DISSERTATION

Video Quality Estimation for Mobile Video Streaming

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften

eingereicht an der Technischen Universitat Wien Fakultät für Elektrotechnik und Informationstechnik, Institut für Nachrichtentechnik und Hochfrequenztechnik

von

Michal Ries
Bratislavská 86, SK-90201 Pezinok
Slowakei

geboren am 31. Dezember 1979 in Bratislava (SK)
Matrikelnummer: 0326485

Wien, im September 2008

.

Begutachter:

Univ. Prof. Dr. Markus Rupp
Institut für Nachrichtentechnik und Hochfrequenztechnik
Technische Universität Wien
Österreich

Doc. Dr. Yevgeni Koucheryavy
Institute of Communication Engineering
Tampere University of Technology
Finland

Abstract

FOR the provisioning of video streaming services it is essential to provide a required level of customer satisfaction, given by the perceived video stream quality. It is therefore important to choose the compression parameters as well as the network settings so that they maximize the end-user quality. Due to video compression improvements of the newest video coding standard H.264/AVC, video streaming for low bit and frame rates is possible while preserving its perceptual quality. This is especially suitable for video applications in 3G wireless networks.

Mobile video streaming is characterized by low resolutions and low bitrates. The commonly used resolutions are Quarter Common Intermediate Format (QCIF,176x144 pixels) for cell phones, Common Intermediate Format (CIF, 352x288 pixels) and Standard Interchange Format (SIF or QVGA, 320x240 pixels) for data-cards and palmtops (PDA). The mandatory codec for Universal Mobile Telecommunications System (UMTS) streaming applications is H.263 but the 3GPP release 6 already supports a baseline profile of the new H.264/AVC codec. The appropriate encoder settings for UMTS streaming services differ for various streaming content and streaming application settings (resolution, frame and bit rate).

In the last years, several objective metrics for perceptual video quality estimation were proposed. The proposed metrics can be divided into two main groups: human vision model based video metrics and metrics based on empirical modeling. The complexity of these methods is quite high and they are mostly based on spatial features, although temporal features better reflect the perceptual quality especially for low-rate videos. Most of these metrics were designed for broadband broadcasting video services and do not consider mobile video streaming scenarios.

The goal of the presented research is to estimate video quality of mobile video streaming at the user-level (perceptual quality of service) for a large set of possible codec settings in 3G network and for a wide range of video content. Measures were derived that do not need the original (non-compressed) sequence for the estimation of quality, because such reference-free measures reduce complexity and at the same time broaden the possibilities of the quality prediction deployment. New reference-free approaches are presented for quality estimation based on motion characteristics. Moreover, this thesis provides a detailed comparison of recently proposed models for video quality estimation.

Kurzfassung

DIE Einführung der dritten Mobilfunkgeneration ermöglichte durch die höheren Datenübertragungsraten die Anwendung von Multimedia Diensten. Echtzeit-Dienste, wie zum Beispiel Videostreaming und Videotelephonie, stellen hierbei für Mobilfunksysteme eine besondere Herausforderung dar, wegen der hohen Empfindlichkeit von Videodiensten gegenüber visuellen Störungen. Der H.264/AVC Videostandard ist durch seine hohe Effizienz besonders geeignet für die Codierung dieser Dienste. Er bietet eine sehr gute Videoqualität für Übertragungen bei niedrigen Bildwiederhol- und Bitraten.

Mobile Videostreaming Anwendungen zeichnen sich durch ihre niedrigen Auflösungen und niedrige Bitraten aus. Gängige Auflösungen sind QCIF (176x144 Pixel) für Mobiltelefone, CIF (352x288 Pixel) und SIF (320x240 Pixel) für Datenkarten und Palmtops.

Der UMTS Standard schreibt als Minimum H.263 als verpflichtenden Codec für alle Terminals vor. Viele Terminals für Release 6 unterstützen auch das Basisprofil des neuen H.264 Codecs. Die entsprechenden Encodereinstellungen für UMTS-Streamingdienste unterscheiden sich abhängig vom Inhalt und Anwendung (Auflösung, Bildrate, Bitrate).

Die Schätzung der vom Benutzer erfahrenen Videoqualität basierte in den letzten Jahren auf mehreren objektiven Parametern. Diese Metriken lassen sich in zwei Lager unterteilen. Die einen Metriken haben empirische Modelle als Grundlage, die anderen basieren auf Modellen für die menschliche visuelle Wahrnehmung.

Die Komplexität dieser Methoden ist besonders hoch und sie sind sehr abhängig von räumlichen Merkmalen. Es hat sich jedoch gezeigt, dass bei der Verwendung mobiler Terminals zeitliche Merkmale die vom Benutzer wahrgenommene Qualität wesentlich besser widerspiegeln. Die meisten dieser Metriken wurden vor dem Hintergrund von Breitband - Videodiensten entworfen und passen daher nicht zu den niedrigen Datenraten einer mobilen Anwendung.

Ziel der vorgestellten Forschung war es einen guten Schätzer für die Videoqualität mobiler Dienste zu finden. Das Modell basiert auf der auf Benutzerebene wahrgenommenen Qualität und kann eine große Anzahl von heute gebräuchlichen Codeceinstellungen in einem 3G Netz abdecken. Weiters ist die vorgeschlagene Metrik referenzfrei, das heißt die Bewertung der Qualität kann ohne die Kenntnis der Originalsequenz erfolgen. Dadurch wird die Komplexität deutlich reduziert und gleichzeitig die möglichen Anwendungsgebiete, in denen diese Metrik zum Einsatz kommen kann, erweitert. Abschliessend stellt diese Arbeit einen umfangreichen Vergleich des verfügbaren und neuen Modells für verschiedene Szenarien dar.

Acknowledgement

"Perfect works are rare, because they must be produced at the happy moment when taste and genius unite."

François-René de Chateaubriand

THERE is a large number of colleagues and friends that substantially contributed to the success of my thesis.

I address particular thanks to my supervisor Prof. Dr. Markus Rupp for providing me a challenging, instructive and highly positive PhD-experience. I greatly appreciate his full support and confidence in any situation and thank him for his interest and good discussions.

I would also like to thank mobilkom austria AG for technical and financial support of this work[1] as well as for their valuable feedback.

Finally, I would like to thank my family members, especially my wife, Martina and my daughter Lucia for supporting and encouraging me to pursue my dissertation work.

Vienna, September 2008

Michal Ries

e-mail: mries@nt.tuwien.ac.at

[1]The views expressed in this thesis are those of author and do not necessarily reflect the views within mobilkom austria AG.

Contents

Chapter 1

Introduction

Contents

1.1 Motivation

MOBILE multimedia streaming applications are becoming more and more popular although the perceptual video quality is limited for such low bit rates, frame rates and resolutions. Therefore, it is essential to provide a required level of customer satisfaction, given by the perceived video stream quality for video streaming services. It is important to choose the compression parameters as well as the network settings so that they maximize the end-user quality. Due to video compression improvements of the newest video coding standard H.264/AVC [1], video streaming for low bit and frame rates is possible while preserving its perceptual quality. This is especially suitable for video applications in mobile wireless networks (3G, WLAN ...).

Video streaming is a one-way quasi real-time data transport, where the content is consumed (viewed/heared/read) while it is being delivered. To compensate jitter (variance of the end-to-end delay) at the receiver, a portion of the received data is buffered in a play-out buffer. In the case of video streaming, the video content is rendered on the screen with the signalized frame rate, making the inter-packet arrival time variations invisible for the user. Therefore, the end-user quality for video streaming does not depend on the absolute end-to-end delay (as long as it is kept in the order of seconds). Thus, video streaming is usually referred to as a quasi real-time service. Moreover, mobile video streaming is characterized by low resolutions and low bitrates. The Universal Mobile Telecommunications System (UMTS) release 4 (implemented by the first UMTS network elements and terminals) provides a maximum data rate of 1920 kbit/s shared by all users in a cell, release 5 (emerging) offers up to 14.4 Mbit/s in downlink (DL) direction for High Speed Downlink Packet Access (HSDPA). The availability of such data rates initiated the launch of new services, out of which the real-time services are the most challenging from the provider point of view. The commonly used resolutions are QCIF for cell phones, CIF and SIF for data-cards and palmtops (PDA). The mandatory codec for UMTS streaming applications is H.263 but the 3GPP release 6 [2] already supports a baseline profile of the new H.264/AVC codec [3]. The appropriate encoder settings for UMTS streaming services differ for various streaming content and streaming application settings (resolution, frame and bit rate) as is demonstrated in [4], [5], [6], [7], [8].

Conventional subjective quality measurement methods [9] involve presenting a test subject with one or more video clips. The subject is asked to evaluate the quality of the test clips, e.g. by recording the perceived degradation of the clip compared with a reference video clip. A typical test requirement is to determine the optimum choice from a set of alternative versions of a video clip, e.g. versions of the same clip encoded with different codecs, alternative strengths of a post-processing filter and alternative trade-offs between encoder settings for a given bit rate. In this type of scenario, each of the alternative versions of a video clip must be viewed and graded separately by the test subjects, so that the time taken to carry out a complete test increases linearly with N, the number of alternatives to be tested. In some cases (e.g. choosing a preferred trade-off between frame-rate and image quality), there is a large number of possible outcomes and the test designer is faced with the choice between running a very large number of tests in order to obtain a fine-grained result or limiting the number of tests at the expense of discretizing the result [13]. Moreover, the subjective testing is extremely man-power and time consuming.

In the last years, several objective metrics for perceptual video quality estimation were proposed. The proposed metrics can be divided into two main groups: human vision model based video metrics [14], [15], [16], [17] and metrics based on empirical modeling [18], [19], [20], [21]. The complexity of these methods is quite high and they are mostly based on spatial features, although temporal features better reflect perceptual quality especially for low-rate videos. Most of these metrics were designed for broadband broadcasting video services and do not consider mobile video streaming scenarios.
The goal of this thesis is to estimate the video quality of mobile video streaming at the user-level (perceptual quality of service) for a large set of possible codec settings in 3G networks and for a large set of content types. The focus is given at measures that do not need the original (non-compressed) sequence for the estimation of quality, because such reference-free measures reduce the complexity and at the same time broaden the possibilities of the quality prediction deployment. Moreover, the objective measures of video quality should be simple enough to be calculated in real-time at the receiver side.
This thesis addresses the design of reference free video quality metrics in mobile environments. The whole chain of metric design regards the definition of mobile streaming scenarios, a subjective test methodology, the selection and statistical evaluation of objective parameters and the estimator design. The proposed video quality metrics are applicable for quality monitoring in mobile networks.

1.2 Video streaming in UMTS network

STREAMING refers to the ability of an application to play synchronized media streams like audio and video streams in a continuous way while those streams are being transmitted to the client over a packet data network. The streaming applications are usually on-demand and offer a live information delivery service (e.g. music, cartoons, panorama and news on-demand). Streaming over fixed Internet Protocol (IP) networks is already one of the most significant applications today. For the third generation (3G) systems, the 3G Packet-Switched Streaming service (PSS) [38] fills the gap between 3G Multimedia Messaging Service (MMS) and conversational services. PSS enables mobile streaming applications, in which the protocol and terminal complexity is lower than for conversational services, in contrast to a streaming terminal that requires media input devices, media encoders [2], and more complex protocols.
The usual way of transport of video streaming over IP packet networks assumes an Real-Time Protocol (RTP) together with Real-Time Control Protocol (RTCP) feedback on the application/session layer and a User Datagram Protocol (UDP) on the transport layer [35], [36]. In contrast to the Transmission Control Protocol (TCP), UDP does not provide any Automatic Repeat reQuest (ARQ) mechanism to perform retransmissions. It only provides a checksum to detect possible errors. The checksum is typically calculated over a rather large packet to avoid a rate increase due to packet headers.
In contrast to that, interactive and background services having a non-real-time nature (e.g. web browsing, file transfer, e-mail) can make use of retransmission mechanisms, e.g. provided by TCP — packetloss is compensated by delay.
Simple streaming services include a basic set of streaming control protocols, transport protocols, media codecs and a scene description protocol. In this simple case it is already possible to provide multimedia

streaming services. A mobile user obtains a Universal Resource Identifier (URI) to specific content that is suitable to the terminal. This URI may come from a World Wide Web (WWW) browser, Wireless Application Protocol (WAP) browser, or is set manually. It specifies a streaming server and the address of the content on that server. An application that establishes the multimedia session obtains the data from a Session Description Protocol (SDP) file. The SDP file contains the description of the session (session name, author ...), the type of media to be presented, and the bit rate of the media. Moreover, the SDP file may be delivered in a number of ways. Usually, it is provided in a link inside the Hyper-Text Mark-up Language (HTML) page that the user downloads via an embedded tag. It may also be directly obtained by typing it as an URI or through the Real-Time Streaming Protocol (RTSP), signaling via the described method. In case of the streaming delivery option of the MMS service, the MMS user agent receives a modified MMS message with SDP file from the MMS relay or server.
The session establishment is the process in which the browser or the mobile user invokes a streaming client to set up the session against the server. The User Equipment (UE) is expected to have an active PDP (Packet Data Protocol) context in accordance with [42] or an other type of radio bearer that enables IP packet transmission at the start of session establishment signaling. The client may be able to ask for more information about the content. The client shall initiate the provisioning of a bearer with appropriate Quality of Service (QoS) for the streaming media. Sessions containing only non-streamable content such as a Synchronized Multimedia Integration Language (SMIL) file, still images and text to form a time-synchronized presentation do not require the SDP file in session establishment. The streaming service is set up by sending an RTSP SETUP message for each media stream chosen by the client. This returns UDP and/or TCP port etc. to be used for the respective media stream. The client sends a RTSP PLAY message to the server that starts to send one or more streams over the IP network. This case is illustrated in **Figure 1.1**.

The minimal requirement for provisioning of a streaming service includes at least a content server and a streaming client. A streaming server is located behind the Gi interface. Additional components like portals, profile servers, caching servers and proxies located behind the Gi interface can be involved as well to provide additional services or to improve the overall service quality (see **Figure 1.2**). Portals are servers allowing convenient access to streamed media content. For instance, a portal can offer content browsing and search facilities. In the simplest case, it is a Web/WAP-page with a list of links to streaming contents. The content itself is usually stored on content servers, which can be located elsewhere in the network. User and terminal profile servers are used to store user preferences and terminal capabilities. This information can be used to control the presentation of streamed media content to a mobile user.

1.2.1 Streaming protocols and codecs

PSS clients and servers support an IP-based network interface for the transport of session control and media data. Control and media data are sent using TCP/IP [22] and UDP/IP [23]. An overview of the protocol stack can be found in **Figure 1.3**.
The Internet Engineering Task Force (IETF) RTP [24], [42] provides a means for sending real-time or streaming data over UDP [24]. The encoded media is encapsulated in RTP packets with media specific

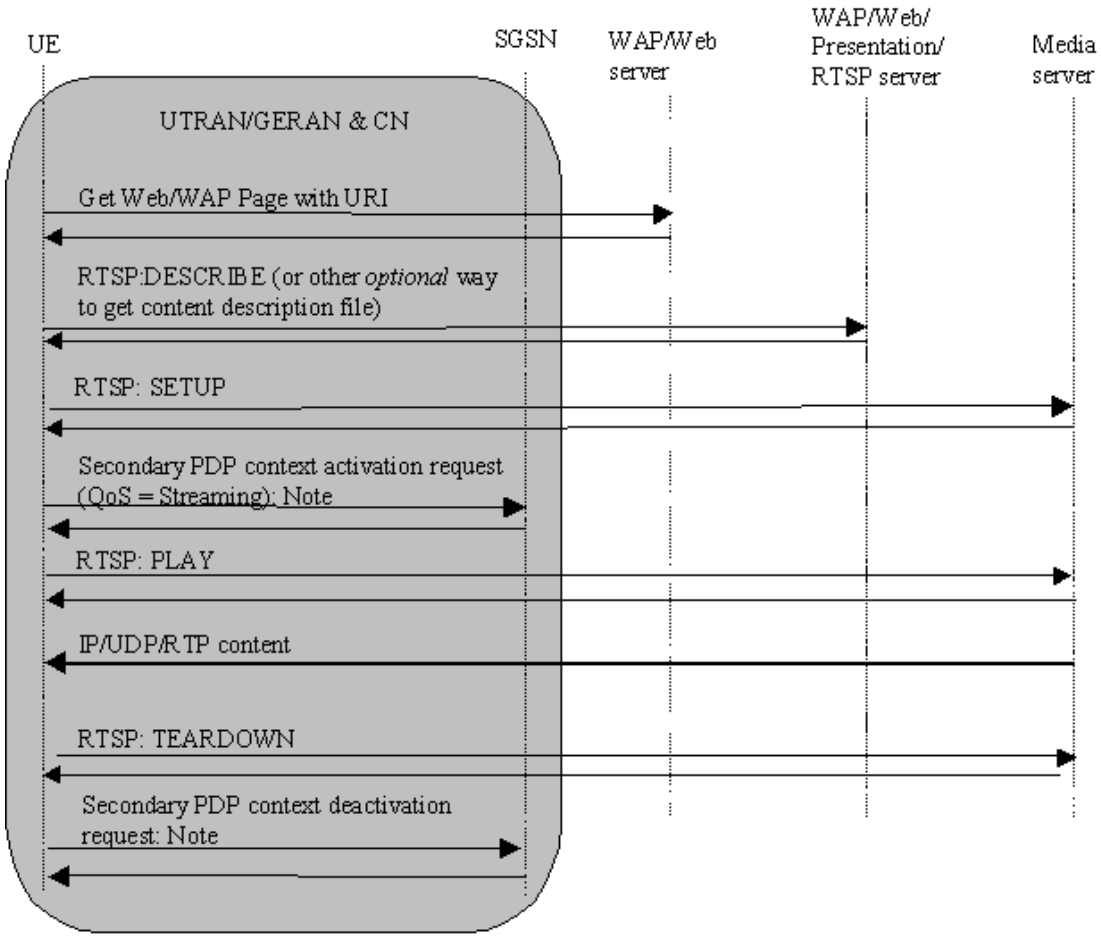

Figure 1.1: Schematic view of a simple streaming session [38].

RTP payload formats. RTP payload formats are defined by [24], [42]. RTP also provides feedback about the transmission quality by RTCP [42]. The continuous media (speech, audio and video) flows are encapsulated in the following protocol stacks RTP/UDP/IP of RTP/TCT/IP (**Figure 1.3**). Moreover, following RTP payload formats are supported for RTP/UDP/IP of RTP/TCT/IP transport:

- AMR (Adaptive Multirate Codec) narrow band speech codec RTP payload format according to [25]. A PSS client is not required to support multi-channel sessions;
- AMR wide band speech codec RTP payload format according to [25]. A PSS client is not required to support multi-channel sessions;
- MPEG-4 AAC (Advanced Audio Coding) audio codec RTP payload format according to RFC 3016 [26];
- MPEG-4 video codec RTP payload format according to RFC 3016 [26];
- H.263 [22] video codec RTP payload format according to RFC 2429 [27].

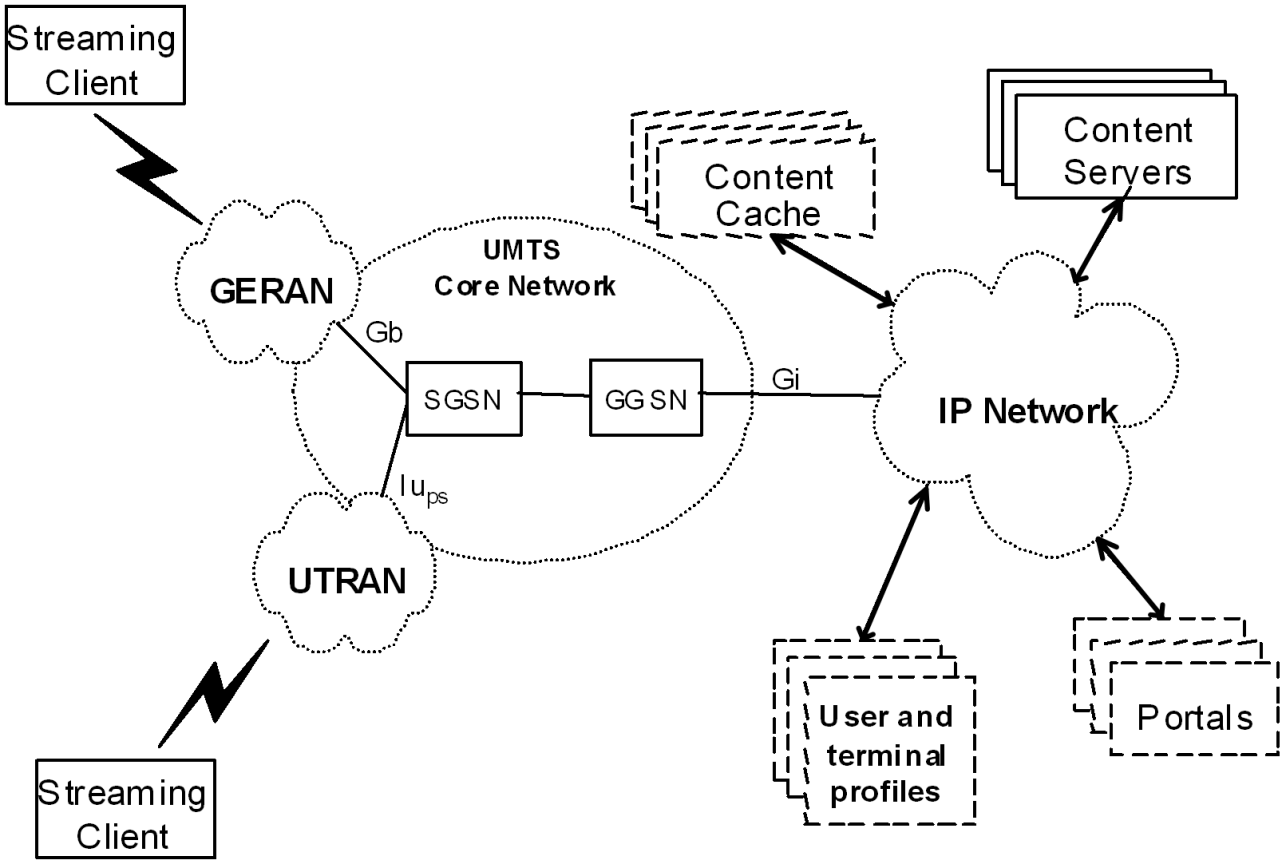

Figure 1.2: Network elements involved in a 3G packet switched streaming service [38].

The following speech, audio and video codecs are supported for UMTS streaming. For audio encoding the following codecs are supported:

- AMR speech codec shall be supported for narrow-band speech [28]. The AMR wideband speech codec [29] shall be supported when wideband speech working at 16 kHz sampling frequency is supported.
- AAC Low Complexity (AAC-LC) object type audio codec [30] should be supported. The maximum sampling rate to be supported by the decoder is 48 kHz. The channel configurations to be supported are mono (1/0) and stereo (2/0).
- AAC Long Term Prediction (AAC-LTP) audio codec may be supported too.
  When a server offers an AAC-LC or AAC-LTP stream with the specified restrictions, it shall include the "profile-level-id" and "object" parameters.

The video codec H.263 [31] profile 0 level 10 is mandatory for the PSS. In addition, PSS supports:

- H.263 [30] profile 3 level 10 decoder,
- MPEG-4 visual simple profile Level 0 decoder, [32] and [33],
- H.264 baseline profile.

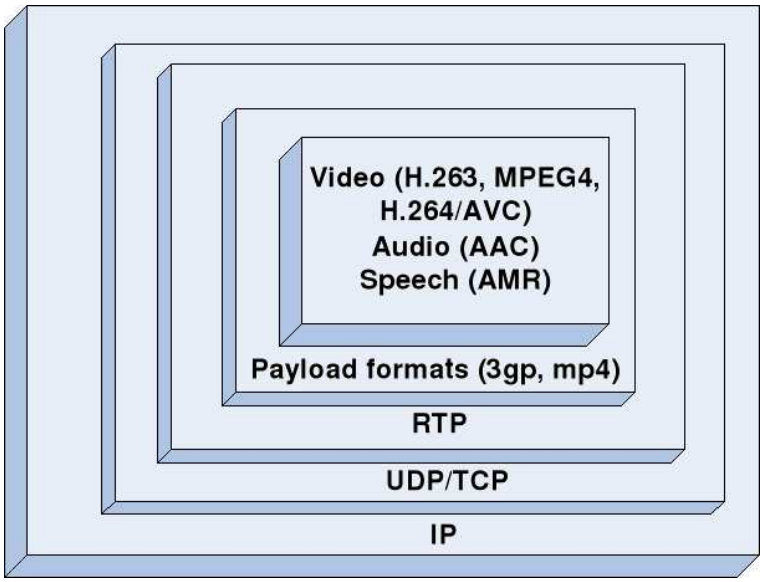

Figure 1.3: Overview of the protocol stack.

Conformance	Suffix	Content
Release 6	.3gp	AMR and hint track
Release 6	.3gp	2 tracks H.263 and 2 hint tracks
Release 6, 5, 4	.3gp	H.263, AMR and hint tracks
3fp file, also conforming to mp4		
Release 4, 5 and mp4	.3gp	MPEG-4 video
mp4 file, also conforming to 3gp		
Release 5 and mp4	.mp4	MPEG-4 video and AAC

Table 1.1: Conformance of 3gp and mp4 streaming formats.

There are 3gp and mp4 [33] streaming formats supported in UMTS streaming. The 3gp standard is defined by 3GPP [2], for creation, delivery, and playback of multimedia over wireless networks. It enables sharing multimedia files between varieties of mobile devices, handsets and PDAs. The 3gp format is based on the ISO base file format. This allows mixing different types of media (see Table **1.1**), video, audio and text. The 3gp file format can encapsulate the following video end audio formats: H.263, H.264, MPEG-4, AMR, AAC and timed text.

1.2.2 Structure of video payload

The Network Abstraction Layer (NAL) of the video encoder encapsulates the slice output of the Video Coding Layer (VCL) encoder into Network Abstraction Layer Units (NALU). Each NALU contains the compressed video data coming from the video coding layer and provides additional non-VCL (Video Coding Layer) information, such as sequence and picture parameters, access unit delimiter, filler data,

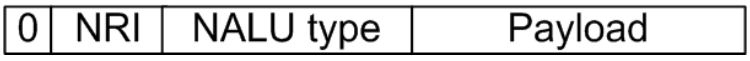

Figure 1.4: Format of a NALU and its header.

Supplemental Enhancement Information (SEI), display parameters, picture timing etc.
All data related to a video stream are encapsulated in NALUs in a way most appropriate for a particular network. The format of a NALU is shown in **Figure 1.4**. The first byte of each NALU is a header byte, the rest are the data. The first bit of the NALU header is a zero (forbidden) bit. The following two NRI (NAL Reference Identification) bits signalize the importance of the NALU for reconstruction purposes. The next five bits indicate the NALU type corresponding to the type of data being carried in that NALU, allowing 32 types of NALUs. These are classified in two categories: VCL NALUs and non-VCL NALUs. The NALU types from one to five are VCL NALUs and contain data related to the output of VCL — slices. Each encoded slice is also attached a header containing information related to that slice.

NALUs with a NALU type indicator value higher than five are non-VCL NALUs carrying information like SEI, sequence and picture parameter set, access unit delimiter etc. Depending on a particular delivery system and scheme, some non-VCL NALUs may or may not be present in the stream containing VCL NALUs. For example, NALU type seven carries the Sequence Parameter Set (SPS), defining profile, resolution and other properties of the whole sequence; type eight carries the Picture Parameter Set (PPS), containing type of entropy coding, slice group and quantization properties. These sequence and picture level data can be sent asynchronously and in advance of the media stream contained in the VCL NALUs. An active SPS remains unchanged troughout a coded video sequence. An active PPS remains unchanged within a coded picture. In order to be able to change picture parameters such as picture size without the need to transmit parameter set updates synchronously to the slice packet stream, the encoder and decoder can maintain a list of more than one SPS and PPS. Each slice header contains then a codeword that indicates the SPS and PPS in use.

The NALUs can easily be encapsulated into different transport protocols and file formats, such as MPEG-2 transport stream, RTP (Real-Time Protocol), MPEG-4 and 3gp file formats. For transmission over mobile networks, a VCL slice is encapsulated in RTP according to [37]. The RTP payload specification supports different packetization modes. In the simplest mode a single NALU is transported in a single RTP packet, and the NALU header serves as an RTP payload header. In non-interleaved mode, several NALUs of the same picture can be encapsulated into the same RTP packet. In interleaved mode several NALUs belonging to different pictures can be encapsulated into the same RTP packet. Moreover, NALUs do not have to be sent in their decoding order. Both the non-interleaved and interleaved modes also allow for fragmentation of a single NALU into several RTP packets.

1.3 Principles of video coding and coding artifacts

UNCOMPRESSED video requires large amounts of bandwidth and storage space. Moreover, the end-user cost in wireless networks is typically proportional to available bandwidth and transmitted data volumes. Therefore, the videos transmitted over wireless networks are compressed with very effective and lossy compression algorithms. For mobile video streaming, the following video compression standards are used today: H.263 [39] standardized by International Telecommunication Union (ITU), MPEG-4 part 2 [32] standardized by International Organization for Standardization (ISO) Motion Picture Expert Group (MPEG), and the emerging, the newest H.264 [1] (known also as Advanced Video Coding (AVC) and MPEG-4 part 10), standardized by the Joint Video Team (JVT) of experts from both ISO/IEC (International Electrotechnical Commission) and ITU. The principles of the compression for all mentioned codecs are very similar.
In this thesis, the focus is given on H.263 and H.264/AVC codecs, which are designed for a multitude of applications; the structure of their bitstreams may vary significantly. To avoid the implementation of all possible stream structures by each specification-conform decoder, *profiles* and *levels* were defined. A profile is a subset of the capabilities including the entire bitstream syntax; a level is a specified set of constraints imposed on values of the syntax elements in the bitstream. Levels allow for standard-compliant low-complexity encoder and decoder implementations. The different profiles and level parameters describe the capabilities of the coder. Several preferred mode combinations for operation are defined and structured into profiles of support.
In this thesis H.263 up to profile 3 level 10 and the H.264/AVC baseline profile designed for low complexity and low rate applications are investigated. At the beginning the research was focused on H.263 (cf. Section 4.6), because it is a freely available mandatory codec in UMTS networks. Later the research attention was concentrated on H.264/AVC (cf. Section 4.5) because of its efficiency. Moreover, the processing power of user terminals increases the ability to play H.264 encoded content.

1.3.1 Video and color sampling

The initial step in video processing is sampling in temporal, spatial and color domain. Sampling in the temporal domain provides a number of pictures per second according to the frame rate, and sampling in the spatial domain provides a number of points (pixels) in each of the pictures according to picture resolution. The color sampling refers to color space (gray scale, RGB ...) and number of bits used to represent the color of a single pixel (color depth).

Frame Rate (FR), or frame frequency, refers to unique consecutive images called frames produced within one time unit. The frame rate is most often expressed in frames per second (fps) or alternatively, in Hertz (Hz). In low-rate and low-resolution applications the frame rate is reduced before the actual transmission to save data rate. Frame rate reduction can be performed by decimating the frame rate by a factor F — leaving each Fth frame while removing the rest. A typical example is mobile video streaming or call/conferencing with usual frame rates decimated by $F = 2,3,4$ or even 5. Other frame rates can be obtained by interpolation and subsequent decimation.
Each frame consists of *pixels*. Pixels of intensity pictures (black-and-white) are scalar values; pixels of color pictures are represented by coordinates within the relevant color space. The captured RGB

Abbreviation	Size	Description
VGA	640×480	Video Graphics Array
QVGA or SIF	320×240	Quarter Video Graphics Array, called also Standard Interchange Format (SIF)
Q2VGA	160×120	
CIF	352×288	Common Intermediate Format (quarter of resolution 704×576 used in PAL)
QCIF	176×144	Quarter Common Intermediate Format

Table 1.2: Typical picture resolutions in pixels, used for mobile video services.

picture is thus represented by three $N \times M$ color component matrices consisting of q-bit long numbers (usually $q = 8$). In Table **1.2** the most common resolutions for mobile video are summarized. Since the human visual system is less sensitive to color than to luminance (brightness), bandwidth can be minimized by storing more luminance detail rather than color detail. At normal viewing distances, there is no perceptible loss incurred by sampling color details at a lower rate. In video systems, this is achieved by using the color difference components. The signal is divided into a luminance (denoted as Y, called shortly 'luma') and two color difference (chrominance) components, denoted as U and V (or Cb and Cr, respectively), called shortly 'chroma'. The YUV signals are created from an original RGB (red, green and blue) source as follows [40]. The weighted values of R, G and B are added together to produce a single Y signal, representing the overall brightness, or luminance, of that spot:

$$Y = k_r \cdot R + k_g \cdot G + k_b \cdot B, \tag{1.1}$$

where k_r, k_g and k_b are weighting factors, with normalization: $k_b + k_r + k_g = 1$. ITU-R recommendation BT.601 [41] defines $k_r = 0.299$ and $k_b = 0.114$.

