Slovenská technická univerzita v Bratislave

Fakulta elektrotechniky a informatiky

Katedra rádioelektroniky

Time-variant video quality evaluation for mobile networks

Šk. rok: 2005/2006

Bc. Matej Závodský

Ing. Olivia Nemethová, Prof. Dr. Markus Rupp, TU Wien

doc. Ing. Vladimír Kudják, PhD. STU Katedra rádioelektroniky

2. strana Zadanie DP

Hodnotenie časovo závislej kvality video signálu v mobilných sieťach

Študijný odbor:	Elektronika
Autor:	Bc. Matej Závodský
Vedúci diplomovej práce:	doc. Ing. Vladimír Kudják, PhD.

Máj 2006

Používaním video aplikácií, s nízkymi bitovými a obrazovými rýchlosťami cez paketovú bezdrôtovú sieť, narastá potreba merania video kvality. Výpadkami paketov, sa video kvalita stáva značne časovo závislá. Subjektívne vyhodnocovanie kvality sa hlavne stanovuje metrikami založenými na rozdielnosti medzi degradovanou a referenčnou sekvenciou. Tieto metriky sa nazývajú referenčné. Nereferenčné metriky stanovujú iba obmedzený kvalitívny odhad. V tejto práci bola použitá referenčná, časovo variantná metrika založená na spojitosti medzi PSNR (odstup signál šum) a spriemerovaným hodnotením názorov respondentov. Študovali sme jednoduché periodicky sa vyskytujúce ľudské reakcie na artefakty a použili ich na korekciu PSNR a to na dosahovanie lepšieho odhadu objektívnej kvality. Výsledky našej metriky sme porovnali s výsledkami subjektívnych testov. Keďže sa zameriavame na bezdrôtovú komunkáciu, subjektívne testy boli uskutočnené na mobilných telefónoch použitím vlastnej metodológie. Metodológia vychádzala s doporučení ITU-R BT 500. Výsledky ukazujú značné vylepšenie odhadovanej kvality vzhľadom na PSNR.

Time-variant video quality evaluation for mobile networks

Degree Course:	Electrical Engineering
Author:	Bc. Matej Závodský
Supervisor:	doc. Ing. Vladimír Kudják, PhD.

May 2006

With low-rate video services over packet networks, the topic of video quality measuring gained on importance. In the presence of packet loss, the video quality becomes highly time-variant. The subjective quality evaluation is in general correlated with the metrics based on the difference between the degraded sequence and the reference. However, it can differ significantly in some cases, thus it provides limited means for the appropriate quality estimation. In this work we focus on the relation between the peak to signal-to-noise ratio (PSNR) and the mean opinion score for the video streams with time-variant quality. Later, we studied simple human perception rules and used them to correct the PSNR accordingly, to achieve suitable subjective quality estimation, based on objective parameters. We compared the results from our metric with the mean opinion score (MOS) obtained from subjective tests. We focused on a mobile terminal and a video application. Thus, we performed assessments on mobile terminals using our proposed methodology which was derived from recommendation ITU-R BT 500. The results showed considerable improvement of the estimation quality with respect to the PSNR.

Čestné prehlásenie

Prehlasujem, že som túto prácu vypracoval sám a inú ako uvedenú literatúru som nepoužíval.

V Bratislave 11.5.2006

Matej Závodský

CONTENTS

ABBREVIATIONS

ANSI	American National Standards Institute
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
CIE	Commission Internationale de l'Eclairage
CRT	Cathode Ray Tube
CSF	Contrast Sensitivity Function
dB	Decibel
DCT	Discrete Cosine Transform
DMOS	Differential Mean Opinion Score
DPCM	Differential Pulse Code Modulation
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DVD	Digital Versatile Disc
FIR	Finite Impulse Response
HDTV	High-Definition Television
HVS	Human Visual System
IEC	International Electrotechnical Commission
IIR	Infinite Impulse Response
ISO	International Organization for Standardization
ITS	Institute for Telecommunication Sciences
ITU	International Telecommunication Union
JPEG	Joint Picture Experts Group
kb/s	Kilobit per second
LCD	Liquid Crystal Display
MB	Macro Block
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MSE	Mean Squared Error
MV	Motion Vector
NTSC	National Television Systems Committee
NVFM	Normalization Video Fidelity Metric

PAL	Phase Alternating Line
PC	Personal Computer
PDM	Perceptual Distortion Metric
PSNR	Peak to Signal to Noise Ratio
QCIF	Quarter Common Intermediate Format
GOB	Group Of Block
QoS	Quality of Service
PL	Packet Loss
PSC	Picture Start Code
RGB	Red, Green, Blue
SI	Spatial Information
SNR	Signal-to-Noise Ratio
SSCQE	Single Stimulus Continuous Quality Evaluation
TCP/IP	Transmission Control Protocol/Internet Protocol
TI	Temporal Information
TR	Temporal Reference
UMTS	Universal Mobil Telecommunications System
VQEG	Video Quality Experts Group

1 Introduction

Wireless mobile systems of the 3rd generation like universal Mobil telecommunications System (UMTS) offers multimedia communications such as video streaming or video conferencing.

Video streaming is a multimedia service, which is becoming more and more popular. Furthermore, it is expected to unlock new revenue flows to mobile network operators. Typically, users access online video clips by clicking on a hyperlink using their web browser, which results in the browser opening a video player to play the selected clip. With streaming, the video content doesn't need to be completely downloaded, but the client can begin play out the video a few seconds after it has begun receiving parts of the content from the streaming server. Since raw video requires high transmission bitrates, video compression is usually employed to achieve transmission efficiency. Video transmission over wireless is characterized by a specific set of requirements that include low bitrates, small frame sizes, and low frame rates. Finally the content is viewed at short distance on a small liquid crystal display (LCD) screen with a progressive display. Furthermore, wireless transport of real-time video streams over an error prone environment results in packet loss. This phenomenon leads to various kinds of error artifacts and their possible spatial and temporal propagation in the video stream displayed at the user terminal.

The video sequences which are to be assessed exhibit artifacts resulting not only of compression, but also of transmission errors. The character of the artifacts depends on the used compression method (MPEG 4, H.263/4), the position of the error and the decoder implementation (i.e. using error concealment). In order to ensure the required quality of service it is important to monitor the network performance on the user side. An important task related to network monitoring is the end-user perceived quality estimation. During the last decade an extensive research has been carried out in the field. In most cases, research focused on the broadcasting and internet applications. Thus, most recommendations describing the tests methodology [3] and metric synthesis [15] are based on such assumptions.

However, according to what we observed and examined in various tests, the user evaluation of the same video shown on the PC and on the mobile phone can differ. Therefore, the metrics synthesized out of such tests may not reflect the user perception

We focus on wireless communication and video artifacts evaluation, which arise throughout video streaming on wireless channel. We use reference time-variant metric for evaluating the video quality. We compare the results from our metric with mean opinion score (MOS) obtained from subjective tests. We performed the assessments on mobile terminals using our proposed methodology. The results show considerable improvement of the estimation quality with respect to the PSNR.

This thesis is organized as follows: Chapter 2 outlines compression methods and standards. The compression and transmission of digital video entail a variety of characteristic artifacts and distortions, the most common of which are discussed. Chapter 3 reviews recent vision models and the state of the art of quality metrics. In Chapter 4 the sequences selected for evaluation are described as well as the setup of survey performed to obtain the MOS values. Chapter 5 describes evaluation of the data from subjective tests, while chapter 6 describes the video quality metric design and metric performance.

2 Block based video coding

This chapter starts with an overview of video essentials, today's compression methods and standards. The compression and transmission of digital video for UMTS application entail a variety of characteristic artifacts and distortions, the most common of which are discussed here.

2.1 Compression

Visual data in general and video in particular require large amounts of bandwidth and storage space. Uncompressed Quarter Common Intermediate Format (QCIF) resolution video has typical data rates of several hundred Mb/s, for example 10 seconds of a 176 x 144 pixel video sequence played at a framerate of 30 frames per second, corresponds to approximately 180 Mbit of data, or 18 Mbps. In UMTS is maximum speed 2Mbps per cell! The need for data compression is obvious. Evidently, effective compression methods are vital to facilitate handling such data rates. Compression is the removal of redundant information from data. Generic lossless compression algorithms, which assure the perfect reconstruction of the initial data, are used for example in dictionary codes or in run-length encoding. However, the results are far from optimal, because these algorithms only take into account the linear bit stream. When compressing video, two special types of redundancy can be exploited:

- Spatio-temporal redundancy: Typically, pixel values are correlated with their neighbors, both within the same frame and across frames. This allows for using the spatial and temporal prediction methods and differential encoding.
- Psychovisual redundancy: The human visual system is not equally sensitive to all patterns. Therefore, the compression algorithm can discard information that is not visible or not so important to the observer. This is referred to as *lossy compression*.

In analog video, these two types of redundancies are exploited through vision-based color coding and interlacing techniques. Digital video offers additional compression methods, which are discussed in the following.

2.2 Color Coding

Many compression schemes and video standards such as National Television Systems Committee (NTSC), Phase Alternating Line (PAL), or Moving Pictures Experts Group (MPEG), are already based on human vision in the way that color information is processed. In particular, they take into account the nonlinear perception of lightness, the organization of color channels, and the low chromatic acuity of the human visual system.

Conventional television cathode ray tube (CRT) displays have a nonlinear, roughly exponential relationship between frame buffer Red, Green, Blue (RGB) values or signal voltage and displayed intensity. In order to compensate for this, *gamma correction* is applied to the intensity values before coding. It so happens that the human visual system has an approximately logarithmic response to intensity, which is very nearly the inverse of the CRT nonlinearity [4]. Therefore, coding visual information in the gamma-corrected domain not only compensates for CRT behavior, but is also more meaningful perceptually. The growing popularity of flat-screen plasma and LCD monitors, which do not have gamma correction factors associated with electron guns, will help standardize calibration between the display and linear imaging devices

The theory of opponent colors states that the human visual system decorrelates its input into white-black, red-green and blue-yellow difference signals, which are processed in separate visual channels. Furthermore, chromatic visual acuity is significantly lower than achromatic acuity. In order to take advantage of this behavior, the color primaries red, green, and blue are rarely used for coding directly. Instead, *color difference* (chroma) signals similar to the ones just mentioned are computed. In component video, for example, the resulting color space is referred to as $Y\,UV$ or $YC_B\ C_R$, where Y encodes luminance, U or C_B the difference between the blue primary and luminance, and V or C_R the difference between the red primary and luminance.

The low chromatic acuity now permits a significant data reduction of the color difference signals. In digital video, this is achieved by chroma subsampling. The notation commonly used is as follows:

- 4:4:4 denotes no chroma subsampling.
- 4:2:2 denotes chroma subsampling by a factor of 2 horizontally; this sampling format is used in the standard for studio-quality component digital video as defined by ITU-R Recommendation BT.601-5 [9], for example.

- 4:2:0 denotes chroma subsampling by a factor of 2 both horizontally and vertically; it is probably the closest approximation of human visual color acuity achievable by chroma subsampling alone. This sampling format is the most common in motion-Joint Photographic Experts Group (JPEG) or MPEG, e.g. for distribution-quality video.
- 4:1:1 denotes chroma subsampling by a factor of 4 horizontally.

2.3 Interlacing

As analog television was developed, it was noted that flicker could be perceived at certain frame rates, and that the magnitude of the flicker was a function of screen brightness and surrounding lighting conditions. A motion picture displayed in the theater at relatively low light levels can be displayed at a frame rate of 24 Hz. A bright CRT display requires a refresh rate of more than 50 Hz for flicker to disappear. The drawback of such a high frame rate is that the bandwidth of the signal becomes very high. On the other hand, the spatial resolution of the visual system decreases significantly at such temporal frequencies (this is the sharp fall-of range of the CSF in the high spatio-temporal frequency range). These two properties combined gave rise to the technique referred to as *interlacing*.

The concept of interlacing is illustrated in Figure 2.1. Interlacing trades of vertical resolution with temporal resolution. Instead of sampling the video signal at 25 (PAL) or 30 (NTSC) frames per second, the sequence is shot at a frequency of 50 or 60 interleaved fields per second. A field corresponds to either the odd or the even lines of a frame, which are sampled at different time instants and displayed alternately (the field containing the even lines is referred to as the top field, and the field containing the odd lines as the bottom field). Thus the required bandwidth of the signal can be reduced by a factor of 2, while the full horizontal and vertical resolution is maintained for stationary image regions, and the refresh rate for objects larger than one scanline is still suffciently high.

Most of the multimedia codecs for desktop computers handle only progressive video, which is better adapted to computer displays. Compression standards for digital video on the other hand have to support both interlaced and progressive video.

2.6 Digital Video Compression

Digital video is amenable to specialized compression methods. They can be roughly classified into model-based methods, e.g. fractal compression, and waveform-based methods, e.g. wavelet compression. Most of to day's video codecs and standards belong to the latter category and comprise the following stages [10]:

- Transformation: To facilitate the exploitation of psychovisual redundancies, the pictures are transformed to a domain where different frequency ranges with varying sensitivities of the human visual system can be separated. This can be achieved by the discrete cosine transform (DCT) or the wavelet transform, for example.
- Quantization: After the transformation, the numerical precision of the transform coefficients is reduced in order to decrease the number of bits in the stream. The degree of quantization applied to each coefficient is usually determined by the visibility of the resulting distortion to a human observer; high-frequency coefficients can be more coarsely quantized than low frequency coefficients, for example. Quantization is the stage that is responsible for the loss of information.

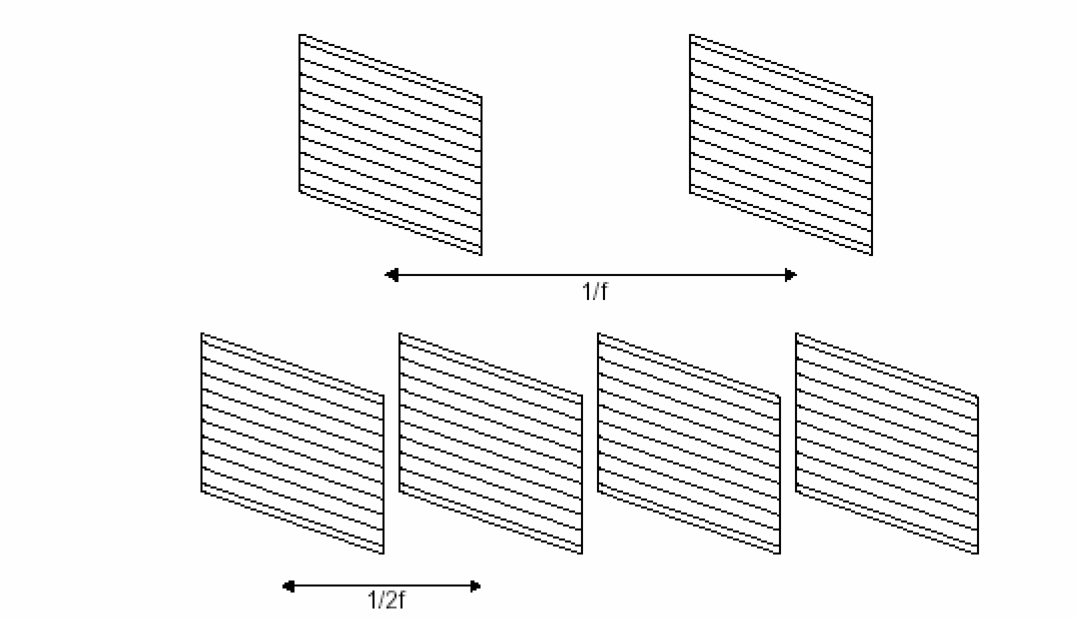


Figure 2.1: Illustration of interlacing. The top sequence is progressive: all lines of each frame are transmitted at the frame rate *f*. The bottom sequence is interlaced: each frame is split in two fields containing the odd and the even lines, respectively. These fields (bold lines) are transmitted alternately at twice the original frame rate.

- Coding: After the data has been quantized into a finite set of values, it can be encoded losslessly by exploiting the redundancy between the quantized coefficients in the bit stream. Entropy coding, which relies on the fact that certain symbols occur much more frequently than others, is often used for this process. Two of the most popular entropy coding schemes are Huffman coding and arithmetic coding [11].

The key aspect of digital video compression is the exploitation of the similarity between successive frames in a sequence instead of coding each picture separately. While this temporal redundancy could be taken care of by a spatio-temporal transformation, a hybrid spatial- and transform-domain approach is of ten adopted instead for reasons of implementation efficiency. A simple method for temporal compression is frame differencing, where only the pixel-wise differences between successive frames are coded. Higher compression can be achieved using motion estimation, a technique for describing a frame based on the content of nearby frames by means of motion vectors. By compensating for the movements of objects in this manner, the differences between frames can be further reduced.

2.5 Standards

The Moving Picture Experts Group (MPEG) is a working group of International Organization for Standardization/ International Electrotechnical Commission (ISO /IEC) in charge of developing international standards for the compression, decompression, processing, and coded representation of moving pictures, audio and their combination. MPEG comprises some of the most popular and widespread standards for video coding. The group was established in January 1988, and since then it has produced:

- MPEG-1, a standard for storage and retrieval of moving pictures and audio, which was approved in November 1992. MPEG-1 is intended to be generic, i.e. only the coding syntax is defined, and therefore mainly the decoding scheme is standardized. MPEG-1 defines a blockbased hybrid Discrete Cosine Transform /Differential Pulse Code Modulation (DCT/DPCM) coding scheme with prediction and motion compensation. It also provides functionality for random access in digital storage media.
- MPEG-2, a standard for digital television, which was approved in November 1994. The video coding scheme used in MPEG-2 is again generic; it is a refinement of the one in MPEG-1. Special consideration is given to interlaced sources. Furthermore, much functionality such as scalability was introduced. In order to keep implementation complexity low for products not requiring all video formats

supported by the standard, so-called “Profiles”, describing functionalities, and “Levels”, describing parameter constraints such as resolutions and bitrates, were defined to provide separate MPEG-2 conformance levels. MPEG-2 video stream is hierarchically structured as illustrated in Figure 2.2 [10]. The sequence is composed of three types of frames, namely intra-coded (I), forward predicted (P), and bidirectionally predicted (B) frames. Each frame is subdivided into slices, which are a collection of consecutive macroblocks. Each macroblock in turn contains 4 blocks of 8×8 pixels each. The DCT is computed on these blocks, while motion estimation is performed on macroblocks.

