications. IETF, July 2003.

[6] B. Braden and D. Clark and S. Shenker. RFC 1633: Integrated Services in the Internet Architecture: an Overview. IETF, June 1994.

[7] Steven Blake, David Black, Mark Carlson, Elwyn Davies, Zheng Wang, and Walter Weiss. RFC 2475: An Architecture for Differentiated Services. IETF, October 1998.

[8] L. Zhang and B. Braden and S. Berson and S. Herzog and S. Jami. RFC 2205: Resource Reservation Protocol. IETF, September 1997.

[9] B. Davie and A. Charny and J.C.R. Bennet and K. Benson and J.Y. Le Boudec and W. Courtney and S. Davari and V. Firoiu and D. Stiliadis. RFC 3246: An Expedited Forwarding PHB (Per-Hop Behavior). IETF, March 2002.

[10] J. Heinanen and F. Baker and W. Weiss and J. Wroclawski. RFC 2597: Assured Forwarding PHB Group. IETF, June 1999.

[11] R. Makkar and I. Lamadaris and J. Salim and N. Seddigh and B. Nandy and J. Babiarz. Empirical Study of Buffer Management Scheme for DiffServ Assured Forwarding PHB. ICCN, May 2000.

[12] S. Wenger. H.264/AVC over IP. *IEEE Trans. Circuits and Systems for Video Technology* , vol. 13(no. 7):645–656, July 2003.

[13] S. Wenger, M.M. Hannuksela, T. Stockhammar, M. Westerlund, and D. Singer. RFC 3984: RTP Payload Format for H.264 Video. IETF, February 2005.

[14] ITU-R. Methodology for the subjective assessment of the quality of television pictures. *ITU-R Rec. BT.500-11*, 2002.

[15] ITU-T. Subjective video quality assessment methods for multimedia applications. *ITU-T Rec. P.910*, October 1999.

[16] Z.Wang and A.C. Bovik and H.R. Sheikh and E.P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Processing*, vol. 13(no. 4):600–612, 2004.

[17] M.H. Pinson and S.Wolf. A New Standardized Method for Objectively Measuring Video Quality. *IEEE Trans. Broadcasting*, vol. 50(no. 3):312–322, July 2004.

[18] A. B. Watson and J. Hu and J.F. McGowan III. DVQ: A digital video quality metric based on human vision. *Journal of Electronic Imaging*, 10(1):20–29, 2001.

[19] F. Xiao. DCT-based Video Quality Evaluation. Stanford University: Final Project for EE392J. 2000, 2000.

[20] VQEG. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment. June 2000. http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/index.php.

[21] VQEG. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment Phase II. August 2003. http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/index.php.

[22] VQEG. Multimedia Group TEST PLAN Draft v1.11. February 2006. http://www.its.bldrdoc.gov/vqeg/projects/multimedia/index.php.

[23] S. Winkler. *Digital Video Quality: Vision Models and Metrics*. John Wiley & Sons Ltd, 1 edition, 2005.

[24] VQEG. Evaluation of new methods for objective testing of video quality: objective test plan. 1998. www.vqeg.org.

[25] Digital Cinema Initiatives, LCC. Standard Evaulation Material (StEM.

[26] Q2S - Centre for Quantifiable Quality of Service in Communication Systems. http://www.q2s.ntnu.no.

[27] Hillestad, O. I. and Libak, B. and Perkis, A. Performance Evaluation of Multimedia Services Over IP Networks. In *IEEE International Conference on Multimedia and Expo (ICME 2005)*, pages 1464–1467, July 2005.

[28] Endace DAG cards. http://www.endace.com.

[29] G. M. Birtwistle. *Demos - A System for Discrete Event Modelling on Simula*. MacMillan, 1979.

[30] Fredrik Solsvik and Stian Michaelsen. Measurements on IP-packets from video sources. Q2S, 2005.

[31] Joint Video Team Reference Software Version 10.2 (JM 10.2). http://iphome.hhi.de/suehring/tml/download/.

[32] Joint Video Team ISO/IEC JTC1/SC29/WG11 og ITU-T SG16 Q.6. Revised H.264/MPEG-4 AVC Reference software manual. http://ftp3.itu.ch/av-arch/jvt-site/2005_04_Busan/JVT-O017.doc.