The U signal is then created by subtracting Y from the blue signal of the original RGB, and a scaling operation; and V by subtracting Y from the red, and then scaling by a different factor. The following formulas convert the RGB color space and YUV.

$$\begin{aligned} Y &= k_r \cdot R + (1 - k_b - k_r) \cdot G + k_b B, \\ U &= \frac{0.5}{1 - k_b} \cdot (B - Y), \\ V &= \frac{0.5}{1 - k_r} \cdot (R - Y). \end{aligned} \tag{1.2}$$

The basic idea behind the YUV format is that the human visual system is less sensitive to high frequency color information (compared to luminance) so that color information can be encoded at a lower spatial resolution. The subsampling scheme is commonly expressed as a three part ratio (e.g. 4:2:2) as shown in **Figure 1.5**. The most common way of subsampling, called 4:1:1 reduces the number of samples in both the horizontal and vertical dimensions by a factor of two, i.e., for four luma pixels there is only one blue and one red chroma pixel. The only color impairments is "color bleeding" (cf. Section 1.3.3), noticeable in between magenta and green color. Hence, the YUV color space subsampling is the first step to data rate reduction. The original bit rate R_{raw}

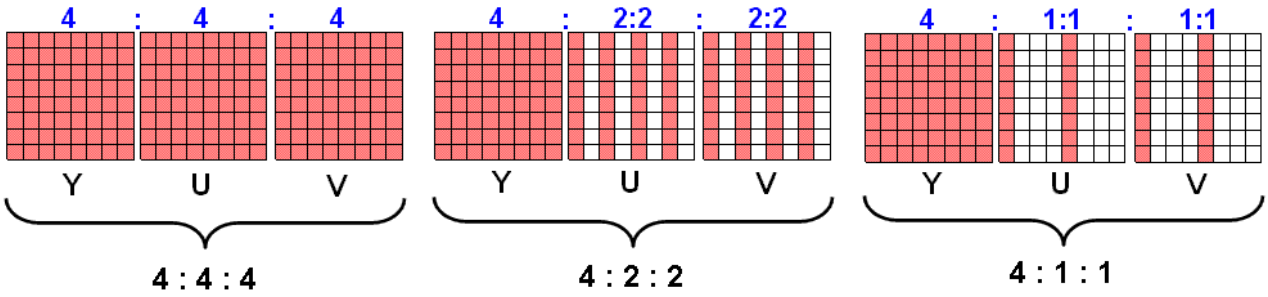

Figure 1.5: YUV — subsampling.

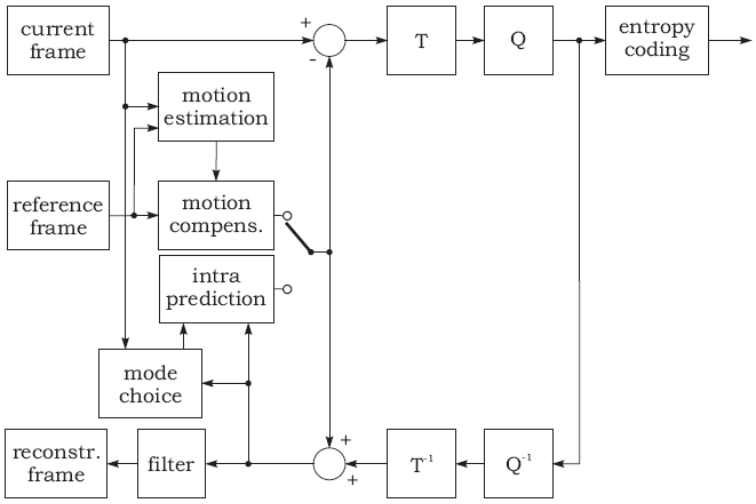

Figure 1.6: Dataflow diagram of H.264/AVC encoder.

of the *raw* (uncompressed) RGB video with frame rate f_r and picture resolution $M \times N$ is given by $R_{\text{raw\_RGB}} = 3 \cdot f_r \cdot M \cdot N \cdot q$; the corresponding raw YUV (4:1:1 or 4:2:0 YUV formats) video only requires rate $R_{\text{raw\_YUV}} = 1.5 \cdot f_r \cdot M \cdot N \cdot q$. For the QCIF resolution with 25 f/s and 8 - bit long numbers ($q = 8$) the necessary bit rate is $R_{\text{raw\_YUV\_QCIF}} = 7.6$ Mbit/s. Although the bit rate is reduced to 50% of the RGB raw video, it is still not feasible for any today's Internet or mobile application. To overcome this, further compression is employed to reduce data rate as described in the following sections.

1.3.2 Compression mechanisms

The newest video coding algorithms support a hybrid of temporal and spatial prediction, together with transform coding. Their dataflow diagram is depicted in **Figure 1.6**. Each frame is split into non-overlapping areas — *macroblocks* (MB) — consisting of 16×16 samples of the luma and 8×8 samples of each of the two chroma components. The macroblocks are organized in *slices*, representing subsets of macroblocks that can be decoded independently.

Frames are called *intra-coded* if they are encoded by means of a *spatial prediction* without using information other than that contained in the picture itself. Typically, the first picture of a video sequence is intra-coded as well as all random access points of the video sequence (pictures that can be fast accessed without decoding previous parts of video sequentially). Each macroblock in an intra-coded frame (called also intra-frame or I frame). It is predicted using spatially neighboring samples of previously coded macroblocks[1]. The encoder performs a mode choice — it decides which and how neighboring samples are used for intra prediction. The chosen intra prediction type is then signalized within the bitstream.

For all remaining pictures of a sequence between random access points, typically *inter-coding* is used, employing temporal prediction from other previously decoded pictures. First, the *motion estimation* of each block is performed by searching the best matching region from the previous or following frame(s). Note that the best match is not searched in the original (uncompressed) block, but rather in the quantized and filtered block. This prevents artifacts during the reconstruction process. The best match is taken as a prediction of the encoded block. Such a prediction is thus called *motion compensated*. Each inter-coded macroblock is a subject to further partitioning into fixed-size blocks (16×16 luma samples corresponding to no partitioning, 16×8, 8×16 or 8×8) used for motion description. Blocks of size 8×8 can be split again into submacroblocks (SMB) of 8×4, 4×8, or 4×4 luma samples. Chrominance parts are segmented correspondingly.

The next advanced feature in recent video coding algorithms is multi-picture motion compensated prediction — more than one previously coded picture can be used as a reference. The accuracy of motion compensation is a quarter of a sample distance. The prediction values at half-sample positions are obtained by applying a one-dimensional six tap Finite Impulse Response (FIR) filter. Prediction values at quarter-sample positions are generated by averaging samples at integer- and half-sample positions. To enable unambiguous reconstruction at the receiver, the *Motion Vector* (MV) between the position of the block within the frame and the position of its best match in the previously encoded frame has to be signalled as well as the mode of segmentation and corresponding reference frame(s). To avoid signalizing of the zero motion vectors and zero residuals in the cases of static picture parts, the *SKIP mode* allows for skipping of signalized number of P/B macroblocks. In SKIP mode neither residuals, nor motion vectors are sent. At the receiver, the spatially corresponding macroblock from the previous frame is taken.

Inter-coded frames are referred to as inter-frames or P and B frames; P being the frames which use for prediction only previous frames, B being the *bi-directionally predicted* frames that use for prediction also successive frames. In H.264/AVC, other pictures can reference B frames for the motion estimation. The substantial difference between P and B macroblocks is that B MBs may use a weighted average of two distinct motion-compensated prediction values for building the prediction signal. The H.264/AVC supports frames with such mixed I,P and B slices. Moreover, P and B slices may even contain some I MBs.

All luma and chroma samples of an MB are either spatially or temporally predicted and the resulting *prediction residuals* (difference between the macroblock samples being encoded and their

[1]Macroblocks that do not have any previously encoded neighbors (e.g. the first MB in picture and MBs at the top slice boundary) are encoded without prediction.

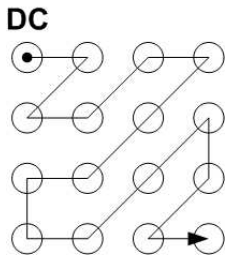

Figure 1.7: Scanning of samples in a 4×4 submacroblock.

prediction) are transformed. Depending on the type of residual data that is to be coded more transform types can be supported:

- a Hadamard transform for the luma Direct Current (DC) coefficients,
- a Discrete Cosine Transformation (DCT); the result of the transformation is a matrix of *coefficients* corresponding to different spatial frequencies. The coefficient corresponding to the lowest frequency is denoted DC, the others are Alternating Current (AC) coefficients.

All coefficients are further quantized. For each macroblock the quantization is controlled by the Quantization Parameter (QP) ranging from zero to 52. The quantization indices are scanned in zig-zag order and finally entropy encoded, together with other signalling information. The zig-zag order used in H.264/AVC is illustrated in **Figure 1.7**.

A macroblock can be coded in one of many possible modes that are enabled, depending on the picture/slice type. The *mode decision* is performed at the encoder, i.e., it is not within the scope of a standard[2]. Additional important gains in coding efficiency become possible if a macroblock mode decision is performed carefully. However, the additional gains can be extracted only at the expense of considerable increase in encoding complexity for example by implementing a Rate-Distortion Optimization (RDO) at the encoder.

If the encoder parameters (QP, MV search area, etc.) are kept during the encoding, then the number of coded bits produced for each macroblock will change depending on the content of the video frame and the mode decision, causing the bit rate of the encoder to vary. This variation in bit rate can cause problems especially in systems where resources are shared (e.g. wireless systems), the resource management cannot efficiently perform the resource allocation. The Variable Bit Rate (VBR) produced by an encoder can be smoothed by buffering the encoded data prior to transmission in a FIFO (First-In/First-Out) buffer, which is emptied at a Constant Bit Rate (CBR) matched to the channel capacity. Another FIFO buffer is placed at the input to the decoder and is filled at the channel bit rate and emptied by the decoder at a variable bit rate since the number of bits to be extracted per frame is varying over frames, but still the frames have to be rendered on the display with a constant frame rate. However, the cost is the buffer storage capacity and delay — the wider the bit rate variation, the larger the buffer size and decoding delay. Another possibility to compensate the VBR is the *rate control* using the quantizer adaptation. However, quantizer changes need to be carefully restricted

[2]Video codec standards define the functions and features of the decoder rather than the encoder.

based on scene complexity, picture types, and coding bit rate to maintain an acceptable end-user quality.
In this thesis, focus is given on the H.263 profile 3 level 10 and H.264/AVC baseline profile. Note that both of them do not support B slices, only I and P frames are possible. Other baseline profile constraints will be discussed later, when necessary for the particular application.

1.3.3 Compression artefacts

The compression algorithms of the various video coding standards are quite similar. Most of them rely on motion compensation and on a block-based DCT with 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 visual distortion, but affect the encoding process indirectly.
A variety of artefacts can be distinguished in a compressed video sequence:

The blocking effect or **blockiness** refers to a block pattern in the compressed sequence (see **Figure 1.8**). It is due to the independent quantization of individual blocks (usually of 8×8 pixels in size) in block-based DCT coding schemes, leading to discontinuities at the boundaries of 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. Recent codecs such as H.264 employ a deblocking filter to reduce the visibility of the artifact.

Figure 1.8: Blocking effect.

Blurring manifests itself as a loss of spatial detail and a reduction of edge sharpness (see **Figure 1.9**). It is due to the suppression of the high-frequency coefficients by coarse quantization.

Jerkiness or Jagged motion refers to object motion disorder. It is usually caused by low temporal resolution or insufficient motion compensation. Typical example is a 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.

Colour bleeding is the smearing of colours between areas of strongly differing chrominance, see **Figure 1.10** a). It results from the suppression of high-frequency coefficients of the chroma components. Due to chroma sub-sampling, colour bleeding extends over an entire macro block.

Figure 1.9: Blurring.

Figure 1.10: a) Colour bleeding b) Slanted lines.

The 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.

Slanted lines often exhibit a staircase effect, see **Figure 1.10**b). This is due to the fact that DCT basis images are best suited to the representation of horizontal 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.

Chrominance mismatch is associated with inaccurate motion compensation. Motion compensation is performed only for luminance component, yet the same motion vector is used for the chroma components.

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 coding differences for 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 (DCT based), many of them are observed with other compression algorithms as well. The forthcoming analysis informs about the tolerance of the user with different artefacts. The relevant artefacts for streaming over DCT-based video coding are blockiness, blurriness and jerkiness.
In wavelet based compression, for example, the transformation is applied to the entire image; therefore, none of the block-related artifacts occur. Instead, blurring and ringing are the mostly appearing distortions.

Essential to be mentioned in this context is error-resilience (combating, e.g. packet delays or losses). With unlimited resources of processing power and storage at the UE the video quality can be improved considerably. On the other hand, due to limitations of the UE, the video quality can be enhanced only to a certain point. Therefore, transmission errors and coding artifacts remain visible.

1.4 Subjective video quality

OTHER than objective video quality, the subjective video quality reflects the subjective perception of individual viewers. The evaluation is performed by a psycho visual experiment and therefore is influenced by the following subjective and objective factors: video content, encoding parameters, usage scenario and network performance. Moreover, objective parameters or QoS parameters [43] are only poorly correlated with subjective quality. The QoS is typically understood as a measure of "the collective effect of service performance which determine the degree of satisfaction of a user of the service" [43]. Service performance is then managed through a set of network parameters and the end user is not aware of network impairments [8]. The subjective video quality belongs to a higher abstraction layer also called Quality of Experience (QoE). The QoE basically relates to a higher abstraction layer when compared to QoS [44], [45], [46], [47] and reflects "the user's perceived experience of what is being presented by the application layer, where the application layer acts as a user interface front-end that presents the overall result of the individual quality of services" [44]. It can be considered as a perceptual layer and an extension to the application layer defined in the OSI Model [48].
Video and audio-visual quality reflects the most significant part of QoE for mobile video services. The estimation of subjective video quality for mobile scenarios is a challenge. Up to today, several methods were proposed for the estimation of video quality. Such methods can be classified as follows:

- The first quality distortion measures are Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR). Both of them poorly reflect the subjective video quality [4], [5]. Nevertheless PSNR is still widely used as reference method for comparing performance of video coding algorithms.
- Metrics based on Human Visual System (HVS) were proposed in recent years [14], [15], [17], [49], [50]. The usage of a metric based on the HVS is expected to be very general in its nature and applicability. These metrics compute a distance measure based on the outputs of a multiple

channel cortical model of the human vision which accounts for known sensitivity variations of the HVS in the primary visual pathway. Moreover, the metrics assume that the multiple channels mediating visual perception are independent of each other. However, recent neuroscience findings and psychophysical experiments [51], [52], [53] have established that there is interaction across the channels and that such interactions are important for visual masking. Thus, in the close future even better HVS models for reliable quality evaluation are expected. The main disadvantage of these HVS models is their high computational complexity.

- Metrics based on a set of objective parameters —[4], [18], [19], [20], [21], [63], [64]— provide a good trade-off between accuracy and complexity. The parameter set consists of quality sensitive objective parameters. This approach is very suitable for quality estimation in scenarios with defined usage, content and video service conditions.

A second classification is possible, depending on the required knowledge of the source material:

- Reference-based metrics [54], [55]: measurements based on the computation of differences between the degraded and the original video sequences. The differences can be used to compute comparative distortion measures, which have a low correlation with the perceived impairment but are easy to extract. The reference is required at the input of the measurement system strongly restricting their applicability.

- Semi-reference-based metrics: measurements obtained by computing a set of parameters on the degraded picture and comparing them with the same parameters computed on the reference picture [14], [15], [18], [64]. Quality indications can be obtained by comparing parameters computed separately on the coded pictures and the reference pictures. These parameters can be distributed in the network at low bit rates to be used when the entire reference signal is not available.

- Reference-free metrics [63], [112]: they do not require any knowledge of the original video source. These metrics find a basic difficulty in telling apart distortions from regular content, which is something that humans can do well by experience. Their biggest advantage is their versatility and flexibility.

The complexity of recently proposed metrics is rather high. Moreover, most of them were proposed for broadband broadcasting and Internet video services. In contrast to those, the proposed approach in this thesis is focused on quality estimation of low bit rate and low resolution videos in mobile environment.

1.5 Outline of the thesis and contributions

THE scope of this thesis is to estimate subjective video quality for mobile video content. Initial step to this research is to perform an extensive video quality survey. The extensive survey has to reflect the test methodology and usage scenario. Moreover, subjective video quality surveys are complex and time-consuming both in their preparation and execution. Therefore, automatic

methods are needed for video quality assessment. An ideal system needs to be able to measure video impairments like a representative sample of human observers do.
The further work continues in investigation of motion and content features of video sequences. This analysis provides a set of content/motion dependent parameters which were successfully mapped on subjective parameters. Finally, video and audiovisual quality estimation methods were proposed for mobile video scenarios.
In the following the organization of this thesis is introduced in detail and the contributions of the author are highlighted. Throughout the document, the publications (co)authored by the author of this thesis are marked by blue color.

Chapter 2 presents usage scenario and methodology for subjective testing of mobile video services. The mobile video streaming domain offers a large choice of parameters, using a variety of proprietary and standardized decoders, players, streamed content and UEs. These factors influence the subjective evaluations [57]. Moreover, the aim of the test methodology is to provide a real world viewing and usage conditions for subjective testing and to make them reproducible. Therefore, it is very important to focus initial investigations at defining the usage scenario and test methodology. ITU-T P.910 [9] and ITU-T P.911 [10] propose even more methodologies. The most suitable experimental method, among those proposed in the ITU-T recommendation, is ACR, also called Single Stimulus Method. This method imitates the real world scenario, because the customers of mobile video services do not have access to original videos (high quality versions).
Moreover, in this thesis an introduction to the state of the art test methodology is given for subjective quality evaluation on handheld devices [4], [56]. In order to emulate real conditions of the mobile video service, all the sequences were displayed on mobile handheld devices [4], [56]. In this single point the proposed methodology is not consistent with ITU-T recommendation. Finally, the most frequent content classes [64] for mobile streaming scenario are defined.

Chapter 3 presents results of extensive surveys on video quality. In some cases (e.g. choosing a preferred trade-off between FR, BR and image quality), there is a large number of possible outcomes and the test designer is faced with the choice between running a very large number of tests in order to obtain a fine-grained result or limiting the number of tests at the expense of discretizing the result [13]. Moreover, the subjective testing is extremely man-power and time consuming. The tests performed within this thesis cover wide ranges of video and audio codecs and their settings. The obtained MOS data for defined content classes and resolutions are presented separately. The obtained MOS results clearly show that video quality is content dependent [4], [100], [93], especially at low bit rates. The video quality surveys allow to estimate which coding parameters should be used, according to the character of their video content, in order to maximize the end-users perceived quality.
Moreover, Willingness-To-Pay (WTP) results are introduced. Technically, WTP has emerged from Subjective or Perceived/Perceptual QoS (PQ) [61] and is solely based on human perception or satisfaction regarding service usability. Determining PQ is typically carried out by surveying a set of persons, which participate in a controlled experiment [62]. There were already complex QoS studies regarding WTP [65] and proposals of WTP utility functions for Internet streaming [66]. Unfortunately, these results are not applicable for mobile streaming due to significantly different

usage scenarios. Within this thesis the most fundamental WTP features are [67] investigated, where we defined WTP as "readiness to pay for the provided quality" of video streaming. Finally, we introduced a single metric with very good performance [67].

Chapter 4 introduces novel methods for estimation of subjective video quality and content classification for low-resolution video sequences as they are typical for mobile video streaming. In the recent period, several objective metrics for perceptual video quality estimation were proposed. The proposed metrics can be subdivided into two dominant groups: human vision model based video metrics [14], [15], [16], [17] and metrics based on empirical modeling [18], [19], [20], [21]. The complexity of these methods is quite high and they are mostly based on spatial features, although temporal features better reflect perceptual quality especially for low-rate videos. Most of these metrics were designed for broadband broadcasting video services and do not consider mobile video streaming scenarios.

Due to content dependent video quality it is necessary to design features which allow temporal content segmentation and content classification of video streams [64]. The temporal content classification was introduced in order to estimate quality for single sequences within a video stream. For this purpose an adaptive metric for scene change detection [84] was developed. The next important approach was to estimate Content Classes (CC) for content specific metrics [4], [64]. For this purpose we investigate the motion and color sequence features and content sensitive parameters. These parameters are inputs for content classification based on hypothesis testing. The proposed content classifier is a robust tool for content classification. Moreover, hypothesis testing allows very fast extension by adding a new content class.

The proposals for quality estimation are trade-offs between applicability, processing demands and prediction accuracy. The aim was to estimate quality at the receiver with reasonable processing complexity. Furthermore, the proposed estimation methods demonstrate that it is possible to predict video quality for wireless video streaming scenario with reference-free video estimators, if the chosen parameters are those that most significantly influence the subjective quality. The relevance of the selected parameters was considered according to results of a multivariate analysis. This knowledge was successfully applied for SIF-H.264 resolution to the proposed estimators based on content adaptive motion parameters which are derived from MV features. Three reference-free estimation methods were proposed. The first method [64] estimates video quality in two steps: the content class is estimated from the original video sequence at the sender side, and then the quality metric is calculated at the receiver with almost zero complexity. The second and the third estimation methods are suitable for stand-alone estimation at the receiver side. The second, the ensemble based metric [94] exhibits a performance similar to the content class based metric. The content classification can be understood as a pre-estimation in order to obtain a more homogeneous set of results within one content class, which allows for more accurate quality estimation. This effect was achieved by introducing cross-validation in ensemble based metrics. Furthermore, the direct motion proposal [63] has a slightly worse estimation performance but allows full reference-free estimation for all content classes. The performance of the introduced video quality metrics shows good agreement between estimated MOS and the evaluation set. The proposed estimation methods for SIF resolution and proposed content clas-

sifier were submitted for a patent [122].
Moreover, for QCIF-H.263 resolution a direct reference-free [4] and an Artificial Neural Network (ANN) [112] - based estimator were proposed. The direct reference-free quality metrics are dedicated to certain content classes. On the other hand, the ANN model is general for all content classes and its training performance is sufficient for video quality estimation.

Chapter 5 focuses on estimating the audiovisual quality of mobile multimedia at the user-level. Several auditory and visual models are often utilized as basis for multi-modal predictions [117], [118]. They consider how audio and video signals are perceived by people. In this way the audio and video signals are perceptually weighted before they are combined in a multi-modal model.
Within this thesis also audiovisual quality was investigated. The scope was to estimate audiovisual quality for mobile streaming services. The audiovisual quality assessments show that audio quality, video quality and sequence character are important factors to determine the overall subjective perceived quality. A mutual compensation property of audio and video can also be clearly seen from the obtained results. In order to predict the audiovisual quality of a multimedia system it is necessary to propose a metric that takes into account both the audio quality and the video. Moreover, the cross-modal interaction between audio and video mode is taken into account [118]. The proposed audiovisual metrics [100] for speech and non-speech content takes into account the cross-modal interaction.

Chapter 6 summarizes the novel achievements within this thesis in the field of test methodology and video quality estimation for mobile video services. Moreover, it provides an outlook to their possible deployment and ends with some general remarks.

Appendix A contains a list of the abbreviations employed throughout this thesis.

Appendix B provides a description of selected objective video parameters defined by ANSI T1.803 [18].

Appendix C presents a brief overview of well known audio quality estimation methods [113] and [115] for subjective audio quality estimation.

Chapter 2

Test methodology

Contents

MOBILE video streaming scenarios are specified by the environment of usage, streamed content, and the screen size of the mobile terminals [56]. Therefore, mobile scenarios are strictly different in comparison with classical TV broadcasting services or broadband IP-TV services. Most of the recommendations for subjective video testing [9], [12], are designed for broadband video services with QCIF resolution or higher and for static scenarios. On the contrary, the mobile scenario is different due to technical conditions and usage. Therefore, the initial research scope focused at the design of a methodology for subjective video testing.
This chapter is dedicated to the design of a subjective test methodology for mobile video services and the description of content features of most frequent content classes. Finally, the results of video quality surveys are presented and analyzed.

2.1 Test methodology

THE aim of the test methodology was to provide a real world viewing and usage conditions for subjective testing and make them reproducible. ITU-T P.910 [9] and ITU-T P.911 [10] propose even more methodologies. The difference between such methods is in utilizing of explicit references and methods that do not use any explicit reference. The non-reference methods Absolute Category Rating (ACR) and Pair Comparison (PC) do not test video system transparency or fidelity. On the other hand the reference methods should be used when testing the fidelity of transmission with respect to the source signal. This is frequently an important factor in the evaluation of high quality systems. The reference method also called Degradation Category Rating (DCR) has long been a key method specified in [12], for the assessment of television pictures whose typical quality represents extreme high quality levels of videotelephony and videoconferencing. The specific comments of the DCR scale (imperceptible/perceptible) are valuable when the viewer's detection of impairment is an important factor. Thus, when it is important to check the fidelity with respect to the source signal, the DCR method should be used. DCR should also be applied for high quality system evaluation in the context of multimedia communications. Discrimination of imperceptible/perceptible impairment in the DCR scale supports this, as well as a comparison with the reference quality.
On the other hand, ACR is easy and fast to implement and the presentation of the stimuli is similar to that of the common use of the systems. Thus, ACR is well-suited for qualification tests. The principal merit of the PC method is its high discriminatory power, which is of particular value when several of the test items are nearly equal in quality. When a large number of items is to be evaluated in the same test, the procedure based on the PC method tends to be lengthy. In such a case an ACR or DCR test may be carried out first with a limited number of observers, followed by a PC test solely on those items which have received about the same rating.

To achieve the most effective methodology for subjective testing of wireless video streaming, the following conditions were defined:

- Viewers do not have access to the test sequences in their original uncompressed form. Only encoded sequences are displayed, a reference-free subjective evaluation is obtained.

- The sequences are presented on a handheld mobile device (**Figures 2.1** and **2.2**).
- The encoding settings reflect a typical UMTS streaming setup (see Tables **3.1** and **3.2**).
- The most frequent streaming content types are displayed.

Figure 2.1: Cell phone with display resolution 176×220 pixels.

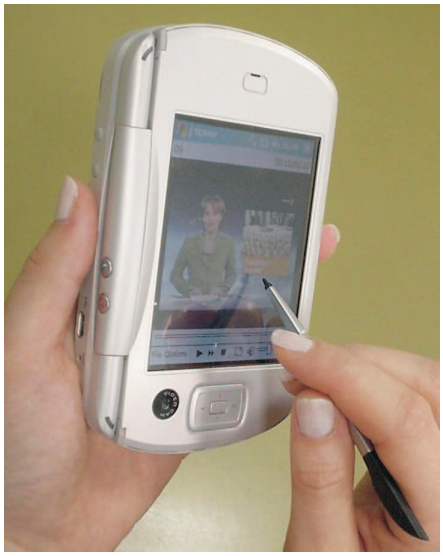

Figure 2.2: PDA with display resolution 480×640 pixels.

2.1.1 Video quality evaluation

The proposed test methodology is based on ITU-T P.910 [9] and adapted to our specific purpose and limitations. For this particular application it was considered that the most suitable experimental method, among those proposed in the ITU-T Recommendation, is ACR, also called Single Stimulus Method. The ACR method is a category judgment in which the test sequences are presented one at a time and are rated independently on a category scale. Only degraded sequences are displayed,

Viewing Distance	20 - 30 cm
Viewing Angle	0°
Room illumination (ambient light level [lux])	low: ≤20 lux at about 30 cm in front of the screen

Table 2.1: Test environment.

and they are presented in arbitrary order. This method imitates the real world scenario, because the customers of mobile video services do not have access to original videos (high quality versions). On the other hand, ACR introduces a higher variance in the results, as compared to other methods in which also the original sequence is presented and used as a reference by the test subjects. The results of quality assessments often depend not only on the actual video quality, but also on other factors such as the total quality range of the test conditions. A description of reference conditions and procedures to produce them is given in Recommendation P.930 [12]. That ITU-T recommendation proposes LCD monitors for subjective testing. However, the mobile video streaming domain offers a large choice of parameters, and uses a variety of proprietary and standardized decoders, players, UEs as opposed to standard broadband video services (IP-TV, DVB-T ...) where the system parameters do not vary so much. Therefore, it is far more difficult to evaluate the quality of multimedia images than those in the broadband video services [57]. Experience shows (see **Figure 2.3**) that this combination strongly influences the final subjectively-perceived picture quality. At **Figure 2.3** comparison of subjective evaluations for semantically identical soccer video content are depicted. The sequences are encoded by a H.264/AVC baseline profile codec. The evaluation at LCD monitors were performed within the formal verification tests on H.264/AVC defined by JVT [58]. Evaluations at PDAs were performed within this thesis (cf. Section 3). The results clearly show that test subjects evaluate much more critical the sequences at LCD monitors.

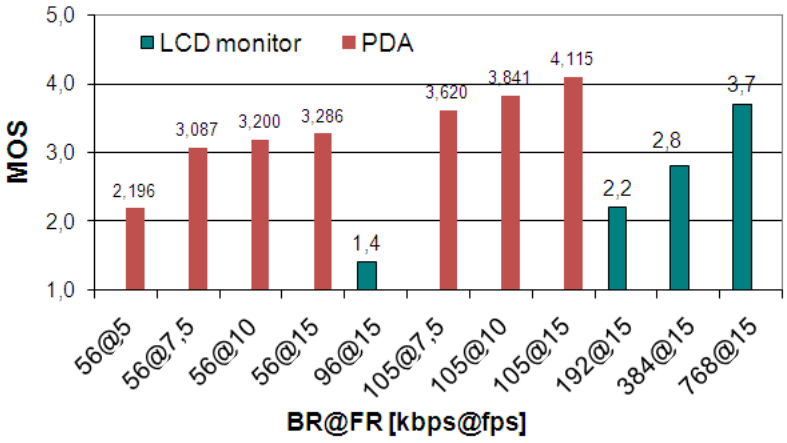

Figure 2.3: Subjective evaluation at LCD monitor and PDA.

In order to emulate real conditions of the mobile video service, all the sequences were displayed on mobile handheld devices. Viewing distance from the phone is not fixed, but selected by the test person. We have noticed that users are comfortable to take UMTS terminal at a distance of 20-30 cm.

Our video quality test design follows these experiences in order to better reflect real world scenarios. The test environment (see Table **2.1**) fulfills all criteria given by ITU-T P.910 [9]. The critical part was to achieve suitable lightning conditions, in order to eliminate UE display reflection.

After each presentation the test subjects were asked to evaluate the overall quality of the sequence shown. In order to measure the quality perceived, a subjective scaling method is required. However, whatever the rating method, this measurement will only be meaningful if there actually exists a relation between the characteristics of the video sequence presented and the magnitude and nature of the sensation that it causes on the subject. The existence of this relation is assumed. Test subjects evaluated the video quality after each sequence in a prepared form using a five grade MOS scale: "5–Excellent", "4–Good", "3–Fair", "2–Poor", "1–Bad". Higher discriminative power was not required, because our test subjects were used to five grade MOS scales (school). Furthermore, a five grade MOS scale offers the best trade-off between the evaluation interval and reliability of the results. Higher discriminative power can introduce higher variations to MOS results.

2.1.2 Audio and audiovisual quality evaluation

In order to keep consistency with the previously described methodology for video quality and the ITU-T Recommendations [10],[11] for audio and audiovisual quality evaluation, the ACR method and a five grade MOS scale was applied. Moreover, for emulating the real word conditions of the UMTS video service all the audio and video sequences were played at UE (Sony Ericsson Z1010). In this singular point the proposed methodology for audiovisual quality testing is not consistent with ITU-T P.911 [10] and ITU-R BS.1534-1 [11]. Furthermore, since one of our intentions is to study the relation between audio quality and video quality, we have decided to take all the tests with a standard stereo headset. During the training session of three sequences the subjects were allowed to adjust the volume level of the headset to a comfortable level. The viewing distance from the phone was not fixed and selected by the test person but we have noticed that all subjects were comfortable to take phone at a distance of 20-30 cm.