- MPEG-4, a standard for multimedia applications, whose first version was approved in October 1998. MPEG-4 addresses the need for robustness in error-prone environments, interactive functionality for content-based access and manipulation, and high compression efficiency at very low bitrates. MPEG-4 achieves these goals by means of an object-oriented coding scheme using so-called “audio-visual objects”, for example a fixed background, the picture of a person in front of that background, the voice associated with that person etc. The basic video coding structure supports shape coding, motion compensation, DCT-based texture coding as well as a zero tree wavelet algorithm.

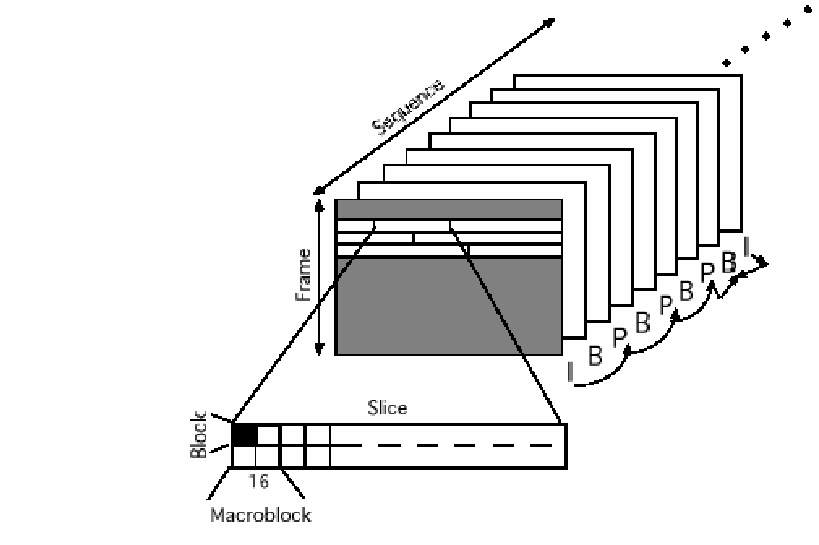


Figure 2.2: Elements of an MPEG-2 video sequence.

- MPEG-7, a standard for content description in the context of audio-visual information indexing, search and retrieval, which is scheduled for approval in September 2001.

- MPEG-21, work on this standard, also called the Multimedia Framework, has just begun. MPEG-21 will attempt to describe the elements needed to build an infrastructure for the delivery and consumption of multimedia content, and how they will relate to each other.
- M- JPEG (Motion JPEG) JPEG stands for Joint Photographic Experts Group. It is also an ISO/IEC working group, but works to build standards for continuous tone image coding. This scheme basically combines a series of JPEG still images that are rapidly displayed to form a moving image.

The next enhanced codec was developed by International Telecommunication Union (ITU-T).

- H 261 It is an ITU standard designed for two-way communication over ISDN lines (video conferencing) and supports data rates which are multiples of 64Kbit/s. The algorithm is based on DCT and can be implemented in hardware or software and uses intraframe and interframe compression. H.261 supports Common Intermediate Format (CIF) and Quater CIF (QCIF) resolutions.
- H 264 uses the latest innovations in video compression technology to provide incredible video quality from the smallest amount of video data. H.264 delivers the same quality as MPEG-2 at a third to half the data rate and up to four times the frame size of MPEG-4 Part 2 at the same data rate. H.264 achieves the best-ever compression efficiency for a broad range of applications, such as broadcast, DVD, video conferencing, video on demand, streaming and multimedia messaging. And true to its advanced design, H.264 delivers excellent quality across a wide operating range, from 3G to HD and everything in between.
- H 263, is based on H.261 with enhancements that improve video quality over modems. It supports CIF, QCIF, SQCIF, 4CIF and 16CIF resolutions. Will be better explain follow.

The standard H 263 which was used for our video sequences is detailed describe thereinafter.

2.5.1 H 263 standard

New standards such as H 263 or MPEG- 4 supported low bitrate communication solution for videoconferencing or streaming media. It is expected that wireless video communication will become a widespread and often used technology.

The coding algorithm of H.263 is similar to that used by H.261, however with some improvements and changes to improve performance and error recovery. The differences between the H.261 and H.263 coding algorithms are listed below. Half pixel precision is used for motion compensation whereas H.261 used full pixel precision and a loop filter. Some parts of the hierarchical structure of the datastream are now optional, so the codec can be configured for a lower datarate or better error recovery. There are now four optional negotiable options included to improve performance: Unrestricted Motion Vectors, Syntax-based arithmetic coding, Advance prediction, and forward and backward frame prediction similar to MPEG called P-B frames.

Source Format: The source encoder operates on non-interlaced pictures with a source format defined in terms of:

- The picture format, as determined by the number of pixels per line, the number of lines per picture, and the pixel aspect ratio
- The timing between the pictures, as determined by the picture clock frequency

In the H.263 video coding standard there are five standardized picture formats, sub-QCIF, QCIF, CIF, 4CIF and 16CIF, which have sizes according to table 2.1. The different formats are based on the CIF format. CIF has 352 pixels per line, 288 lines, a pixel aspect ratio of 12:11, and a picture clock frequency of 30 000/ 1 001, which is approximately 29.97 pictures per second. The time interval between two pictures, or frames, is specified by the frame rate, which is measured in the number of frames per second. The output frame rate from the source encoder is highly dependent on the encoder and the transmission capacity, and is often lower than the information source frame rate. For applications such as video conferencing the frame rate usually falls in the range between 10 and 20 frames per second.

Picture Format	Pixels	Lines
Sub-QCIF	128	96
QCIF	176	144
CIF	352	288
4CIF	704	576
16CIF	1408	1152

Table 2.1.: Picture formats in H.263.

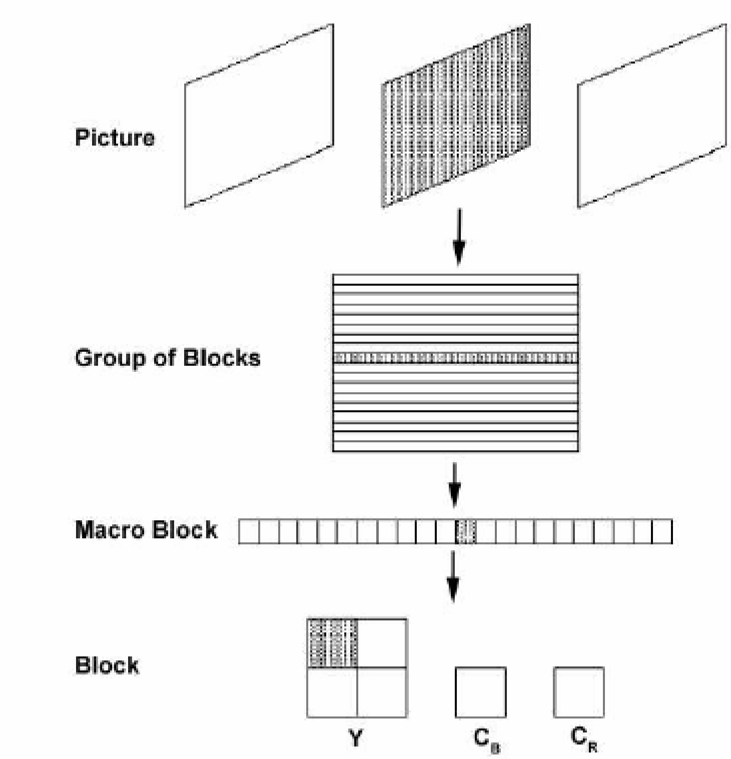


Figure 2.3: Hierarchy of picture segmentation for the CIF picture format.

Picture Segmentation

In order to provide more efficient encoding the pictures are segmented in smaller parts. The division of pictures used in H.263 follows the hierarchy shown in figure 2.3.

- **Picture** The picture constitutes the highest entity in the picture segmentation hierarchy. The picture data consists of a picture header followed by either Group of Blocks (GOBs) or slices. The picture header begins with the Picture Start Code (PSC) and the temporal reference (TR) which makes it possible to keep the encoder and the decoder synchronized on the picture level.
- **Group of Blocks** Each picture is divided in either GOBs or slices. Each GOB contains the picture information for one row in the picture. A GOB is comprised of

16 x k lines, where k depends on the number of lines in the employed picture format. If the CIF format is used each GOB will consist of 16 lines and thus, each picture will have 18 GOBs. The GOBs are numbered from the top to the bottom row, beginning with zero for the uppermost GOB. The GOB data consists of a GOB header followed by a group of macro blocks. The GOB header begins with the GOB start code (GBSC) and the Group Number (GN). These fields provide synchronisation for each GOB and contain information of its position in the picture.

- **Slices** are similar to GOBs in that they consist of a group of macro blocks. The difference is that slices can have more flexible shape, since slices do not have to begin and end at specific positions in the picture. The slice data is almost equal to the GOB data.
- **Macro Block** (MB) has the size of 16 pixels by 16 lines for the luminance component and 8 pixels by 8 lines for the chrominance components. If the CIF picture format is used, each GOB will consist of 22 macro blocks.
- **Block** If the macro blocks were to be transformed with the DCT algorithm described below, the calculations would be very cumbersome. Therefore the MBs are divided into blocks with the size of 8 pixels by 8 lines. A MB thus consists of four luminance blocks and two corresponding chrominance blocks.

Interactive and Streaming Wireless Profile (Profile 3)

The Version 2 Interactive and Streaming Wireless Profile, designated as Profile 3, is defined to provide enhanced coding efficiency performance and enhanced error resilience for delivery to wireless devices within the feature set available in the second version of [19] (which did not include Annexes U, V, and W). This profile of support is composed of the baseline design plus the following modes:

- **Advanced INTRA Coding** Use of this mode improves the coding efficiency for INTRA macroblocks (whether within INTRA pictures or predictively-coded pictures). The additional computational requirements of this mode are minimal at both the encoder and decoder (as low as a maximum of 8 additions/subtractions per 8 x 8 block in the decoding process plus the use of a different but very similar VLC table in order to obtain a significant improvement in coding efficiency). For these reasons, Advanced INTRA Coding is included in this basic package of support.

- **Deblocking Filter** Because of the significant subjective quality improvement that may be realized with a deblocking filter, these filters are widely in use as a method of post-processing in video communication terminals. Annex J of [12] represents the preferred mode of operation for a deblocking filter because it places the filter within the coding loop. This placement eases the implementation of the filter (by reducing the required memory) and somewhat improves the coding performance over a post-processing implementation. As with the Advanced Prediction mode, this mode also includes the four-motion-vectorper- macroblock feature and picture boundary extrapolation for motion compensation, both of which can further improve coding efficiency. The computational requirements of the deblocking filter are several hundred operations per coded macroblock, but memory accesses and computational dependencies are uncomplicated. This last point is what makes the Deblocking Filter preferable to Advanced Prediction for some implementations. Also, the benefits of Advanced Prediction are not as substantial when the Deblocking Filter is used as well. Thus, the Deblocking Filter is included in this basic package of support.
- **Slice Structured Mode** The Slice Structured mode is included here due to its enhanced ability to provide resynchronization points within the video bitstream for recovery from erroneous or lost data. Support for the Arbitrary Slice Ordering (ASO) and Rectangular Slice (RS) submodes of the Slice Structured mode are not included in this profile, in order to limit the complexity requirements of the decoder. The additional computational burden imposed by the Slice Structured mode is minimal, limited primarily to bitstream generation and parsing.
- **Modified Quantization:** This mode includes an extended DCT coefficient range, modified DQUANT syntax, and a modified step size for chrominance. The first two features allow for more flexibility at the encoder and may actually decrease the encoder’s computational load (by eliminating the need re-encode macroblocks when coefficient level saturation occurs). The third feature noticeably improves chrominance fidelity, typically with little added bit-rate cost and with virtually no increase in computation. At the decoder, the only significant added computational burden is the ability to parse some new bitstream symbols.

2.6 Artifacts

We would like to show the differences between compression artifacts and transmission artifacts in this chapter. In our work we focus on transmission artifacts, and their subjective and objective evaluation. As we mentioned in the introduction the reason why we analyzed transmission artifacts is its frequent occurrence in wireless communication.

2.6.1 Compression Artifacts

The compression algorithms used in various video coding standards are very similar. Most of them rely on a block-based DCT with motion compensation and subsequent quantization of the coefficients. In such coding schemes, compression distortions are caused by only one operation, namely the quantization of the transform coefficients. Although other factors affect the visual quality of the stream, such as motion prediction or decoding buffer size, they do not introduce any distortion per se, but affect the encoding process indirectly through the quantization scale factor.

A variety of artifacts can be distinguished in a compressed video sequence [13]:

- **Blocking effect** or blockiness refers to a block pattern in the compressed sequence. It is due to the independent quantization of individual blocks (usually of 8×8 pixels in size) in blockbased DCT coding schemes, leading to discontinuities at the boundaries of a adjacent blocks. The blocking effect is often the most prominent visual distortion in a compressed sequence due to the regularity and extent of the pattern.

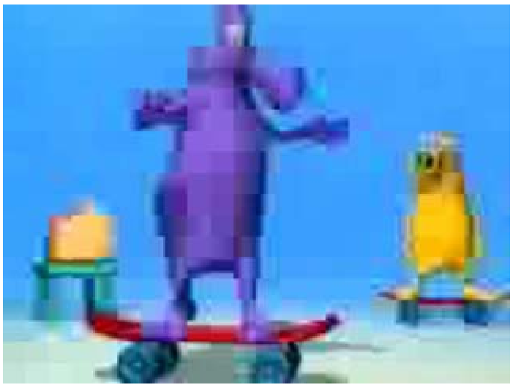


Figure 2.4: Picture from compressed sequence with blocking effect.

- **Blurring** manifests itself as a loss of spatial detail and a reduction of edge sharpness. It is due to the suppression of the high-frequency coefficients by coarse quantization.

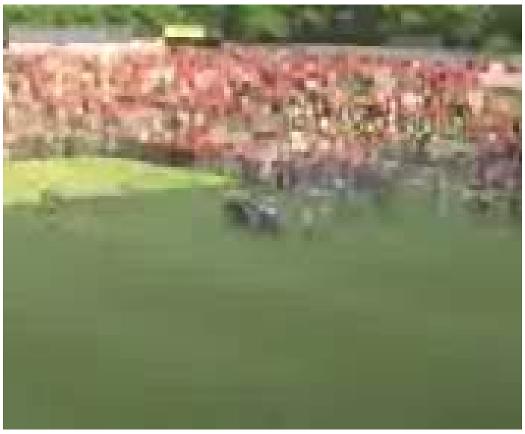


Figure 2.5: Picture from compressed sequence with blurring effect.

- **Color bleeding** is the smearing of the color between areas of strongly differing chrominance. It results from the suppression of high-frequency coefficients of the chroma components. Due to chroma subsampling, color bleeding extends over an entire macroblock.
- **DCT basis image effect** is prominent when a single DCT coefficient is dominant in a block. At coarse quantization levels, this results in an emphasis of the dominant basis image and the reduction of all other basis images.
- **Staircase effect** Slanted lines of ten exhibit the staircase effect. It is due to the fact that DCT basis images are best suited to the representation of horizon tal and vertical lines, whereas lines with other orientations require higher-frequency DCT coefficients for accurate reconstruction. The typically strong quantization of these coefficients causes slanted lines to appear jagged.
- **Ringing** is fundamentally associated with Gibbs' phenomenon and is thus most evident along high-contrast edges in otherwise smooth areas. It is a direct result of quantization leading to high-frequency irregularities in the reconstruction. Ringing occurs with both luminance and chroma components.
- **False edges** are a consequence of the transfer of block-boundary discontinuities due to the blocking effect from reference frames into the predicted frame by motion compensation.
- **Jagged motion** can be due to poor performance of the motion estimation. Block-based motion estimation works best when the movement of all pixels in a macro block is identical. When the residual error of motion prediction is large, it is coarsely quantized.

- **Chrominance mismatch** Motion estimation is often conducted with the luminance component only, yet the same motion vector is used for the chroma components. This can result in *chrominance mismatch* for a macro block.
- **Mosquito noise** is a temporal artifact seen mainly in smoothly textured regions as luminance/ chrominance fluctuations around high-contrast edges or moving objects. It is a consequence of the varied coding of the same area of a scene in consecutive frames of a sequence.
- **Flickering** appears when a scene has high texture content. Texture blocks are compressed with varying quantization factors over time, which results in a visible flickering effect.
- **Aliasing** can be noticed when the content of the scene is above the Nyquist rate, either spatially or temporally.

While some of these effects are unique to block-based coding schemes, many of them are observed with other compression algorithms as well. In wavelet-based compression, for example, the transform is applied to the entire image, therefore none of the block-related artifacts occur. Instead, blurring and ringing are the most prominent distortions.

Video streaming uses a family of transport protocols, namely, User Datagram Protocol (UDP), Real-time Transport Protocol (RTP) and Real-Time Control Protocol (RTCP). UDP is a lower-layer transport protocol while RTP and RTCP are upper layer transport protocols. UDP is employed because it provides timely delivery of packets. However, UDP does not guarantee packet delivery.

2.6.2 Transmission Errors

A very important and of ten overlooked source of impairments is the transmission of the bit stream over a noisy channel. The physical transport can take place over a wire or wireless, where some transport protocol such as Asynchronous Transfer Mode, Transmission Control Protocol/Internet Protocol or User Datagram Protocol (ATM, TCP/IP or UDP) ensures the transport of the bit stream and digitally compressed video is typically transferred over a packet-switched network. Video streaming uses a family of transport protocols, namely, UDP, Real-time Transport Protocol (RTP) and Real-Time Control Protocol (RTCP). UDP is a lower-layer transport protocol while RTP and RTCP are upperlayer transport protocols. UDP is employed because it provides timely delivery of packets. However, UDP does not guarantee packet delivery.

The bit stream is transported in packets whose headers contain sequencing and timing information. This process is illustrated in Figure 2.6. Streams can carry additional signaling information at the session level. A variety of protocols are used to transport the audio-visual information, synchronize the actual media and add timing information [14]. Most applications require the streaming of video, i.e. it must be possible to decode and display the bit stream in real time.