[33] MPEG4IP. http://mpeg4ip.sourceforge.net.

[34] O.I. Hillestad, O. Jetlund, and A. Perkis. RTP-based Broadcast Streaming of High Definition H.264/AVC Video: an Error Robustness Evaluation. Packet Video 2006, 2006.

[35] Bjørnar Libæk. SimTraceTools. http://www.q2s.ntnu.no/ libak.

[36] VQEG, Video Quality Experts Group. http://www.vqeg.org.

Appendix **A**

PVS results

Table A.1 shows the calculated metrics for each video clip used in the objective and the subjective evaluation.

Table A.1: Subjective and Objective Results

SRC	PVS	PLR(%)	Network	MOS	DMOS	PSNR(Y)	SSIM	VQM MSU	VQM NTIA
1	01	0,00	Best Effort	4,50	5,00	40,05	0,956	0,949	0,112
1	02	0,11	Best Effort	4,48	4,98	39,71	0,955	1,178	0,121
1	03	0,96	Best Effort	3,36	3,86	38,49	0,949	1,759	0,176
1	04	2,06	Best Effort	3,10	3,60	37,24	0,941	2,111	0,254
1	05	2,45	Best Effort	3,00	3,50	37,95	0,940	1,753	0,229
1	06	3,04	Best Effort	2,62	3,12	35,94	0,936	2,712	0,313
1	07	3,90	Best Effort	2,26	2,76	35,49	0,930	2,687	0,342
1	08	5,06	Best Effort	1,52	2,02	34,62	0,899	2,895	0,409
1	09	7,41	Best Effort	1,81	2,31	33,55	0,904	3,290	0,452
1	10	10,06	Best Effort	1,31	1,81	32,07	0,875	4,034	0,523
1	11	0,00	DiffServ	4,50	5,00	40,05	0,956	0,949	0,112
1	12	0,10	DiffServ	4,57	5,07	39,85	0,956	1,099	0,120
1	13	1,03	DiffServ	2,62	3,12	38,22	0,944	1,926	0,166
1	14	2,01	DiffServ	2,07	2,57	37,29	0,935	2,221	0,240
1	15	2,53	DiffServ	2,43	2,93	37,08	0,937	2,246	0,229
1	16	2,96	DiffServ	1,57	2,07	35,85	0,925	2,813	0,267
1	17	3,98	DiffServ	2,07	2,57	35,71	0,920	2,728	0,282
1	18	5,01	DiffServ	1,55	2,05	33,87	0,912	3,485	0,326
1	19	7,56	DiffServ	1,57	2,07	34,12	0,898	3,541	0,345
1	20	10,01	DiffServ	1,33	1,83	33,27	0,875	3,686	0,385
2	21	0,00	Best Effort	4,43	5,00	41,22	0,964	0,628	0,138
2	22	0,10	Best Effort	4,55	5,12	41,10	0,964	0,695	0,139
2	23	1,06	Best Effort	3,55	4,12	40,58	0,960	0,851	0,150
2	24	1,98	Best Effort	2,88	3,45	39,50	0,950	1,061	0,218
2	25	2,52	Best Effort	2,45	3,02	39,14	0,944	1,278	0,258
2	26	3,02	Best Effort	2,79	3,36	39,13	0,945	1,245	0,212
2	27	3,96	Best Effort	2,50	3,07	39,54	0,948	1,105	0,216