2.1.3 Subjective testing

At the beginning of each test round a trial run was presented with three sequences. The subjective quality of these trial sequences varied substantially, in order to offer the test subject initial experience with subjective quality evaluation. The contents and audio or video coding artifacts of these sequences were similar to video sequences. The subjective evaluation of trial run sequences was not taken into account in the statistical analysis of the test results. After this trial run the test sequences were presented in an arbitrary order, the only conditions being that two clips of the same content, though differently degradated, must not appear in succession, and that consecutive sequences must not have identical bit rate and frame rate. If, on the contrary, all the versions of one sequence were displayed in succession, subjects would perform a degradation rating rather than an absolute rating. Since the intension is the subjective evaluation on different sequences of different contents relatively to each other. Therefore, it is important to alternate the sequences.
Duration of clips was approximately 10 seconds. The length was not identical in all the clips, because

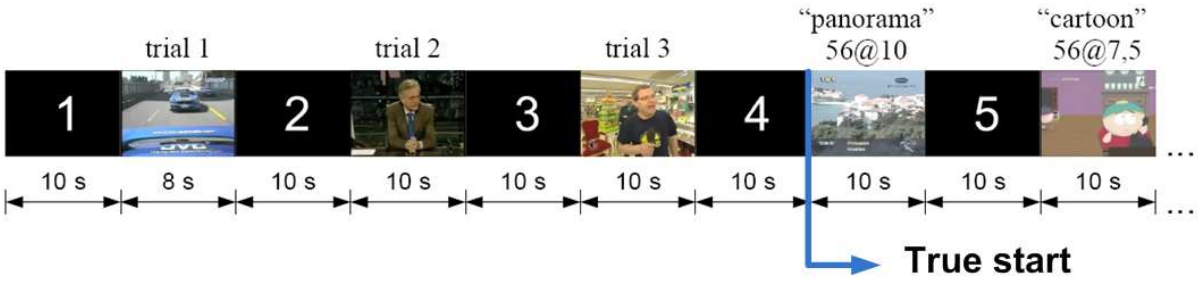

Figure 2.4: Time pattern of the video quality survey.

the sequences were adjusted to a scene cut, in order to keep the contents consistent. The voting time to respond to questions was also set to 10 seconds. A still image showing the order number of the following sequence —white big digits on a black background— was displayed during this voting time between sequences, in order to guide the test person through the questionnaire. The succession of clips was presented using a playlist, to ensure that the subject did not have to interact with the device, and could be fully concentrated in his viewing and evaluation task. The use of a playlist assures a level of homogeneity in the viewing conditions for all the viewers, as the presentations cannot be stopped, the voting time is fixed, and each test sequence is viewed only once before it is evaluated. The time pattern for the presentation of the clips is illustrated by **Figure 2.4**, showing that after third initial trials the actual evaluation process starts.

2.2 Source materials

ALL the original sequences were formatted to CIF or SIF resolutions. Source material in a higher resolution were converted to SIF resolution. For mobile video streaming the most frequent contents with different impact on the user perception were defined, resulting in the following seven classes:

Figure 2.5: Snapshot of typical content class News (CC1).

- **Content Class News (CC1)**: The first content class includes sequences with a small moving region of interest (face) on a static background. The movement in the Region of Interests (ROI) is mainly determined by eyes, mouth and face movements. The ROI covers up to approximately 15% of the screen surface.

Figure 2.6: Snapshot of typical content class Soccer (CC2).

- **Content Class Soccer (CC2)**: This content class contains wide angle camera sequences with uniform camera movement (panning). The camera is tracking small rapid moving objects (ball, players) on the uniformly colored (typically green) background.

Figure 2.7: Snapshot of typical content class Cartoon (CC3).

- **Content Class Cartoon (CC3)**: In this content class object motion is dominant; the background is usually static. The global motion is almost not present due to its artificial origin (no camera). The movement object has no natural character.

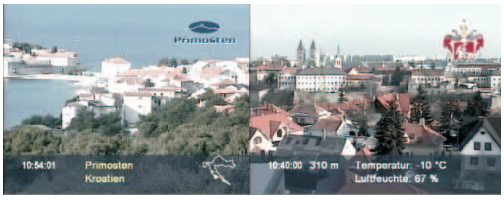

Figure 2.8: Snapshot of typical content class Panorama (CC4).

- **Content Class Panorama (CC4)**: Global motion sequences taken with a wide angle panning camera. The camera movement is uniform and in a single direction.

- **Content Class Video clip (CC5)**: The content class contains a lot of global and local motion or fast scene changes. Scenes shorter than three seconds are also associated to this content class.

- **Content Class Video call (CC6)**: Well-known professional test sequence, which contains a monologue of a man moving his head dynamically and at the end of the sequence there is

Figure 2.9: Snapshot of typical content class Video clip (CC5).

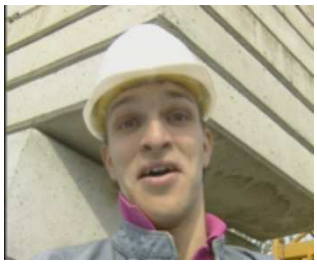

Figure 2.10: Snapshot of typical content class Video call (CC6).

a contiguous scene change. This sequence contains a lot of local and global movement. The "foreman" sequence is a typical scenario for a video call.

Figure 2.11: Snapshot of typical content class Traffic (CC7).

- **Content Class Traffic (CC7)**: The "traffic" sequence is obtained by a static traffic camera. The camera is static and slowly moving cars can be observed.

For subjective video testing on different sets of sequences (see Table **2.2**) QCIF and CIF resolutions were chosen, in order to obtain the most representative set for each resolution. Content classes CC1, CC 2, and CC4 were used for subjective testing in both QCIF and SIF resolution. Content classes CC3 and CC5 only in SIF resolution and content classes CC6 and CC7 only with QCIF resolution. In [9] the measure of spatial and temporal perceptual information is used to characterize a video sequence. The Spatial Information (SI) measurement reflects the complexity (amount of edges) of still pictures. SI is based on the Sobel filter, that is applied to each luminance frame F_n at time instance n. After that the standard deviation over the pixels is computed. The maximum value within the

Resolution	CC1	CC2	CC3	CC4	CC5	CC6	CC7
QCIF	×	×		×		×	×
SIF	×	×	×	×	×		

Table 2.2: Sequence sets for subjective tests.

whole sequence represents the spatial information:

$$SI = \max_{time_n} \left\{ \text{std}_{space_{i,j}} \left[\text{Sobel}\big(F_n(i,j)\big) \right] \right\}. \tag{2.1}$$

The temporal perceptual information measurement is based upon the motion difference feature. For every time instance n, the luminance pixel values difference is counted:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j). \tag{2.2}$$

Temporal Information (TI) is computed as a maximum over time of the standard deviation over space:

$$TI = \max_{time_n} \left\{ \text{std}_{space_{i,j}} \left[M_n(i,j) \right] \right\}. \tag{2.3}$$

In the following **Figure 2.12** SI and TI values of the original sequences are depicted. As it can be seen the spatial and temporal features of the chosen content types vary significantly.

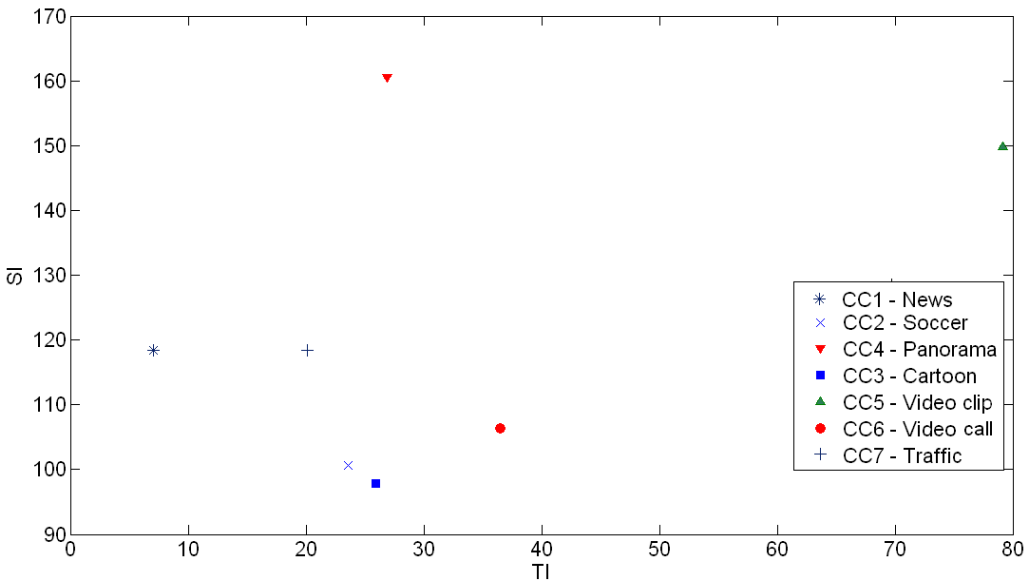

Figure 2.12: Spatial and temporal features of original test sequences.

Chapter 3

Results of subjective quality tests

Contents

THIS chapter describes the results of subjective video tests. These subjective tests are performed for mobile scenario, but for different video codecs, resolutions and Content Classes (CCs). The ACR method was used to capture MOS for a wide range of video sequences. Content was selected to cover a representative range of coding complexity and content types. The subjective MOS data was obtained according the subjective test methodology described in Chapter 2. This content was then encoded at a variety of BRs and FRs and resolutions to represent high and low medium qualities. The obtained subjective data was analyzed for each resolution and CC. Furthermore, the data was checked for consistency. The preliminary look at the obtained data shows, that the subjective video quality is strongly content dependent, especially for lower BR. This feature can be seen in tests on QCIF and SIF resolution. Furthermore, the results were used for the design of subjective quality estimators.

3.1 Subjective quality tests on QCIF resolution and H.263 codec

ALL sequences were encoded with H.263 profile 3 and level 10. For subjective quality testing we used combinations of bit rates and frame rates shown in Table **3.1**. In total, there were 60 encoded test sequences. Six sequences were excluded due to significant visual impairments in spatial and temporal domain.

FR [fps]/BR [kbps]	18	44	80
5	CC1, CC2, CC4, CC6, CC7	CC1, CC2, CC4, CC6, CC7	CC1, CC2, CC4, CC6, CC7
7.5	CC1, CC2, CC4, CC6, CC7	CC1, CC2, CC4, CC6, CC7	CC1, CC2, CC4, CC6, CC7
10	CC1, CC2, , CC7	CC1, CC2, CC4, CC6, CC7	CC1 CC2, CC4, CC6, CC7
15	CC1	CC1, CC2, CC4, CC6, CC7	CC1 CC2, CC4, CC6, CC7

Table 3.1: Tested combinations of frame rates and bit rates for QCIF resolution.

To obtain the MOS values, we worked with 38 paid test persons. The chosen group ranged different ages (between 17 and 30), gender, education and different experience with image processing. Three runs of each test were taken. In order to avoid a learning effect we made a break of half an hour between the first and the second run, and a break of two weeks between the second and the third run. There were not really noticeable differences between the first two runs and the third run, performed two weeks after.

Finally, the 95% Confidence Interval (CI) was calculated for the obtained data set, in order to comfirm consistency of the data set. The value $\overline{x_i}$ corresponds to the average MOS of the obtained data set averaged over all runs of subjective evaluations for one encoding setting of a particular sequence. The 95% CI is given by:

$$\mathrm{CI} = [\overline{x_i} - \delta_i, \overline{x_i} + \delta_i]\,, \tag{3.1}$$

where:

$$\delta_i = 1.96\frac{S_i}{\sqrt{W}}, \tag{3.2}$$

where W is the number of MOS values obtained by all runs of subjective evaluations for one encoding

setting of a particular sequence and S_i is the standard deviation for each presentation given by:

$$S_i = \sqrt{\sum_{i=1}^{W} \frac{(\overline{x_i} - x_i)^2}{W - 1}}. \tag{3.3}$$

3.1.1 Results for QCIF resolution and H.263 codec

In the further processing of our data we have rejected sequences which were evaluated with individual standard deviation higher than one. Following this rule, we excluded 2.23% of the tests results. The impact of the correction was negligible. The MOS was obtained by averaging over all the remaining evaluations. The average size $2\delta_i$ (over all tested sequences) of the 95% confidence intervals (3.1) was computed as well and is $\delta_i = 0.15$ on the five grade MOS scale.

3.1.2 Survey results for content class News (CC1)

The survey was performed on one sequence with twelve different encoding settings (see Table **3.1**). The obtained MOS results for CC1 depicted in **Figure 3.1** show very high MOS ratings even at extremely low bit- and frame-rates. This is caused by the static character of the sequence (see **Figure 2.12**) that allows very high compression in the temporal domain. Furthermore, the viewers are mainly focusing on the region of the newscaster's face, representing approximately 15% of the surface. Results above MOS grade 4 were achieved at all tested setting encoded at BR 44 kbps and higher. The top grades were achieved for the encoding combination 80@10 kbps@fps and 80@15 kbps@fps.

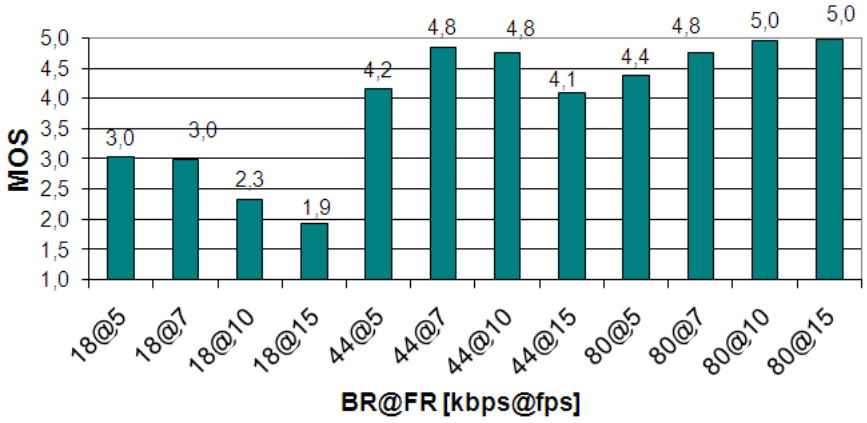

Figure 3.1: MOS results for the content class News in QCIF resolution.

3.1.3 Survey results for content class Soccer (CC2)

Subjective quality of one soccer sequences encoded with eleven different settings (see Table **3.1**) was investigated. For soccer sequences insufficient subjective results below MOS grade 3 were obtained. Moreover, for BR lower than 44 kbps the subjective quality was annoying (see in **Figure 3.2**). The first reason of such critical evaluation is that subjective quality requirements for this content type are

higher. Furthermore, high codec compressions and low video resolutions lead to ball impairment and loss of lines as well as loss of the ball itself.

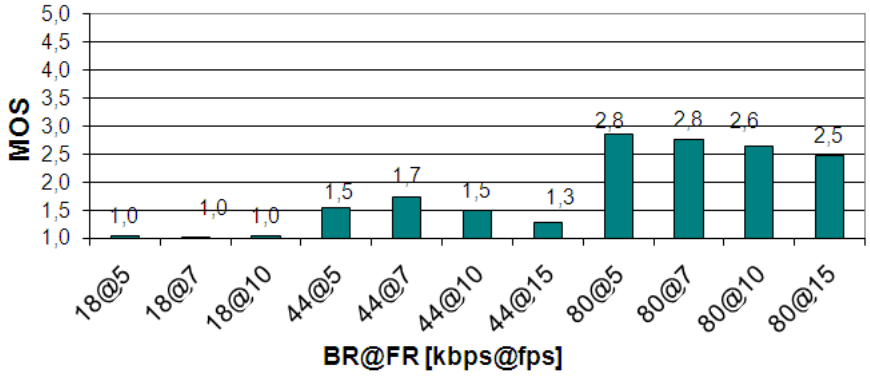

Figure 3.2: MOS results for the content class Soccer in QCIF resolution.

3.1.4 Survey results for content class Panorama (CC4)

The survey was performed on one panorama sequence with ten different encoding settings (see Table **3.1**). In panorama sequences better MOS values were obtained for low FR as are depicted in **Figure 3.3**. The best subjective ratings increase with decreasing BR, the best MOS grades were obtained for sequences encoded at 5 fps. The explanation of this so called "panorama effect" [60] is that the human vision can interpolate the uniform camera movement and test subjects were more sensitive on spatial details.

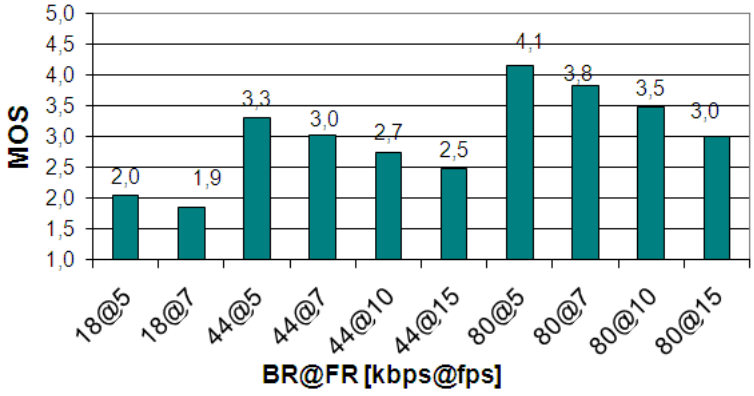

Figure 3.3: MOS results for the content class Panorama in QCIF resolution.

3.1.5 Survey results for content class Video call (CC6)

Ten codec settings were encoded in total for CC6 (see Table **3.1**). The video call (Foreman) sequence achieves significantly better MOS values at BR of 80 kbps. This sequence contains a lot of local and global movement and spatial details. The codec cannot compress this sequence as effective as CC1

because there is less redundant information in spatial and temporal domain in the Foreman sequence.

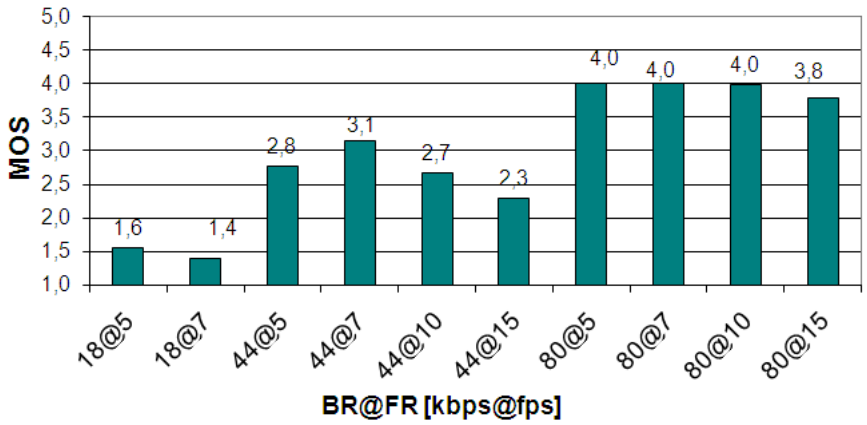

Figure 3.4: MOS results for the content class Video call in QCIF resolution.

3.1.6 Survey results for content class Traffic (CC7)

Subjective quality of one traffic sequences encoded with eleven different settings (see Table **3.1**) was investigated. The traffic sequence has a relative static character. For the test subjects a reasonable trade-off between continuous car movement and spatial details were more important. The subjective results for the traffic sequence depicted in **Figure 3.5** show very high MOS at FR 7.5 fps for BR 44 and 80 kbps.

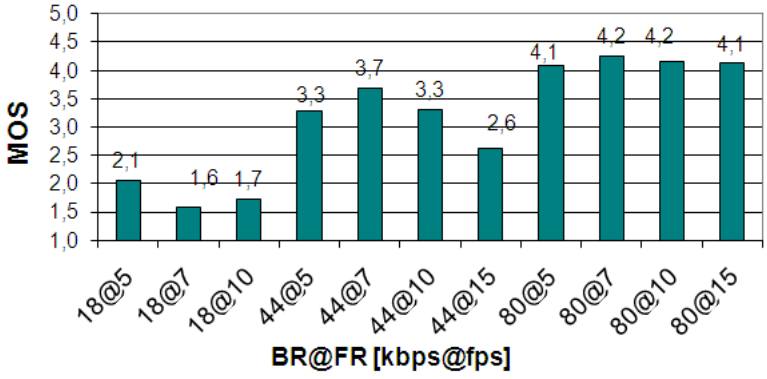

Figure 3.5: MOS results for the content class Traffic in QCIF resolution.

3.2 Subjective quality tests on SIF resolution and H.264/AVC codec

FOR the tests in SIF resolution all sequences were encoded with the H.264/AVC baseline profile 1b. For subjective quality testing frame and bit rate combinations shown in Table **3.2** were used.

In total there were 39 combinations.

FR [fps]/BR [kbit/s]	24	50	56	60	70	80	105
5	CC1, CC3, CC4	CC5	CC1, CC2, CC3, CC4				CC1
7.5	CC1, CC3, CC4		CC1, CC2, CC3, CC4	CC5	CC5		CC1, CC2,CC5
10	CC1, CC3		CC1, CC2, CC3, CC4		CC5	CC5	CC1, CC2, CC5
15	CC1		CC1, CC2			CC5	CC1, CC2, CC5

Table 3.2: Tested combinations of frame rates and bit rates for SIF resolution.

To obtain MOS values, we worked with 36 test persons for two different sets of test sequences. The first set was used for the design of a metric and the second for the evaluation of the metric performance. The training set test was carried out with 26 test persons and the evaluation test set was carried out with 10 test persons. The training and evaluation tests were collected from different sets of the five video classes. The chosen group of test persons ranged different ages (between 20 and 30), gender, education and different experience with image processing. Two runs of each test were taken. In order to avoid a learning effect, we made a break of half an hour between the first and the second run.

3.2.1 Results for SIF resolution and H.264/AVC codec

The raw ratings obtained in the subjective quality survey were scanned for unreliable results. Votes from one viewer to a certain sequence that differ two or more MOS grade from the first to the second run were considered unreliable and therefore rejected. In total, 12.3% of the results were rejected. In average the MOS values were modified maximally in 0.04 units, or 1.2%. This modification, though, had almost no influence, and thus its average effect on the test global mean score was negligible —0.005 points, or 0.16%—. Subsequently, the MOS was obtained by averaging over all the remaining votes. The average size (over all tested sequences) of the 95% confidence intervals $2\delta_i$ (3.1) was computed as well and is $\delta_i = 0.27$ on the five grade MOS scale.

3.2.2 Survey results for content class News (CC1)

The MOS results for the CC1 "News" scene (see snapshot in **Figure 2.5**) are depicted in **Figure 3.6**. The subjective quality test included 12 differently degraded versions of this video sequence. The news sequence is the most static of all test sequences. The highly static character of these test clips can also be seen in the low values of the TI parameter (see **Figure 2.12**), the lowest of all test sequences. Only a small part of the surface is in movement. Viewer are concentrated at the face of the newscaster, carrying the biggest part of the visual information contained in the clip. The four versions compressed at 24 kbps received poor scores and all the clips down-sampled at 5 fps obtained subjective MOS scores below 3. The motion in the face of the speaker is visibly slowed down and looks very unnatural at such low frame rate and bit rate, what turns out to be a very annoying effect for the viewer. The highest scores of MOS 3.8 are obtained by the configuration BR@FR = 105@7.5 kbps@fps , closely followed by 105@10 kbps@fps, 56@10 kbps@fps whith MOS 3.7 and 105@15 kbps@fps MOS 3.6, respectively. Very

interesting is the fact that the viewer seems to notice no difference in quality between the combination 56@10 and 105@10 kbps@fps, which both received very positive evaluations.

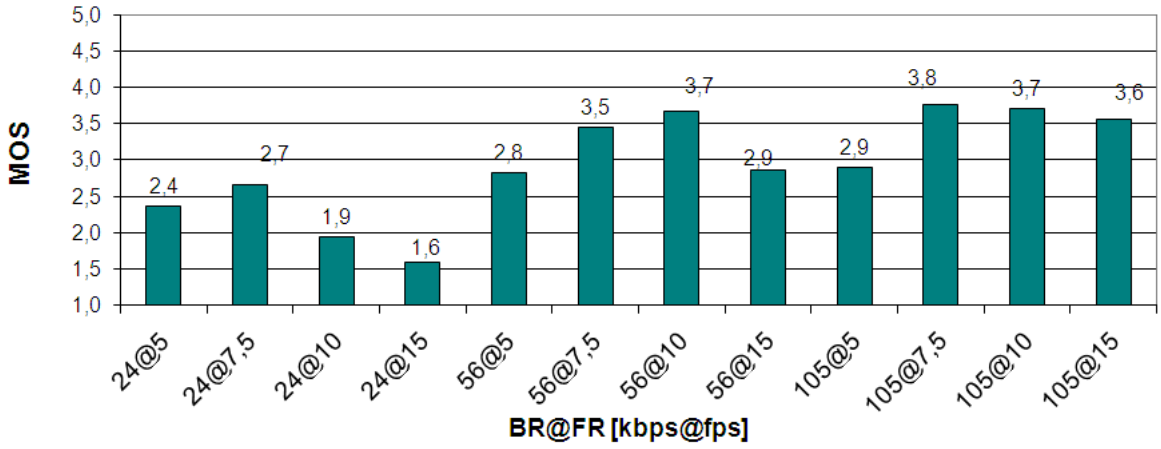

Figure 3.6: MOS results for the content class News.

3.2.3 Survey results for content class Soccer (CC2)

The "Soccer" scene (snapshot in **Figure 2.6**) is the most dynamic of the test sequences, although the TI measurement (see **Figure 2.12**) is the second lowest after the "News" sequence. This is caused by the green playground which covers more than two thirds of the screen surface. The MOS results for the "Soccer" clip are presented in **Figure 3.7**. A set of seven compressed versions of this sequence was included in the subjective test. Bit rates below 56 kbps were not used in the survey as their quality was found to be clearly too poor due to extreme blurriness, which causes the football and the lines on the playground to disappear from the image. The seven encoded versions were evaluated rather positively by the test subjects: only the configuration 56@5, in which the motion is very jerky, receives a MOS below 3. Increasing FR had always a positive effect on the perceived quality. This means that, on the contrary to what happens with other content types (especially the "News" case) in the "Soccer" sequence viewers prefer smoothness of motion rather than static quality. The best results are obtained with the configuration 105@15 with a MOS of 4.1 (0.8 points higher than the best score at 56 kbps). In conclusion, we can say that data rates below 56 kbps or frame rates lower that 7.5 fps do not seem to be appropriate to compress sequences of these characteristics. Moreover, the encoding setting with higher FR are better evaluated.

3.2.4 Survey results for content class Cartoon (CC3)

The "Cartoon" sequence (snapshot in **Figure 2.7**) is an animation. In contrast to the other four test sequences, the motion in this clip was not captured by a camera, but is artificially created. For all the six encoding configurations, the best subjective quality results were obtained (see **Figure 3.8**). In view of the results, we can say that a sequence of these characteristics can be compressed at the very low data rate of 24 kbps, obtaining a "good" perceived quality. At 56 kbps the static quality of

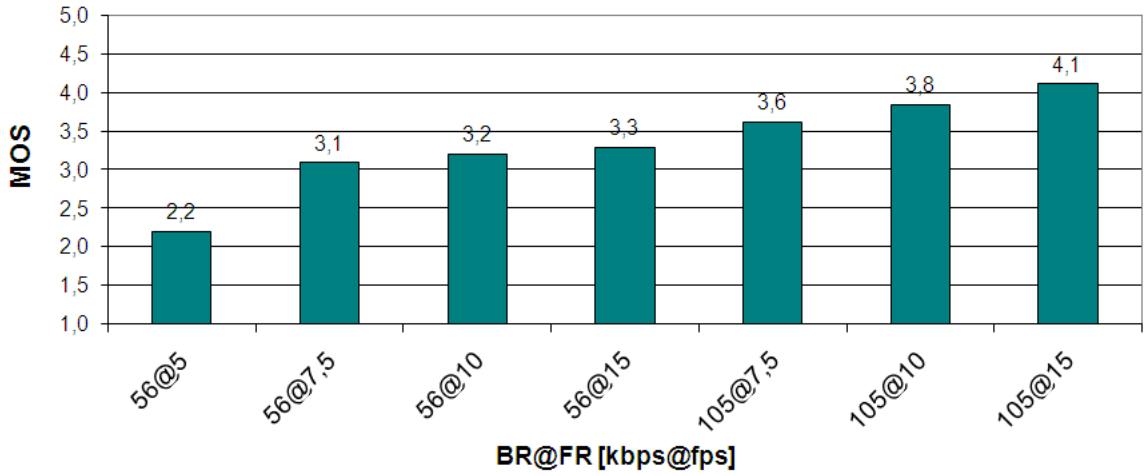

Figure 3.7: MOS results for the content class Soccer.

the images is very good and does not get perceptibly worse with the increasing frame rate. Therefore, at this data rate, the viewer's quality perception improves with the frame rate and the configuration 56@10 receives the highest score: 4.4 MOS grade, which is the best score reached in the survey.

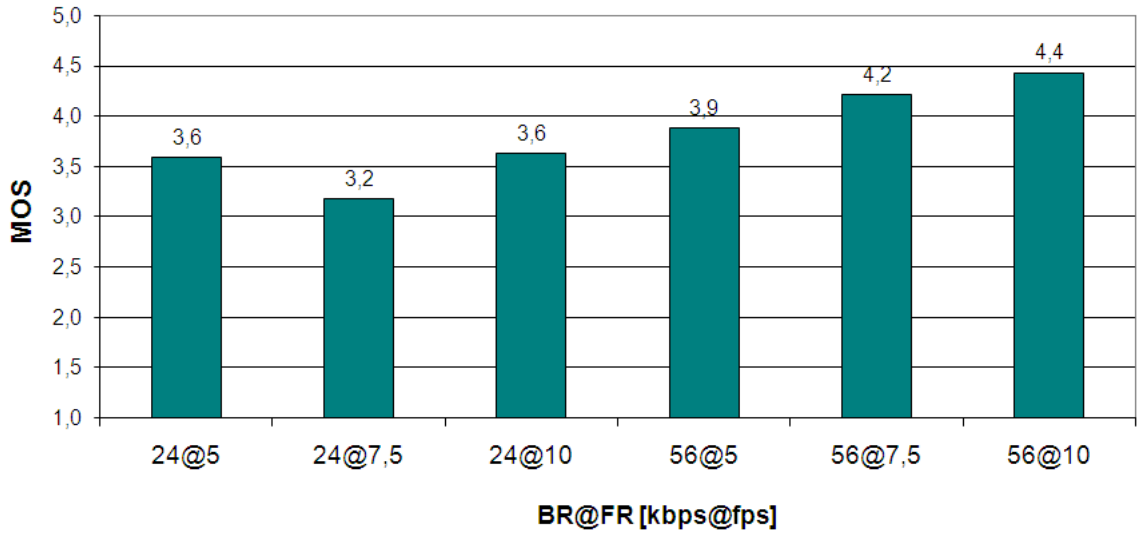

Figure 3.8: MOS results for the content class Cartoon.

3.2.5 Survey results for content class Panorama (CC4)

The "Panorama" sequence (see snapshot in **Figure 2.8**) is characterized by smoothness and uniformity of the motion. The TI values (see **Figure 2.12**) of this clip are the second highest after the "Video clip." The "Panorama" scene can be very effectively compressed due to the uniformity of the camera movement and the lack of object motion. It was considered unnecessary to include in the evaluation compressions of more than 56 kbps, as this data rate already allowed for a very good video quality.

For the very low bit rate of 24 kbps and frame rate 7.5 reaches "Panorama" an acceptable 2.8 MOS grade. The degraded clips at 56 kbps received a much better evaluation. At this data rate, the MOS values obtained by "Panorama" are only surpassed by those of the "Cartoon" clip. The configuration 56@7.5 reaches 4.1 MOS points, only 0.3 points below the maximum of the survey. **Figure 3.9** shows that there is no need for compressing above 56 kbps due to high video codec efficiency at this BR. The fact that the frame rate 7.5 fps is favored over 10 fps indicates that the user gives in this case priority to the static quality.