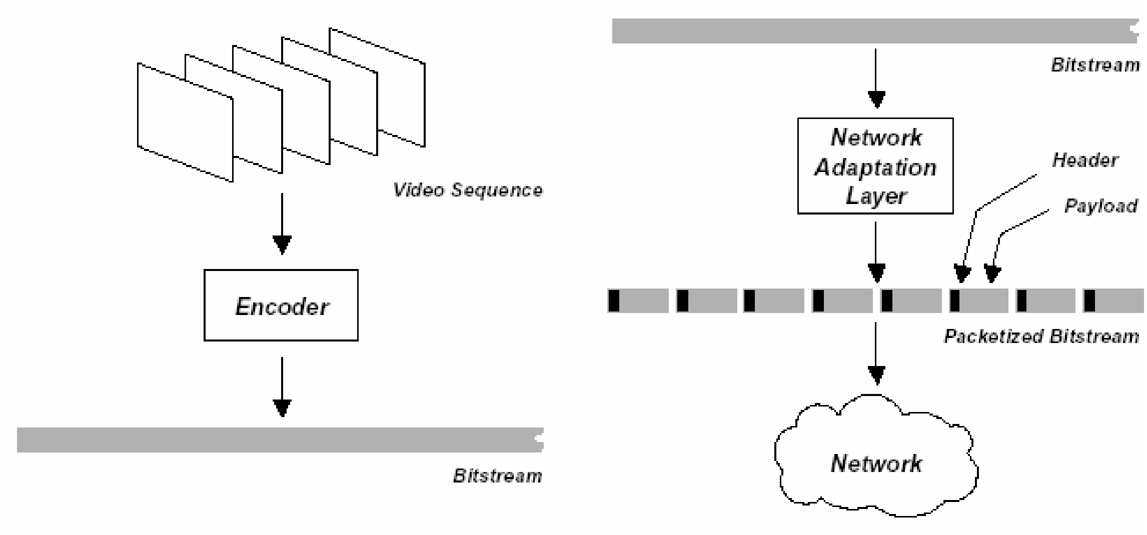


Figure 2.6: Illustration of a video transmission system. The video sequence is first compressed by the encoder. The resulting bit stream is packetized in the network adaptation layer, where a header containing sequencing and synchronization data is added to each packet. The packets are then sent over the network of choice.

Two different types of impairments can occur when transporting media over noisy channels. Packets can be lost, or they can be delayed to the point where they are not received in time for decoding. In [28] assume that no packet losses, errors or congestion occur on either the Internet or the UMTS core network. The transfer delay introduced by the Internet and the UMTS core network is constant throughout the entire video streaming duration. Hence, the quality of the streaming video is solely attributed to the UMTS radio interface [28]. To the application, both have the same effect: part of the media stream is not available, thus packets are missing when they are needed for decoding.

Such losses can affect both the semantics and the syntax of the media stream. When the losses affect syntactic information, not only the data relevant to the lost block are corrupted, but also any data that depend on this syntactic information. For example, an MPEG macroblock that is damaged through the loss of packets corrupts all following macroblocks until an end of slice is encountered, where the decoder can resynchronize. This spatial loss propagation is due to the fact that the DCT coefficient of a macroblock is differentially predicted between macroblocks and resets at the beginning of a slice. Furthermore, for each of these corrupted macroblocks, all blocks that are predicted from these by motion estimation will be damaged as well, which is referred to as temporal loss propagation. Hence the loss of a single macroblock can affect the stream up to the next intra-coded frame. These loss propagation phenomena are illustrated in Figure 2.7.

The effect can be even more damaging when global data are corrupted. An example of this is the timing information in an MPEG stream. The system layer specification of MPEG imposes that the decoder clock be synchronized with the encoder clock via periodic refresh of the program clock reference sent in some packet. Too much jitter (end-to-end delay) on packet arrival can corrupt the synchronization of the decoder clock, which can result in highly noticeable impairments. The visual effects of such losses vary a lot among decoders depending on their ability to deal with corrupted streams. Some decoders never recover from certain errors, while others apply clever concealment techniques such as early synchronization and spatial or temporal interpolation in order to minimize these effects [14].

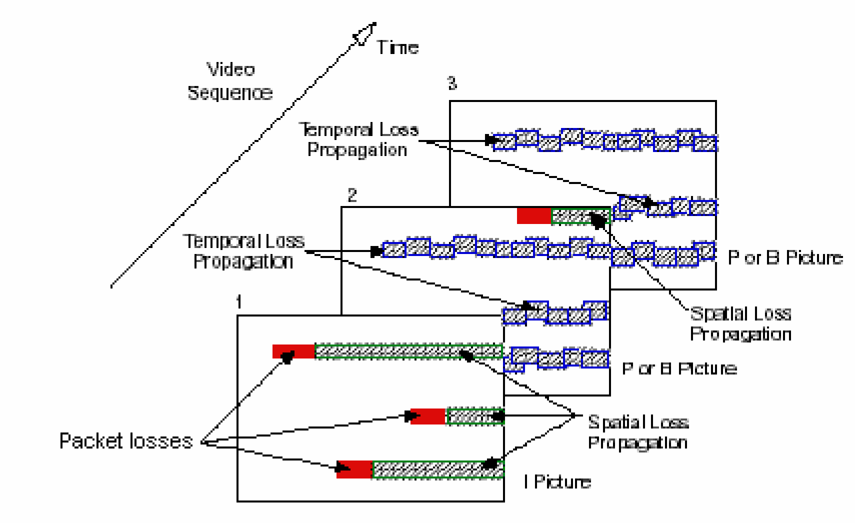
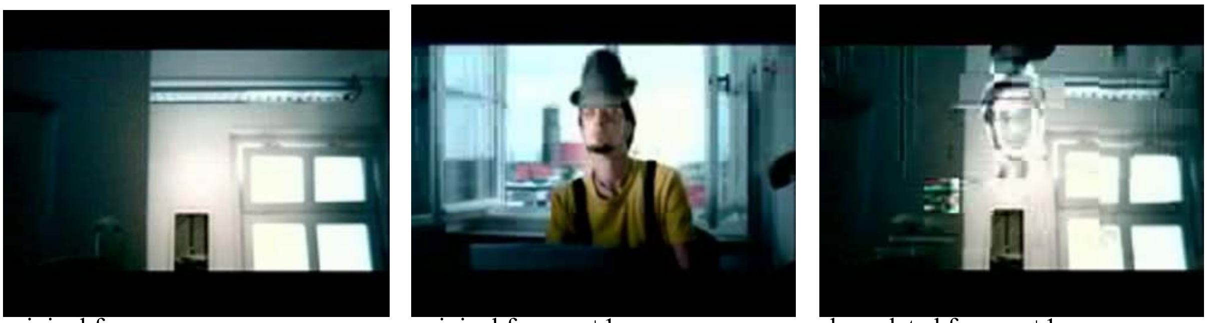


Figure 2.7: Spatial and temporal propagation of losses in an MPEG-compressed video sequence. The loss of a single macroblock causes the inability to decode the data up to the end of the slice. Macroblocks in neighboring frames that are predicted from the damaged area are corrupted as well.

A variety of artifacts can be distinguished in a transmitted video sequence:

- **Aliasing** occurs if there is a packet loss during the rapid scene change (especially cut) and results in aliasing of two frames from different scenes. However, if there is only a slow local movement in the new scene, the user does not necessarily note that there should have been a scene change. This effect we call invisible aliasing (see Figure 2.8).



original frame n   original frame n+1   degradated frame n+1

Figure 2.8: On the first two pictures are following frames from original video, and on last picture is following frame, but with lost I frame. These pictures are from test sequences coded in H.263 standard.

- **Spread** is temporal propagation of error. P frame is lost, and moving macroblocks remain from previous frames. Seen in figure 2.9.

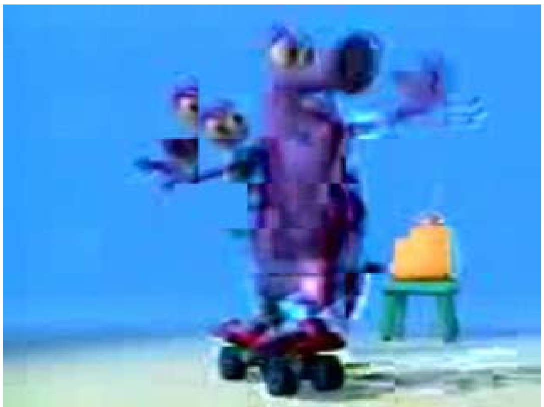


Figure 2.9: Frame with spreading artifact. Cartoon figure has eyes from previous frame

- **Local distortions**: smaller artifacts of single MB or several of them. See in figure 2.10.

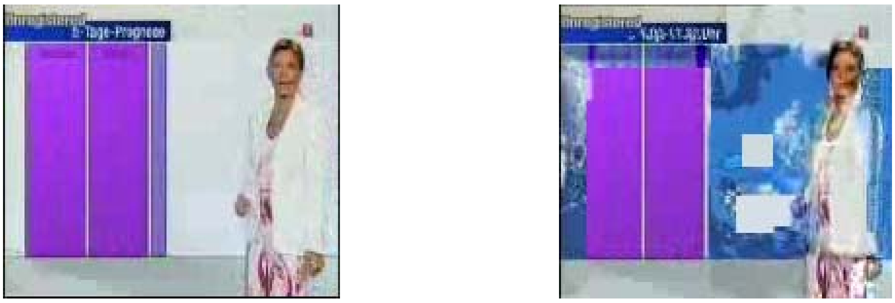


Figure2.10: Local distortions

2.6.3 Other Impairments

Aside from compression artifacts and transmission errors, the fidelity of digital video sequences can be affected by any pre- or post-processing stage in the system. These include:

- conversions between the digital and the analog domain;
- chroma subsampling, which was discussed in Section 2.2;
- frame rate and spatial reduction ;
- de-interlacing, i.e. the process of creating a progressive sequence from an interlaced one [16].

.

3 Perceptual visual quality metrics

This chapter reviews the history and the state of the art of visual quality metrics, from simple pixel-based metrics such as MSE and PSNR to the advanced vision-based metrics proposed during recent years. Most of the metrics were design for compression artifacts evaluation.

3.1 Pixel-Based Metrics

The mean squared error (MSE) and the peak signal-to-noise ratio (PSNR) are the most popular difference metrics in image and video processing [1]. The MSE is the mean of the squared differences between the gray-level values of pixels in two pictures or sequences *F* and R.

$$MSE[n] = \frac{1}{C \cdot X \cdot Y} \sum_{c=1}^{C} \sum_{x=1}^{X} \sum_{y=1}^{Y} [F_n^{(c)}(x,y) - R_n^{(c)}(x,y)]^2 \qquad (3.1)$$

where: X · Y is picture's size

C is number of color components

The PSNR in decibels is defined as:

$$PSNR = 10 \log_{10} \frac{m^2}{MSE} \qquad (3.2)$$

where *m* is the maximum value that a pixel can take (e.g. 255 for 8-bit images). Note that MSE and PSNR are well-defined only for luminance information; once color comes into play, there is no agreement on the computation of these measures.

Technically, MSE measures image difference, whereas PSNR measures image fidelity, i.e. how closely an image resembles a reference image, usually the uncorrupted original. The popularity of these two metrics is rooted in the fact that minimizing the MSE is equivalent to maximum likelihood estimation for independent measurement errors with normal distribution. Besides, computing MSE and PSNR is very easy and fast. Because they are based on a pixel-by-pixel comparison of images, however, they only have a limited, approximate relationship with the distortion or quality perceived by the human visual system. In certain situations the subjective image quality can be improved by adding noise and thereby reducing the PSNR. Dithering of color images with reduced color depth, which adds noise to the image to remove the perceived banding caused by the color quantization, is a common example of this. Furthermore, the visibility of distortions depends to a great extent on the image background, a property known as masking. Distortions are often much more disturbing in relatively smooth areas of an image than in texture regions with a lot of activity, an effect not taken into account by pixel-based metrics. Therefore the perceived quality of images with the same PSNR can actually be very different. A number of additional pixel-based metrics are discussed in [17]. They found that although some of these metrics can predict subjective ratings quite successfully for a given compression technique or type of distortion, they are not reliable for evaluations across techniques. Another study [18] concluded that even perceptual weighting of MSE does not give consistently reliable predictions of visual quality for different pictures and scenes. But we did not find evaluation of pixel based metric for low resolutions, low bitrate video with transmission errors.

In our work we will try present, that with modified pixel based metric we can obtain fairly good results for estimating the distortions caused by transmission errors.

3.2 Single-Channel Models

The first models of human vision adopted a single-channel approach. Single-channel models regard the human visual system as a single spatial filter, whose characteristics are defined by the contrast sensitivity function. The output of such a system is the filtered version of the input stimulus, and detect ability depends on a threshold criterion.

In [19] is the first computational model of vision was designed to predict pattern sensitivity for foveal vision. It is based on the assumption that the cortical representation is a shift invariant transformation of the retinal image and can thus be expressed as a convolution. In order to determine the convolution kernel of this transformation, Schade [19] carried out psychophysical experiments to measure the sensitivity to harmonic contrast patterns. From this CSF the convolution kernel for the model can be computed, which is an estimate of the psychophysical line spread function. Schade’s [19] model was able to predict the visibility of simple stimuli but failed as the complexity of the patterns increased. In [20] the first image quality metric for luminance images was developed. They realized that simple pixel-based distortion measures were not able to accurately predict the quality difference s perceived by observers. On the basis of psychophysical experiments on the visibility of gratings, they inferred some properties of the human visual system and came up with a closed-form expression for the contrast sensitivity as a function of spatial frequency, which is still widely used in HVS-models. The input images are filtered with this CSF after lightness nonlinearity. The squared difference between the filter outputs for the two images is the distortion measure. It was shown to correlate quite well with subjective ranking data. Despite its simplicity, this metric was one of the first works in engineering to recognize the importance of applying vision science to image processing.

In [22] is the first video quality metric was developed. It is based on a spatio-temporal model of the contrast sensitivity function using an excitatory and an inhibitory path. The two paths are combined in a nonlinear way, enabling the model to adapt to changes in the level of background luminance. Masking is also incorporated in the model by means of a weighting function derived from the spatial and temporal activity in the reference sequence. In the final stage of the metric, an *Lp*-norm of the masked error signal is computed over blocks in the frame whose size is chosen such that each block covers the size of the foveal field of vision. The resulting distortion measure was shown to outperform MSE as a predictor of perceived quality.

Single-channel models and metrics are still in use today because of their relative simplicity and computational efficiency, and a variety of extensions and improvements have been proposed. However, they are intrinsically limited in prediction accuracy. They are unable to cope with more complex patterns and cannot account for empirical data from masking and pattern adaptation experiments.

These data can be explained quite successfully by a multi-channel theory of vision, which assumes a whole set of different channels instead of just one. The corresponding multi-channel models and metrics are discussed in the next section.

3.3 Multi-Channel Models

Multi-channel models assume that each band of spatial frequencies is dealt with by an independent channel. The CSF is just the envelope of the sensitivities of these channels. Detection occurs independently in any channel when the signal in that band reaches a threshold criterion.

Watson [32] also was the first to outline the architecture of a multi-channel vision model for video coding. The model partitions the input into achromatic and chromatic opponent-color channels, into static and motion channels, and further into channels of particular frequencies and orientations. Bits are then allocated to each band taking into account human visual sensitivity to that band as well as visual masking effects. In contrast to the spatial model for images, it has never been implemented and tested, however.

In [30] is proposed a number of video quality metrics based on multichannel vision models. The Moving Picture Quality Metric (MPQM) is based on a local contrast definition and filters for the spatial decomposition, two temporal mechanisms, as well as a spatio-temporal contrast sensitivity function and a simple intra-channel model of contrast masking. A color version of the MPQM based on an opponent color space was presented as well as a variety of applications and extensions of the MPQM [30]. Due to the MPQM's purely frequency-domain implementation of the spatio-temporal filtering process and the resulting huge memory requirements, it is not practical for measuring the quality of sequences with duration of more than a few seconds, however. The Normalization Video Fidelity Metric (NVFM) by [30] avoids this shortcoming by using a steerable pyramid transform for spatial filtering and discrete time-domain filter approximations of the temporal mechanisms. It is a spatio-temporal extension image distortion metric and implements inter-channel masking through an early model of contrast gain control. Both the MPQM and the NVFM are of particular relevance here because their implementations are used as the basis for the metrics developed in the framework of this thesis.

3.4 Specialized Metrics

Metrics based on multi-channel vision models such as the ones presented in Section 3.3 are the most general and potentially the most accurate ones is in [1]. However, quality metrics need not necessarily rely on sophisticated general models of the human visual system; they can exploit a priori knowledge about the compression algorithm and the pertinent types of artifacts (cf. Section 3.3) using ad-hoc techniques or specialized vision models. While such metrics are not as versatile, they normally perform well in a given application area. Their main advantage lies in the fact that they of ten permit a computationally more efficient implementation.

Watson [30] recently extended the DCTune metric to video. In addition to the spatial sensitivity and masking effects considered in DCTune, this so-called Digital Video Quality (DVQ) metric relies on measurements of the visibility thresholds for temporally varying DCT quantization noise. It also models temporal forward masking effects by means of a masking sequence, which is produced by passing the reference through a temporal low-pass filter.

In [2] is developed a video quality assessment system that is based on a combination of three low-level features. These features were selected empirically from a number of candidates so as to yield the best correlation with subjective data for a given test set. The features compute the spatial and temporal activity using spatial gradients and frame difference s and are combined linearly to a measure of video quality.

In [25] is developed another video distortion metric that uses reduced reference information in the form of low-level features extracted from spatio-temporal blocks of the sequences. These features were selected empirically from a number of candidates so as to yield the best correlation with subjective data. First, horizontal and vertical edge enhancement filter's are applied to facilitate gradient computation in the feature extraction stage. The resulting sequences are divided into spatio-temporal blocks. A number of features measuring the amount and orientation of activity in each of these blocks are then computed from the spatial luminance gradient. To measure the distortion, the features from the reference and the distorted sequence are compared using a process similar to masking.

In [29] is presented a measurement tool for MPEG video quality. It first computes the perceptual impairment in each frame based on contrast sensitivity and masking with the help of spatial filtering and Sobel-operators, respectively. Then the PSNR of the masked error signal is calculated and normalized. The interesting part of this metric is its second stage, a cognitive emulator, which simulates higher-level aspects of perception. This includes the delay and temporal smoothing effect of observer responses, the nonlinear saturation of perceived quality, and the asymmetric behavior with respect to quality changes from bad to good and vice versa. This metric is one of the few models targeted at measuring the temporally varying quality of video sequences. While it still requires the reference as input, the cognitive emulator was shown to improve the predictions of subjects' SSCQE ratings.