SRC	PVS	PLR(%)	Network	MOS	DMOS	PSNR(Y)	SSIM	VQM MSU	VQM NTIA
2	28	5,10	Best Effort	2,05	2,62	38,21	0,940	1,449	0,262
2	29	7,56	Best Effort	1,31	1,88	36,47	0,925	2,193	0,288
2	30	9,98	Best Effort	1,38	1,95	35,06	0,904	2,210	0,381
2	31	0,00	DiffServ	4,43	5,00	41,22	0,964	0,628	0,138
2	32	0,10	DiffServ	3,93	4,50	41,03	0,964	0,734	0,145
2	33	1,12	DiffServ	3,02	3,60	40,23	0,958	1,006	0,159
2	34	2,08	DiffServ	2,93	3,50	39,61	0,954	1,224	0,183
2	35	2,56	DiffServ	2,67	3,24	39,40	0,951	1,184	0,183
2	36	3,00	DiffServ	2,69	3,26	39,72	0,953	1,169	0,180
2	37	4,00	DiffServ	2,14	2,71	38,75	0,941	1,320	0,204
2	38	5,06	DiffServ	1,74	2,31	37,59	0,934	1,781	0,244
2	39	7,54	DiffServ	1,64	2,21	37,15	0,928	1,870	0,245
2	40	10,02	DiffServ	1,29	1,86	36,19	0,909	2,324	0,280
3	41	0,00	Best Effort	4,79	5,00	41,23	0,971	0,738	0,104
3	42	0,11	Best Effort	4,38	4,60	41,04	0,971	0,884	0,114
3	43	1,02	Best Effort	3,50	3,71	40,20	0,967	1,079	0,160
3	44	2,10	Best Effort	2,26	2,48	37,94	0,959	2,006	0,302
3	45	2,46	Best Effort	2,60	2,81	37,30	0,957	2,119	0,346
3	46	2,99	Best Effort	1,95	2,17	37,72	0,958	2,063	0,321
3	47	4,06	Best Effort	1,60	1,81	35,50	0,945	2,434	0,445
3	48	5,00	Best Effort	1,43	1,64	36,33	0,941	2,122	0,403
3	49	7,43	Best Effort	1,38	1,60	32,83	0,918	3,286	0,605
3	50	9,99	Best Effort	1,10	1,31	32,68	0,892	3,228	0,600
3	51	0,00	DiffServ	4,79	5,00	41,23	0,971	0,738	0,104
3	52	0,10	DiffServ	4,76	4,98	41,19	0,971	0,750	0,106
3	53	0,92	DiffServ	3,71	3,93	40,35	0,969	1,173	0,144
3	54	2,01	DiffServ	3,05	3,26	39,43	0,965	1,401	0,194
3	55	2,59	DiffServ	2,21	2,43	38,91	0,962	1,577	0,206
3	56	3,01	DiffServ	2,14	2,36	38,98	0,963	1,709	0,201
3	57	3,92	DiffServ	2,24	2,45	38,40	0,962	2,008	0,209
3	58	5,05	DiffServ	2,36	2,57	38,77	0,958	1,505	0,239
3	59	7,43	DiffServ	2,14	2,36	37,63	0,955	2,034	0,259
3	60	9,99	DiffServ	1,45	1,67	35,58	0,943	2,775	0,363

Appendix B

Instructions to subjects

Subjective Test of Video Quality

1. Introduction

Thank you very much for participating in this research project, which is related to the visual quality of video when transported in error-prone environments, like e.g. the Internet.

You will be participating in what is called a subjective test, in which you, based on your own personal taste, will judge and rate the visual quality of 57 video clips that are presented on the computer screen in front of you. Each clip is between 8 an 12 seconds long, and after seeing each clip you will have 10 seconds to give your rating in the attached form (see pages 2, 3 and 4). If you, in the middle of the test, need to see a clip one more time, please tell the test organizer as soon as possible.

The rating is done using a five point measurement scale called Mean Opinion Score (MOS), represented by the categories "Excellent", "Good", "Fair", "Poor" and "Bad". Your scores should reflect your own personal taste and judgment. Your task is to judge the visual quality of the images in the video – not the cinematic content.

During the test, we would like you to sit at the located chair, and keep your head reasonably close to the marked point. This is because the presented video might look different from different positions, and we would like everyone to judge the videos from the same positions. You can of course move around on the chair to stay comfortable.

2. Terms of participation and privacy

Your participation is completely voluntarily, and you may withdraw from the test at any time. This will not have any negative consequences for you, and your details (and partially submitted test response) will be deleted. There is no predicted risk associated with participating in this test. The data you provide can in no way be connected to your name or any other personal details that you provide, at a later stage. The submitted scores will only be reported as a statistical average among the entire group of test participants.