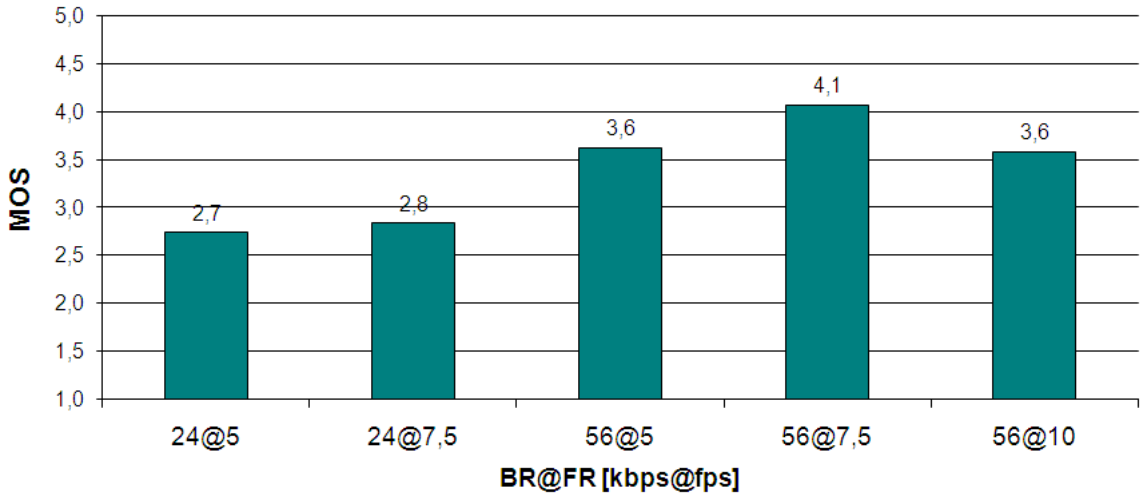

Figure 3.9: MOS results for the content class Panorama.

3.2.6 Survey results for content class Video clip (CC5)

The MOS results obtained by the "Video clip" (snapshot in **Figure 2.9**) are presented in **Figure 3.10**. Due to the presence of scene cuts and fast movement, this sequence's TI value (see **Figure 2.12**) is much higher than those of all the other test sequences. This suspects that this sequence cannot be easily compressed without loss. In total, nine different compression configurations were included in the experiment. Versions below 70 kbps received poor opinion scores are affected by severe blurriness. The combinations with FR 5 fps turns out to be annoyingly jerky and unnatural. Interesting is the fact that the viewers evaluate all configurations at 70 and 80 kbps almost identically and rather positively, showing no clear preference for a particular frame rate. The configuration 70@10 kbps@fps seems to be an acceptable trade-off between required data rate and perceived quality. The clips encoded at the highest rate 105 kbps have very good acceptance, but again no conclusions can be extracted about the most appropriate frame rate.

3.3 Summary of survey results

THE most important outcome of our tests is that human visual perception of video content is strongly determined by the character of the observed sequence. As can be seen at **Figures 3.11**

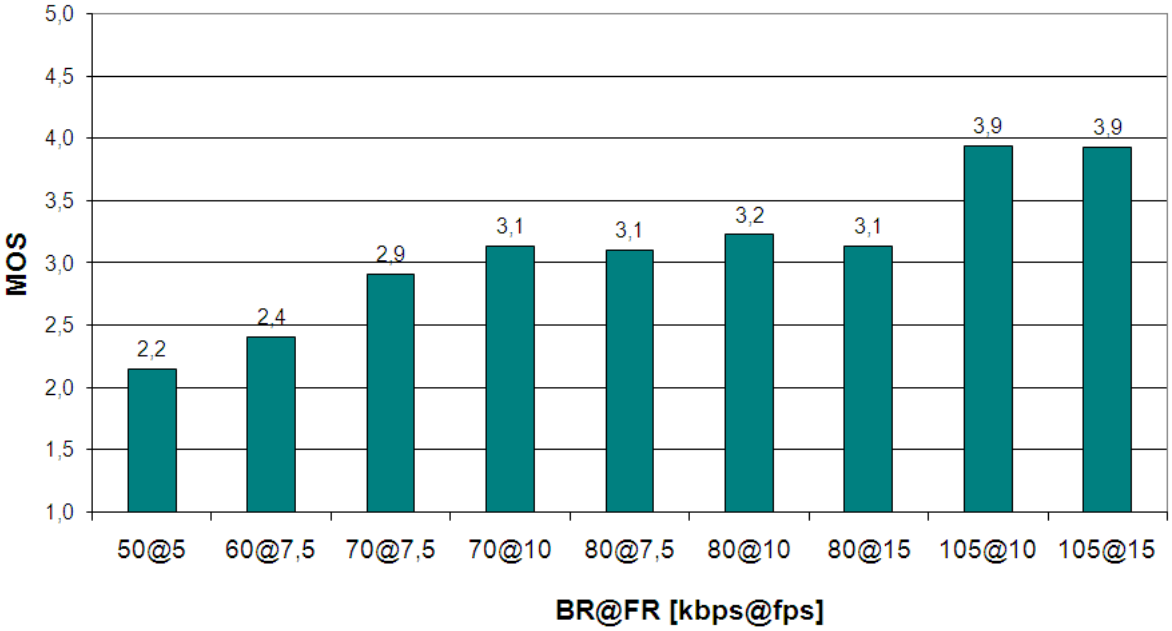

Figure 3.10: MOS results for the content class Video clip.

and **3.12**, the measured subjective video quality is strongly content dependent, especially at low BR and resolution. The difference between two contents can result in up to 3 MOS grades for the QCIF resolution and 1.6 MOS grades for the SIF resolution.

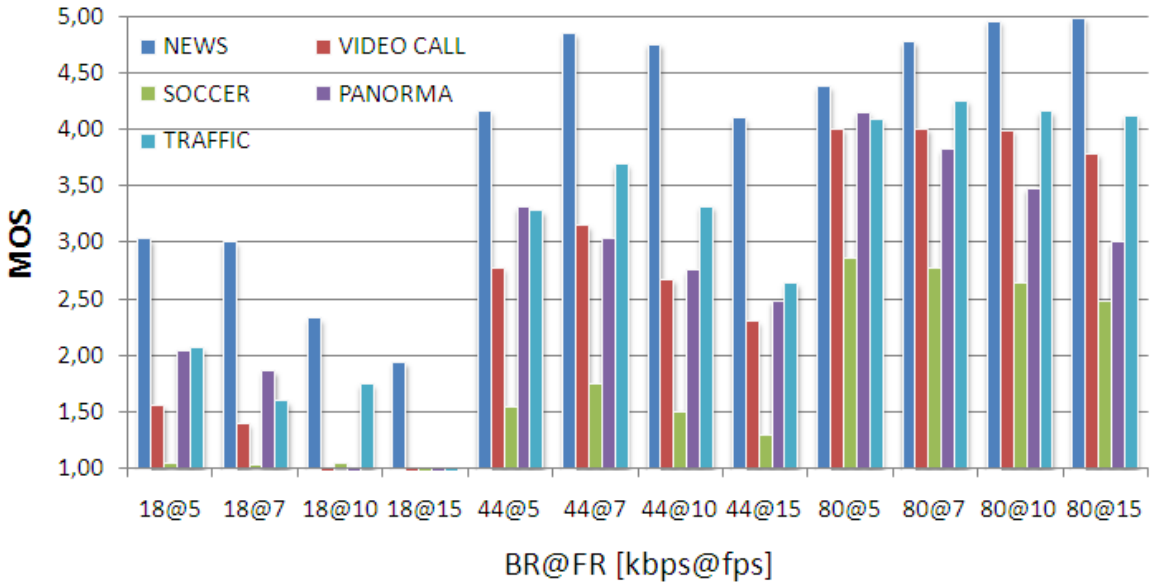

Figure 3.11: MOS for all the tested sequences for QCIF Resolution.

The subjective assessment for the QCIF resolution (see **Figure 3.11**) shows that the test subjects prefer almost for all CCs frame rates 7.5 and 10 fps. The only exception is CC4 due to the "Panorama effect" [60] for which the preferable FR is 5 fps. Moreover, it can be seen that lower FRs do not

typically decrease the subjective quality at a QCIF resolution.

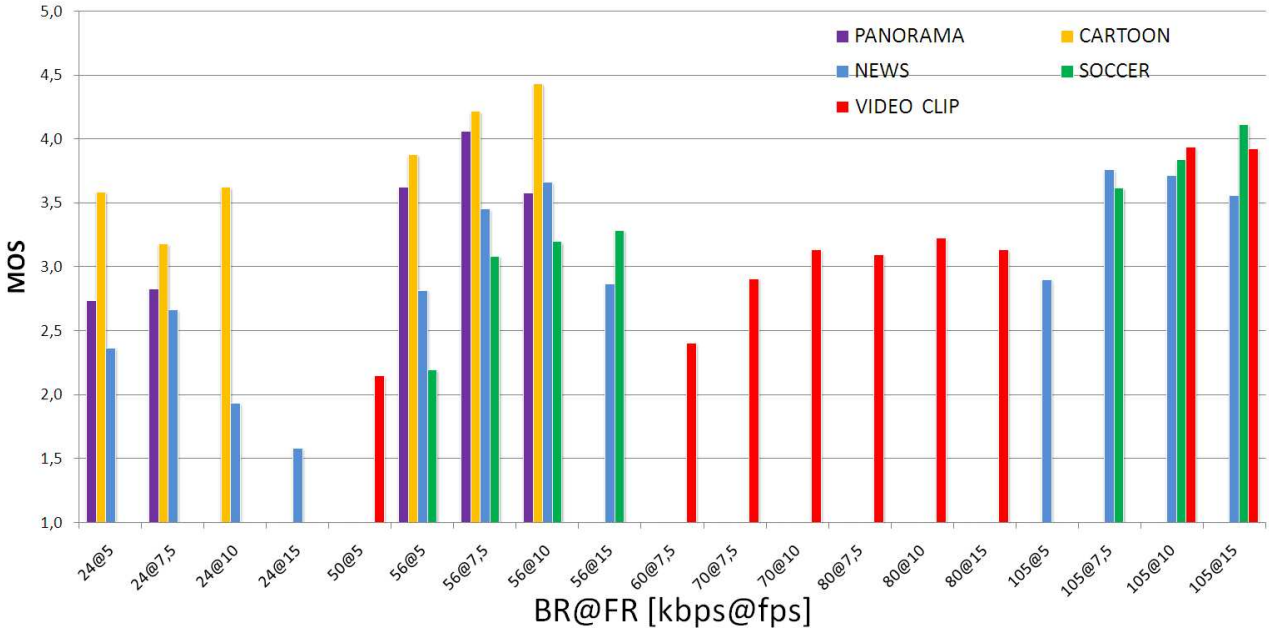

Figure 3.12: MOS for all the tested sequences for SIF Resolution.

The subjective assessment of the SIF resolution (see **Figure 3.12**) shows high quality variation at lower BRs, but a quality saturation at bitrates over 100 kbps can also be observed. For the "News" sequence, the highest score is obtained by the configuration BR@FR=105@7.5 kbps@fps, closely followed by 105@10 kbps@fps and 56@10 kbps@fps. Very interesting is the fact that the viewers seem to notice no difference in quality between the combination 56@10 kbps@fps and 105@10 kbps@fps, which both receive very positive evaluations. The most dynamic sequence "Soccer" received the best evaluation at 105 kbps. An increasing frame rate in soccer videos has always a positive effect on the perceived quality, which is in contrast with other content types, specially to the more static "News" case. In the "Soccer" sequence viewers prefer smoothness of motion rather than static quality.
Moreover, the video quality surveys allow to estimate which coding parameters should be used, according to the character of the video content, in order to maximize the end-user's perceived quality. Furthermore, the results can be used for determining the most suitable trade-off settings in terms of bit-rate and subjective quality requirements. Table **3.3** contains the suggested trade-off settings for low bit-rates for each content class. For these results we selected the lowest bit-rate that scored above 3 MOS grade and chose, for this bit-rate, the frame-rate configuration that obtained the best score.

3.4 Willingness to pay in relation to delivered quality

THE following Willingness To Pay (WTP) results were obtained together with the video quality survey described above. The WTP was evaluated with a simple binary question:

Content Class	QCIF		SIF	
	BR@FR [kbps@fps]	$MOS_{trade-off}$	BR@FR [kbps@fps]	$MOS_{trade-off}$
1	44@7.5	4.8	56@10	3.7
2	—	—	56@15	3.3
3			56@7,5	4.1
4	44@5	3.3	24@10	3.6
5			70@10	3.1
6	44@7.5	3.1		
7	44@7.5	3.7		

Table 3.3: The best trade-off encoding settings.

"*If you should pay for this video sequence, would you be satisfied with the video quality?*"

The question was designed in order to define a relation between WTP (as is defined) and subjective quality of video streaming. Technically, WTP has emerged from Subjective or Perceived/Perceptual QoS (PQ) [61] and is solely based on human perception or satisfaction regarding service usability. Determining PQ is typically carried out by surveying a set of persons, which participate in a controlled experiment [62]. Furthermore, the human visual perception of multimedia content is determined by the character of the observed sequence [63], [64]. The sequence character reflects motion characteristics (content type, video motion features, spatial information) [63], [64]. The recent trends show that the perceptual video QoS is defined by a set of intrinsic QoS parameters [61] as well as audio and video parameters. There were already complex QoS studies regarding WTP [65] and proposals of WTP utility functions for Internet streaming [66]. Unfortunately, these results are not applicable for mobile streaming due to significantly different usage scenarios. Furthermore, the consumers do not have a clear vision how much and what for they are willing to pay for. These conditions make it almost impossible to objectively perform a classical WTP survey in order to define the maximum amount of money that may be contributed by an individual to equalize the utility change. Therefore, the most fundamental WTP [67] features are investigated. The WTP is defined as **readiness to pay for the provided quality** of video streaming.
Figure 3.13 clearly shows a linear dependence between MOS and WTP. Moreover, the behavior is not dependent on the codec or the resolution.

These features allow us to estimate WTP with subjective video quality. The following metrics in a mobile scenario which returns the percentage of customers that are ready to pay for the provided video quality are obtained from the QCIF test scenario by a linear regression:

$$\mathrm{WTP}_{QCIF} = 29.952 \cdot \mathrm{MOS} - 45.948, \tag{3.4}$$

and from the SIF test scenario:

$$\mathrm{WTP}_{SIF} = 33.943 \cdot \mathrm{MOS} - 46.671. \tag{3.5}$$

In order to evaluate this regression the Pearson correlation factor separately for both resolutions was calculated. The performance of metrics (see Equations (3.4) and (3.5)) is over 96%. Moreover, the

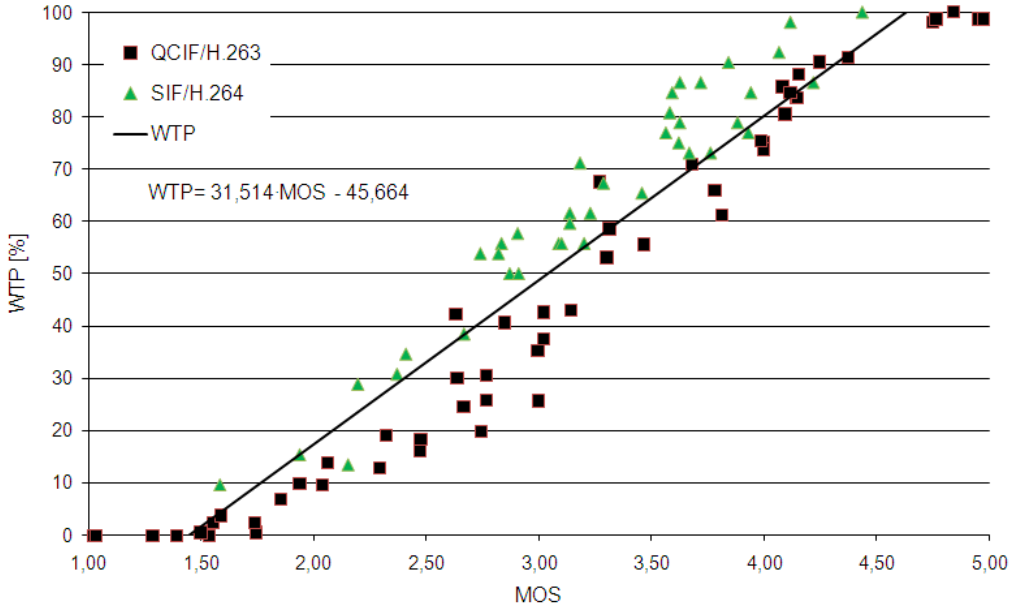

Figure 3.13: Relation between MOS and WTP.

visual inspection shows that QCIF and SIF results are very well correlated. Finally, one universal metric for both resolutions is proposed by simple linear regression over all measurements:

$$\mathrm{WTP} = 31.514 \cdot \mathrm{MOS} - 45.664. \tag{3.6}$$

For the joined test case the proposed metric is independent from codec, resolution and content. The performance of the proposed metric for QCIF resolution is 97.53% and for SIF 96.83%. Finally, for the joint test case 95.73% correlation was achieved.

Chapter 4

Video quality estimation

Contents

4.1 Introduction

THE human visual perception of video content is determined by the character of the observed sequence. It is necessary to determine different content characters/classes or content adaptive parameters because the video content itself strongly influences the subjective quality (cf. Section 3.3). The character of a sequence can be described by the amount of the edges (spatial information) in the individual frames and by the type and direction of camera and object movement (temporal information). The constant BR of the video sequence is shared by the number of frames per second. Higher frame rates at constant BR result in a lower amount of spatial information in individual frames and possibly in some compression artifacts.
In the literature the focus is given mainly on the spatial information [20], [21]. Such approaches come mainly from the quality estimation of still images [68], [69]. However, especially in small resolutions and after applying compression, not only the speed of movement (influencing at most the compression rate) but also the character of the movement plays an important role in the user perception. Therefore, in this thesis the focus is given on the motion features of the video sequences that determine the perceived quality.
In this chapter the design of content classifiers and video quality estimators for different content classes, codecs and resolutions are described. Since each shot of a sequence can have a different content character, a scene change detection is required as a pre-stage of content classification and quality estimation. Moreover, the quality estimation is based on content sensitive video parameters.

4.2 Temporal segmentation

THE temporal segmentation of a video into its basic temporal units - so called shots - is of great importance for a number of applications today. Video indexing techniques, which are necessary for video databases, rely on it. It is also necessary for the extraction of high level semantic features. Moreover, it provides information for video preprocessing, compression codecs and error concealment techniques. Temporal segmentation is also a prerequisite in the process of video quality estimation. A shot is a series of consecutive video frames taken by one camera. Two consecutive shots are separated by a shot boundary which can be abrupt or gradual. While an abrupt shot boundary (=cut) is generated by simply attaching one shot to another without modifying them like in **Figure 4.1**, a gradual shot boundary is the result of applying an editing effect to merge two shots. For the purpose of video quality estimation it is sufficient to detect scene cuts because abrupt shot boundaries are the most frequent and gradual shot boundaries (dissolve, fades or wipes) usually not leading to a content class change.
Many types of scene change detection schemes were proposed in literature. Scene change can be described by a similarity measure between two consecutive frames. When this measure reveals a big enough change defined by a threshold, a scene change is declared. However, a fixed threshold value does not perform well for all videos mainly due to the variety of content character. The key problem is to obtain an optimal value for such threshold. If it is set too high, there is a high probability that some cuts remain undetected. If it is too low, the detection scheme produces false detections. In video

streams of diverse CC types both cases can appear simultaneously.

Figure 4.1: Abrupt scene change.

A variety of techniques were proposed for shot boundary detection in digital video. A pairwise comparison checks each pixel in one frame with the corresponding pixel in the next frame [71]. In order to overcome the detection problem, a double threshold (high – low) was proposed to eliminate missed scene changes and dismiss false ones [76]. Although it improved the efficiency, results are not sufficient, especially in real-world videos with high motion like sport games. In addition to this method, a function-based lowering of the threshold, after a scene change was used to decay from high to lower threshold [77]. This technique was used to avoid false detections close to a real scene change, assuming that scene changes cannot occur immediately after each other. However, in most of these methods an optimal threshold (or two thresholds) had to be determined for each video in advance. The Likelihood ratio approach compares blocks of pixel regions [71]. Net comparison breaks the frame into base windows [72]. The color histogram method compares the intensity or color histograms between adjacent frames [74]. Model based comparison uses the video production system as a template [73]. Edge detection segmentation looks for entering and exiting edge pixels [75]. Other methods were proposed to find automatically an optimal static threshold e.g. using histogram differences [78], entropy [79] or the Otsu method [80], all having the disadvantage of a static threshold and therefore not being suitable for real-time applications. A truly dynamic threshold is presented in [81], where the input data are filtered by a median filter and then a threshold is set using the filtered output and standard deviation. However, it is not suitable for real-time applications, as the median filter uses future frames as well. A different approach for variable bit rate video is presented in [82], where the bit-rate used in each frame is the change metric. It uses statistical properties of the metric values in a single shot, together with the shots length, to define a threshold.

4.2.1 Sum of absolute differences

The Sum of Absolute Differences (SAD) is a widely used, extremely simple video quality metric, applied for block-matching in motion estimation for video compression. It works by taking the absolute value of the difference between each pixel in the original block and the corresponding pixel in the block being used for comparison. These differences are summed to create a simple metric of block similarity. Furthermore, SAD is a very suitable parameter for shot-boundary detection because SAD provides a clear distinctive measure of the video temporal behavior [83], [84].

The SAD is calculated between two consecutive frames (n) and $(n+1)$ and is computed as follows:

$$SAD_n = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} |F_n(i,j) - F_{n-1}(i,j)|\,, \tag{4.1}$$

where F_n is the n-th frame of size $N \times M$, i and j denote the pixel coordinates. As can be seen at **Figure 4.4**, SAD reflects the sequence motion characteristics as well as indicates shot-boundaries. Furthermore, the high peaks usually refer to the shot boundaries and gradual SAD changes refer to dynamic scenes or to cinema tricks. The SAD proportion between the dynamic scenes or to cinema tricks and shot boundaries is not constant. These features are investigated in the next section.

4.2.2 Analysis of scene change boundaries for different content types

To decide whether a shot boundary has occurred, it is necessary to set a threshold, or thresholds for the similarity between adjacent frames. The SAD values above this threshold are considered as cuts, while values below this threshold are ignored. To accurately segment videos of various content types, it is necessary to balance the following three - apparently conflicting points:

- robust scene change detection in all content types and resolutions,
- prevent detection of false shot boundaries by setting a sufficiently high threshold level,
- detect all shot boundaries, by setting a sufficiently low threshold level.

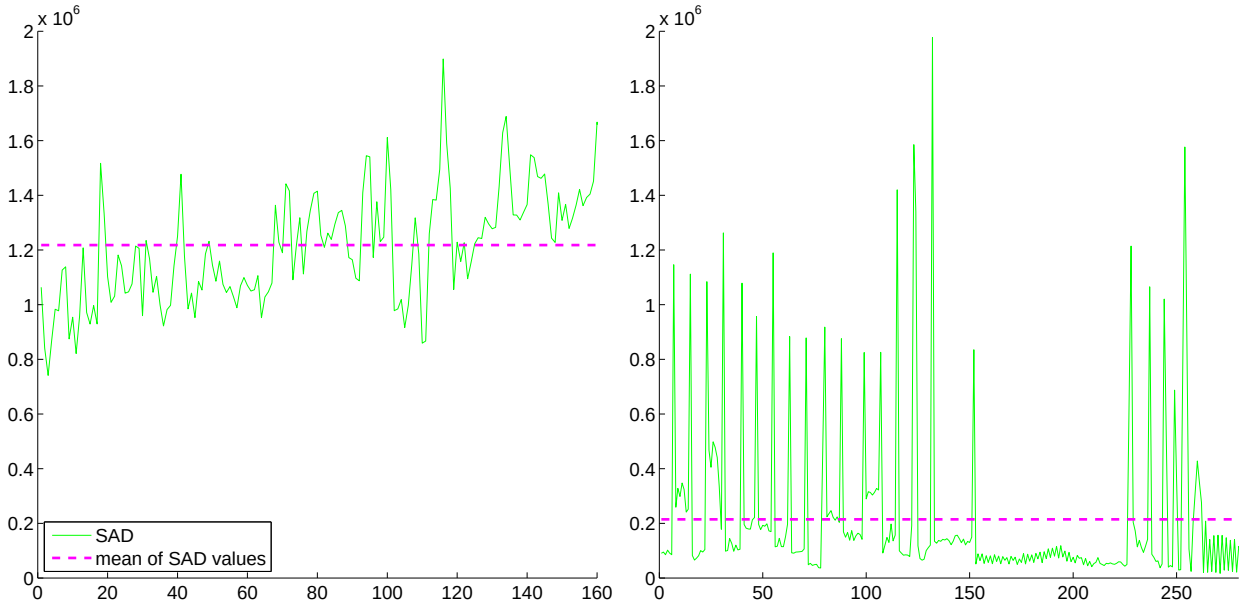

Figure 4.2: At the left side the temporal SAD behavior of the "car race" sequence is shown on the right side the temporal SAD behavior of the cinema trailer.

The sequence of SAD values that is computed, has statistical properties that are worth exploiting in the effort to detect scene changes. The scene changes are simply detected by a high SAD value (usually peak). However, high SAD values can also be found during rapid movement, sudden light changes and transition effects such as zoom in/out, dissolve etc. Moreover, the value level of a scene change is usually not constant. A scene break in which both scenes have similar background does not give a peak as high as if they had different ones. Furthermore, the maximal differences of two consecutive frames are much higher for SIF resolution than for QCIF resolution due to the different amount of pixels. The "car race" sequence (see **Figure 4.3**) containing a lot of local and global movement is taken as an example. This particular sequence leads to very high SAD values and its variations and mean values. Moreover, this sequence does not contain any cut. On the contrary, a second example is a cinema trailer which contains 19 cuts. The temporal character of both sequences can be seen at **Figure 4.2**. The green line refers to the SAD values and the dashed magenta line to the mean of the SAD values. As a consequence of the above listed conditions, there should be an optimal threshold level for all content types and resolutions. Unfortunately, it can be clearly seen at **Figure 4.2** that it is not possible to define a universal fixed threshold, due to significantly different motion characteristics of the various content classes. Consequently, a thresholding function is needed which will be able to adapt to the character of the scene without the need for a previous input. Therefore, the next investigation is focused on variable threshold settings based on local statistical features.

4.2.3 Dynamic threshold boundaries

Since the sequence can contain different scenes - shots with different characteristics, each sequence was segmented first by a scene change detection based on a dynamic threshold [83]. For this purpose the method was adopted to all content types [64].

The thresholding function is based on statistical features of the local sequence. The higher accuracy

Figure 4.3: Right snapshot of "Car race" sequence and left snapshot of cinema trailer.

was achieved by introducing 10 forecoming and 10 upcoming frames into averaging. The SAD is calculated between two frames (n) and $(n+1)$. Moreover, the empirical mean m_n and the standard deviation σ_n are computed for a sliding window $[n-N, n+N, N=10]$:

$$m_n = \frac{1}{2N+1} \sum_{K=n-N}^{K=n+N} \mathrm{SAD}_K \tag{4.2}$$

and

$$\sigma_n = \sqrt{\frac{1}{2N} \sum_{K=n-N}^{K=n+N} (\mathrm{SAD}_K - m_K)^2}. \tag{4.3}$$

Equations (4.2) and (4.3) are used for defining the variable threshold function:

$$T_n = a \cdot m_n + b \cdot \sigma_n. \tag{4.4}$$

The constants a, b were tuned in order to obtain the best performance for all content types. The constant a was set in order to avoid wrong scene change detections like as in case of intense motion scenes; but on the other hand, the detector can miss some low-valued, difficult scene changes. The constant b was tuned in order to prevent from detecting intense motion as a scene change as can be seen in **Figure 4.4**. The peaks above the dynamic threshold function (blue line) are recognized as cuts. The scene change detector works with both precision and recall higher than 97%.

This implementation differs from [83] in introducing additional future frames ($K = n...n + N$).

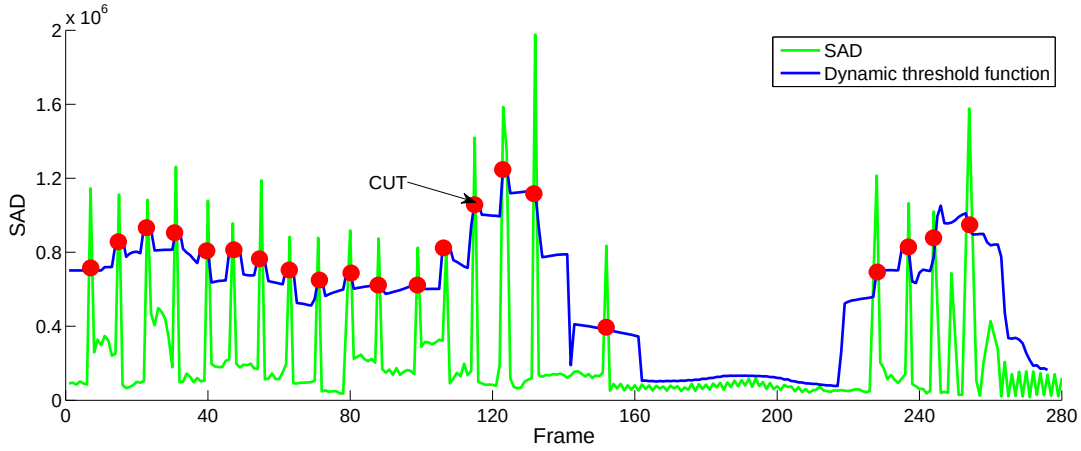

Figure 4.4: Performance of the proposed dynamic threshold function (4.4) on a sequence with multiple cuts (•).

Processing of upcoming frames increases the processing delay ($\frac{N}{\mathrm{FR}}$), but on the other hand improves accuracy and allows the method to be applied onto all content classes.

4.3 Video content classification

THE character of a sequence can be described by the amount of edges in the individual frames, SI, by the type and direction of movement, TI and color features. Moreover, the video quality estimation can be significantly simplified by content recognition as an independent issue. An automatic content recognition module is able to distinguish between a discrete set of classes can then be used in combination with an individual metric to estimate subjective quality of all typical video contents.

4.3.1 SI and TI sequence features

As an initial step for content classification SI and TI sequence features were investigated. SI (2.1) and TI (2.3) quantify still picture complexity and the motion character, respectively. Moreover, these parameters play a crucial role in determining the amount of video compression that is possible, and consequently, the level of impairment that is suffered when the scene is transmitted over a constant BR digital transmission service channel. The aim was to find a correlation among semantically related shots and a combination of SI and TI values.

Figure 4.5: Snapshots of investigated content classes: Talk show, Comedy, Cartoon, Nature.

Further investigation was performed on videos grouped to semantically compact content classes. The SI and TI values were computed only between two cuts and the videos were in original quality (uncompressed, FR = 25 fps). These sequences were segmented out of pool of more than 40 video sequences. The results are depicted at **Figure 4.6**.For this experiment sequences with (semantically) identical shots within one content class were chosen on purpose. Furthermore, the investigated content classes were semantically heterogeneous. The following content classes were chosen: Cartoon, Comedy, Talk show, Nature (see **Figure 4.5**) as well as Soccer and Cinema trailer.

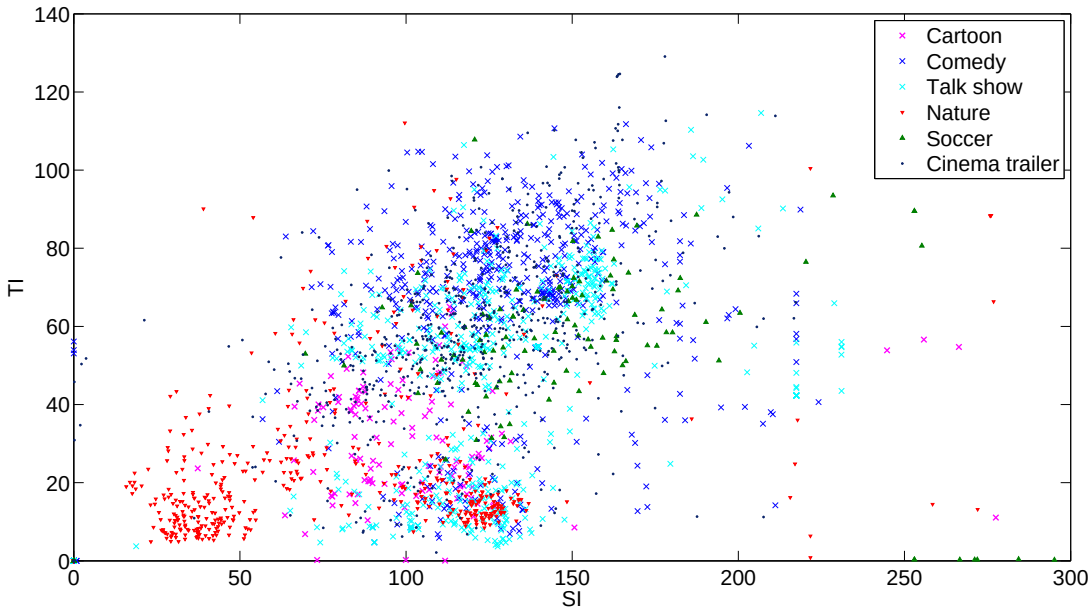

Figure 4.6: SI-TI diagram of investigated content classes.