3.5 Model Based on HVS Characteristics

The structure of the vision model [5] is shown in Figure 3.1. After conversion of the input to opponent-colors space, each of the resulting three components is subjected to a spatio-temporal perceptual decomposition, yielding a number of perceptual channels. They are weighted according to contrast sensitivity data and subsequently undergo a contrast gain control stage. Both the reference sequence and the processed sequence are used as input to the model and go through identical stages. Finally, all the sensor differences are combined into a distortion measure. Each of these stages is explained in more detail below.

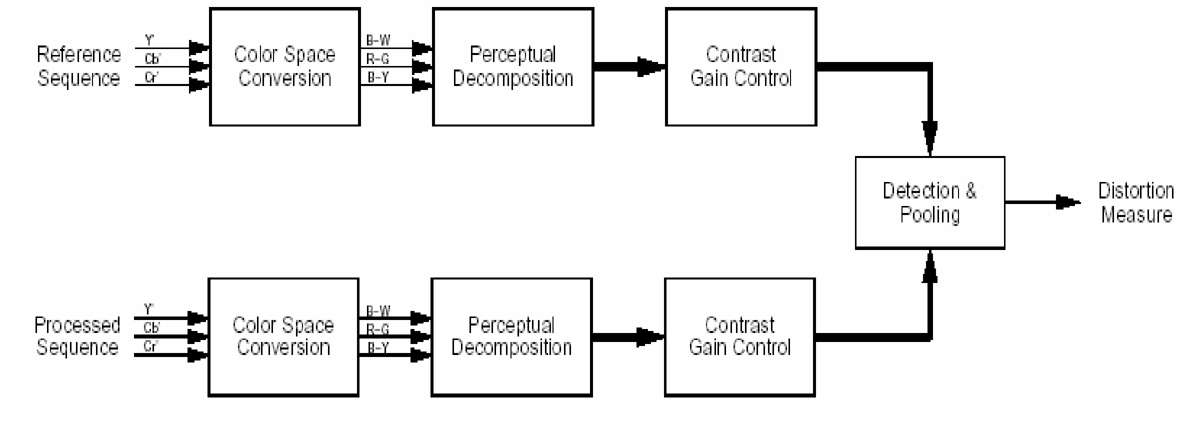


Figure 3.1: Block diagram of the perceptual distortion metric. After conversion to opponent-colors space, each of the resulting three components is subjected to a perceptual decomposition, yielding several perceptual channels. Subsequently they undergo weighting and contrast gain control, after which all the sensor differences are combined into a distortion measure.

Color Space Conversion: Y'C'$_B$C'$_R$ is defined in ITU-R Recommendation 601.13 Y' encodes luminance, C'$_B$ the difference between the blue primary and luminance, and C'$_R$ the difference between the red primary and luminance. Conversion from Y'C'$_B$C'$_R$ to opponent-colors space requires a number of transformations as illustrated in Figure 3.2.

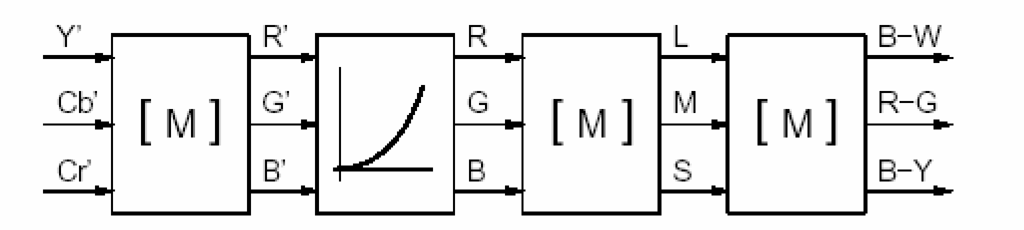


Figure3.2: Color space conversion from component video Y'C'$_B$C'$_R$ to opponent-colors space.

Perceptual Decomposition: The perceptual decomposition is performed first in the temporal and then in the spatial domain. While this separation is not entirely unproblematic, these two domains can be consolidated in the fitting process as described below.

The temporal filters used in this distortion metric are based on the temporal mechanisms with derivatives of the following impulse response function:

$$h(t) = e^{-\left(\frac{\ln(t/\tau)}{\sigma}\right)^2} \qquad (3.3)$$

The temporal mechanisms are modeled by two Infinite Impulse Response (IIR) filters. The low-pass filters are applied to all three color channels, but the band-pass filter is applied only to the luminance channel in order to reduce computing time. This simplification is based on the fact that color contrast sensitivity is rather low for higher frequencies.

The spatial Mechanisms- the decomposition in the spatial domain is carried out by means of the steerable pyramid transform. This transform decomposes an image into a number of spatial frequency and orientation bands; its basis functions are directional derivative operators. For use within a vision model, it has the advantage of being rotation-invariant and self-inverting, and it minimizes the amount of aliasing in the subbands.

Contrast Gain Control Stage: contrast gain control model can be extended to color and to sequences as follows: Let *a* = *a*(*t, c, f, θ, x, y*) be a coefficient of the perceptual decomposition in temporal channel *t*, color channel *c*, frequency band *f*, orientation band *θ*, at location *x, y*. Then the corresponding sensor output *s* = *s*(*t, c, f, θ, x, y*) is computed as

$$s = k\frac{a^P}{b^2 + a^q * h} \qquad (3.4)$$

The excitatory path in the numerator consists of power-law nonlinearity with exponent *p*. The inhibitory path in the denominator controls the gain of the excitatory path. In addition to a nonlinearity with a possibly different exponent *q*, filter responses are pooled over different channels by means of a convolution with the pooling function *h* = *h*(*t, c, f, θ, x, y*). In its most general form, this pooling operation may combine coefficients from the dimensions of time, color, temporal frequency, spatial frequency, orientation, space, and phase.

Detection and Pooling: The pooling stage combines the elementary differences between the sensor outputs **s** = **s**(*t, c, f, θ, x, y*) for the reference (**s**0) and the processed sequence (**s**1) over several dimensions:

$$\Delta s = \sqrt[\beta]{\sum |s_0 - s_1|^\beta}. \qquad (3.5)$$

In principle, any subset of dimensions can be used for this summation, depending on what kind of result is desired.

In general, the development of computational HVS-models itself is still in its infancy, and many issues remain to be solved. Most importantly, more comparative analysis of different modeling approaches is necessary. Human visual perception is highly adaptive, but also very dependent on certain parameters such as color and intensity of ambient lighting, viewing distance, media resolution, and others. It is not possible to design HVS-models that try to meticulously incorporate all of these parameters.

Our focus, low resolutions, low bitrate video with transmission errors until recently, the issue was not addressed in a sufficient manner, and existing metrics were not applied to deal with the problematic either.

4 Tests of subjective quality

In order to be able to design reliable visual quality metrics, it is necessary to understand what “quality” means to the viewer. One method to determine the viewer’s quality opinion is obtained by MOS. To obtain MOS we performed the subjective perceptual quality tests, the users evaluated the quality changing in time. Thus, we determine viewer’s opinion in any time point during the video sequence. This information will be helpful during the Metric Design.

Obtaining the MOS curves can be subdivided into the two steps: The first step was selection of the appropriate sequences. A viewer’s enjoyment when watching video depends on many factors. One of the most important is of course program content and material. We needed sequences with different parameters and content. The second step was preparation of the testing conditions (laboratory), choose a suitable hardware and software corresponding to the internationally standardized testing methodology. The methodology for performing and evaluating of the time-variant video quality tests is described in Recommendation ITU-R BT.500 [3].

4.1 Test Material

For the subjective video quality tests we selected five video sequences to cover a wide range of typical content for mobile application, such as news or sports. Furthermore, they were selected to contain various temporal and spatial characteristics such as flat areas, object and camera motions, and faces landscapes. Consequently, the scenes span a wide range of coding complexity.

The “News” video sequence contains various different scenes with both, static and moving camera (moderator reading news, scenes illustrating the news content). The scenes are usually separated by scene cuts, in three cases by fast zooming-in/zooming-out scene change. The video sequence ”Football” is a typical soccer match, containing wide-angle panning camera as well as close-up scenes, separated from each other by cuts. The “Cartoon” sequence contains two different short cartoons. Scene separation is performed mostly by transitions, in some cases by cuts or wipes. The “Town” sequence is a picture guide over the town. Different scenes are separated by slow transitions (dissolving).

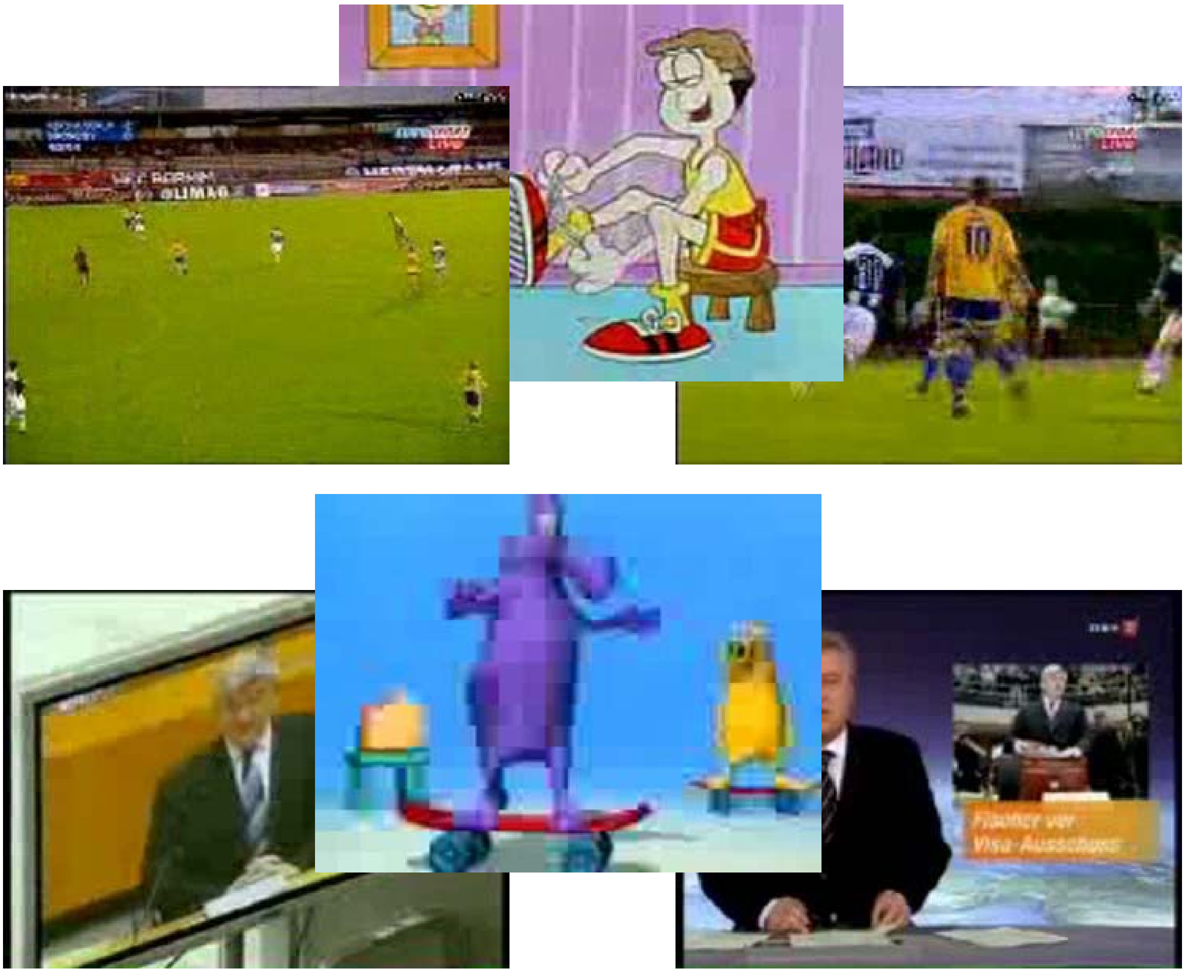


Figure 4.1: Screenshots of some video test sequences used in the scrutinize: “News”, “Football” and “Cartoon”.

All sequences were encoded using H.263 profile 0 and level 10. For subjective quality testing we used combinations of bit rates and frame rate of 7.5 fps as shown in Table 4.1 together with the number of cuts in these sequences and the packet loss (PL). We used PL because it is coincident with wireless network, like 3rd generation mobile telecommunication system (3G). Contrary to BER (bit error rate), the bit errors do not necessarily lead to dropped slices/frames or delays in the decoded video [5]. The PSNR in the Table 4.1 determines the overall degradation of the video sequence; it is averaged over the time, test persons and their runs.

# Sequence	Name	Scenes	Bit rate	Packet loss	PSNR [dB]
1	News	15	56kb/s	8%	22,96
2	Football	7	105kb/s	1%	36,88
3	Town	39	80kb/s	5%	25,14
4	Cartoon	25	44kb/s	10%	23,83

Table 4.1: Test sequences and their parameters.

4.2 Test Setup

Subjective testing for visual quality assessment has been formalized in ITU-R recommendation BT.500-10 [3], which suggests standard viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods. The three most commonly used procedures are the following:

- **Double Stimulus Continuous Quality Scale (DSCQS)** The presentation sequence for a DSCQS trial is illustrated in Figure 4.2(a). Viewers are shown multiple sequence pairs consisting of a “reference” and a “test” sequence, which are rather short (typically 10 seconds). The reference and test sequence are presented twice in alternating fashion, with the order of the two chosen randomly for each trial. Subjects are not informed which is the reference and which is the test sequence. They rate each of the two separately on a continuous quality scale ranging from “bad” to “excellent” as shown in Figure 4.2(b). Analysis is based on the difference in rating for each pair, which is calculated from an equivalent numerical scale from 0 to 100. This differencing removes a lot of the subjectivity with respect to scene content and experience.
- **Double Stimulus Impairment Scale (DSIS).** The presentation sequence for a DSIS trial is illustrated in Figure 4.3(a). As opposed to the DSCQS method, the reference is always shown before the test sequence, and neither is repeated. Subjects rate the amount of impairment in the test sequence on a discrete five-level scale ranging from “very annoying” to “imperceptible” as shown in Figure 4.3(b).
- **Single Stimulus Continuous Quality Evaluation (SSCQE)** The introduction of digital television compression will produce impairments to the picture quality which are scene-dependent and time-varying. Even within short extracts of digitally-coded video, the quality can fluctuate quite widely depending on scene content, and impairments may be very short-lived. Conventional ITU-R methodologies alone are not sufficient to assess this type of material. Furthermore,

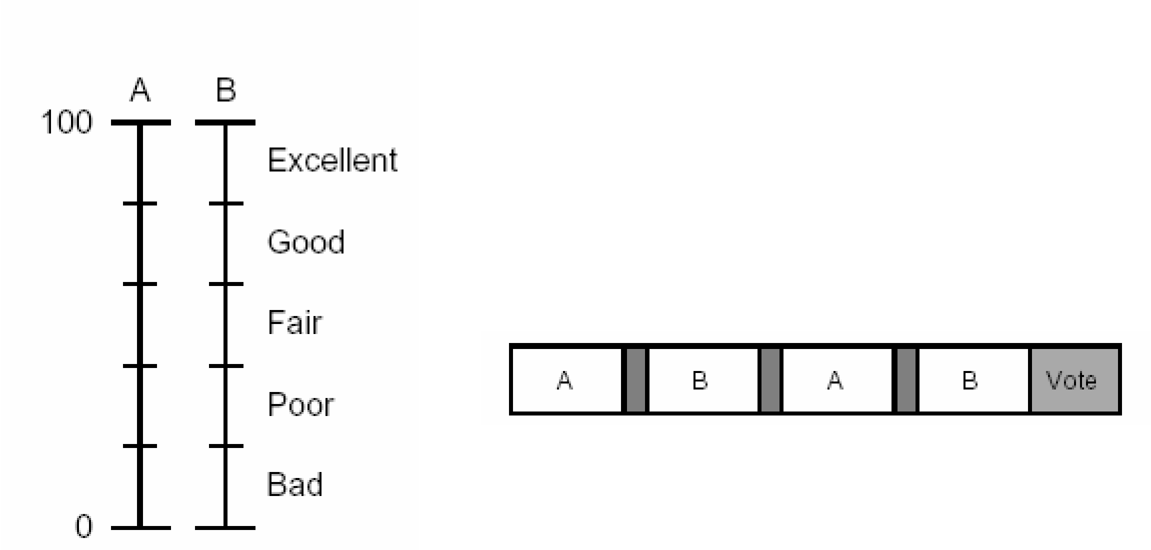
the double stimulus method of laboratory testing does not replicate the SS home viewing conditions.(In SS methods, a single image or sequence of images is presented and the assessor provides an index of the entire presentation.) It was considered useful, therefore, for the subjective quality of digitally coded video to be measured continuously, with subjects viewing the material once, without a source reference.

As a result, the following new SSCQE technique has been developed and tested [3].

Recording device and set-up [3]:

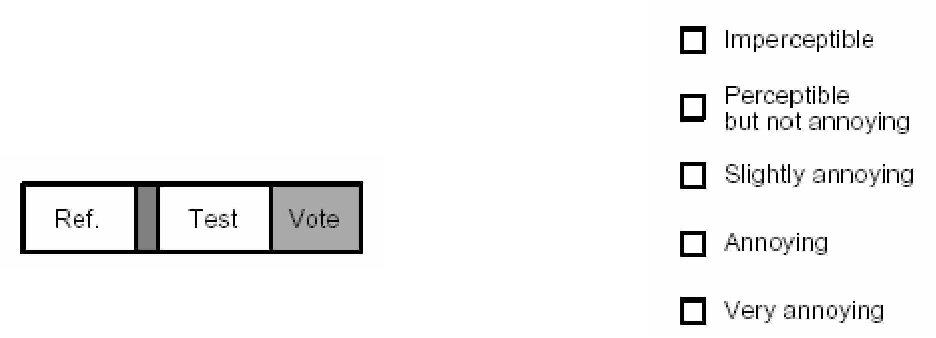
An electronic recording handset connected to a computer used for recording the continuous quality assessment from the subjects. This device has the following characteristics:

- slider mechanism without any sprung position,
- linear range of travel of 10 cm,
- fixed or desk-mounted position,
- samples recorded twice a second.