3. Personal details

Name: _____ (Optional, see section 2 above)

Department: IME __ SVT __ AB __ Other: _____

Age: _____

Gender: Male __ Female __

Subject number: _____ (To be filled in by test organizer)

Appendix C

Conversion from 1080i to 720p

A procedure for converting an interlaced 1440 by 1080 pixel video sequence to a progressive 1280 by 720 pixel video sequence was described in section 3.2. This is done by removing the bottom field, before upsampling by a filter matrix is performed. The filter is an interpolation filter which extends the height of a frame by a ratio of 3:4.

The filter matrix[1] has the following filter coefficients:

$$\begin{bmatrix} 0 & 32 & 0 & 0 & 0 & 0 \\ -1 & 8 & 28 & -3 & 0 & 0 \\ 0 & -3 & 19 & 19 & -3 & 0 \\ 0 & 0 & -3 & 28 & 8 & -1 \\ 0 & 0 & 0 & 0 & 32 & 0 \end{bmatrix} * \frac{1}{32}$$

The matlab code used for the conversion is listed below:

```
% Video files: raw YUV 4:2:0 sampled video (YV12 planar format)
fpin = fopen('CLIP2A_1080i_60Hz_800f.yuv','rb');  %Original input sequence
fpout= fopen('CLIP2A_720p_30Hz_800f.yuv', 'wb');  %Converted output sequence

% number of frames to be read
nof = [1 800];
CurrSize = [1440 1080];
TargetSize = [1280 720];
yTop = zeros(size(1), size(2));
yBot = zeros(size(1), size(2));

% Filter matrix
A = [ 0 32  0  0  0  0 ;
     -1  8 28 -3  0  0 ;
      0 -3 19 19 -3  0 ;
      0  0 -3 28  8 -1 ;
      0  0  0  0 32  0]/32;

% read frames individually in a for loop.
for k=nof(1):nof(2)
```

[1] The coefficients are provided by Dr. Gisle Bjøntegaard, TANDBERG

```
      [y]=fread(fpin,[size(1),size(2)],'uchar');         %Read in luminance.
      [u]=fread(fpin,[(size(1)/2),(size(2)/2)],'uchar'); %Read in the chrominance
      [v]=fread(fpin,[(size(1)/2),(size(2)/2)],'uchar'); %Read in the chrominance
      yTop = y(:,1:2:size(2)-1)'; %Use only top fields
      uTop = u(:,1:2:size(2)/2-1)';
      vTop = v(:,1:2:size(2)/2-1)';

      % Luminance interpolation
      x = zeros(546,1440); %540 + 6 samples due to filter
      My = zeros(720,1440);  %padding
      x(2:541,:) = yTop;
      for c=1:1440
         j=1;
         for i=2:3:540
             My(j:j+4,c) = A*x(i-1:i+4,c);
             j = j+4;
         end
      end

      %Chroma interpolation
      xu = zeros(276,720);
      xv = zeros(276,720);
      Mu = zeros(360,720);
      Mv = zeros(360,720);
      xu(2:271,:) = uTop;
      xv(2:271,:) = vTop;
      for c=1:720
         j=1;
         for i=2:3:270
             Mu(j:j+4,c) = A*xu(i-1:i+4,c);
             Mv(j:j+4,c) = A*xv(i-1:i+4,c);
             j = j+4;
         end
      end

      % Write output
      fwrite(fpout, My(1:720,81:1360)', 'uchar');
      fwrite(fpout, Mu(1:360,41:680)', 'uchar');
      fwrite(fpout, Mv(1:360,41:680)', 'uchar');
end;
fclose('all');
```

Appendix **D**

Cross-traffic encoded video sequences

The following sequences are use to create the aggregated cross-traffic. Their corresponding packet flows were used once or several times in the cross-traffic, inserted with a uniform delay between 1 and 4 seconds. All the sequences are encoded with the H.264/AVC reference encoder.

Sequence	Length(seconds)	Rate (kbps)	Resolution (pixels)	Framerate (FPS)
IceCity	27.800	3166	704x576	30
Lillestrøm	21.640	1112	352x288	25
Paris	35.500	792	352x288	30
SpanishNews_News	41.800	1342	352x288	25
SpanishNews_football	37.000	1907	352x288	25
StEM excerpt #1	99.916	1170	352x288	24
StEM excerpt #2	83.333	1486	352x288	24
StEM excerpt #3	100.791	1070	352x288	24