In total more than 400 sequences were investigated belonging to these six classes. The preliminary look at the results shows big diversity of the values even within one content class; there are not typical regions for semantically identical content classes. Moreover, the regions are semantically overlapping. Further statistical analysis (see Table **4.1**) shows very high standard deviation in proportion to SI and TI mean values for all CCs.

CC	SI_{mean}	SI_{std}	TI_{mean}	TI_{std}
Cartoon	100.20	37.49	30.38	15.07
Comedy	125.30	39.31	64.15	26.83
Talk show	123.30	40.16	45.61	27.04
Nature	78.18	48.76	23.28	20.57
Soccer	142.20	67.12	52.46	27.72
Cinema trailer	118.90	34.92	58.27	23.73

Table 4.1: Statistical results from obtained SI and TI values.

According to SI and TI values, the most dynamic sequences are Comedy, Cinema trailer and Soccer as was expected. Surprisingly, the highest SI value was achieved for Soccer sequences; it was caused due to complexity of significantly high surface of the grand stand. Finally, it can be concluded that SI and TI values poorly map the content semantic and that another approach has to be used for content classification.

4.4 Content sensitive features

Especially at low resolutions and BRs after applying compression, not only speed and the amount of the movement but also the character of the movement plays an important role for the user perception. Moreover, the user content perception is also determined by color features of the sequence. Therefore, in this work the motion and color features of the video sequences within one shot are investigated that determine the perceived quality.

4.4.1 Motion vector extraction

The block based motion compensation techniques are commonly used in inter-frame video compression (supported also by H.263 and H.264) in order to reduce temporal redundancy. The difference between consecutive frames in a sequence is predicted from a block of equal size in the previous frame which serves as reference frame. The blocks are not transformed in any way apart from being shifted to the position of the predicted block. This shift is represented by a Motion Vector (MV). This technique was used to analyze the motion characteristics of the video sequences.

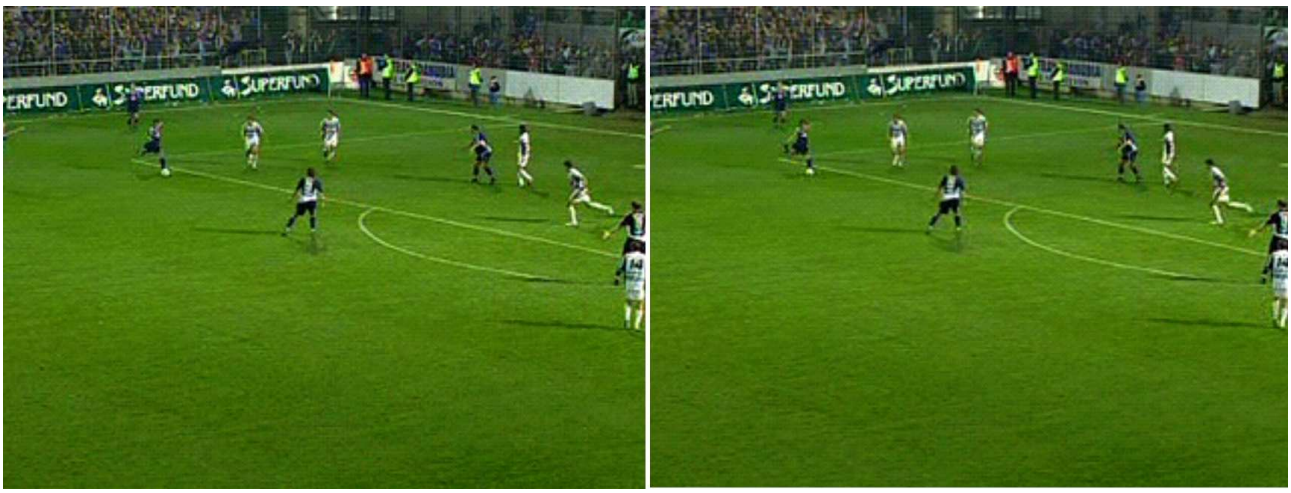

Figure 4.7: Snapshot of two successive frames - Soccer sequence.

The block from the current frame for which a matching block is sought, is known as the *target block*. The relative difference in the locations between the matching block and the target block is known as the MV. If the matching block is found at the same location as the target block then the difference is zero, and the motion vector is known as *zero vector*.

The difference between target and matching block increases (approximately linearly) with the size of the blocks. Thus, smaller blocks better describe the actual motion in the frame. On the other hand an increase of the objective accuracy does not always imply a better performance. If the blocks are selected too small, the resulting MVs do not reflect anymore the motion as it is perceived by a viewer [89]. Due to the unavoidable presence of noise in video sequences, and the characteristics of the human

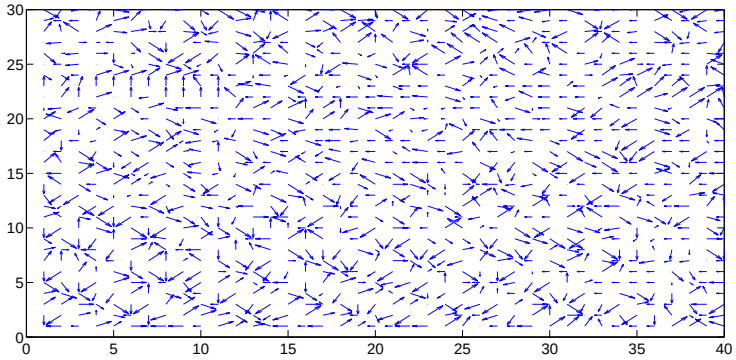

Figure 4.8: MV field obtained with blocks of 8x8 pixels without DCT filtering.

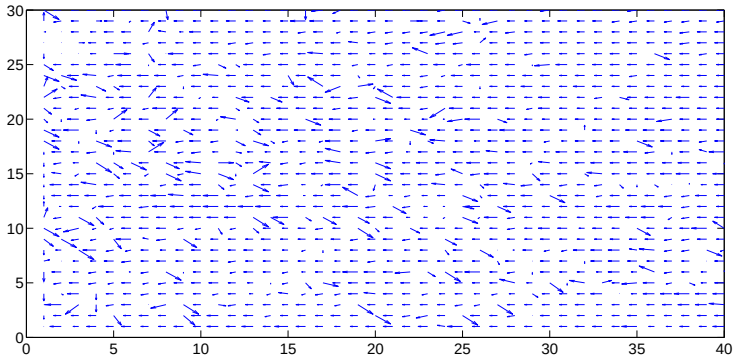

Figure 4.9: MV field obtained with blocks of 8x8 pixels with DCT filtering.

visual system, it happens that movement is detected although a human observer does not see it. Such behavior is not suitable for the purpose. After several trials with videos of different character, a pixel block size of 8×8 was selected due to a good trade-off for QVGA resolution sequences. The 320×240 pixels are divided into 30×40 blocks (see **Figure 4.8**), which gives a total number of 1200 MVs per frame.

The second part of the process, and the most time and resource consuming one, is block matching. Each block in the current frame is compared to a certain search region in the past frame in order to find a matching block. This operation is performed only on the luminance component of the frame. A matching criterion has to be used to quantify the similarity between the target block and the candidate blocks. Because of its simplicity and good performance, the sum of absolute differences (SAD) was used and computed as the pixel-wise sum of the absolute differences between the two blocks being compared:

$$\mathrm{SAD}_{n,m} = \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} \left| B_n(i,j) - B_m(i,j) \right|, \tag{4.5}$$

where B_n and B_m are the two blocks of size $N \times M$, and i and j denote pixel coordinates. If more than one SAD minimum is detected, priority is given to the matching block the position of which is most similar to that of the target block, or equivalently, to the MV of smallest size.

The presence of noise, lighting variation, and the existence of multiple local minima in the SAD distribution in the video sequences causes that the system detects movement although a human observer does not perceive it. Such effect introduces significant deviation to further usage of MV features for content classification and quality estimation. The vector fields found after performing the described exhaustive search on raw luminance frames do not always represent the true motion. A good example can be seen at the MV field in **Figure 4.8** made out of two consecutive soccer frames (see **Figure 4.7**). This problem is not new, the possible solution to erroneous motion detection is the use of smoothing techniques [86]. Smoothing techniques can be applied to the motion vectors that can detect erroneous MVs and suggest alternatives. The alternative motion vectors can be used in place of those suggested by the block matching algorithm. Moreover, this technique is not suitable for the purpose due to adding considerable complexity to the motion estimation algorithms and smoothing can cause the wrong detection of small moving objects.
The next option is to apply a binary mask to the motion field in order to exclude the low-textured areas from computation. A convolution mask is run in each macro-block of the uncompressed image, and those macro-blocks that have too few edge pixels (pixels with a gradient above a given threshold) are considered low-textured [87]. The drawback of such approach is that the uncompressed original images need to be available.
Another approach is MV correction based on the pattern-like image analysis [88] which exploits SAD distribution. This method does not provide sufficient accuracy for all content types and adds considerable processing complexity.
Therefore, it was necessary to develop a new, low complexity, and reference-free method. The most suitable approach was to analyze the AC components of the DCT coefficients and introduce additional low-pass filtering before the block matching [89]. The matching algorithms are applied using 8×8 pixel blocks and then the first ten coefficients of the inverse DCT are extracted. It can be observed in the example at **Figure 4.7** where two successive frames are depicted . The comparison of the MV field before filtering in **Figure 4.8** and MV after filtering in **Figure 4.9** shows considerable improvement.

4.4.2 Extraction of motion sequence parameters

The extracted and filtered MVs allow the further analysis of the motion (motion features) in the sequence. The static or dynamic character of a sequence is one of the main causes for the differences in perceived quality [64]. This investigation leads to a classification not only in terms of "static sequences" and "dynamic sequences", but also to deeper understanding of these aspect and to determine typical levels of quantity of movement for every content class. The overall amount of movement, or equivalently, the lack of movement in a frame, can be easily estimated from the proportion of blocks with zero vectors, that is, blocks that do not move from one frame to the other. Therefore, the average proportion of static blocks in a sequence of frames is very useful when it comes to distinguishing contents with typical different "levels" of overall movement.
The length of the MV indicates how far the block has moved from one frame to the next, and its angle tells us in which direction this movement occurred. Therefore, the mean size of the MVs in a frame or sequence of frames is an indicator of how fast the overall movement happens. Moreover, detecting a main direction of movement, that corresponds to a big proportion of MVs pointing in the same

direction, is a valuable information. Thus, it can be assumed that the analysis of the distribution of sizes and angles of the MVs can provide substantial information about the character of the motion in the sequence [89].
A set of statistical MV features were investigated in order to study their level of significance for further content classification. As initial step, the motion features were analyzed throughout the sequence in order to observe their temporal behavior. The following statistical and resolution independent features of MVs within one shot (over all the frames of the analyzed sequence) were investigated:

1. **Zero MV ratio N_z:**
   Percentage of zero MVs in a frame. It is the proportion of the frame that does not change at all (or changes very slightly) between two consecutive frames. It usually corresponds to the background if the camera is static within one shot.

2. **Mean MV size M:**
   Proportion of mean size of the MVs within one frame normalized to the screen width, expressed in percentage. This parameter determines the amount of the global motion.

3. **Mean non-zero MV size n:**
   Proportion of mean size of the non-zero MVs within one frame normalized to the screen width, expressed in percentage. This parameter determines the amount of the global motion.

4. **Intra-frame MV standard deviation F:**
   Standard deviation of MV sizes in the frame. MV standard deviation is expressed as a percentage of the MV mean size in the frame. This feature is sensitive at dynamic motion changes.

5. **Intra-frame standard deviation of non-static blocks f:**
   Standard deviation of the sizes of non-zero MVs, expressed as a percentage of the mean size of these MVs in the frame. It is a measure of the uniformity or dynamic changes of the moving regions.

6. **Uniformity of movement d:**
   Percentage of MVs pointing in the dominant direction (the most frequent direction of MVs) in the frame. For this purpose, the granularity of the direction is 10 degrees.

7. **MV direction uniformity of non-static blocks B:**
   Percentage of MVs pointing in the main direction in the moving regions (zero MV are excluded).

A Principal Component Analysis (PCA) [90] was carried out on the obtained data set. PCA is a suitable method for viewing a high-dimensional set of data in considerably fewer dimensions. For this purpose, it is necessary to calculate and plot the scores of k sample members using the first few principal components as axes. PCA can be described as follows:
As scalar variate ($\mathbf{y}_j$) is a linear combination $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + ... + a_p\mathbf{x}_p$ of original variates $\mathbf{x}_i$ (boldface $\mathbf{x}_i$ elements represent row elements of matrix $\mathbf{X}$). In this particular case data matrix $\mathbf{X} = \{X[p, n]\}$

is represented as:

$$\mathbf{X} = \begin{bmatrix} N_{z_1} & \dots & N_{z_p} \\ M_1 & \dots & M_p \\ n_1 & \dots & n_p \\ F_1 & \dots & F_p \\ f_1 & \dots & f_p \\ d_1 & \dots & d_p \\ B_1 & \dots & B_p \end{bmatrix}. \tag{4.6}$$

Thus, it can be more concisely written as $\mathbf{y} = \mathbf{Xa}$, where $\mathbf{a}^T = (a_1, a_2, ..., a_p)$ and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$. The coefficients of $\mathbf{a}_j^T$ are given by the elements of the eigenvector corresponding to the j-th largest eigenvalue l_j of the covariance matrix $\mathbf{S}$ of dimension $n \times n$:

$$\mathbf{S} = \frac{1}{p-1} \sum_{i=1}^{p} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T. \tag{4.7}$$

The j-th principal component is a linear combination $\mathbf{y} = \mathbf{Xa}$ which has the greatest sample variance for all $\mathbf{a}_j^T$ satisfying $\mathbf{a}_j^T \mathbf{a}_j = 1$ and $\mathbf{a}_j^T \mathbf{a}_i = 0$ $(i < j)$.

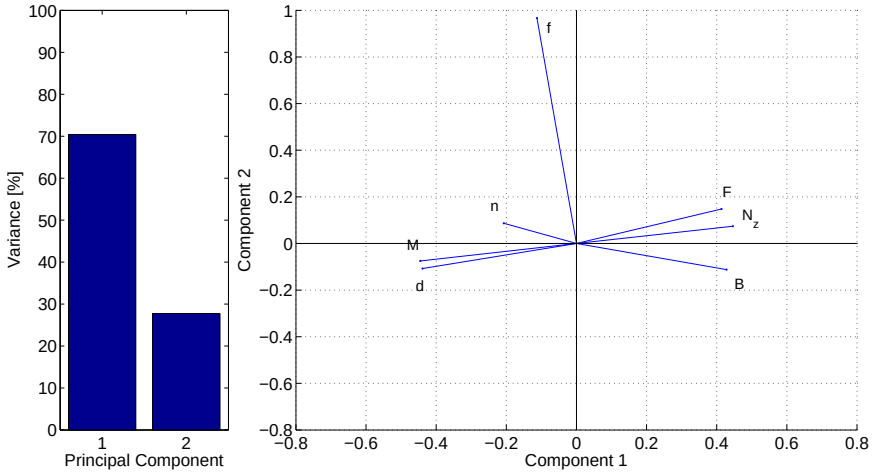

Figure 4.10: Visualization of PCA results for content class News (CC1).

The first two components proved to be sufficient for an adequate modeling of the variance of the data. The first principal component is the linear combination $\mathbf{y}_1 = \mathbf{Xa_1}$ of the original variables which has maximum variance among all such linear combinations. The second principal component $\mathbf{y}_2 = \mathbf{Xa_2}$ has the second largest variance among all such linear combinations.

Finally, the biplot technique [91] was applied for representing the first two principal components by means of parameter vectors. The biplot provides a useful tool of data analysis and allows the visual appraisal of the structure of large data matrices. It is especially revealing in principal component analysis, where the biplot can show inter-unit distances and indicate clustering of units as well as display variances and correlations of the variables.

To approximate biplot representation involves constructing the rank-2 approximation $\widetilde{\mathbf{S}}$ to covariance

(original) matrix $\mathbf{S}$ provided by the singular value decomposition (covariance matrix $\mathbf{S}$ is symmetric) given by:

$$\mathbf{S} = \sigma_1 \mathbf{v}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{v}_2 \mathbf{v}_2^T + \ldots + \sigma_k \mathbf{v}_k \mathbf{v}_k^T, \tag{4.8}$$

then the 2- rank approximation of $\mathbf{S}$ is:

$$\widetilde{\mathbf{S}} = \sigma_1 \mathbf{v}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{v}_2 \mathbf{v}_2^T. \tag{4.9}$$

$$\widetilde{\mathbf{S}} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix}.$$

If the vector $\mathbf{v}_i$ are given by $\mathbf{v}_i = (v_{i1}, v_{i2}, ..., v_{in},)$ for $i = 1, 2$, then we have:

$$\widetilde{\mathbf{S}} = \begin{pmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \\ \vdots & \vdots \\ v_{1m} & v_{2m} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & v_{1n} \\ v_{21} & v_{22} & v_{2n} \end{pmatrix}.$$

To obtain a biplot, it is first necessary to write $\widetilde{\mathbf{S}}$ as a product of matrices $\mathbf{G}\mathbf{G}^T$, where $\mathbf{G}$ is an $(m \times 2)$ matrix. This can be done by a simple factorization:

$$\widetilde{\mathbf{S}} = \mathbf{G}\mathbf{G}^T, \tag{4.10}$$

$$\widetilde{\mathbf{S}} = \begin{pmatrix} v_{11}\sqrt{\sigma_1} & v_{21}\sqrt{\sigma_2} \\ v_{12}\sqrt{\sigma_1} & v_{22}\sqrt{\sigma_2} \\ \vdots & \vdots \\ v_{1m}\sqrt{\sigma_1} & v_{2m}\sqrt{\sigma_2} \end{pmatrix} \begin{pmatrix} v_{11}\sqrt{\sigma_1} & v_{12}\sqrt{\sigma_1} & v_{1n}\sqrt{\sigma_1} \\ v_{21}\sqrt{\sigma_2} & v_{22}\sqrt{\sigma_2} & v_{2n}\sqrt{\sigma_2} \end{pmatrix}.$$

Finally, the end point of the parameter vectors represents $(v_{1i}\sqrt{\sigma_1},\ v_{2i}\sqrt{\sigma_2})$. The parameter vectors show variances and correlations of the defined video quality parameters. The variance of the first two components is over 90%. Moreover, the PCA results (see **Figures 4.10**, **4.11**, **4.12**, **4.13**, **4.14**) show a sufficient influence of the investigated parameters on the data set within each content class.

The visualization of PCA results shows that all parameter vectors have approximately similar influence on the data set. Therefore, it was not possible to set hard decision criteria for selection of video quality parameters according to PCA results. The chosen parameters have low computational complexity and good distribution over the most significant subspace of PCA within all content classes.

According to PCA results and computational complexity, the following three parameters were chosen:

- Zero MV ratio (N_z),
- The mean non-zero MV (n),
- The uniformity of movement (d).

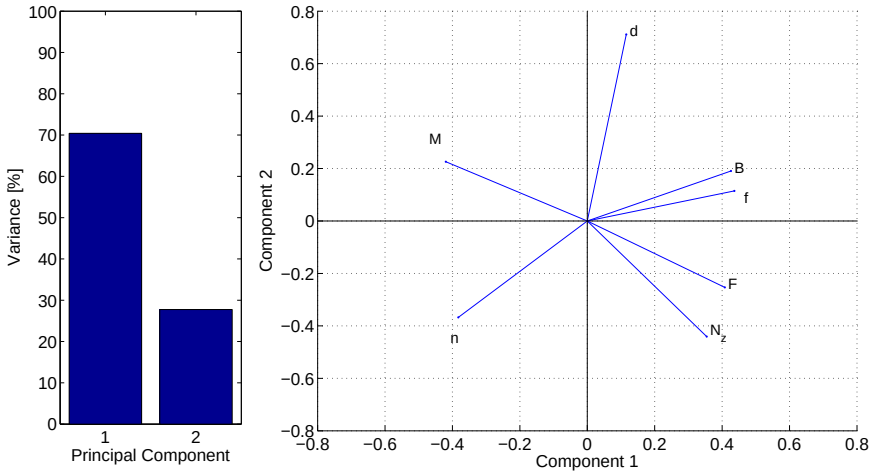

Figure 4.11: Visualization of PCA results for content Soccer (CC2).

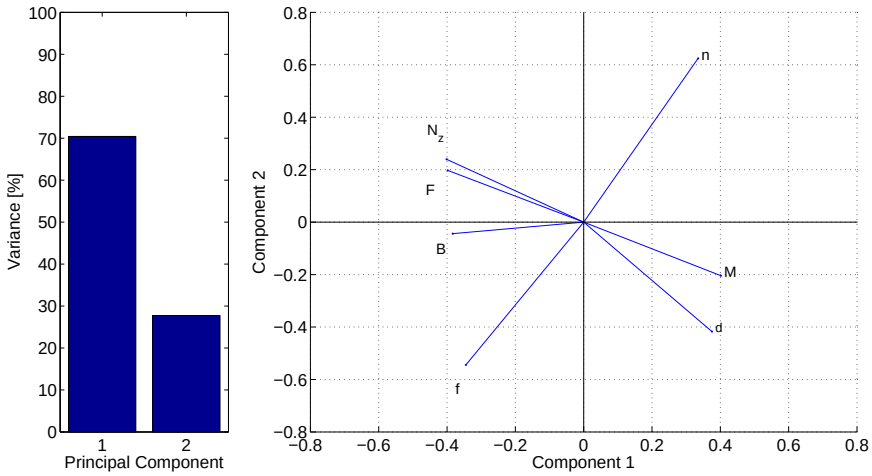

Figure 4.12: Visualization of PCA results for content Cartoon (CC3).

Finally, the selected content sensitive parameters reflect the static part of the surface, the significance of the movement in the non-static part and the dominant motion direction.

In order to increase the accuracy of the proposed content classifier it was necessary to define content sensitive features for detecting movement in horizontal direction as well as green color . Therefore, an additional feature for detecting the amount of horizontal movement was defined.

8. **Horizontalness of movement** h**:**
   Horizontalness is defined as the percentage of MVs pointing in horizontal direction. Horizontal MVs are from intervals $\langle -10; 10\rangle$ or $\langle 170; 190\rangle$ degrees.

This feature allows enhanced detection of panorama and the most accurate detection of soccer sequences. As can be seen at **Figure 4.15** in Panorama and Soccer, the horizontal movement is

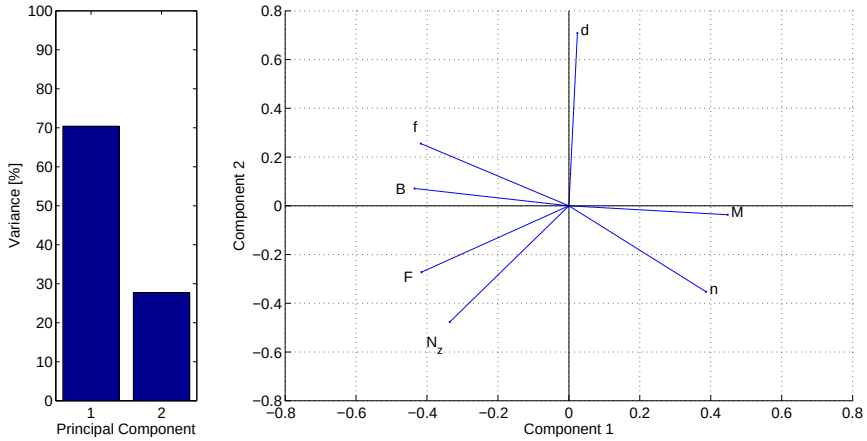

Figure 4.13: Visualization of PCA results for content Panorama (CC4).

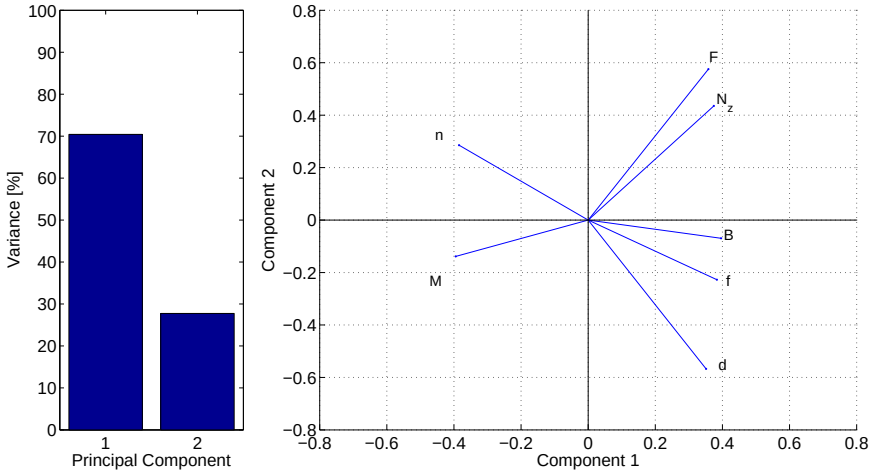

Figure 4.14: Visualization of PCA results for content Video clip (CC5).

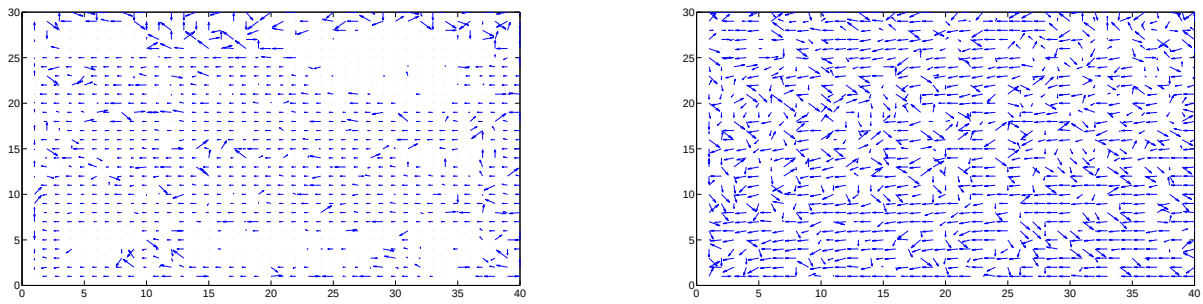

Figure 4.15: MV field of typical Panorama (left) and Soccer (right) sequences.

dominantly presented.

Soccer sequences for example contain a lot of varying green colors while cartoon sequences exhibit discrete saturated colors. For this purpose the color distribution of CCs was investigated. Color histograms provide additional information about the spatial sequence character because in different types of contents, the density and magnitude of colors differ as well. This characteristic has important consequences to the compression and transmission artifacts. Therefore, the following parameter was analyzed:

9. **Greenness g:**
   Define greenness as percentage of green pixels in a frame. For this purpose the RGB color space was down sampled to two bits per color component, resulting in 64 colors. Five colors out of the 64 colors cover all variation of the green color.

For calculating color histograms sequences were converted to RGB color space and down-sampled to 64 colors. The color bins were regularly spaced. As can be seen at **Figure 4.16** the five green color bins proved to be an effective element in the detection of the green sequence.

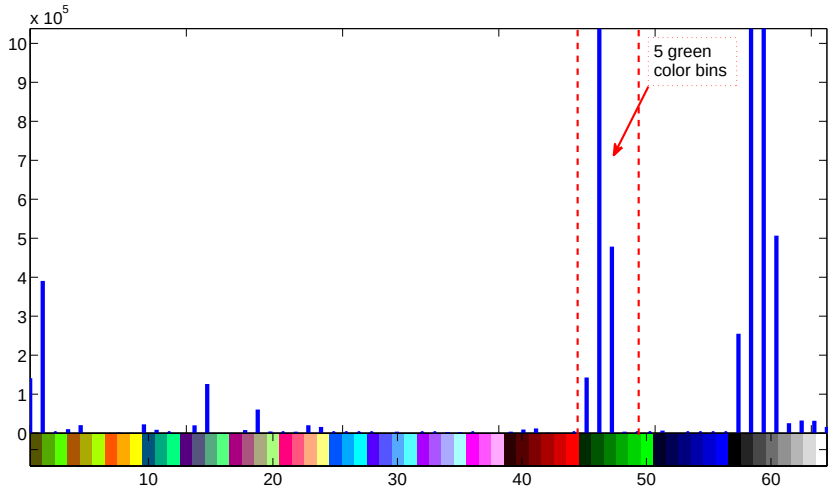

Figure 4.16: Color histogram of Soccer sequence.

4.4.3 Hypothesis testing and content classification

The content classification is based on the above defined parameters. Due to the extensive set of objective parameters, a statistical method was used for data analysis and content classification. This excludes content classifying based on threshold which is a limited and not accurate method for evaluating larger data sets.
A statistical method based on hypotheses testing was introduced. Each of the described content classes is determined by unique statistical features of motion and color parameters (see **Figure 4.17**). Due to their unique statistical features of well defined content classes it is not necessary to perform M-ary

hypothesis testing and it is sufficient to formulate a null hypothesis (H0) for each content class based on these statistical features separately. The obtained Empirical Cumulative Distribution Functions (ECDF) from the typical set of sequences for each content class show substantial differences (see **Figure 4.17**). From the next investigation it results that it is very difficult to determine single parametric distribution model representation from the obtained ECDFs. For this purpose a hypotheses testing method allowing for defining non-parametric, distribution free H0 hypotheses was of interest.
For the hypothesis evaluation a method is needed capable of working with empirical (sample) distributions. For this purpose the most suitable is the non-parametric Kolmogorov-Smirnov (KS) test [92]. The KS test is used to determine whether two underlying probability distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution, in either case based on finite samples. The two-sample KS test is one of the most used and general non-parametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.
From the typical set of sequences for each content class the ECDFs are obtained. The model ECDFs were derived from a set of 142 typical sequences. Each content class is described with five model ECDFs (zero MV ratio (N_z), mean MV size (n), uniformity of movement (d), horizontalness of movement (h), greenness (g :)), which correspond to their H0 hypothesis, respectively. Furthermore, it is necessary to find the maximal deviation ($D_{CC\ max}$) within one content class for all parameters (for each model ECDF). If $Q_n(x)$ is the model ECDF and $Q(x)$ is the ECDF of the investigated sequence. D_n is the maximal difference between $Q_n(x)$ and $Q(x)$:

$$D_n = \max_x \|Q_n(x) - Q(x)\| . \tag{4.11}$$

The content class estimation is based on a binary hypothesis test within the four content classes (CC1 — CC4). With the KS test the ECDFs of the investigated sequence and all model ECDFs of the first four content classes are compared. The KS test compares five ECDF (of defined MV or color parameters) of defined content classes specified by the H0 hypothesis with all five ECDFs of the investigated content.

$$D_n \leq D_{CC\ max}. \tag{4.12}$$

$D_{CC\ max}$ reflects maximal deviation from model ECDF within defined content class (see Table **4.2**). If the D_n obtained for the tested CC, is smaller than $D_{CC\ max}$ for each parameter (4.12), then the sequence is expected to match this CC.

$D_{CC\ max}$	N_z	n	d	h	g
News	0.4120	0.6952	0.4708	0.0467	0.9406
Soccer	0.8410	0.7234	0.8226	0.7654	0.1912
Cartoon	0.7548	0.9265	0.7093	0.4198	0.9362
Panorama	0.9716	0.9034	0.9113	0.0000	0.9017

Table 4.2: Statistical results from obtained SI and TI values.