(b) Rating scale **(a)** Presentation sequence

Figure 4.2: DSCQS method. The reference and the test sequence are presented twice in alternating fashion **(a)**. The order of the two is chosen randomly for each trial, and subjects are not informed which is which. They rate each of the two separately on a continuous quality scale ranging from “bad” to “excellent” **(b)**.



(a) Presentation sequence **(b)** Rating scale

Figure4.3: DSIS method. The reference and the test sequence are shown only once **(a)**. Subjects rate the amount of impairment in the test sequence on a discrete five-level scale ranging from "very annoying" to "imperceptible" **(b)**.

These three methods generally have different applications. DSCQS is the preferred method when the quality of test and reference sequence is similar, because it is quite sensitive to small differences in quality. The DSIS method is better suited for evaluating clearly visible impairments such as artifacts caused by transmission errors. Both DSCQS and DSIS method share a common drawback, however: Changes in scene complexity, statistical multiplexing or transmission errors can produce substantial quality variations that are not evenly distributed over time; severe degradations may appear only once every few minutes. The standard DSCQS and DSIS methods with their single rating are not suited to the evaluation of such long sequences because of the regency phenomenon, a bias in the ratings toward the final 10 - 20 seconds due to limitations of human working memory[1]. Furthermore, it has been argued that the presentation of a reference as well as the repetition of the sequences in the DSCQS method puts the subjects in a situation to removed from the home viewing environment by allowing them to become familiar with the material under investigation. SSCQE has been designed with these problems in mind, as it relates well to the time-varying quality of today's compressed digital video systems. On the other hand, program content tends to have a significant influence on SSCQE scores. Also, SSCQE scores of different tests are harder to compare because of the lack of a reference.

We decided to adopt the SSCQE as a basis for our tests, respecting following modifications:

- In addition to the original SSCQE method recommendations, we portray original (compressed) sequence (without transmission artifacts) to the observers, without letting them know about it.
- Our video sequences duration is only about 1 minute.
- Video sequences with typical content for mobile applications, are played back on mobile terminal to the observers.

Herewith we wanted to obtain only observers reaction to transmission artifacts (see Chapter 2.6.2) and not to the compression artifacts (see Chapter 2.4). This will be more comprehensively explained in Chapter 3.

We did the subjective assessment of the video quality with 38 paid test persons. They are non-experts, in the sense that they are not directly concerned with digital picture quality as a part of their normal work. The observers were all university students (50 % of them from the Technical University and 50 % from the University of Economics). We chose the group to reflect the major part of the video services customers. The age range of all the test persons was from 20 to 30 years. The test persons are evaluating more critically if the video is played-back at the PC (Personal Computer) [7]. Therefore, our test sequences were played-back on the UMTS mobile terminal Sony-Ericsson Z1010. The screen specification of Sony-Ericsson Z1010 is: LCD with 65.536 colors and the display resolution 176x220. The Sony-Ericsson Z1010 can be seen in the Figure 4.6. As a response capturing interface we choose sliding bar the sliding bar tool seems to be the most appropriate for time-variant video quality testing. This is given by its simple design, obvious functionality and comfortable size. The test persons presented clear preferences for the sliding bar tool that in turn also best correspond to the design and usability [8]. The picture of sliding bar can be seen in Figure 4.4. Signal from the interface device was captured by “Aestethic Quality Recorder” software written at the Institute of Communications and Radio-Frequency Engineering, Vienna University of Technology [27]. Screenshot of the program is illustrated in Figure 4.5. Synchronization between mobile phone and software was provided by assessor. The overall testing environment can be seen in Figure 4.7.

The display distance was chosen by the test person according to his/her convenience. All tests were performed in the same laboratory. The interval between each two samples was 150 ms, the quality scale was between 0 (lowest quality) and 255 (highest quality). With each test person we performed three runs: the second run took place one hour after the first run, the third run followed two weeks later.

Results from the subjective assessment are presented in the next chapter.

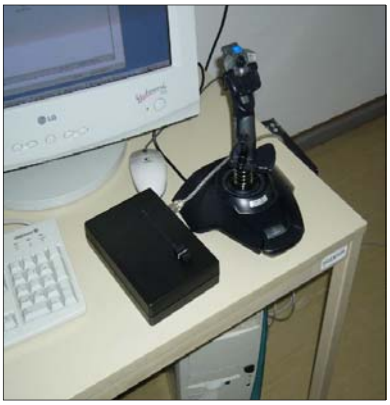
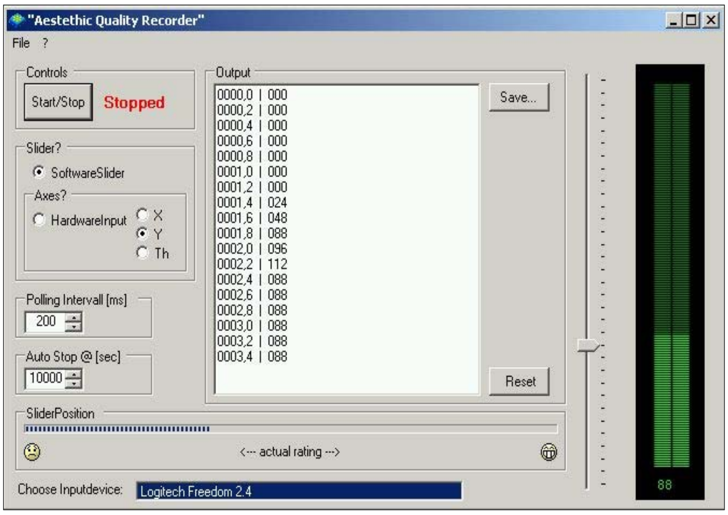


Figure 4.4: User interfaces to capture the evaluation: joystick, sliding bar



Stopped

● SoftwareSlider

○ HardwareInput ○ X ● Y ○ Th

Polling Intervall [ms] 200

Auto Stop @ [sec] 10000

```
0000,0 | 000
0000,2 | 000
0000,4 | 000
0000,6 | 000
0000,8 | 000
0001,0 | 000
0001,2 | 000
0001,4 | 024
0001,6 | 048
0001,8 | 088
0002,0 | 096
0002,2 | 112
0002,4 | 088
0002,6 | 088
0002,8 | 088
0003,0 | 088
0003,2 | 088
0003,4 | 088
```

Choose Inputdevice: Logitech Freedom 2.4

Figure 4.5: Screenshot of a capture program controlled by a standard mouse or any of the user interface devices software

Figure 4.6: Sony-Ericsson Z1010

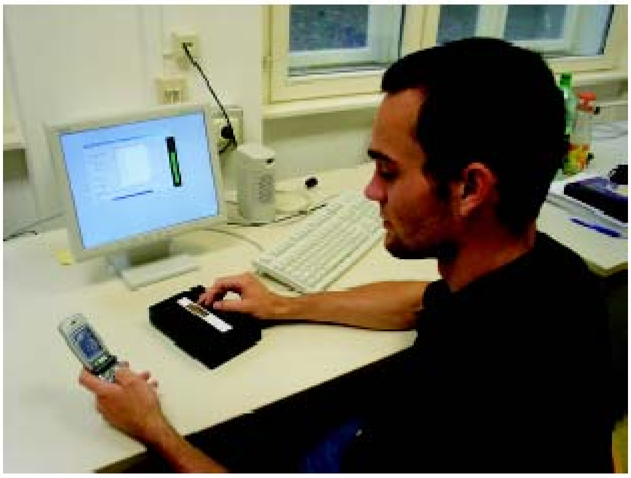


Figure 4.7: Testing environment: the test person holding the mobile phone, evaluating with the slider.

5 Tests evaluation

In [3] data processing from subjective assessment of the quality and data presentation is described as well. We had this on mind, but we also needed to adapt the data evaluation method on our modified testing method as will be shown in the following.

A test consisted of $L = 114$ presentations. Each presentation had two test conditions ($J=2$) applied to one of a six test sequences ($K=6$). Each combination of test sequence was repeated three times ($R=3$).

The first step of the analysis of the results is the calculation of the mean score $\bar{u}_{\mathrm{jkr}}$ for each of the presentations:

$$\bar{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \qquad (5.1)$$

where:

u_{ijkr} : score of observer i for test condition j, sequence/image k, repetition r

N: number of observers, in our case $N=38$.

In the Figure (5.1) mean score for sequence “News” with PL (curve MOS_deg) and also without PL (curve MOS_ori) is shown

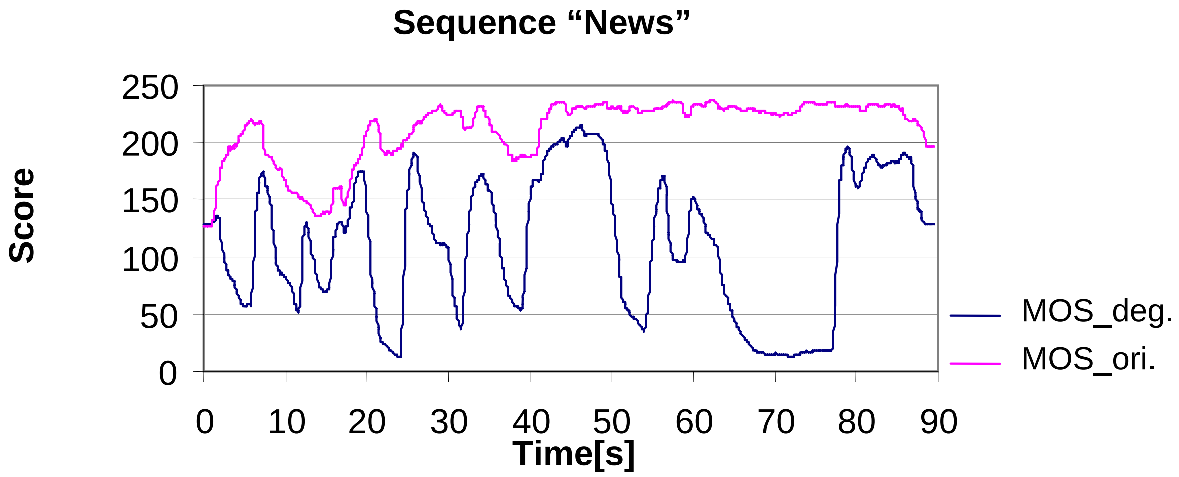


Figure 5.1: Average value for sequence “News“

By comparing single runs of the same observer or observers between them we found, that there were big differences between them. We calculated standard deviation to find observers who have produced votes significantly distant from the average score.

The standard deviation for each presentation, S_{jkr}, is given by:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\bar{u}_{jkr} - u_{ijkr})^2}{(N-1)}} \qquad (5.2)$$

In Figure (5.2) mean score of degraded sequence „ Cartoon “ and standard deviation of all observers are presented. Maximum value of 180 is 101.38.

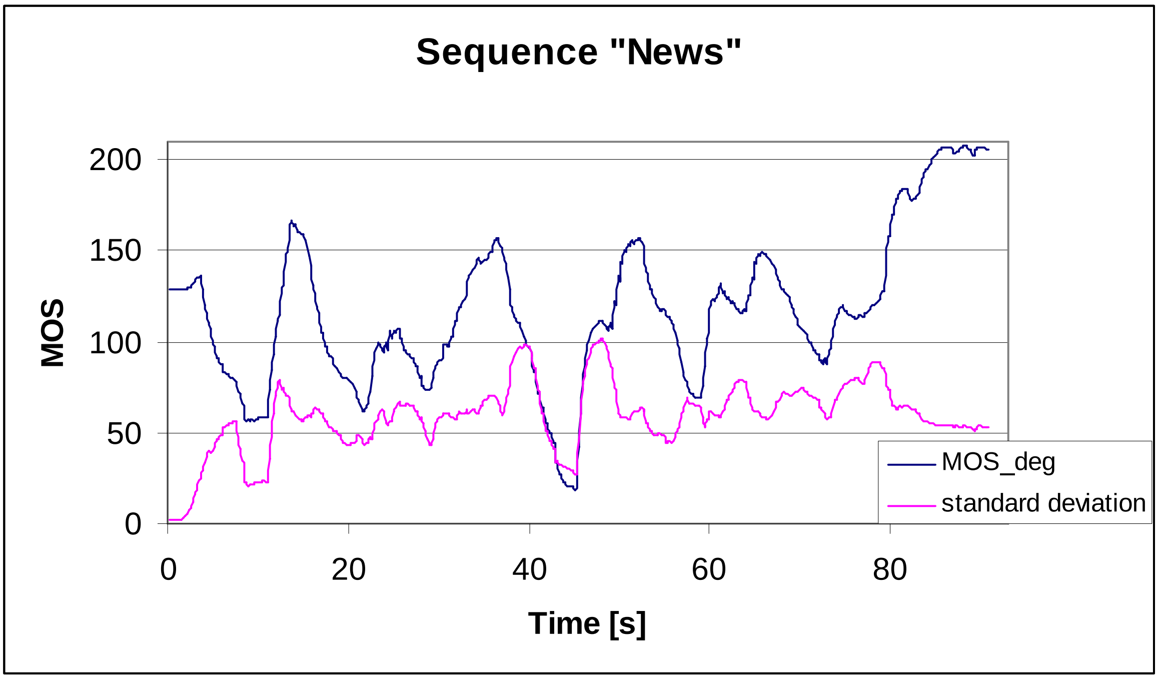


Figure 5.2: Average value and standard deviation of MOS for degraded sequence “News”

For verification the consistency of our data we calculated 95 % confidence interval for all sequences. We performed 95 % screening as recommended in [3] and we used the next algorithm.

It must be first ascertained whether this distribution of scores for each time window of each test configuration is “normal” or not, using the β_2 test. If β_2 is between 2 and 4, the distribution may be considered as “normal”. Then the process applies for each time window of each test configuration as mathematically expressed hereafter. For each time window of each test configuration and using the votes u_{ijkr} of each observer, the mean *jklr* *u* standard deviation, S_{jklr}, and the coefficient, β_{2jklr}, are calculated. β_{2jklr} is given by:

$$\beta_{2jklr} = \frac{m_4}{(m_2)^2} \qquad \text{with} \qquad m_x = \frac{\sum_{n=1}^{N} (u_{njklr} - \bar{u})^x}{N} \qquad (5.3)$$

For each observer, *i*, find *Pi* and *Qi*, i.e.:

```
for j, k, l, r = 1, 1, 1, 1 to J, K, L, R

        if $2 \le \beta_{2jklr} \le 4$, then:
                if $u_{ijkr} \ge \bar{u}_{jklr} + 2S_{jklr}$        then $P_i = P_i + 1$
                if $u_{ijkr} \le \bar{u}_{jklr} + 2S_{jklr}$        then $Q_i = Q_i + 1$
        else:
                if $u_{ijkr} \ge \bar{u}_{jklr} + \sqrt{20}S_{jklr}$        then $P_i = P_i + 1$
                if $u_{ijkr} \le \bar{u}_{jklr} + \sqrt{20}S_{jklr}$        then $Q_i = Q_i + 1$
        if $\frac{P_i}{J \cdot K \cdot L \cdot R} > X\%$ or $\frac{Q_i}{J \cdot K \cdot L \cdot R} > X\%$    then reject observer i
```

with: *N* : number of observers

J : number of time windows within a test combination of test condition and sequence

K : number of test conditions

L : number of sequences

R : number of repetitions.

This process allows discarding observers who have produced votes significantly distant from the average scores.

We find out there was no observer to be discarded. As a quality indicator of the subjective data, the distribution of the 95 % confidence intervals is shown in Figure 5.3. Where $\delta_{jkr} = 1.96\frac{S_{jkr}}{\sqrt{N}}$ for MOS_deg test sequence “News”.

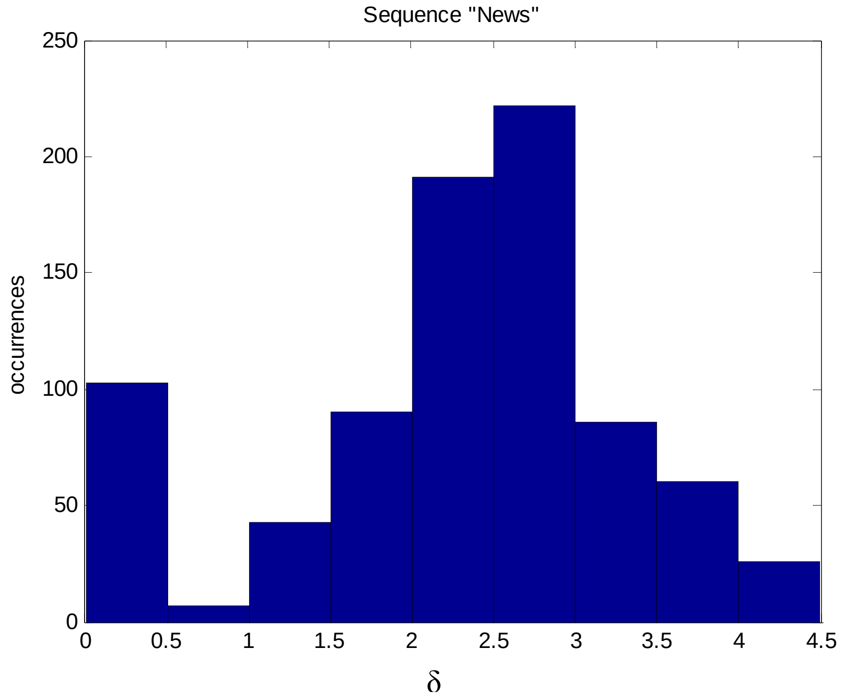


Figure 5.3.: Distribution of 95% confidence intervals for MOS_deg test sequence "News".