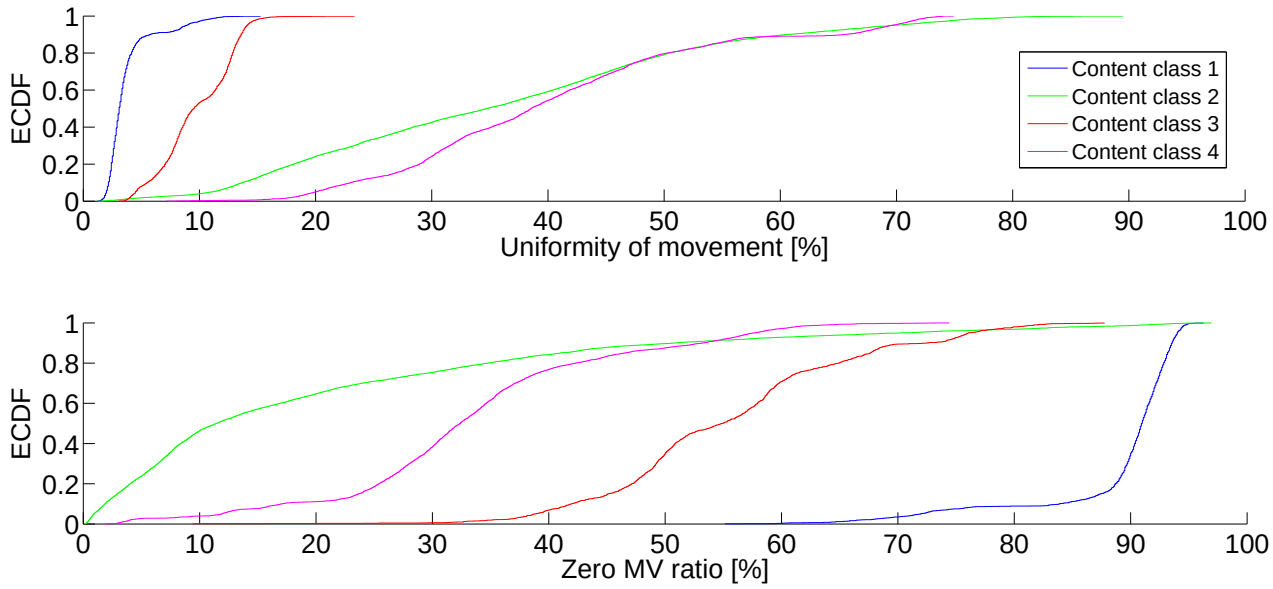

Figure 4.17: Model ECDF of zero MV ratio and uniformity of movement.

If the ECDFs of the investigated sequence have not a fit with any of the first four content classes, the content classifier decides for the remaining content class number five. The classifier estimates the content at transmitter side from the original sequence.
The performance of the content classifier was evaluated with two parameters. **False detection** reflects the ratio of improper detection of a content class, in the case when investigated sequences belong to any **other** content class. **Good match** reflects the ratio of successful classification of investigated sequences, when investigated sequences belong to any of the first four classes. Note, the sequences contain almost only cuts and no gradual changes. The scene change detector was sensitive on gradual shot boundaries (dissolve, fades or wipes). 786 sequences were tested to evaluate the performance of the content classifier, 98% were classified correctly. The achieved precision of the content classifier is shown in Table **4.3**, what is a satisfying result for further quality estimation.

Content class	False detection [%]	Good match [%]
1	0	97
2	0	100
3	5.6	92
4	0	100

Table 4.3: The evaluation results of content classifier.

4.5 Video quality estimation for SIF-H.264 resolution

IN this section, three methods for quality estimation in SIF resolution are presented. The proposed methods are focused on reference free video quality estimation. The character of motion is determined by the amount and direction of the motion between two scene changes. The first method estimates video quality in two steps. First, a content classification with character sensitive parameters is carried out [64]. Finally, based on the content class, frame rate and bitrate, the video quality is estimated in a second step. The following method thus presents the design of a quality metric based on content adaptive parameters, allowing for content dependent video quality estimation. The second method estimates the quality with one single universal metric. In contrast to those two, the third method exploits the estimation ensemble of models. The performance of the proposed method is evaluated and compared to the ANSI T1.801.03 metric. The results show that the motion-based approach provides powerful means of estimating the video quality for low resolution video streaming services.

4.5.1 Content based video quality estimation

The estimation is based only on the compressed sequence without the original (uncompressed) sequence. If estimation is performed on the receiver side, the information about the content class needs in parallel to be signalized with the video stream (see **Figures 4.18** and **4.19**). Such measurement setup allows for continuous real time video streaming quality measurement on both sides: user and provider. The video quality is estimated after content classification (cf. Section 4.3 and **Figure 4.19**) within one cut (cf. Section 4.2).

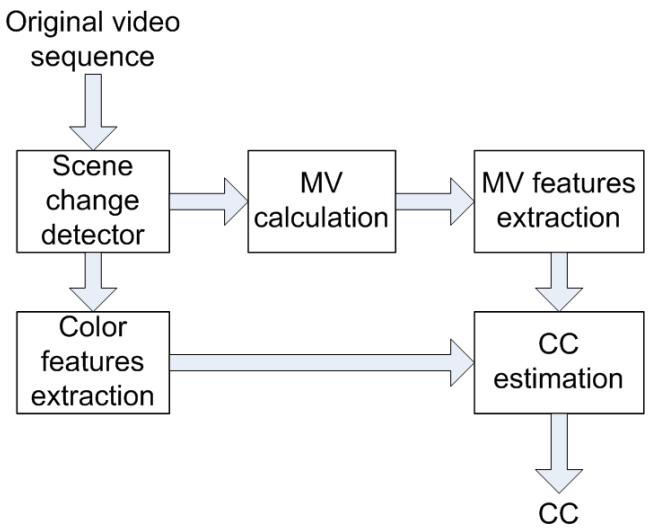

Figure 4.18: Content classifier.

Due to limited processing power of the user equipment it was necessary to identify low complexity objective parameters. In order to keep the complexity as low as possible the most suitable parameters are already provided: FR and BR. These parameters are the codec compression settings and signalized during the initiation of the streaming session, requiring no computational complexity for estimation

Coeff.	CC 1	CC 2	CC 3	CC 4	CC 5
A_{CC}	4.0317	1.3033	4.3118	1.8094	1.0292
B_{CC}	0	0.0157	0	0.0337	0.0290
C_{CC}	−44.9873	0	−31.7755	0	0
D_{CC}	0	0.0828	0.0604	0.0044	0
E_{CC}	−0.5752	0	0	0	−1.6115

Table 4.4: Coefficients of metric model for all content classes (CC).

as they are known at both transceiver and receiver.

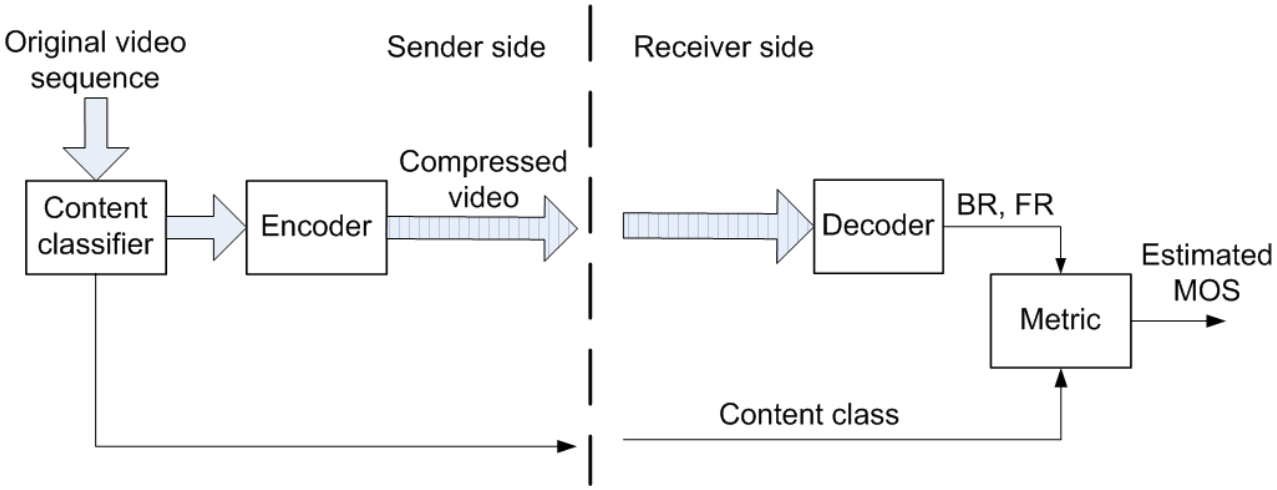

Figure 4.19: Content based video quality estimation.

The proposed low complexity metric is thus simply based on two objective parameters (BR and FR) for each content class:

$$\widehat{\mathrm{MOS}} = \mathrm{f}(\mathrm{BR}, \mathrm{FR}, \mathrm{CC}). \tag{4.13}$$

A general model is proposed with linear and hyperbolic elements (see (4.14)). The coefficients vary substantially for each content class. Typically, some of them even take a zero values. On the other hand, rather good correlation was achieved with one offset and two non-zero coefficients (see Table **4.4**).

$$\widehat{\mathrm{MOS}} = A_{CC} + B_{CC} \cdot \mathrm{BR} + \frac{C_{CC}}{\mathrm{BR}} + D_{CC} \cdot \mathrm{FR} + \frac{E_{CC}}{\mathrm{FR}}. \tag{4.14}$$

The metric coefficients were obtained by a linear regression of the proposed model with the training set (MOS values averaged over two runs of all 26 subjective evaluations for particular test sequence). The performance of the subjective video quality estimation compared to the subjective quality data is summarized in Tables **4.6** and **4.7** and shown in **Figure 4.22**. Further ahead, the obtained correlations with the evaluation set show very good performance of the proposed metric for all content classes except for Cartoon (CC3), containing two and three dimensional cartoon movies. The proposed metric has weak estimation performance for the three dimensional cartoon sequences.

4.5.2 Quality estimation based on content sensitive parameters

In this section the focus is given on the motion features of the video sequences. The motion features can be used directly as an input into the estimation formulas or models as shown in **Figure 4.20**. Both possibilities were investigated in [63] and [94], respectively.
The investigated motion features concentrate on the motion vector statistics, including the size distribution and the directional features of the motion vectors (MV) within one sequence of frames between two cuts. Zero MVs allow for estimating the size of the still regions in the video pictures. That, in turn, allows to analyze MV features for the regions with movement separately. This particular MV feature makes it possible for distinguishing between rapid local movements and global movement.

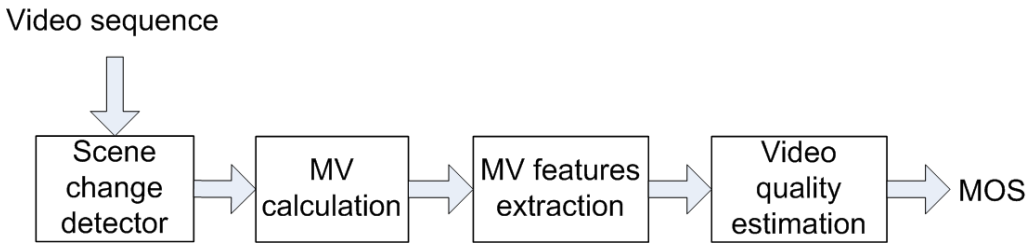

Figure 4.20: Video quality estimation based on content adaptive parameters.

Extraction of content adaptive features

The aim was to define measures that do not need the original (non-compressed) sequence for the estimation of quality because this reduces the complexity and at the same time broadens the possibilities of the quality prediction deployment. Furthermore, the size distribution and the directional features of the MVs were analyzed within one sequence in between two cuts. The details regarding MV extraction can be found in Section 4.4.1. The still and moving regions were analyzed separately. The size of the still region was estimated by the amount of zero MV vectors. That allows to analyze MV features separately for regions with movement. This particular MV features make it possible to detect rapid local movements or the character of global movements. The content sensitive parameters and BR were investigated. For this purpose motion sequence parameters (cf. Section 4.4.2) were reinvestigated. Furthermore, it was necessary to investigate the influence of these motion parameters and the BR on investigated contents. For this purpose we used a PCA analysis [90]. The PCA was carried out to verify further applicability of the motion characteristics, BR for metric design. It turned out that the first two components proved to be sufficient for an adequate modeling of the variance of the data. The variance of the first component is 60.19% and second 18.20%. The PCA results (see **Figure 4.21**) show sufficient influence of most significant parameters on the data set for all content classes.
The following features of MV and BR represent the motion characteristics:

- **Zero MV ratio within one shot Z:**
  The percentage of zero MVs is the proportion of the frame that does not change at all (or changes very slightly) between two consecutive frames averaged over all frames in the shot. This feature detects the proportion of a still region. The high proportion of the still region refers to a very static sequence with small significant local movement. The viewer attention is focused mainly on this small moving region. The low proportion of the still region indicates uniform global movement and/or a lot of local movement.

- **Mean MV size within one shot N:**
  This is the percentage of mean size of the non-zero MVs normalized to the screen width. This parameter determines the intensity of a movement within a moving region. Low intensity indicates the static sequence. High intensity within a large moving region indicates a rapidly changing scene.

- **Ratio of MV deviation within one shot S:**
  Percentage of standard MV deviation to mean MV size within one shot. A high deviation indicates a lot of local movement and a low deviation indicates a global movement.

- **Uniformity of movement within one shot U:**
  Percentage of MVs pointing in the dominant direction (the most frequent direction of MVs) within one shot. For this purpose, the resolution of the direction is 10^o. This feature expresses the proportion of uniform and local movement within one sequence.

- **Average $\overline{BR}$:**
  This parameter refers to the pure video payload. The $\overline{BR}$ is calculated as an average over the whole stream. Furthermore, the parameter $\overline{BR}$ reflects a compression gain in spatial and temporal domain. Moreover the encoder performance is dependent on the motion characteristics. The $\overline{BR}$ reduction causes a loss of the spatial and temporal information what is usually annoying for viewers.

The perceptual quality reduction in spatial and temporal domain is very sensitive to the chosen motion features, making these very suitable for reference free quality estimation because a higher compression does not necessarily reduce the subjective video quality (e.g. in static sequences).

4.5.3 Direct motion based quality estimation

The initial approach is to design a universal metric based on content sensitive parameters [63]. The subjective video quality is estimated with five objective parameters. Additional investigated objective parameters do not improve the estimation performance. On the other hand, reducing objective parameters decreases significantly the estimation accuracy. The proposed model reflects the relation of objective parameters to the MOS. Furthermore, the mix-term (S · N) reflects the dependence of the movement intensity (N) and its motion character (S). Finally, one universal metric is proposed for all contents based on the defined motion parameters Z, S, N, U and the BR:

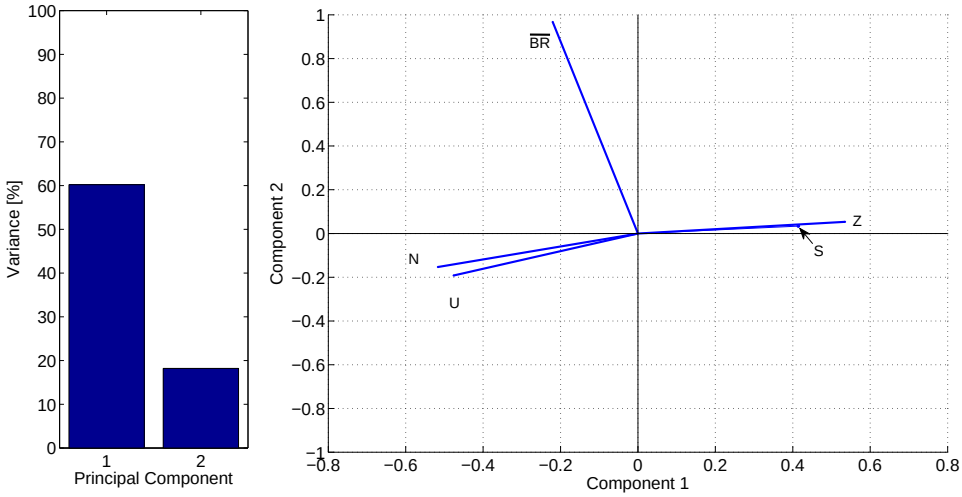

Figure 4.21: Visualization of PCA results.

Coeff.	
a	4.631
b	8.966×10^{-3}
c	8.900×10^{-3}
d	-5.914×10^{-2}
e	0.783
f	-0.455
g	-5.272×10^{-2}
h	8.441×10^{-3}

Table 4.5: Coefficients of estimation model.

$$\widehat{\mathrm{MOS}} = \begin{array}{llll} a & +b \cdot \overline{BR} & +c \cdot \mathrm{Z} & +d \cdot \mathrm{S}^e + \\ +f \cdot \mathrm{N}^2 & +g \cdot \ln(\mathrm{U}) & +h \cdot \mathrm{S} \cdot \mathrm{N} & . \end{array} \tag{4.15}$$

The metric coefficients (see Table **4.5**) were obtained by a regression of the proposed model with the training set (MOS values averaged over two runs of all 26 subjective evaluations for a particular test sequence). To evaluate the quality of the fit of the proposed metrics for the data, a Pearson correlation factor [85] was used. The metric model was evaluated with MOS values from the evaluation set (MOS values averaged over two runs of all 10 subjective evaluations for the particular test sequence). The performance of this proposed model was not satisfying (see Tables **4.6** and **4.7**). In the following so obtained estimates were used as input for a further design of an ensemble based video quality estimator.

4.5.4 Ensemble based quality estimation

In order to obtain higher prediction accuracy, ensemble based estimation was investigated. Ensemble based estimators average the outputs of several estimators in order to reduce the risk of an unfortunate selection of a poorly performing estimator. The very first idea was to use more than one classifier for estimation comes from the neural network community [95]. In the last decade research in this field has expanded in strategies [96] for generating individual classifiers, and/or the strategy employed for combining the classifiers.

The aim is to train a defined ensemble of models with a set of four motion sensitive objective parameters (Z, N, S, U) and $\overline{BR}$. The ensemble consists of different model classes to improve the performance in regression problems. The theoretical background [97] of this approach is that an ensemble of heterogeneous models usually leads to reduction of the ensemble variance because the cross terms in the variance contribution have a higher ambiguity. A data set with input values (motion sensitive parameters and $\overline{BR}$) x and output value (MOS) y with a functional relationship is considered , where e is an estimation error:

$$y = f(x) + e. \tag{4.16}$$

The weighted average $\bar{f}(x)$ of the ensemble of models is defined as follows:

$$\bar{f}(x) = \sum_{k=1}^{K} w_k f_k(x), \tag{4.17}$$

where $f_k(x)$ denotes the k-th individual model and the weights w_k sum to one ($\sum_k w_k = 1$). The generalization (squared) error $q(x)$ of the ensemble is given by:

$$q(x) = (y(x) - \bar{f}(x))^2. \tag{4.18}$$

According to [97], the error can be decomposed as follows:

$$q(x) = \bar{q}(x) - \bar{a}(x). \tag{4.19}$$

This assumption allows us to neglect the mixed terms of the following equation where the average error $\bar{q}(x)$ of the individual model is:

$$\bar{q}(x) = \sum_{k=1}^{K} w_k (y(x) - f_k(x))^2, \tag{4.20}$$

and the average ambiguity $\bar{a}(x)$ of the ensemble is:

$$\bar{a}(x) = \sum_{k=1}^{K} (f_k(x) - \bar{f}(x))^2. \tag{4.21}$$

- A consequence of (4.19) is that the ensemble generalization error $q(x)$ is always smaller than the expected error of the individual models $\bar{q}(x)$.

- The previous equations (4.16) — (4.21) require that an ensemble should consist of well trained but diverse models in order to increase the ensemble ambiguity.

This prerequisite was applied to an ensemble of universal models. In order to estimate the generalization error and to select models for the final ensemble a cross-validation scheme for model training [98] was used. These algorithms increase the ambiguity and thus improve generalization of a trained model. Furthermore, an unbiased estimator of the ensemble generalization error was obtained. The cross-validation works as follows:

- The data set is divided in two subsets and the models are trained on the first set.
- The models are evaluated on the second set, the model with the best performance becomes an ensemble member.
- The data set is divided with light overlapping with previous subsets into two new subsets and the models are trained on the first set.
- The cross-validation continues until the ensemble has a desired size. The best trade-off between ensemble complexity and performance was achieved for an ensemble of six estimators.

The final step in the design of an ensemble based system is to find a suitable combination of models. Due to outliers and overlapping in data distribution of the data set, it is impossible to propose a single estimator with perfect generalization performance. Therefore, an ensemble of many classifiers was designed and their outputs were combined such that the combination improves upon the performance of a single classifier. Moreover, classifiers with significantly different decision boundaries from the rest of the ensemble set were chosen. This property of an ensemble set is called diversity. The above mentioned cross-validation introduces model-diversity, the training on slightly different data sets leads to different estimators (classifiers). Additionally, diversity was increased by using two independent models. Furthermore, in cross validation classifiers with worse correlation than 50% on the second set were automatically excluded.
As the first estimation model, we chose a simple nonparametric method, the k-Nearest Neighbor rule (kNN) with adaptive metric [98]. This method is very flexible and does not require any preprocessing of the training data. The kNN decision rule assigns to an unclassified sample point the classification of the nearest sample point of a set of previous classified points. Moreover, a locally adaptive form of the k-nearest neighbor was used for classification. The value of k is selected by cross validation.
As the second method an Artificial Neural Network (ANN) was used. A network with three layers was proposed; input, one hidden and output layer using five objective parameters as an input and estimated MOS as output. Each ANN has 90 neurons in the hidden layer. As a learning method Improved Resilient Propagation (IRPROP+) with back propagation [99] was used. IRPROP+ is a fast and accurate learning method in solving estimation tasks for the data set. Finally, the ensemble consists of two estimation models kNN and ANN and six estimators, three kNN and three ANN.
The performance of the ensemble based estimator is the best out of the proposed ones. The results show a very good agreement between estimated and evaluated MOS values (see Tables **4.6** and **4.7**).

Metric	Pearson correlation
Content based	0.8303
Direct motion based	0.8190
Ensemble based	0.8554
ANSI	0.4173

Table 4.6: Metrics prediction performance on evaluation set by Pearson correlation.

Metric/Content type	CC 1	CC 2	CC 3	CC 4	CC 5
Content based	0.93	0.97	0.99	0.90	0.93
Direct motion based	0.85	0.98	1.00	0.71	0.95
Ensemble based	0.93	0.97	0.77	0.91	0.97
ANSI	0.63	0.85	0.95	0.93	0.97

Table 4.7: Metric prediction performance on defined CC by correlation on evaluation set.

4.5.5 Performance of the video quality estimators

To validate the performance of the proposed metric, the Pearson (linear) correlation factor [85] was applied:

$$r = \frac{(\mathbf{x} - \overline{\mathbf{x}})^T(\mathbf{y} - \overline{\mathbf{y}})}{\sqrt{((\mathbf{x} - \overline{\mathbf{x}})^T(\mathbf{x} - \overline{x}))((\mathbf{y} - \overline{\mathbf{y}})^T(\mathbf{y} - \overline{\mathbf{y}}))}}, \tag{4.22}$$

Here, the vector $\mathbf{x}$ corresponds to the **average** MOS values of the evaluation set (averaged over two runs of all obtained subjective evaluations for particular test sequence and one encoding setting) for all tested encoded sequences and $\overline{\mathbf{x}}$ corresponds to average over $\mathbf{x}$. Vector $\mathbf{y}$ corresponds to the prediction made by the proposed metric and $\overline{\mathbf{y}}$ corresponds to average over $\mathbf{y}$. The dimension of $\mathbf{x}$ and $\mathbf{y}$ refers to amount of tested sequences. In order to provide a detailed comparison, the performance of the ensemble based estimator [94] was compared with the content class based [64] and direct motion based [63] estimator as well as the ANSI metric [18] on the evaluation set. The depicted results for the Pearson correlation factor in Table **4.6** reflect the goodness of fit (see **Figure 4.22**) with the independent evaluation set for all content types together.

This correlation method only assumes a monotone relationship between the two quantities. A virtue of this form of correlation is that it does not require the assumption of any particular functional form in the relationship between data and predictions. The results in Table **4.6** clearly show a good agreement between the obtained and the estimated values for all proposed metrics. In addition, the goodness of the fit on different content classes (see Table **4.7**) was investigated. The best performance over all content classes is provided by ensemble and content class based metrics. A fair performance was obtained by the motion based metric and very poor performance by the ANSI metric.

The ensemble based metric shows a performance similar to the content class based metric. The content classification can be understood as a kind of pre-estimation in order to obtain a more homogeneous set of results within one content class, which allows for more accurate quality estimation.

This effect was achieved by introducing cross-validation in ensemble based metrics. The direct motion based metric suffers from weak estimation performance for news and cartoon sequences in comparison to the other content classes.
The weak performance of the ANSI metric shows that this metric is not suitable for a mobile streaming scenario. The usage of the mobile streaming services influences the subjective evaluation. Therefore, a universal metric like ANSI is not suitable for the estimation of mobile video quality. Only for higher MOS values which occur at high bitrates ($\geq$ 90 kbps) the ANSI metric performs comparable to the proposed metrics (see **Figure 4.22**). A closer look on the ANSI metric performance shows that the ANSI metric provided a good fit for CC3, CC4, CC5 and poor performance only for CC1 and CC2. However, the ANSI metric requires the knowledge of a reference video (original) and is the most complex estimator.

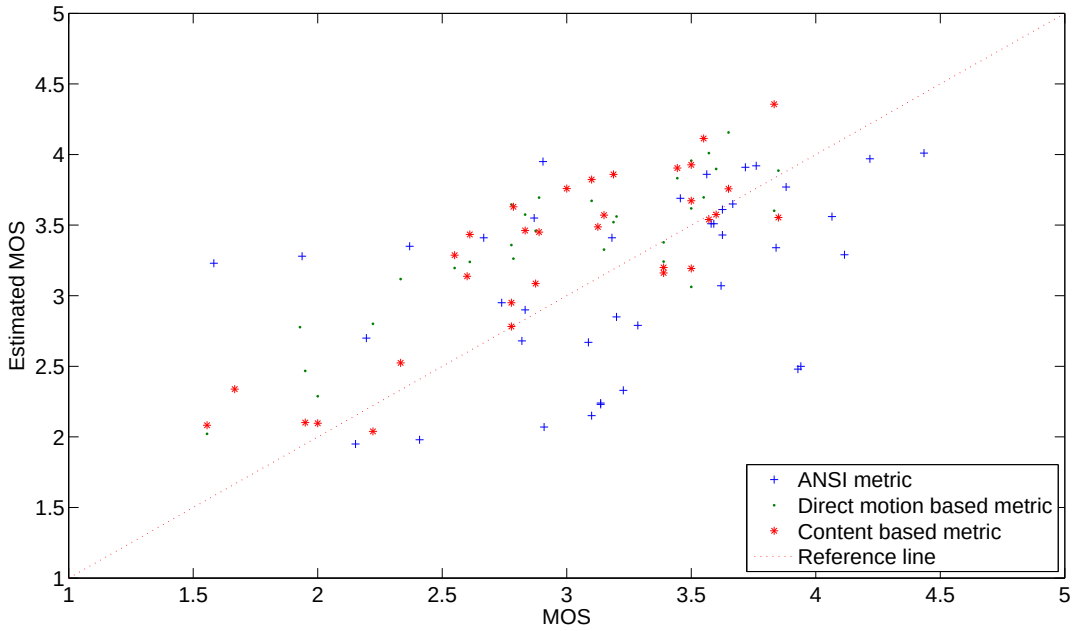

Figure 4.22: Estimated vs. subjective MOS results.

4.6 Video quality estimation for QCIF-H.263 resolution

IN this section a low-complexity reference-free estimation of visual perceptual quality for QCIF resolution is presented, based on a combination of a small set of the most important objective parameters - compression settings and content features [4]. To achieve this, the chosen objective parameters are mapped on obtained MOS by an extensive survey.
For QCIF resolution and mobile environment it is possible to find simpler estimates achieving the same performance as the already known ones [18] and [19]. Therefore, the focus is given on the quality sensitive measures that would not need the original (non-compressed) sequence for the estimation of quality, because this reduces the complexity and at the same time broadens the possibilities of the quality prediction deployment.

4.6.1 Quality sensitive parameter set

Nine objective video parameters were investigated with various computational complexity:

- *sigain*,
- *siloss*,
- *hvgain*,
- *hvloss*,
- SI_{13},
- SI,
- TI,
- BR,
- FR.

The first five of the objective video parameters (*sigain*, *hvloss*, *siloss*, *hvgain* and SI_{13}) are also recommended in [18]; further parameter details are described in Appendix B. The first two of them are *sigain* and *siloss* measuring the gain and the loss in the amount of spatial activity, respectively. If the codec operates through an edge sharpening or enhancement, a gain in the spatial activity is obtained, that is an improvement in the video quality of the image. On the other hand, when a blurring effect is present in an image, it leads to a loss in the spatial activity. The other two parameters, *hvgain* and *hvloss* measure the changes in the orientation of the spatial activity. In particular, *hvloss* reveals if horizontal and vertical edges suffer of more blurring than diagonal edges. The parameter *hvgain* reveals if erroneous horizontal and vertical edges are introduced in the form of blocking or tiling distortions. These parameters are calculated over the space-time (S-T) regions of original and degraded frames. The S-T regions are described by the number of pixels horizontally, vertically and by the time duration of their region. An S-T region corresponds to 8×8 pixels over five frames. The complexity to calculate these parameters is rather high. Please, note that these parameters require the knowledge of the original sequence.

The fifth ANSI parameter SI_{13} is a reference-free measure of overall spatial information, since images were preprocessed using the 13×13 Sobel filter masks. It is calculated as the standard deviation over an S-T region of $R(i, j, t)$ samples, i and j being the coordinates within the picture displayed in time t. The result is clipped at the perceptibility threshold P [18]:

$$SI_{13} = \max_{time_t} \left\{ \mathrm{std}_{space_{i,j}} \Big[R(i,j,t) \Big] \right\} \Big|_P : i, j, t \in \{\text{S-T region}\}. \tag{4.23}$$

This feature is sensitive to the changes in the overall amount of spatial activity within a given S-T region. For instance, localized blurring produces a reduction in the amount of spatial activity, whereas noise produces an increase of it. Well-known reference-free parameters [9] describing the video sequence character are also SI (2.1) and TI (2.3). Finally, the codec compression settings FR and BR were investigated, requiring no computational complexity for estimation as they are known at both sender and receiver. The video sequences were encoded on constant BR, therefore it can be assumed that $BR = \overline{BR}$.

For the reduction of the dimensionality of the data set while retaining as much information as possible, the PCA [90] was used. The PCA was carried out to determine the relationship between MOS and the objective video parameters and to identify the objective parameters with the lowest mutual correlation, allowing us to propose a compact description of the data set. The PCA was performed for all content classes separately. In this case the first two components proved to be sufficient for an adequate modeling of the variance of the data, because the total variance of the first two components was at least 81% for all content classes. Variability describes the percentual part of data sets variance that is covered by the variance of particular component. Each of the nine parameters and MOS is

represented in the **Figures 4.23**, **4.26**, **4.24**, **4.25**, **4.27** by a vector. The direction and length of the vector indicates how each parameter contributes to the two principal components in the graph.

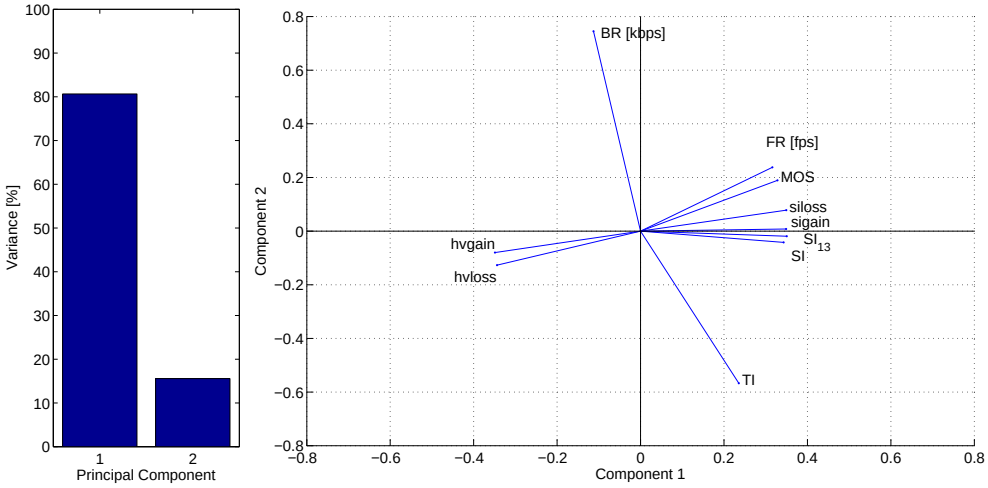

Figure 4.23: Visualization of PCA results for content News (CC1).