As we mentioned in the Chapter 4.2, we did time-variant test to obtain MOS as a reaction to transmissions artifacts. But there was ability that the observers will evaluate also compression artifacts. Therefore we made time-variant test also with sequences, without transmission artifacts and than we reduced this value in sequences with transmission errors. We used formula (5.4) to obtain corrected MOS.

$$MOS_{ik} = (MOS\_\deg)_{ik} - (\max(MOS\_ori) - (MOS\_ori)_i)_k \qquad (5.4)$$

where MOS_deg is average of sequence with PL (depredated) counted by (5.1)

MOS_ori is average of sequence without PL (original) counted by (5.1)

max (MOS_ori) is the highest value from MOS_ori

The MOS for test sequence "News" is in Figure 5.4. We printed also histogram and empirical CDF (Cumulative Distribution Function) for the same sequence. There are depicted in Figures 5.5 and 5.6, respectively

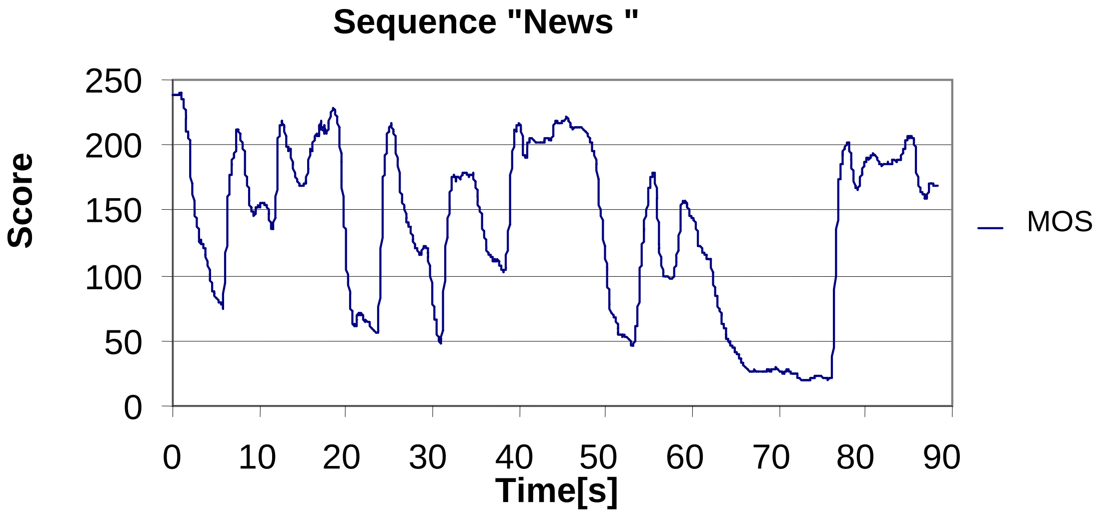


Figure 5.4: Mean opinion score for sequence “News”

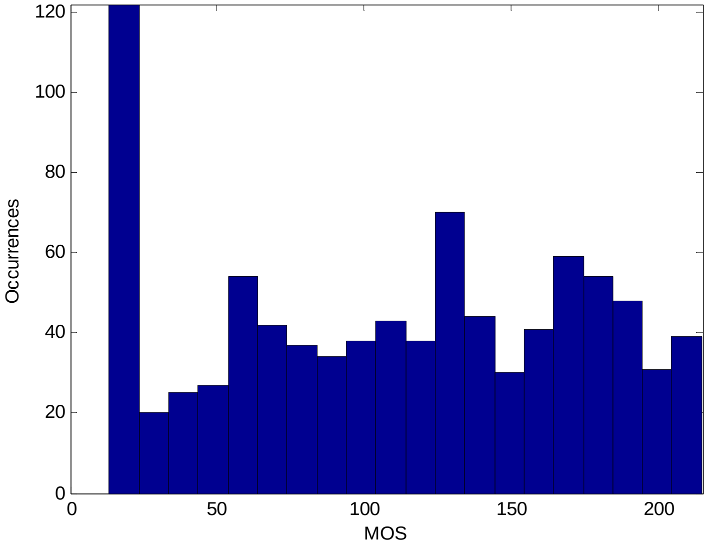


Figure 5.5: Distribution of MOS for sequence “News”

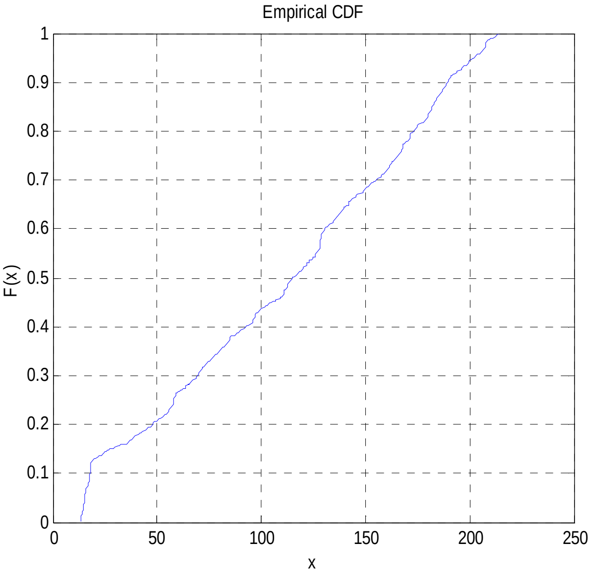


Figure 5.6: Empirical Cumulative distribution function for sequence “News” where F(x) = (Number of observations <= x)/(Total number of observations)

We have final MOS curve of each video sequence. It shows users perceptive sensibility to video quality in time. The data obtained in our experiments will be used to estimate MOS. The MOS obtained with the subjective testing for visual quality will help us to understand viewers’ reaction to transmission artifacts and this information will be very helpful during metric design.

6 Metric design

As we mentioned in Chapter 3 an objective quality assessment is possible. Most of them are to evaluate compression artifacts or noise in video. There exist [5] also metrics to evaluate transmission errors in Internet communication, but we pay attention to evaluate artifacts in digital video (with low resolution and low bitrate) incurrence by transmission errors in wireless networks. In the era of a rapid expansion of the 3G networks in many countries, this phenomenon represents a common problem. Streaming video is one of the most popular advantages of this network.

Until recently, the issue was not addressed in a sufficient manner, and existing metrics were not applied to deal with the problematic either. Nevertheless, two basic image objective evaluations exist, pixel-based (such as MSE and PSNR), and vision-based quality metrics have been developed. Realizing this deficiency, novel specific quality metrics for video processing applications are to be designed and analyzed.

We choose to explore the pixel base one in fact, that it is simple and with compare with vision-based does not need a difficult HVS simulations.

The first step is to adapt the PSNR curve to the scale of MOS. To avoid infinite values of PSNR resulting from zero MSE if there was no error, we perform clipping of the PSNR to the value of 48dB, corresponding to MSE = 1. According to our experience with subjective tests, people do not perceive the improvements resulting from higher PSNR.

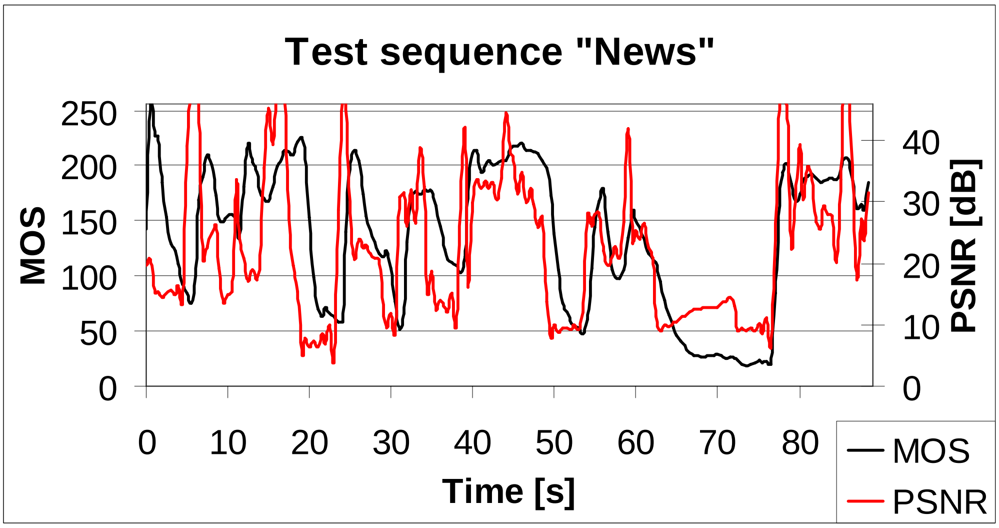


Figure 6.1.: MOS and PSNR with clipping to the value of 48dB for test sequence "News"

The PSNR curve needs to be smoothed and subsampled. Smoothing is necessary due to the human reaction time and the memory of the human eye; the PSNR curve contains a lot of sharp edges. To smooth the curve, we used moving average (formula 6.1). The moving average definition: Given a sequence $\{a_i\}_{i=1}^{N}$, an n –moving average is a new sequence $\{s_i\}_{i=1}^{N-n+1}$ defined from the a_i by taking the average of subsequences of n terms $s_i = \frac{1}{n}\sum_{j=i}^{i+n-1} a_j$ [31]

$$PSNR\_f[i] = \frac{1}{n}\sum_{j=i}^{i+n-1} PSNR[j] \qquad (6.1)$$

We set size of moving window relevant to 1.5 sec. The size of moving window corresponded to the human reaction time. After the clipping and smoothing, the PSNR curve is normalized in the MOS scale by scaling factor *a* (formula 6.2). The scaling factor *a* have the form:

$$a = max\ (MOS)\ /\ max\ (PSNR) \qquad (6.2)$$

a = 5.3

We obtain the MOS estimation PSNR_f based on the PSNR by Formula 6.3:

$$MOSest_{PSNR\_f} = a{*}PSNR\_f \qquad (6.3)$$

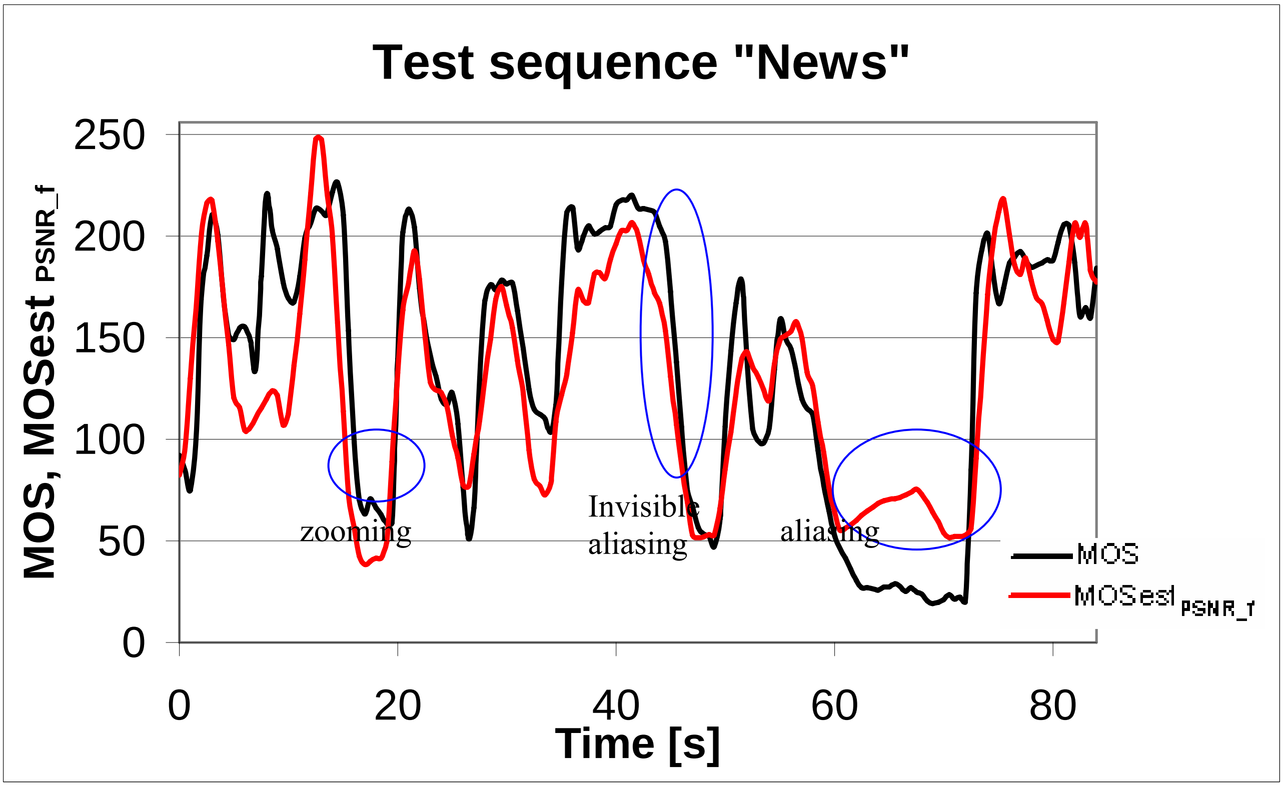


Figure 6.2: MOS and smoothed and sampled PSNR ($MOSest_{PSNR\_f}$) for test sequence "News"

The measured MOS curve is usually steepest when ascending than descending. This is caused by the fact, that in many cases, the user needs time to recognize an impairment from the real content. On the contrary, as soon as the error free key frame arrives, it is a clearly visible improvement of the meanwhile distorted stream. Especially critical are user to the effect we call ”aliasing”. It occurs if there is a packet loss during the rapid scene change (especially cut) and results in aliasing of two frames from different scenes. However, if there is only a slow local movement in the new scene, the user does not necessarily note that there should have been a scene change. This effect we call invisible aliasing (see Figure 2.6). If the impairment itself is almost as large as the whole screen, the correct part (movement) will be considered as impairment. Invisible aliasing results to a time shift between the measured and estimated MOS. Gradual scene changes like zooming or transition can also ”conceal” some errors - it is not clear, what should be the part of the new scene and therefore end-user is not as critical to the error as is the PSNR based estimation. Table 6.1 summarizes the scenarios at which we systematically observed misalignments.

scenario	condition	to correct
invisible aliasing	low motion scene change	time
aliasing	packet loss in scene cut	value
zooming	packet loss in zooming	value

Table 6.1: Rules for correction of $MOSest_{PSNR\_f}$, based on the human perception.

Having observed these simple rules, we can correct the misalignments. We perform following corrections:

1) In case of invisible aliasing the shift n_s is added. The n_s seconds missing after the shift we replace by the last value.

2) Aliasing is penalized (without gradual changes) by multiplying all values with $k_a \in (0,1)$.

3) Gradual scene changes are compensated by multiplying the minima with $k_g > 1$.

We obtained the parameters n_s = 2s by averaging over our measured values for all sequences. We further obtained the parameters k_a = 0.7 and k_g = 1.5 by linear minimum mean square estimation applied to all tested sequences. The resulting predicted MOS

(MOSest$_{PSNR\_m}$)can be seen in Figure 6.3. Note that our corrections do not correct each misalignment, but in general (for different sequences) the similarity to the MOS curve is considerably increased. After analyzing more test sequences with different contents and compression parameters the similarity would be much more improved.

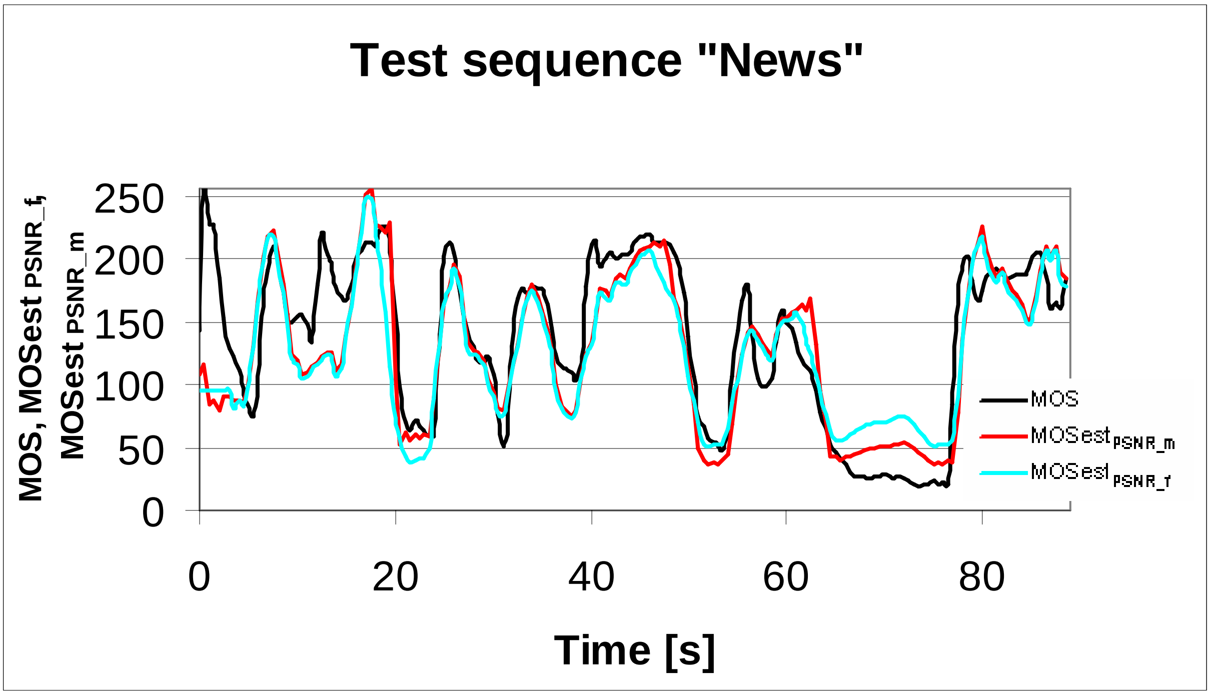


Figure 6.3: MOS, smoothed and sampled PSNR (MOSest$_{PSNR\_f}$) and PSNR after all modifications (MOSest$_{PSNR\_m}$) for test sequence “News”

6.1 Aliasing detection

We need to modify PSNR values during aliasing. Therefore we perform aliasing detection in video. As mentioned in Chapter 2.6.1, aliasing occurs if there is a packet loss during the rapid scene change which results in aliasing of two frames from different scenes. When we compare original frames with degraded frames most of the picture is corrupt. Mathematic definition of aliasing can be: If number of different pixels between original and degraded frame is higher than 1/5, there is an aliasing in the video. We use differential quadratic metric (Formula 6.4) to detect differences between original and degraded sequence for all pixels and all colors.

$$M(i,j)=\sqrt{\sum_{K=R,G,B}\left(O_K(i,j)-D_K(i,j)\right)^2} \qquad (6.4)$$

where O_k is original frame with resolution *i* x *j* for *k* color

D_k is degraded frame with resolution *i* x *j* for *k* color

When was *M(i,j)* higher then 0 we increase the value of the bad pixels and when was number of bad pixel higher than 15 000 for one frame, we know, that there is an aliasing in the frame.

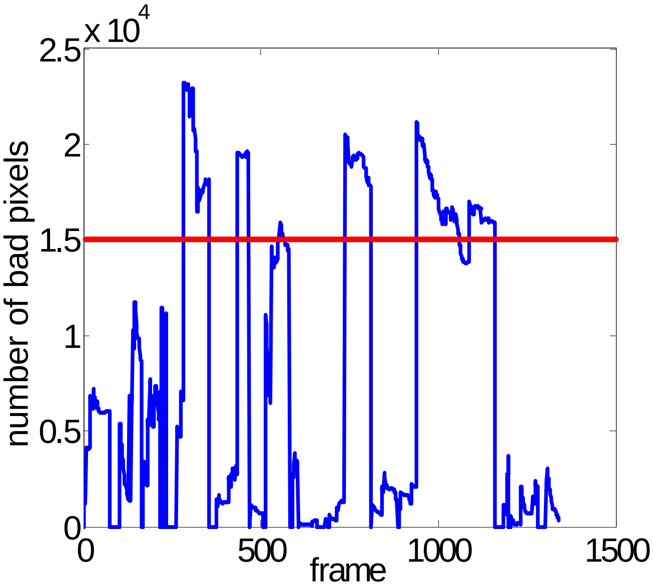


Figure 6.4.: Number of bad pixels in depend on frames for sequence "News"

6.2 Evaluation of metric results

To evaluate the performance of the video quality estimation, we perform the linear (Pearson) correlation R_{lin} to express the prediction accuracy and outlier ratio $R_{2\sigma}$ (Formula 6.5) to check the consistency.

$$R_{2\sigma} = \frac{N_{OUT}}{N} \qquad (6.5)$$

where N is the length of the data vector

N_{OUT} is the number of outliers for which $\left| MOS[i] - MOSest_{PSNR\_m}[i] \right| > 2 * \sigma_{MOS}$

and σ^2_{MOS} is the variance of the MOS

The evaluation of the proposed metric, we perform correlation (1) between the MOS and the PSNR, correlation (2) between MOS and $MOSest_{PSNR\_f}$ and correlation (3) between MOS and $MOSest_{PSNR\_m}$ for different videos and besides for all the video sequences. The results can be found in Table 6.2.

	Correlation (1)	Correlation (2)	Correlation (3)	$R_2\sigma$
Football	0.33	0.65	0.66	12%
News	0.50	0.86	0.89	0%
Town	0.27	0.66	0.67	6%
Cartoon	0.52	0.77	0.80	0%
All sequences	0.42	0.69	0.77	1%

Table 6.2: Goodness of fit for the proposed metric. Correlation (1) is between MOS and PSNR, correlation (2) MOS and smoothed and sampled PSNR and correlation (3) MOS and PSNR after all modifications

In [1], the Perceptual Distortion Metric PDM is presented. The PDM is based on a model of the human visual system (describe in Chapter 3.5). In [1] the PDM was evaluated by means of subjective experiments using different video sequences. The sequences were coded in H. 263 and other standards with and without transmission errors. The correlations for video sequences coded in H263 are less than 70% and correlations for sequences with transmission errors are less than 75 %. As shown our results are comparable or better than the more complicated PDM metric.

As a visual illustration of the effect of the corrections to the PSNR for the ”News” video sequence can be seen on the scatter plots before (figure 6.6) and after the correction (figure 6.5.). It can be seen that the simple corrections are able to compensate the biggest mismatches between the perceived quality and the quality estimated by modified PSNR (scaled, clipped and smoothed). Rest of the test sequences are in figures 6.7 – 6.12. Scatter plot for all sequences are in figure 6.13 (before the corrections) and 6.14 (after the corrections).

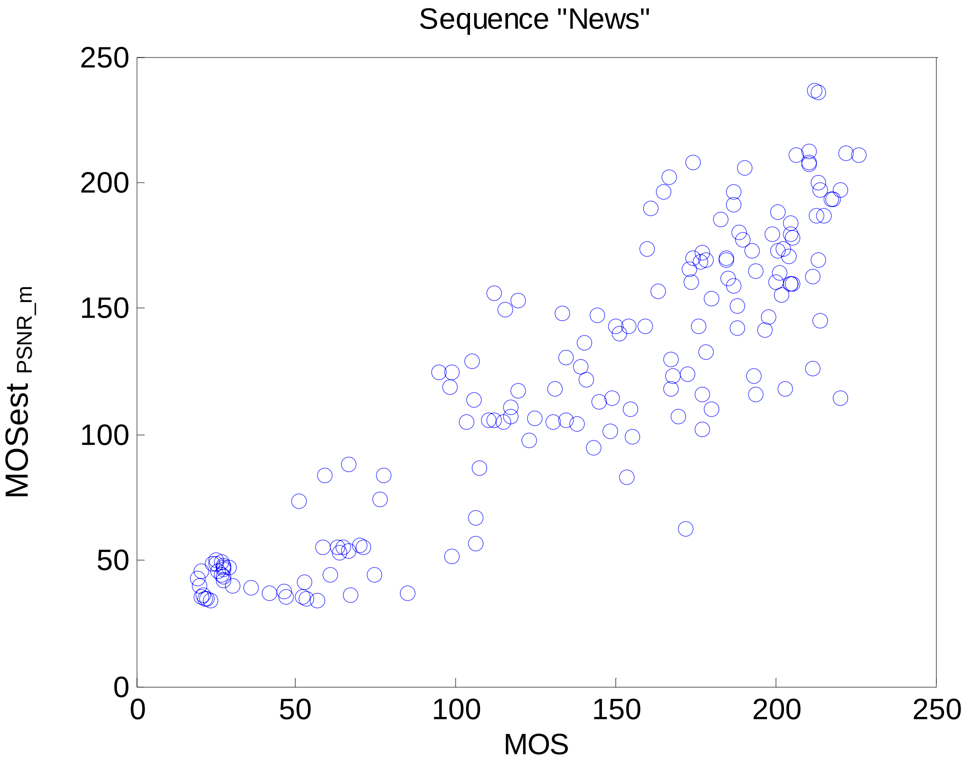


Figure 6.5: Scatter plots of MOS vs. estimated MOS (MOSest$_{\text{PSNR\_m}}$) model predictions for the complete data set. The 0 symbols indicate scores obtained in the low quality quadrants of the subjective test and the 256 symbols indicate scores obtained in the high quality quadrants of the subjective test.

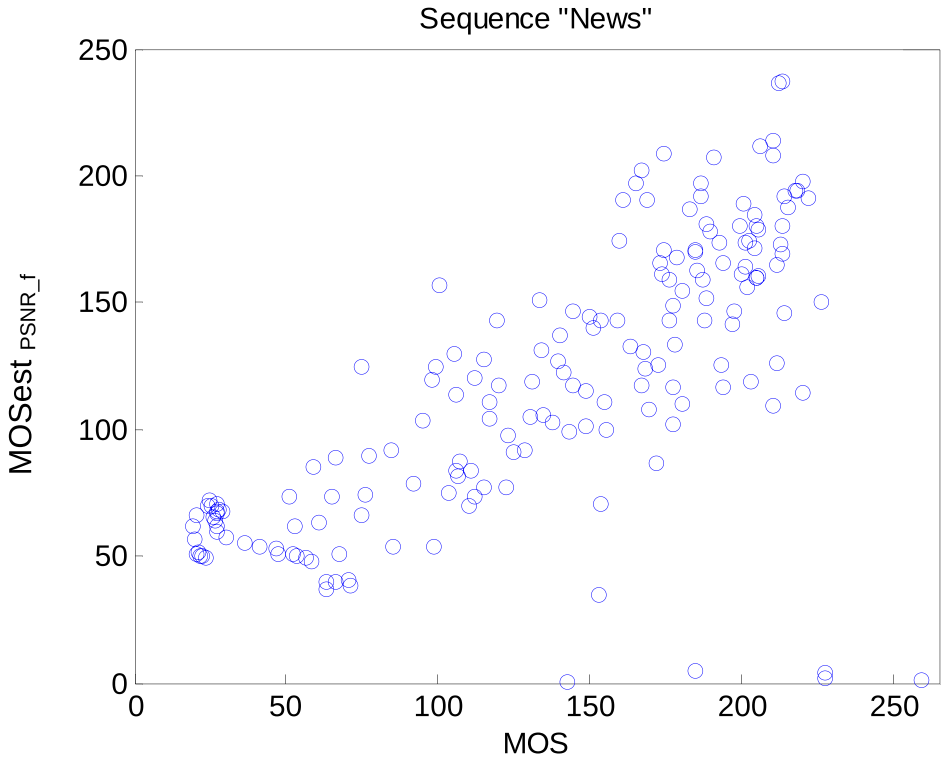


Figure 6.6: Scatter plots of MOS vs. MOSest$_{\text{PSNR\_f}}$ model predictions for the test sequence “News”. The 0 symbols indicate scores obtained in the low quality quadrants of the subjective test and the 256 symbols indicate scores obtained in the high quality quadrants of the subjective test.

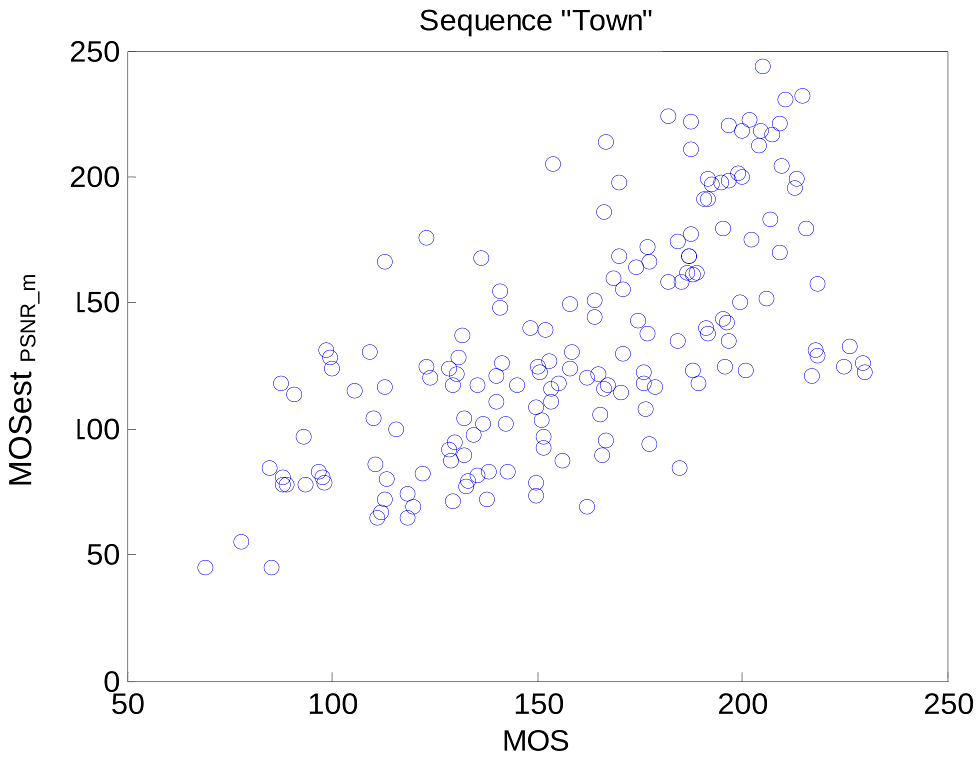


Figure 6.7: Scatter plots of MOS vs. MOSest$_{\mathrm{PSNR\_m}}$ model predictions for the test sequence “Town”

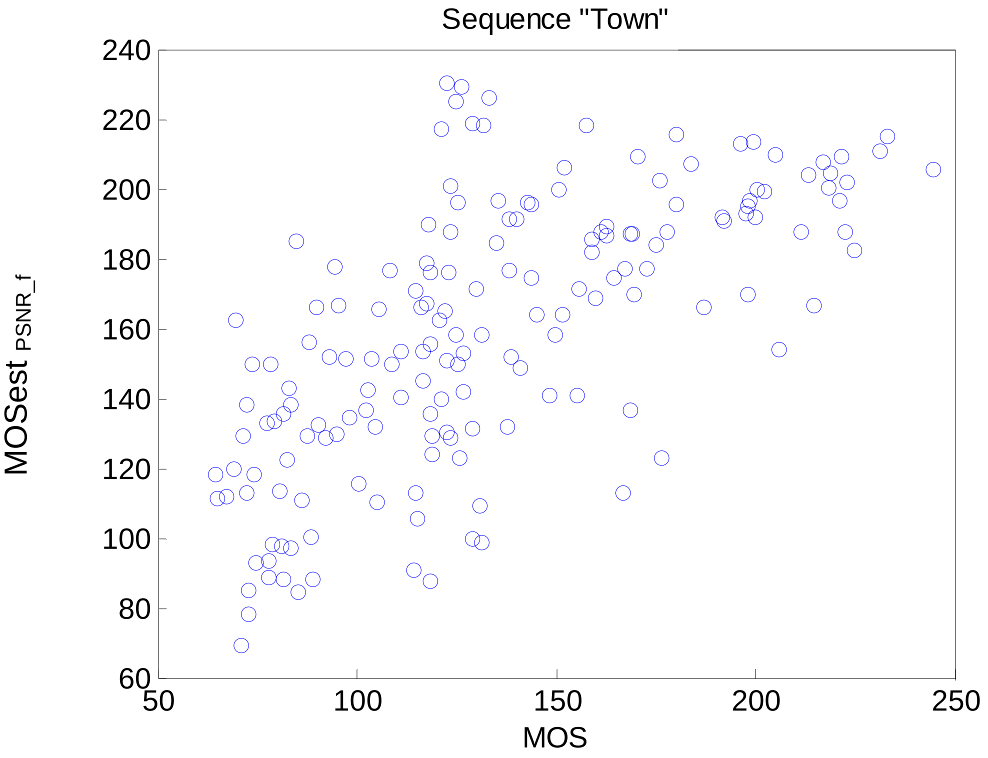


Figure 6.8: Scatter plots of MOS vs. MOSest$_{\mathrm{PSNR\_f}}$ model predictions for the test sequence “Town”

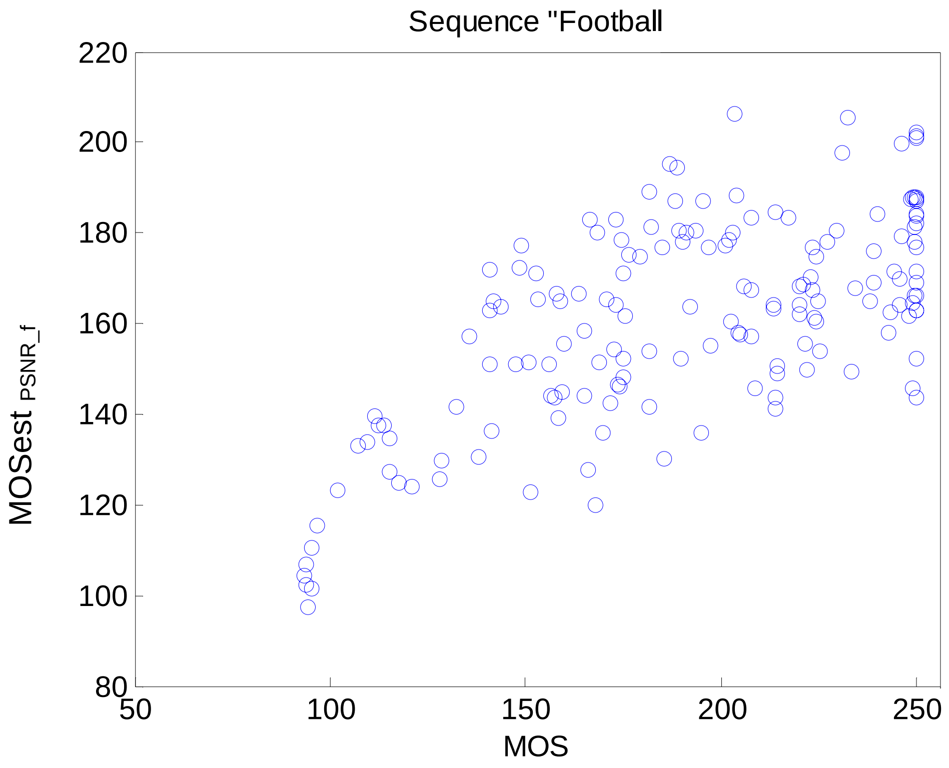


Figure 6.9: Scatter plots of MOS vs. $\text{MOSest}_{\text{PSNR\_f}}$ model predictions for the test sequence “Football”

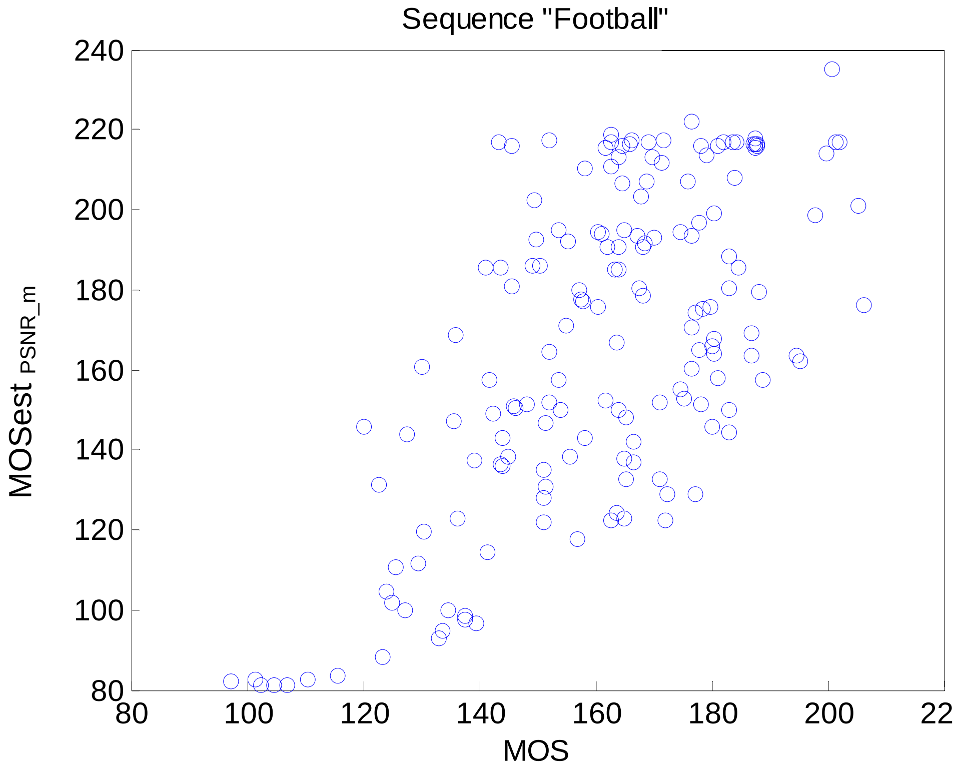


Figure 6.10: Scatter plots of MOS vs. $\text{MOSest}_{\text{PSNR\_m}}$ model predictions for the test sequence “Football”