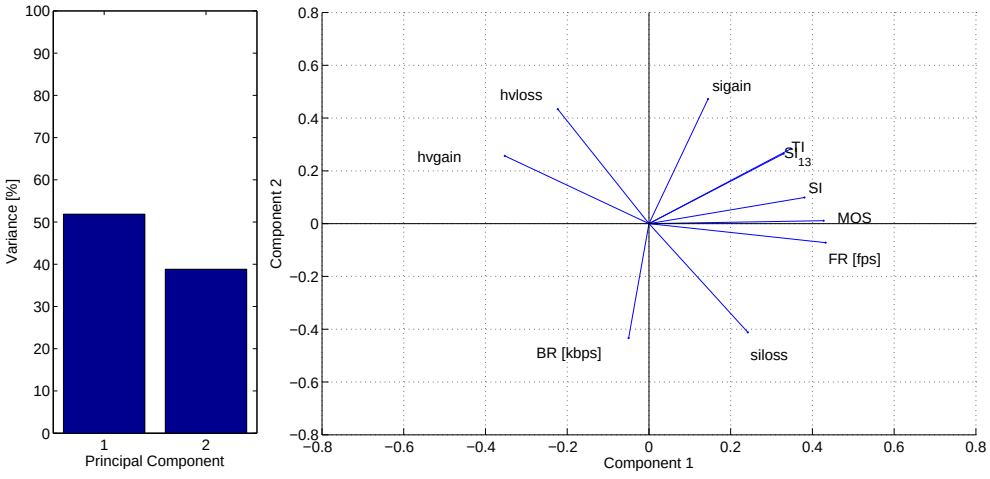

Figure 4.24: Visualization of PCA results for content Soccer (CC2).

According to this analysis, each parameter in the analyzed data set can be assessed concerning its contribution to the overall distribution of the data set. This is achieved by correlating the direction of the maximum spread of each variable in the direction of each principal component axis (eigenvector). A high correlation between PC1 and the investigated parameters indicates that the variable is associated with the direction of the maximum amount of variation in the data set. A strong correlation between the parameters and the PC2 indicates that the variable is responsible for the next largest variation in the data, perpendicular to PC1.

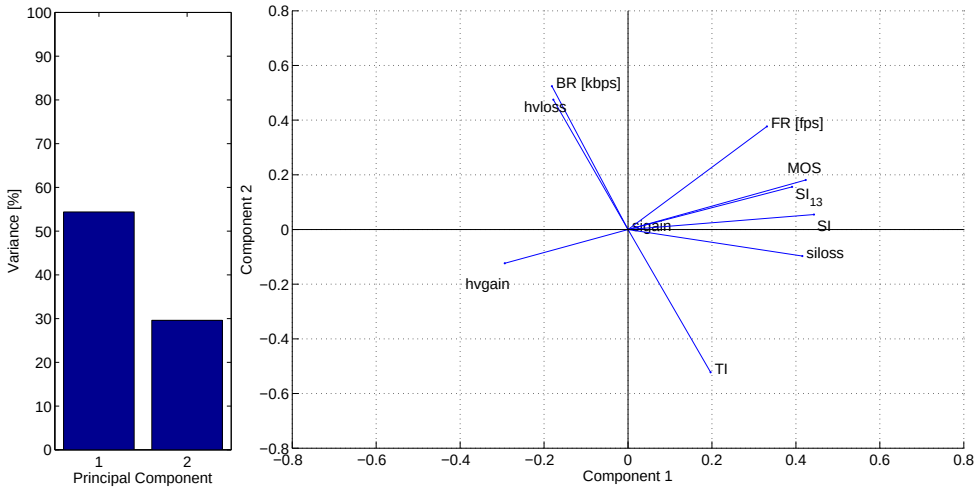

Figure 4.25: Visualization of PCA results for content Panorama (CC4).

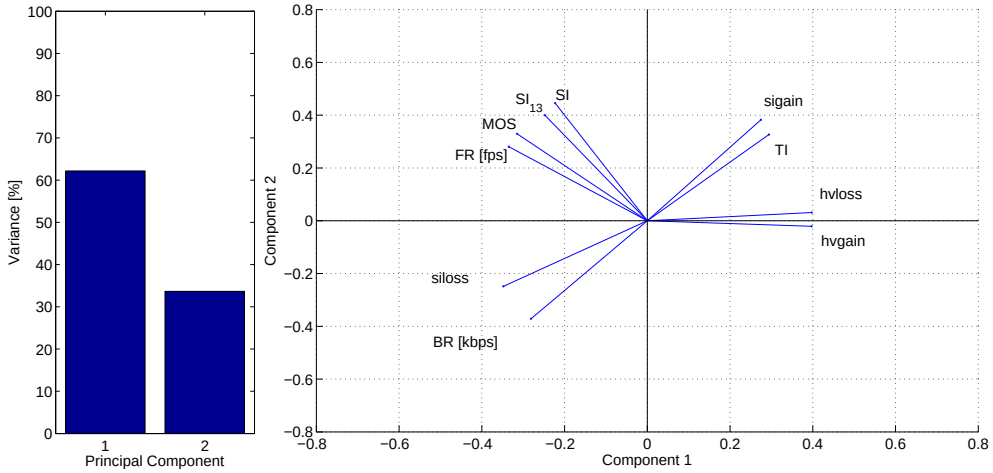

Figure 4.26: Visualization of PCA results for content Videocall (CC6).

Conversely, if a parameter (vector) does not correspond to any principal component axis and its length is small compared to the principal component axis dimensions, this usually suggests that the variable has little or no control on the distribution of the data set. Therefore, the PCA suggests which parameters in the data set are important and which ones is of little consequence. Moreover, the length and direction of parameter vector.

According to the PCA results the most suitable objective parameters are determined for the metric design relevant for all content classes. Surprisingly, the objective parameters with higher complexity (*sigain*, *siloss*, *hvgain*, *hvloss*) did not correlate better as low complexity parameters with PC1 and PC2 for all sequences. PCA results (see **Figures 4.23**, **4.26**, **4.24**, **4.25**, **4.27**) and high complexity of these objective parameters show us that these parameters are not appropriate for metric design in

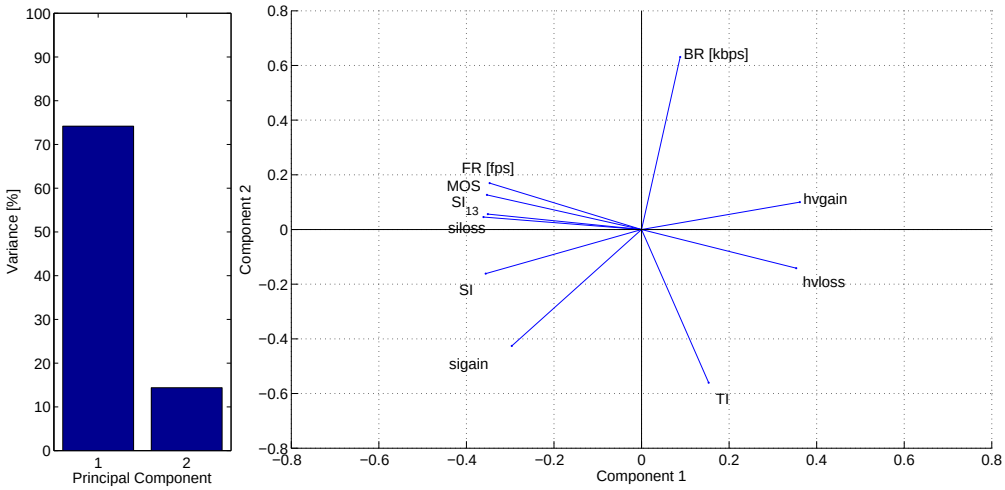

Figure 4.27: Visualization of PCA results for content Traffic (CC7).

our scenario.
After considering all content classes, the complexity of the objective parameters, and correlation between MOS and parameter vectors of video quality parameters. The following parameters were selected:

- FR has almost zero computational complexity and very well correlates with parameter vector MOS over all CCs,
- BR has almost zero computational complexity,
- SI_{13} very well correlates with parameter vector MOS over all CCs.

4.6.2 Direct reference-free quality estimation

Due to different spatial and temporal sequence characteristics of investigated CCs the proposed metric [4] has different coefficient values for each CC. Furthermore, the proposed metric is based on three objective parameters, encoding parameters FR and BR and sequence character parameter SI_{13}, according to the correlation with PC and their complexity. The uniform mathematical model for all content classes was chosen due to its simplicity and rather good fit with the measured data.

$$\begin{aligned} MOS = \quad & K_{CC} && +A_{CC}\cdot \mathrm{BR} && +B_{CC}\cdot \mathrm{FR} \\ & +C_{CC}\cdot \mathrm{SI}_{13} && +D_{CC}\cdot \mathrm{BR}\cdot \mathrm{FR} && +E_{CC}\cdot \mathrm{BR}\cdot \mathrm{SI}_{13} \\ & && +F_{CC}\cdot \mathrm{FR}\cdot \mathrm{SI}_{13} && +G_{CC}\cdot \mathrm{BR}\cdot \mathrm{FR}\cdot \mathrm{SI}_{13}. \end{aligned} \tag{4.24}$$

As initial step the simple linear model was used for metric design. The model contains only the first four elements (K_{CC}, A_{CC}, B_{CC}, C_{CC}) of model (4.24). The results in Table **4.8** show already a satisfying fit. Moreover, this confirms our choice of objective parameters.

In order to design a more accurate estimator the simple linear parameter combination was improved with four mixed terms reflecting all possible mutual combinations selected objective parameters

Coeff.	News	Video call	Soccer	Panorama	Traffic
K_{CC}	−78.4283	3.5970	−6.1850	−12.6834	−11.1982
A_{CC}	−0.0302	0.0411	0.0241	0.0226	0.0322
B_{CC}	0.1382	−0.0371	−0.0117	−0.0688	−0.02701
C_{CC}	0.9252	−0.0288	0.1237	0.1429	0.14415
r	0.9500	0.9600	0.9800	0.9300	0.9000

Table 4.8: Coefficients of linear metric model and correlation of average MOS with so obtained estimation for all content classes.

Coeff.	News	Video call	Soccer	Panorama	Traffic
K_{CC}	−39.0282	50.6525	98.3703	134.2721	−181.0251
A_{CC}	2.1618	1.4769	−1.1826	−3.4122	2.2139
B_{CC}	−3.2939	−4.9962	−13.7067	−33.2325	11.1780
C_{CC}	0.4491	−0.5986	−1.7714	−1.2473	2.0426
D_{CC}	−0.1338	−0.0936	0.1623	0.7717	−0.0859
E_{CC}	−0.0234	−0.0158	0.0219	0.0325	−0.0244
F_{CC}	0.0407	0.0592	0.2479	0.3134	−0.1256
G_{CC}	0.0014	0.0010	−0.0029	−0.0073	0.0010
r	0.9890	0.9970	0.9960	0.9990	0.9740

Table 4.9: Coefficients of improved metric model and correlation of average MOS with so obtained estimation for all content classes.

of (4.24). This extension introduced higher accuracy on cost of additional complexity. Furthermore, in Tables **4.8** and **4.9** it can be clearly seen that the metric model coefficients are different for each content class. This suggests that the choice of the content classes is right and subjective video quality is content dependent.

4.6.3 ANN based quality estimation

In this section the design of an Artificial Neural Network (ANN) with low complexity for the estimation of visual perceptual quality is presented [112], based on a combination of a possibly small set of the objective parameters (compression settings and content features). To achieve this, the neural network was trained with a set of objective and subjective parameters, obtained by an extensive survey. Inputs of proposed ANN are three reference-free measures, two encoding parameters BR, FR, and the sequence character parameter SI_{13} with the estimated MOS as output (see **Figure 4.28**).

In multi-layer networks, with any of a wide variety of continuous nonlinear hidden-layer activation functions, one hidden layer with an arbitrarily large number of units suffices for the "universal approximation" property [101], [102]. According to this knowledge the network with three layers was designed - input, one hidden and output layer (**Figure 4.29**). But there is no theory yet to

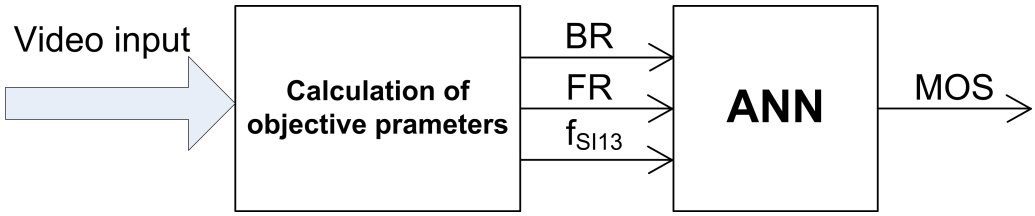

Figure 4.28: ANN for MOS estimation.

determine, how many hidden units are needed to approximate any given function. The best number of hidden units depends in a complex way on: the numbers of input and output units, the number of training cases, the amount of noise in the targets, the complexity of the function or classification to be learned, the architecture, the type of hidden unit activation function, the training algorithm and the regularization [103]. In most situations, there is no way to determine the best number of hidden units without training several networks and estimating the generalization error of each. If there are too few hidden units, high training errors and high generalization errors are obtained due to underfitting and high statistical bias. If there are too many hidden units, low training errors still having high generalization errors are obtained due to overfitting and high variance [104]. The error on the training set is driven to a very small value, but when new data is presented to the network the error increases. The network has memorized the training examples, but it has not learned to generalize new situations. Assume for example, if there is one continuous input X that takes values on the interval (0, 100) and if there is one continuous target $Y = \sin(X)$. In order to obtain a good approximation to Y, about 20 to 25 hidden units with tangents-hyperbolic function are required, although one hidden unit with a sine function would do the job [105]. A possible way how to improve generalization is to have a network that is just large enough to provide an adequate fit because it can approximate a more complex function. In case of a sufficiently small network, it will not have enough power to overfit the data. But it is difficult to know beforehand how large a network should be for a specific application. Another way is to have much more points in a training data set than network parameters, avoiding the chance of overfitting.

A typical recommendation is that the number of weights should be not more than 1/30 of the number of training cases [105]. Such rules are only concerned with overfitting and are at best crude approximations. Also, these rules do not apply when regularization is used. It is true that without regularization, if the number of training cases is much larger (but no one knows exactly how much larger) than the number of weights, overfitting or underfitting appears more often. For a noise-free quantitative target variable, twice as many training cases as weights may be more than enough to avoid overfitting [106].
The lack of training data in this case requires to improve on the generalization. A few training methods (Variable learning rate, Resilient backpropagation, Quasi-Newton algorithm) was tested, but generalization was insufficient. Finally, the methods improving the generalization were applied. This method is called Automated regularization [107], that is a combination of Bayesian regularization [108], [109], and Levenberg-Marquardt training [110]. The weights and biases of the network are assumed to be

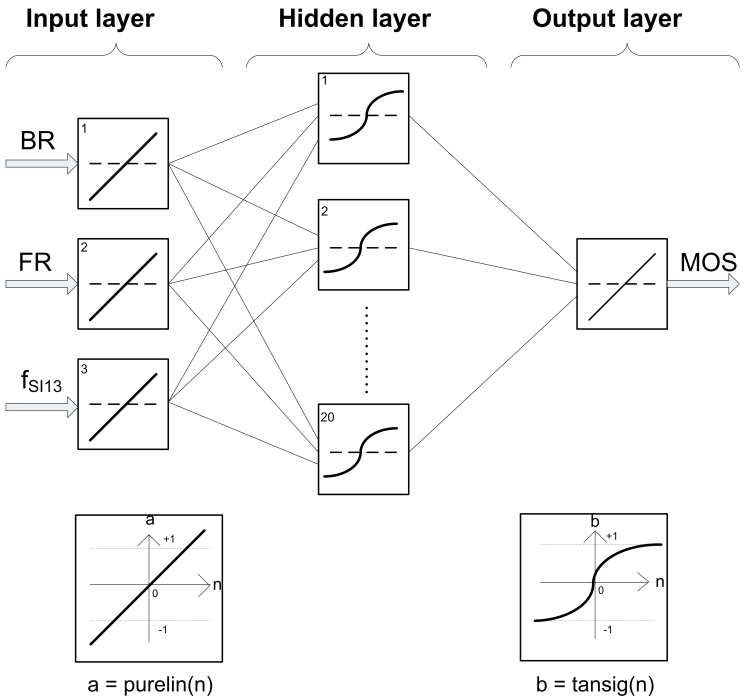

Figure 4.29: Architecture of the proposed three-layer feedforward ANN.

random variables with specified distributions. The regularization parameters are related to the unknown variances associated with these distributions. One feature of this algorithm is that it provides a measure of how many network parameters (weights and biases) are being effectively used by the network. The inputs are scaled and targets so that they fall in the range [-1,1], because this algorithm generally works best when the network inputs and targets are scaled so that they fall approximately in the range. The outputs were converted back into the same units that were used for the original targets.

Once the network weights and biases were initialized, the network is ready for training. During training the weights and biases of the network are iteratively adjusted to minimize the squared error between the network outputs and the target outputs.

Finally, several ANNs were designed, in order to find a trade-off between the ANN's minimal number of neurons in the hidden layer and their accuracy. We used for the training 54 vectors (rows of matrix 4.25) with dimension of four ($\overline{BR}$, FR, SI_{13} and MOS), with three input values and one target value (see **Figure 4.29** and (4.25)).

$$\mathbf{X}^T = \begin{bmatrix} \overline{BR}_1 & FR_1 & SI_{13_1} & MOS_1 \\ \vdots & \vdots & \vdots & \vdots \\ \overline{BR}_{54} & FR_{54} & SI_{13_{54}} & MOS_{54} \end{bmatrix} \tag{4.25}$$

The three layered ANN architecture was introduced with three linear units in input layer and one linear unit in the output layer. The minimal training and generalization error was obtained for hidden layer which consists of 20 tangents-signoid neurons [111].

Moreover, a linear regression analysis between the network response and the corresponding target was

performed. This relationship between estimated (**y**) and target data (**x**) or between predicted and measured subjective MOS is represented in the form:

$$\mathbf{y} = m\mathbf{x} + b, \tag{4.26}$$

where m corresponds to the slope and b to the y-intercept of the best linear regression relating targets to the network outputs. If there is a perfect fit (outputs are exactly equal to targets), the slope is 1, and the y-intercept is 0. In our case we obtain results clearly close to optimum m =0,80215 and b=0,59462. Furthermore the correlation factor (4.22) between estimated and target data (or between predicted and subjective MOS) of the proposed ANN is 90,4% (see **Figures 4.30**).

4.6.4 Performance of the video quality estimators

Note that the direct reference-free quality metric from Section 4.6.2 is not general but dedicated to a certain content class. On the other hand, the ANN model is general for all content classes and training performance is sufficient video quality estimation. Therefore, a better fittng performance of the direct reference-free quality metric was expected.

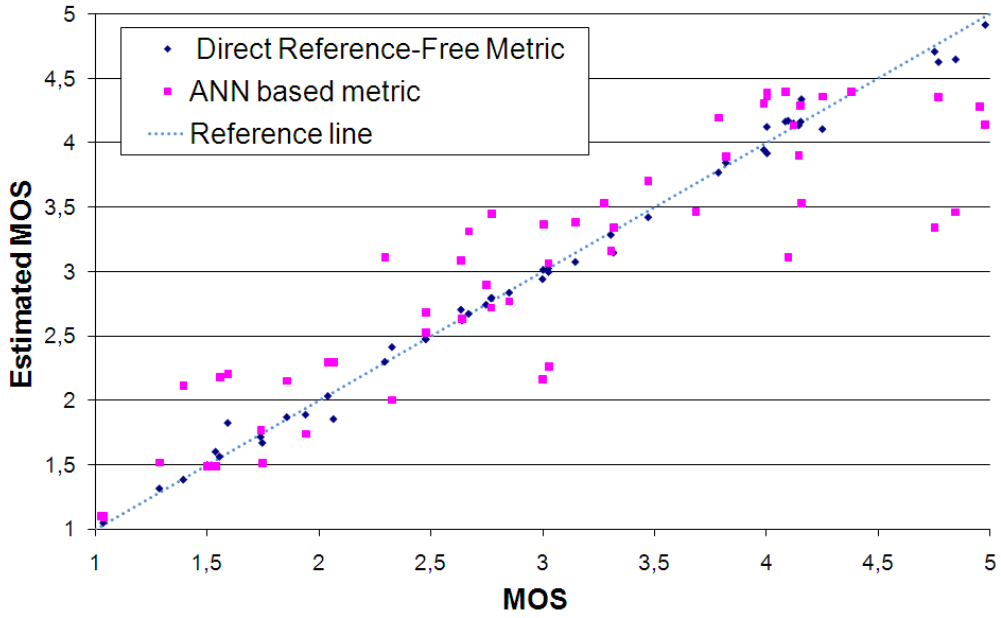

Figure 4.30: Estimated vs. subjective MOS results.

A further metric performance evaluation was executed with two different sequences. The test sequences were approximately ten-seconds long, in order to keep scene integrity. The first sequence was a soccer sequence (CC2) and the second sequence was a typical talking head scenario (CC6). To validate the performance of the proposed metrics, the Pearson (linear) correlation factor (4.22) was applied. The performance of direct reference-free quality metric was 0.96 and the ANN based estimator achieved 0.89 (see Figure **4.30**). Furthermore, the results in Table **4.11** clearly show good agreement between the obtained and the estimated values for both proposed metrics. In addition, the goodness of the fit on different content classes (see Table **4.10**) was investigated. Both metrics have

Metric/Content type	CC 1	CC 2	CC 4	CC 6	CC 7
Direct Reference-Free metric	0.99	0.99	0.99	0.99	0.98
ANN based metric	0.91	0.98	0.97	0.98	0.97

Table 4.10: Metric prediction performance on defined CC by Pearson correlation.

Metric	Pearson correlation
Direct Reference-Free metric	0.96
ANN based metric	0.89

Table 4.11: Metrics prediction performance by Pearson correlation.

over 97% prediction performance over particular content classes. The lower ANN accuracy is the prize for its universality. Moreover, the validation results for significantly different content types show, that our ANN takes also content character into account.

4.7 Summary of video quality estimation

In this section video quality estimators were proposed for the mobile environment. Initially, the research was focused on quality estimation for H.263 codecs and QCIF resolution. This streaming setup was mainly used in mobile environment due to processing limitations of handheld devices and license conditions of the H.263 codec. Moreover, the H.263 codec is mandatory for UMTS video services. The proposed estimators perform very well for this streaming setup.
The recent development in handheld devices and video encoding brought significant improvement in processing power of handheld devices thus allow increasing screen resolution of these devices. Moreover, H.264 codec, which significantly outperform H.263 in coding efficiency, became the state of the art codec for mobile video services. This brought the research attention to H.264 streaming at SIF resolution. The metrics proposed for this scenario show good performance on all investigated content classes. The scenarios for H.264 streaming at QCIF resolution and H.263 streaming at SIF resolution were not investigated because mobile video services are provided only marginally with these settings. Finally, accurate video quality estimators were proposed for the most frequent content types and streaming setups. On the other hand, the drawback of the proposed estimation methods is dependency from video codec and resolution. Moreover, it was not possible to set selection criteria of video quality parameters precisely because the PCA results for QCIF and SIF resolution show similar performance for all investigated video parameters.

Chapter 5

Audiovisual quality estimation

Contents

5.1 Introduction

THE majority of the recently proposed metrics for quality estimation assumes only one continuous medium, either audio [113], [114], [115] or video (cf. Section 1.4). The focus of this chapter is to estimate the quality of mobile multimedia in form of audio and video signals at the user-level. In order to predict the audiovisual quality of a multi-media system it is necessary to propose a metric that takes into account both the audio and the video quality [100]. When assessing a multi-modal system, one cannot model it only as a simple combination of mono-modal models, because the pure combination of audio and video models does not give a robust perceived-quality performance metric [116]. So, it is imperative to take into account the cross-modal interaction between audio and video modes. In this particular case the cross-modal interaction was introduced by proposing independent models for speech and non-speech scenarios including mixed-terms.
Various auditory and visual models have served as basis for multi-modal predictions [117], [118]. They consider how the audio and video signals are perceived by people. In this way the audio and video signals are perceptually weighted before they are combined in a multi-modal model. But the multi-modal model must also account for cross-modal interactions as well as task influences in order to give a task related perceived performance metric. Perceptual interaction is indeed dependent on the nature of the task undertaken. In multi-media systems video and audio modes not only interact, but there is even a synergy of component media. It is a human ability to make up for the lack of information from one sense with the other senses. A similar observation may be made on a multi perceptual model. In video telephone system, for instance, even if the quality of the video stream is somewhat low, the voice stream with good quality can compensate for the degradation of the overall perceptual quality, and vice versa [118]. In other words, we expect that different media compensate for each other from a perceptual point of view.
As initial step video and audio objective parameters describing the character of the sequence are investigated. Due to the complexity of this task only the objective parameters of known audio and video quality estimation models were investigated. Furthermore, according to their relevance for our scenario the most suitable were chosen. Finally, a model for audiovisual quality evaluation is proposed.

5.2 Audio and audiovisual subjective quality tests

THE intention of this section is to investigate the relation between perceived audio quality and video quality. Moreover, the focus was given on impact of encoding algorithms for video and audio and encoder settings on perceived audiovisual quality. For audiovisual quality tests only three sequences were selected due to the size of the combination set. The first two sequences Cinema trailer and Video clip belong to content class CC5. The main difference between them is in their audio part. In the cinema trailer the music is only accompanied to video and no voice is present. In the Video clip instrumental music with voice is present in the foreground. The next sequence belongs to content class CC6, in which the only audio material is a speech monologue.
From the obtained results for audiovisual quality it can be seen that audio quality, video quality as well as sequence character are important factors to determine the overall subjective perceived quality.

Content	Codec combination	Audio BR [kbps]	Video BR [kbps]	MOS_{av}
Video call	H.263/AMR	7.9	97	3.1
Cinema Trailer	MPEG4/AAC	16	59	3.8
Video Clip	MPEG4/AAC	24	51	3.9

Table 5.1: The best trade-off audiovisual quality.

A mutual compensation property of audio and video can also be clearly observed from our results at sequences encoded below 75 kbps. This effect was more dominant for cinema trailer and video clip contents. In the video call content the MOS_{av} is more influenced by the audio quality than the video quality. In Table **5.1** the best encoding settings are selected for audiovisual quality if minimal BR is considered according to following equation:

$$\min_{enc.\ param.} \{BR\} \text{ s.t. } MOS_{av} \geq 3 \tag{5.1}$$



Figure 5.1: Snapshots of selected sequences for audiovisual test: Cinema trailer (left), Video clip (middle), Video call (right).

5.2.1 Audio subjective quality tests

In order to investigate the mutual compensation property between audio and video the audio tests were performed independently. Moreover the performance of Advanced Audio Coding (AAC) and Adaptive Multi-Rate (AMR) audio codecs on speech and music contents was investigated. The same audio content as for the audiovisual tests was used. In total 26 audio encoding settings (see Table **5.2**) were selected for subjective testing.

As can be seen in **Figures 5.2** and **5.3** the video call (CC6) content and the Cinema trailer/Video clip content (CC5) are perceived differently. This behavior is certainly due to the fact that the audio material is speech in the Video call, whereas it is music in the Video clip and in the Cinema trailer.

It can be clearly seen that the AMR codec operates very efficiently in the Video call (see **Figure 5.2**). For the Video call audio content high MOS audio values (MOS_a) for very low BR 5.9 kbps were obtained. Further improvement by increasing the BR is not considered. The mean value of MOS_a is 3.5 using the AMR codec. On the contrary, the AAC codec performs in dependence on the bit rate. The MOS_a is smaller than 2.5 for BRs 8 kbps and 16 kbps, whereas it is higher than 4 for BRs 24 kbps or 32 kbps.

AAC	8 kbps	clips 1,2,3
audio bit rate	16 kbps	clips 1,2,3
	24 kbps	clips 1,2,3
	32 kbps	clips 1,2,3
	48 kbps	clips 1,2
AMR	5,9 kbps	clips 1,2,3
audio bit rate	7,9 kbps	clips 1,2,3
	10,2 kbps	clips 1,2,3
	12,2 kbps	clips 1,2,3

Table 5.2: encoding settings for audio content. Clip 1 = Cinema trailer, Clip 2 = Video clip, Clip 3 = Video call.

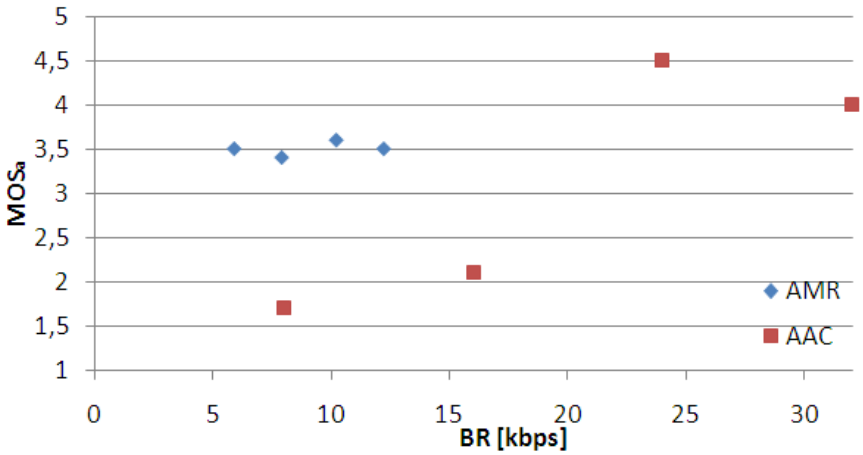

Figure 5.2: MOS_a results for the Video call audio content.

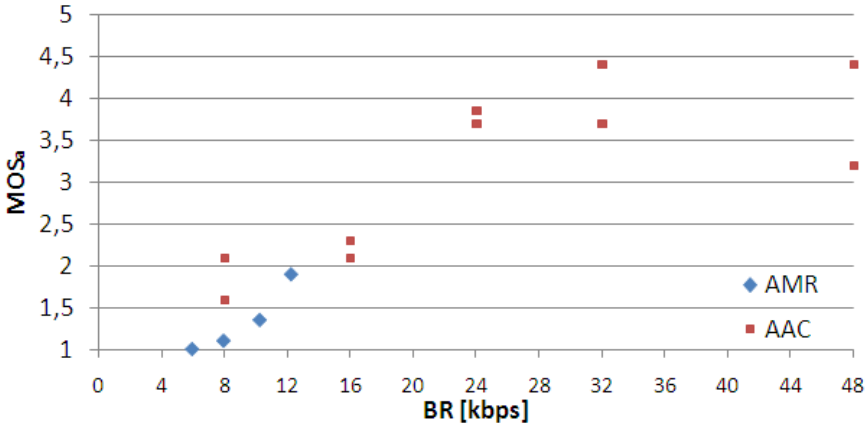

Figure 5.3: MOS_a results for the Cinema trailer/Video clip audio content.

The AMR codec performs inadequate for the cinema trailer/video clip video content. The obtained results are below 2.2 MOS_a grade. Moreover, for the investigated audio content the subjective audio quality improves with increasing BR. For BRs above 24 kbps the subjective quality is higher than 3 MOS_a grades.

5.2.2 Audiovisual subjective quality tests

H.263 and MPEG-4 video codecs, and AMR and AAC as audio codecs were chosen for audiovisual tests. In total there were 102 encoding combinations (see Table **5.3**) tested. The test methodology described in Section 2.1.2 was followed. The video FR was set to 7.5 fps according to previous experiences (cf. Section 3.1.1). The FR 7.5 fps provides the best trade-off for QCIF resolution between spatial and temporal video features. To evaluate the subjective perceptual audio and audiovisual quality a group of 20 people was chosen. The chosen group ranged different ages (between 17 and 30), gender, education and experience. The sequences were presented in an arbitrary order, with the additional condition that the same sequence (even differently degraded) did not appear in succession. In the further processing of data results we have rejected the sequences which were evaluated with individual variance higher than one. In total there were 7% of the obtained results rejected. Two rounds of each test were taken. The duration of one test round was about 40 minutes.