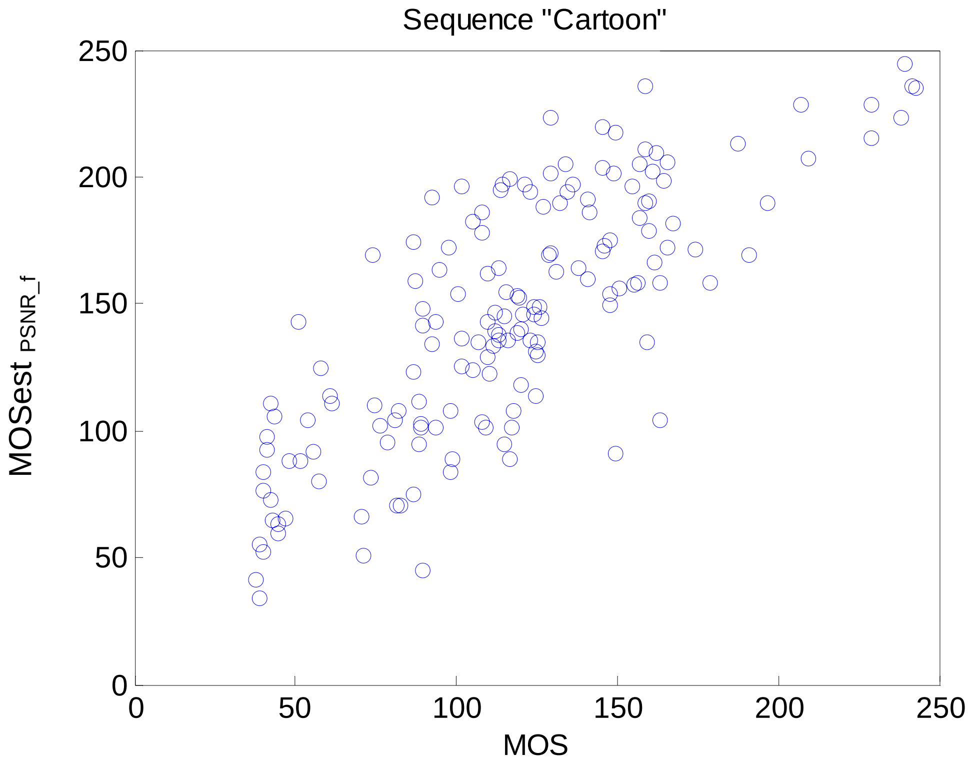


Figure 6.11: Scatter plots of MOS vs. MOSest$_{\text{PSNR\_f}}$ model predictions for the test sequence “Cartoon”

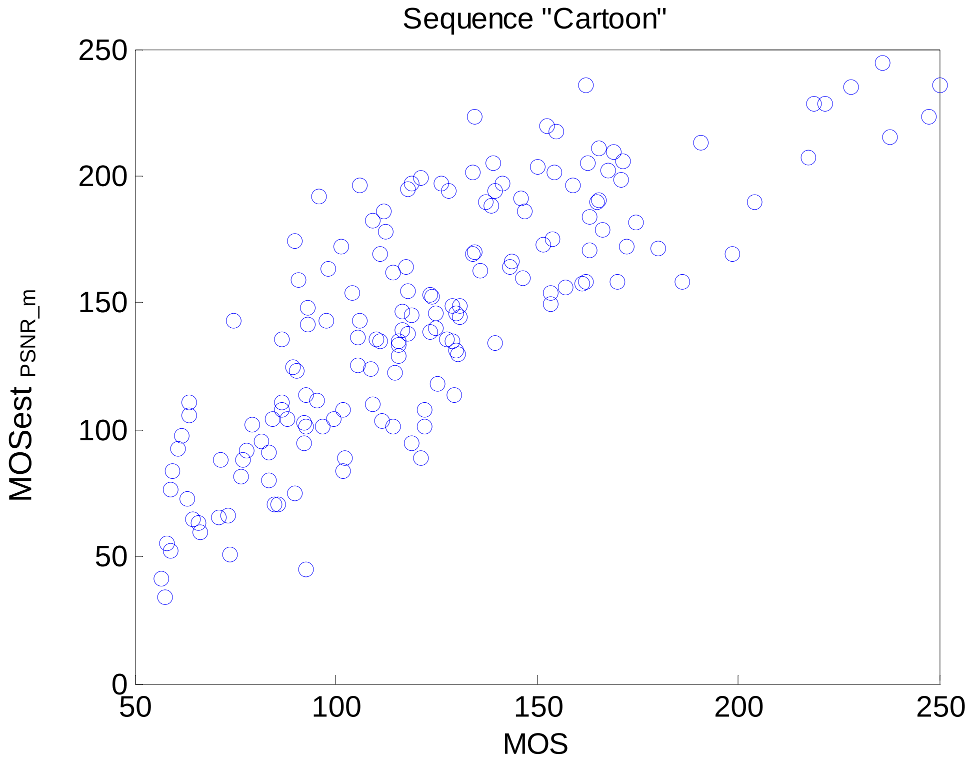


Figure 6.12: Scatter plots of MOS vs. MOSest$_{\text{PSNR\_m}}$ model predictions for the test sequence “Cartoon”

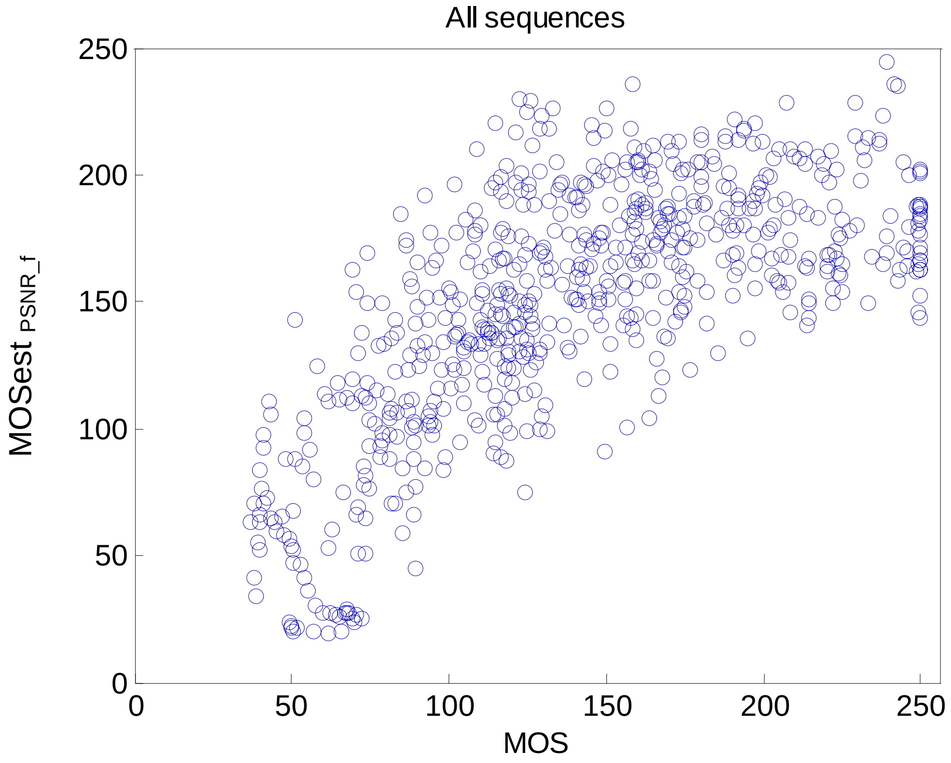


Figure 6.13: Scatter plots of MOS vs. MOSest$_{PSNR\_f}$ model predictions for the all test sequences

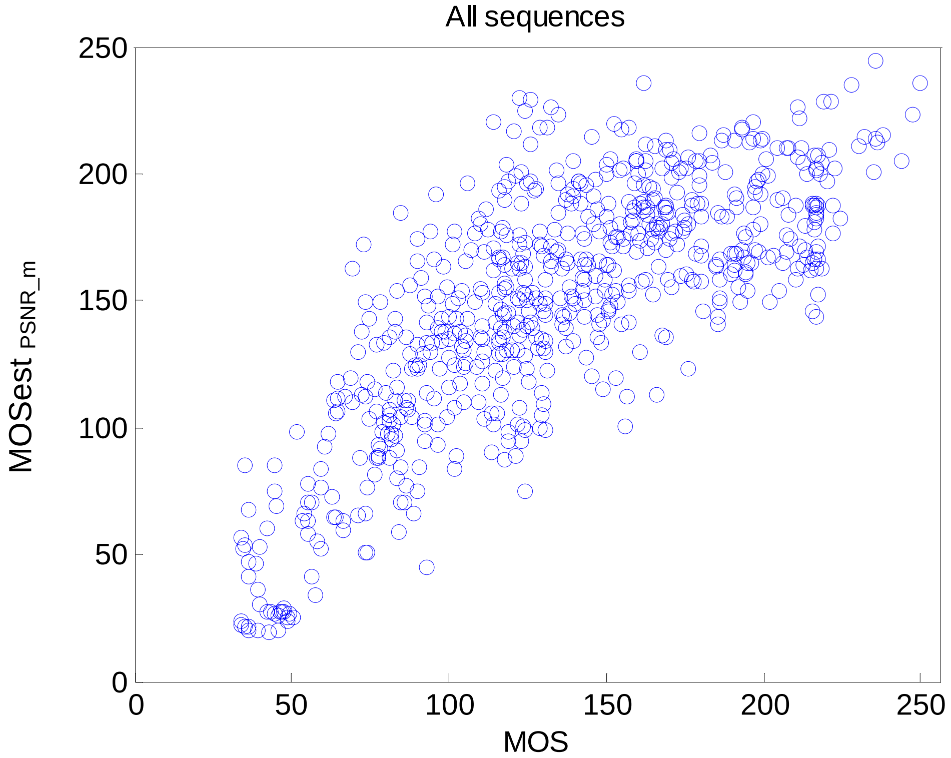


Figure 6.14: Scatter plots of MOS vs. MOSest$_{PSNR\_m}$ model predictions for the all test sequences

7 Conclusion

The foundations of digital imaging systems were discussed. Image and video coding standards exploit certain properties of the human visual system to reduce bandwidth and storage requirements. Compression as well as transmission errors lead to artifacts and distortions affecting video quality. Guaranteeing a certain level of quality has thus become an important concern for 3G networks providers. However, perceived quality depends on many different factors. It is inherently subjective and can only be described statistically. Existing visual quality metrics were reviewed. Pixel-based metrics such as MSE and PSNR are popular despite their inability to reliably predict perceived quality across different scenes and distortion types. Many vision-based quality metrics have been developed, especially over the last few years that try to improve on prediction accuracy. Nevertheless, they still leave much to be desired. Comparative analyses are rare, and so far no video quality metric for the low rate and low resolution video sequences, displayed on the mobile terminal has been found that is able to replace subjective testing.

In this work we presented a PSNR-based estimator of the time-variant video quality for the low rate and low resolution video sequences, displayed on the mobile terminal. To obtain the mean opinion score, we performed subjective perceptual quality tests on UMTS mobile terminals using several video sequences with various data rates and packet loss probabilities. For these tests and obtaining MOS out of the user evaluations, we used a methodology adapted to the testing on mobile terminals, different to the methods recommended for the tests on PC screens. Using the proposed methodology, we were also able to obtain a consistent set of data. We mapped the resulting MOS curve on the appropriately scaled and smoothed PSNR ($MOSest_{PSNR\_f}$) and analyzed the differences. PSNR provided 27-52% correlation with the measured data. Several differences turned out to be caused by the simple human perception rules. After summarizing them, we proposed a new reference-based video quality metric – the PSNR corrected using simple rule-based algorithm and we obtain 66%-89% correlation with the measured data. We evaluated the performance of our estimator by checking its accuracy and consistency. The results show a considerable improvement compared to the PSNR.

The future tasks in this area are to perform tests including more different content types, more data rate and error rate combinations. This would allow for investigating the dependency between the error rate and the appropriate PSNR scaling.

References

[1] S. Winkler, "Vision models and Quality Metrics for Image processing applications", thesis PhD, Lausanne, December 2000.

[2] A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, S. Wolf, "An objective video quality assessment system based on human perception". In SPIE Human Vision, Visual Processing and digital display IV, vol. 1913, pp. 15-26, San Jose (CA), February 1993.

[3] ITU-R Recommendation BT.500-11: "Methodology for the subjective assessment of the quality of television pictures", International Telecommunication Union, Geneva, Switzerland, 2002.

[4] C. Poynton, "The rehabilitation of gamma," In Proceedings of SPIE Human Vision and Electronic Imaging, vol. 3299, pp. 232–249, San Jose, CA, 1998.

[5] S. Winkler, R. Campos, "Video Quality Evaluation for Internet Streaming Applications," Signal Processing Laboratory Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland, 2003.

[6] ITU-T, "Recommendation P.910: Subjective video quality assessment methods for multimedia application", International Telecommunication Union, 1999.

[7] O. Nemethova, M. Ries, M. Zavodsky, M. Rupp, "PSNR-Based Estimation of Subjective Time-Variant Video Quality for Mobiles," Proc. of MESAQIN 2006, Prag, Czech Republic, June 2006.

[8] O. Nemethova, M. Ries, A. Dantcheva, S. Fikar, M. Rupp, "Test Equipment for Time-Variant Subjective Perceptual Video Quality Testing with Mobile Terminals," in Proc. of International Conference on Human Computer Interaction (HCI 2005), Phoenix, USA, November, 2005.

[9] ITU-R "Recommendation BT.601-5: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios," ITU, Geneva, Switzerland, 1995.

[10] P.N. Tudor, "MPEG-2 video compression," Electronics & Communication Engineering Journal 7(6):257–264, 1995.

[11] K. Sayood, "Introduction to Data Compression," Morgan Kaufmann Publishers, 2nd edn., 2000.

[12] ITU-T Recommendation H.263, "Video coding for low bit rate communication," ITU, Geneva, Switzerland, 1998.

[13] R.A. Young, "Oh say, can you see?," The physiology of vision. In proceedings of SPIE Human Vision, Visual Processing and Digital Display, vol. 1453, pp. 92–123, San Jose, CA, 1991.

[14] I. Dalgic, H. Fang, "Comparison of H.323 and SIP for internet telephony signaling," In Proceedings of SPIE Multimedia Systems and Applications, vol. 3845, Boston, MA, 1999.

[15] Video Quality Experts Group (VQEG), "Final Report from The VQEG on the Validation of Objective Models of Video Quality Assessment," March 2000, available in http://www.vqeg.org/.

[16] G. de Haan, E.B. Bellers, "Deinterlacing – an overview," Proceedings of the IEEE 86(9), pp. 1839–1857, 1998.

[17] A.M. Eskicioglu, P.S. Fisher, "Image quality measures and their performance," IEEE Transactions on Communications 43 (12), pp. 2959–2965, 1995.

[18] H. Marmolin, "Subjective MSE measures," IEEE Transactions on Systems, Man, and Cybernetics 16(3), pp. 486–489, 1986.

[19] O.H. Schade, “Optical and photoelectric analog of the eye,” Journal of the Optical Society of America 46(9), pp. 721–739, 1956.

[20] J.L. Mannos, D. J. Sakrison, “The effects of a visual fidelity criterion on the encoding of images,” IEEE Transactions on Information Theory 20(4), pp. 525–536, 1974.

[21] O.D. Faugeras, “Digital color image processing within the framework of a human visua model,” IEEE Transactions on Acoustics, Speech and Signal Processing 27(4), pp. 380–393, 1979.

[22] F.X.J. Lukas, Z.L. Budrikis, “Picture quality prediction based on a visual model,” IEEE Transactions on Communications 30(7), pp. 1679–1692, 1982.

[23] X. Tong, D. Heeger, C. J. van den Branden Lambrecht, “Video quality evaluation using STCIELAB,” In Proceedings of SPIE Human Vision and Electronic Imaging, vol. 3644, pp. 185–196, San Jose, CA, 1999.

[24] X. Zhang, B.A. Wandell, “A spatial extension of CIELA B to predict the discriminability of colored patterns,” In Society for Information Display Symposium Digest, vol. 27, pp. 731–735, 1996.

[25] S. Wolf, M.H. Pinson, “Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system,” In Proceedings of SPIE Multimedia Systems and Applications, vol. 3845, pp. 266–277, Boston, MA, 1999.

[26] S. Winkler, “A Perceptual Distortion Metric for Digital Color Video,” Signal Processing Laboratory Swiss Federal Institute of Technology 1015 Lausanne, Switzerland, 1999.

[27] http://www.nt.tuwien.ac.at/research/mobile-communications/theses-and-practica/finished-practica/comparison-of-equipment-for-temporal-video-quality-evaluation-equipment/

[28] A. Lo, G. Heijenk, I. Niemegeers, “Evaluation of MPEG-4 Video Streaming over UMTS/WCDMA Dedicated Channels,“ Delft University of Technology , 2600 GA Delft, The Netherlands, 2005.

[29] K.T. Tan, M. Ghanbari, D. E. Pearson, “An objective measurement tool for MPEG video quality,“ Signal Processing 70(3), pp. 279–294, 1998.

[30] C. J. van den Branden Lambrecht, “Color moving pictures quality metric,“ In Proceedings of the International Conference on Image Processing, vol. 1, pp. 885–888, Lausanne, Switzerland, 1996.

[31] Eric W. Weisstein, “Spencer's Formula,” From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/SpencersFormula.html

[32] A. B. Watson, “Perceptual-components architecture for digital video,“ In Journal of the Optical Society of America A 7(10), pp. 1943–1954, 1990.