		Video + Audio BR		
		56 kbps	75 kbps	105 kbps
AAC	8 kbps	clips 1,2,3		
	16 kbps	clips 1,2,3	clips 1,2,3	
audio bit rate	24 kbps	clips 1,2	clips 1,2,3	clips 1,2,3
	32 kbps		clips 1,2	clips 1,2,3
	48 kbps			clips 1,2
AMR	5,9 kbps	clips 1,2,3	clips 1,2,3	clips 1,2,3
audio bit rate	7,9 kbps	clips 1,2,3	clips 1,2,3	clips 1,2,3
	10,2 kbps	clips 1,2,3		
	12,2 kbps		clips 1,2,3	clips 1,2,3

Table 5.3: Encoding settings for audiovisual test 36 combinations; for Cinema trailer (clip 1) and Video clip (clip 2), 30 combinations for Video call (clip 3).

5.2.2.1 Video call test results

Video calls contain lower amounts of spatial and temporal information leading to a loss of the critical ability of the subjective judgment. We can also observe in **Figures 5.4**, **5.5**, **5.6**, **5.7** (color code serves for better visualization of the results) that in the Video call scenario the MOS_{av} depends more on the audio quality than on the video quality; the first reason is because in this case the audio information is more important than the video; the second reason for this is that the video encoding efficiency does not significantly influence the audiovisual quality due to low structural complexity of

video call content. Therefore, the obtained MOS_{av} results show that the most suitable combination for the Video call scenario is a combination of a H.263 video codec with AMR audio codec. The best MOS_{av} trade-offs between audiovisual quality and BR were achieved for the combinations H.263/AMR and MPEG4/AAC at 105 kbps with MOS_{av} grades around 3. If the amount of processing power is considered, the combination H.263/AMR is more suitable due to its simplicity in comparison to the MPEG4/AAC combination.

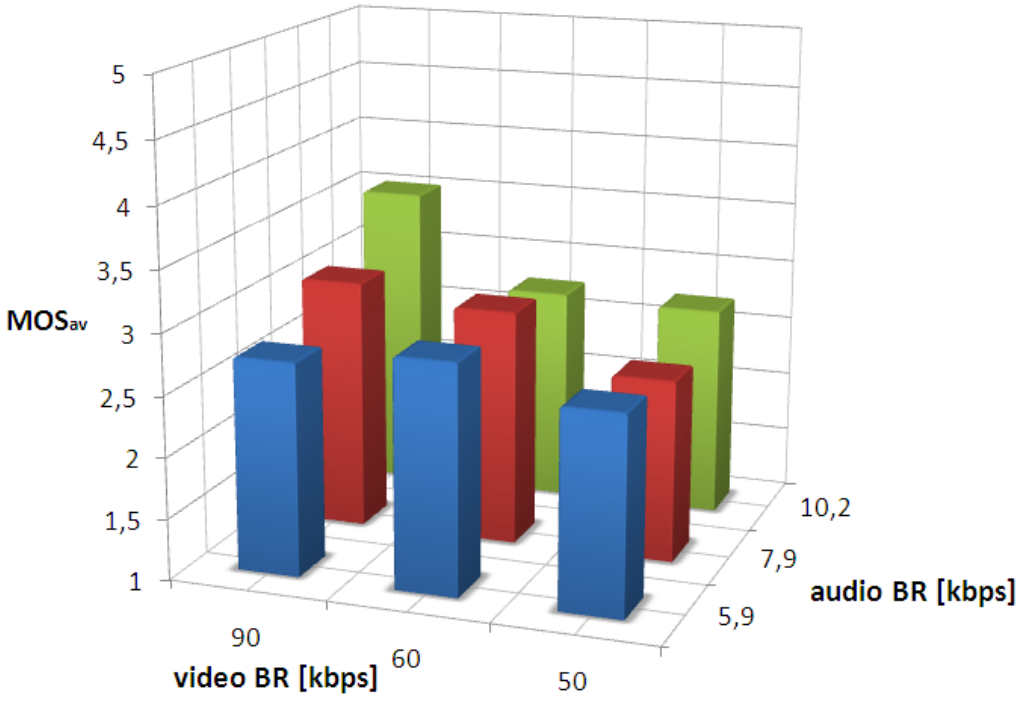

Figure 5.4: MOS_{av} results for the Video call content - codecs combination H.263/AMR.

5.2.2.2 Video clip and Cinema trailer test results

For Cinema trailer and Video clip we obtain better MOS results with MPEG-4 than with H.263 (see **Figures 5.8**, **5.9**, **5.10**, **5.11**, **5.12**, **5.13**, **5.14** — color code serves for better visualization of the results) because these sequences contain a lot of spatial and temporal changes (fast camera movements, scene cuts, zoom out/in). It was evident that the AMR codec cannot achieve sufficient results for music and non speech audio content. The audiovisual subjective quality judgment is below 2.8 MOS_{av} (see **Figures 5.8**, **5.12**, **5.9**). Moreover, the obtained results for all combinations with the H.263 codec are below MOS_{av} grade 3. It can be assumed that for high structured and dynamic contents H.263 and AMR codecs are not suitable. For fast movement sequences the mutual compensation effect can be clearly observed. The MOS_{av} is significantly more influenced by audio quality for the lowest bit rate (56 kbps). It is caused by a loss of spatial information due to the compression, where higher audio quality can compensate the lower video quality. On the other hand MOS_{av} is not strongly influenced by audio quality for the higher bit rates (75, 105 kbps).

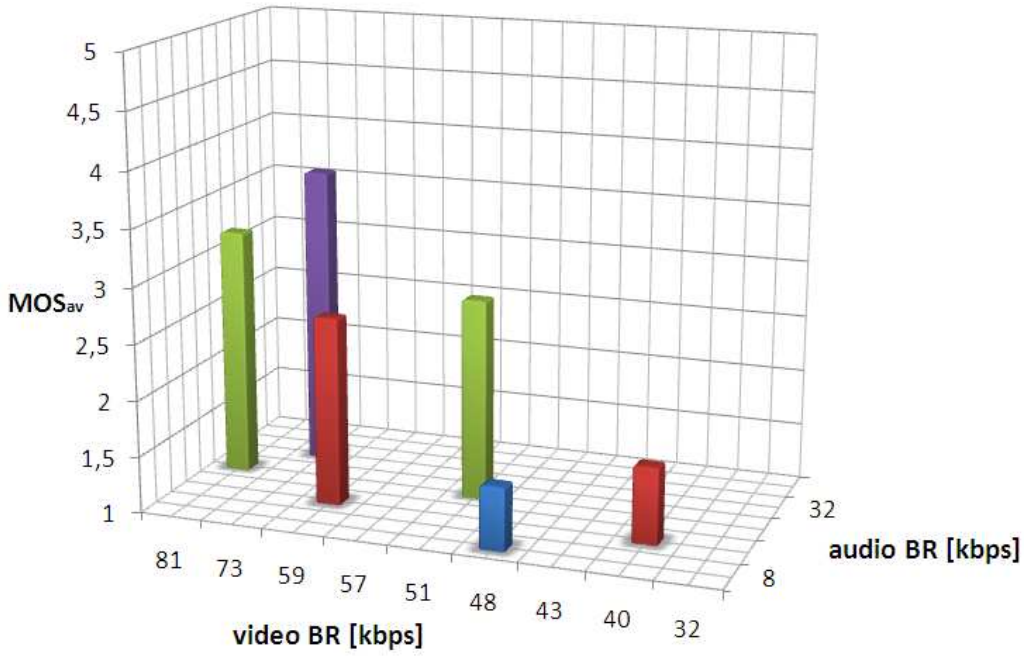

Figure 5.5: MOS_{av} results for the Video call content - codecs combination H.263/AAC

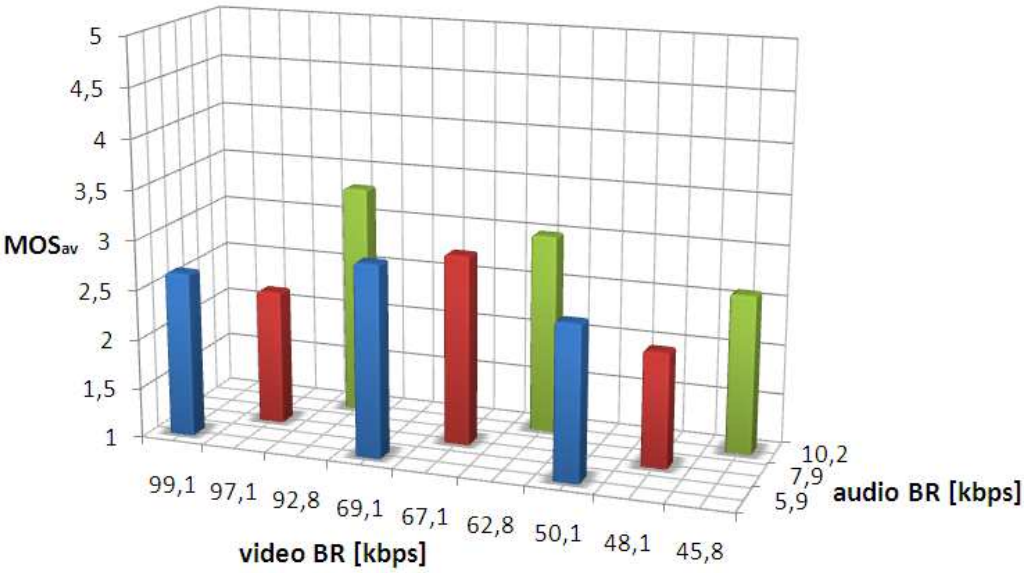

Figure 5.6: MOS_{av} results for the Video call content - codecs combination MPEG4/AMR.

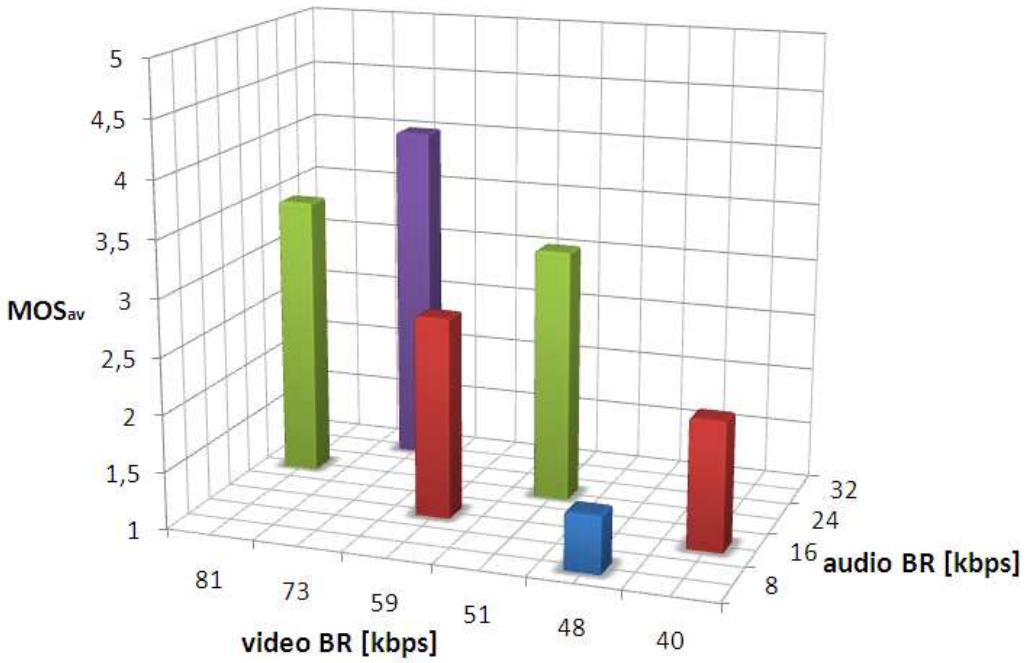

Figure 5.7: MOS_{av} results for the Video call content - codecs combination MPEG4/AAC.

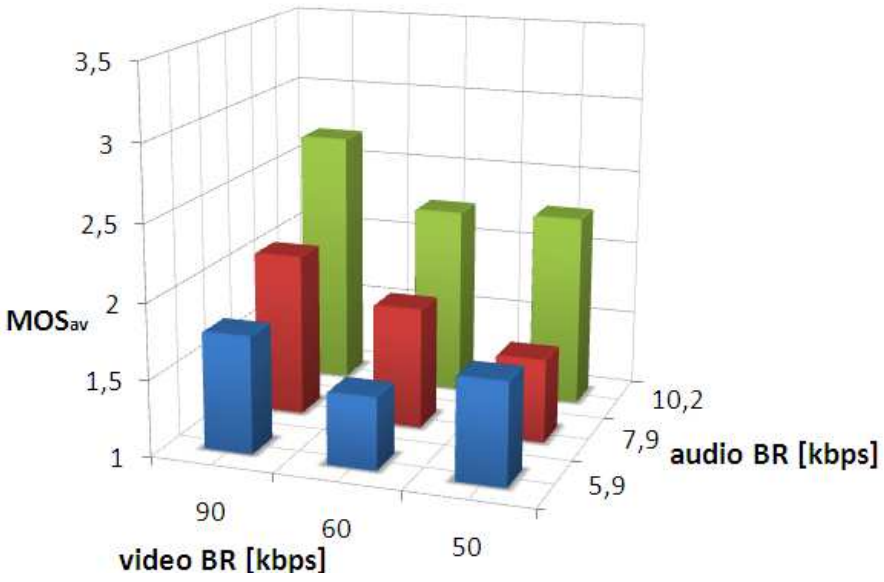

Figure 5.8: MOS_{av} results for the Cinema trailer content - codec combinations H.263/AMR.

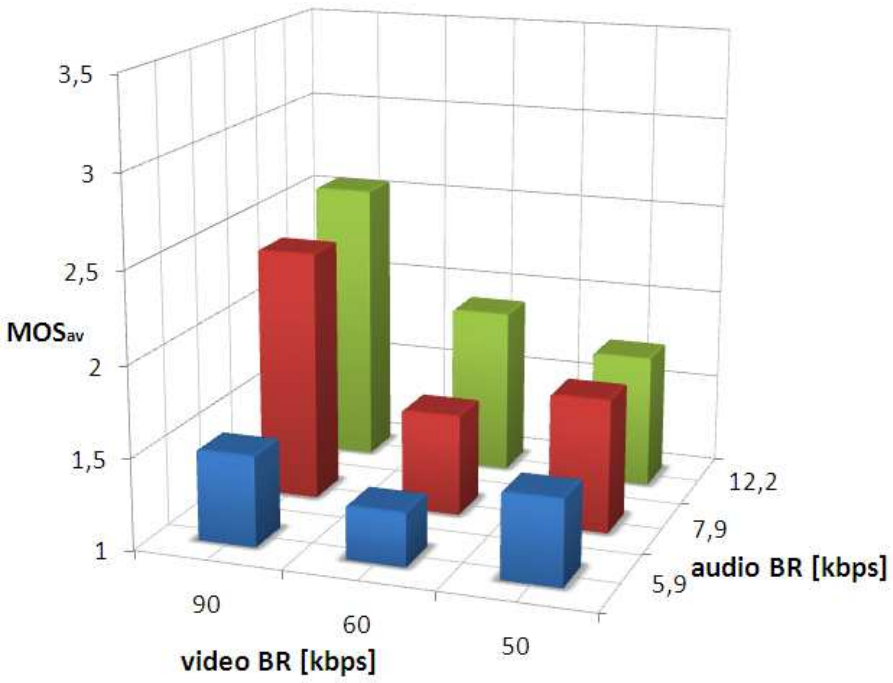

Figure 5.9: MOS_{av} results for the Video clip - codecs combination H.263/AMR.

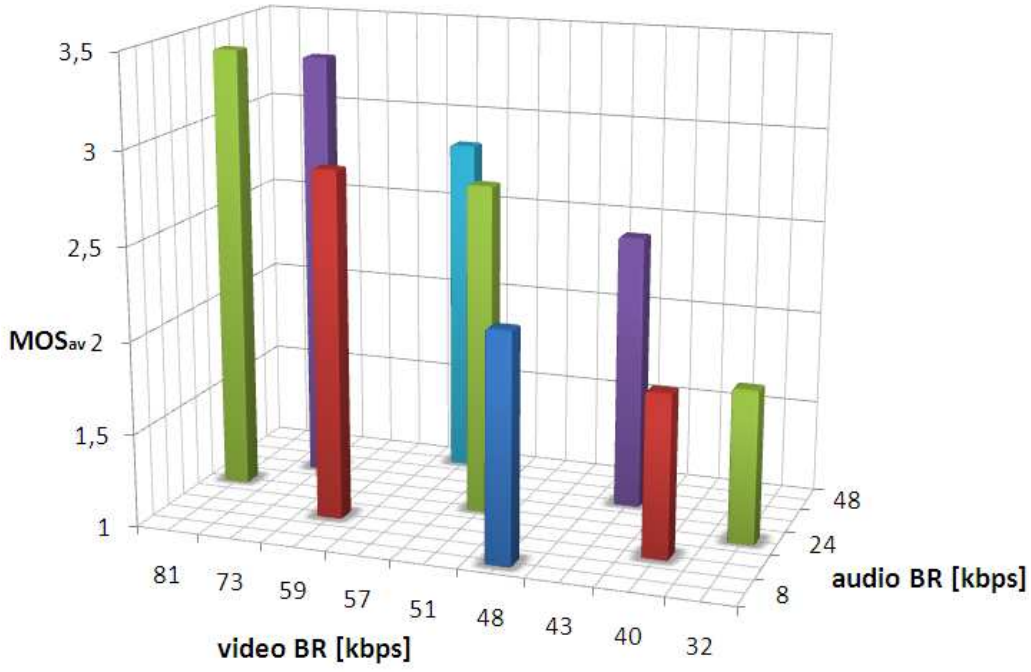

Figure 5.10: MOS_{av} results for the Cinema trailer content - codecs combination H.263/AAC.

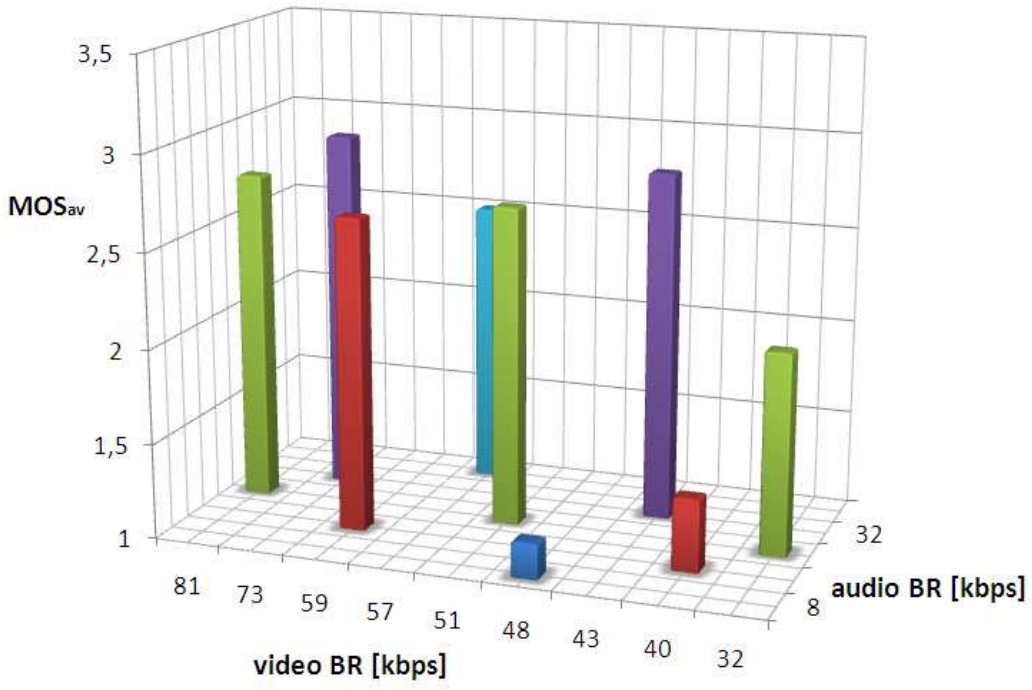

Figure 5.11: MOS_{av} results for the Video clip - codecs combination H.263/AAC.

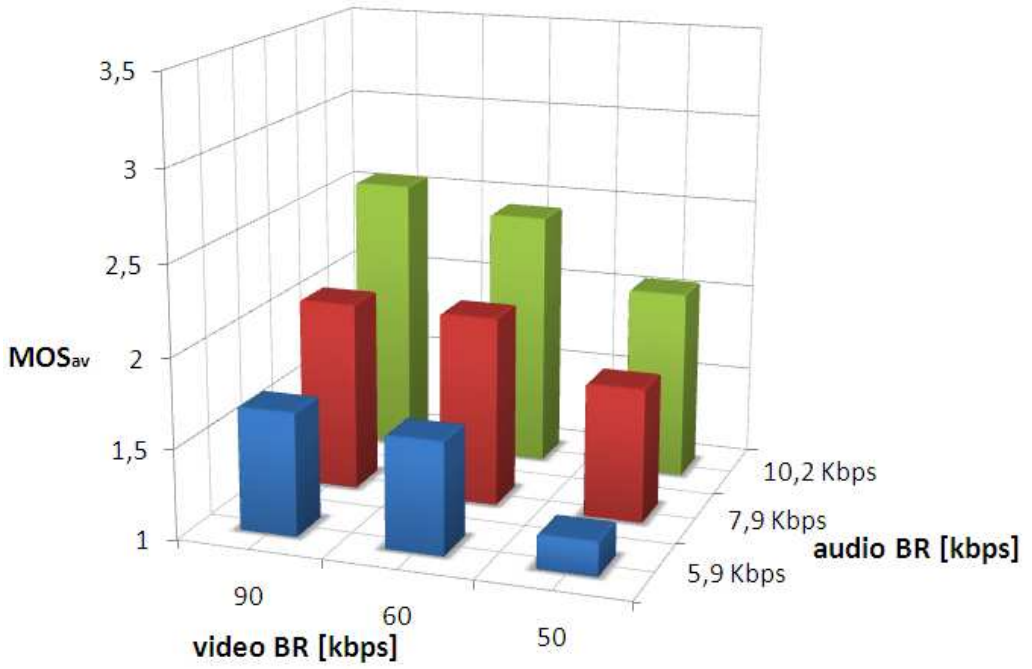

Figure 5.12: MOS_{av} results for the Cinema trailer content - codecs combination MPEG4/AMR.

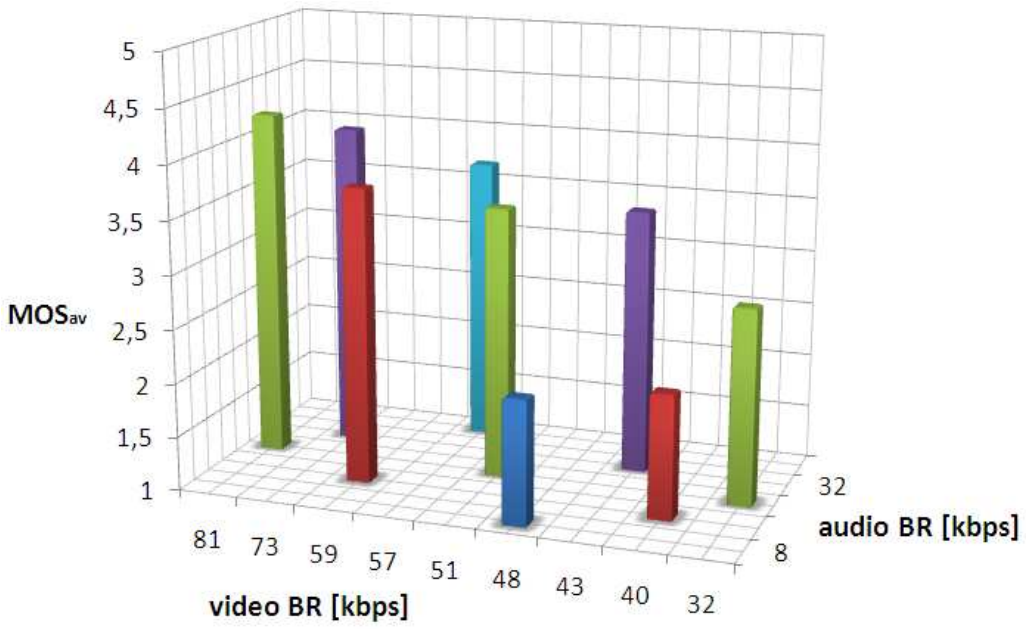

Figure 5.13: MOS_{av} results for the Cinema trailer content - codecs combination MPEG4/AAC.

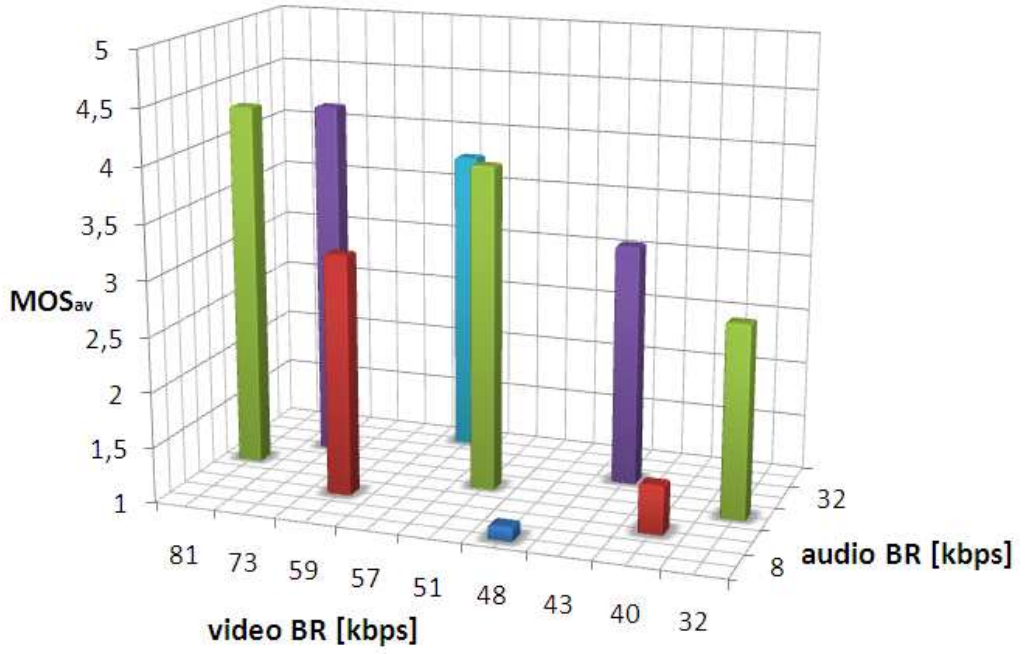

Figure 5.14: MOS_{av} results for the Video clip - codecs combination MPEG4/AAC.

5.3 Audio and video parameters

DUE to the complexity of audiovisual quality estimation the already known audio and video metrics were investigated and used for further audiovisual quality estimation. For video quality estimation the ANSI T1.803 [18] metric was investigated and for audio and speech quality estimation the metrics in [113] and [115] were investigated. Finally, the metrics were tuned for a mobile scenario according to an obtained audiovisual quality survey.

5.3.1 Video parameters

First, objective video parameters defined by the ANSI standard T1.803 [18] were investigated. The ANSI standard in [18] defines seven objective parameters based on spatial, temporal and chrominance properties of video streams. In order to decrease their complexity a reduced ANSI metric was proposed. The focus was given at the four most relevant ones:

- $sigain$,
- $hvloss$,
- $siloss$,
- $hvgain$,

because in the former study [119], [120], these parameters are the most significant for quality estimation in this particular scenario.
The relevance of these parameters was explained and investigated in Section 4.6.1 and further parameter details are described in Appendix B. Thus, the following video quality metric based only on the four most important parameters [119], [120] was proposed:

$$\begin{aligned} MOS_v = \; 5 \; & +0.8388 \cdot siloss \quad -2.3876 \cdot hvloss \\ & -0.9932 \cdot hvgain \quad +9.3664 \cdot sigain|^{0.56}, \end{aligned} \tag{5.2}$$

where the positive parameter $sigain$ is clipped at an upper threshold of 0.56 (if $sigain > 0.56$ than set $sigain = 0.56$) which indicates the maximum improvement of the video quality observed in the encoded sequences [18]. It is thus a shortened ANSI metric with the four most relevant objective video parameters for mobile video scenario. Furthermore, the output of the ANSI metric was mapped to the five grade MOS scale.

5.3.2 Audio parameters

Different subjective evaluation was noticed between the subjective evaluation of speech and music in audio and in the audiovisual survey. Therefore, it was necessary to design two independent metrics for speech and for music. For speech quality evaluation we have adopted the Auditory Distance (AD) parameter, according to [115]. It measures the dissimilarities between the original and the compressed speech signals. The main components in the AD metric are the delay estimation between the two input vectors of speech samples, the perceptual transformation and the distance measure (see cf. Appendix C) [115]. Designing an audio quality metric we have noticed a difference in the subjective audio evaluation when the sequences are encoded with the codec AMR or AAC. The maximal correlation between our quality prediction and the measured MOS_a (MOS audio) is obtained

by linear regression with a translation of the speech audio metric in the two cases of AMR (5.3) and AAC coding (5.4).

$$MOS_a^{AMR} = -6.996AD^2 + 10.95AD + 1.165, \tag{5.3}$$

$$MOS_a^{AAC1} = -6.996AD^2 + 10.95AD + 0.370. \tag{5.4}$$

AD is here normalized between 0 and 1; further details can be found in (C.7). The reason for this translation is due to the operation of the codecs. The AAC codec utilizes a wider range of frequencies; thus it degrades objectively the signal less than AMR. However, the subjective audio evaluation is higher for AMR. Indeed AMR is a codec designed for speech. It degrades the signal in a way that human ears do not perceive it. Therefore, although the objective degradation is stronger for AMR, the subjective speech evaluation is higher. The fit of proposed metrics according to the Pearson correlation factor (4.22) is 98% for the AMR (5.3) and 84% for the AAC (5.4) metric.

For music quality evaluation were used a more suitable audio metric according to [113]. The original and the encoded streams were splited into 32 ms frames with 50% overlap [113]. Successively each frame of the two signals was transformed in the perceptual Bark frequency scale [113]. In this way we obtain both temporal and frequency information of the original and the encoded signals. According to the internal representation of audio signals in the human auditory system, the signals are elaborated through Zwicker's law [121] that takes into account how the human ears perceive sound loudness. The first parameter [113], Integrated Frequency Distance (IFD), measures how much the powers of the original and of the encoded signals diverge. The IFD is the integrated difference between the non compressed audio signal and the compressed one (see cf. Appendix C). The other two parameters, denoted as D_n and DA_n (disturbance indicators) [113] (see cf. Appendix C), consider how much the presence of noise and the loss of time-frequency components influence the audio quality.
The resulting music audio metric is a linear combination of the parameters IFD, D_n and DA_n obtained by linear regression:

$$MOS_a^{AAC2} = 3.1717 + \frac{4.8809}{IFD} + 0.3562 \cdot D_n + 0.0786 \cdot DA_n. \tag{5.5}$$

This metric exhibits 91% for AAC codec correlation (4.22) with the subjective evaluation.

5.4 Audiovisual model

IN the Video call scenario the MOS_{av} (MOS audiovisual) depends more on the audio quality than on the video quality, because in this case the audio information is more important than the video. Therefore, the obtained MOS_{av} results show that the most suitable combination for the "video call" scenario is a combination of H.263 video codec with the AMR audio codec (**Figure 5.4**). It makes no sense to use an AAC codec, because AAC needs higher throughput for the same perceptual quality performance of human speech. For Cinema trailer and Video clip scenarios we obtain better MOS results with MPEG-4 than with the H.263 (see **Figures 5.13** and **5.14**) because these sequences contain a lot of spatial and temporal changes (fast camera movements, scene cuts, zoom out/in). It was

Model	K_{cc}	A_{cc}	V_{cc}	AV_{cc}	A'_{cc}	V'_{cc}	correlation
MOS^{I}_{av}	-0.4934	0.5420	0.4327	/	/	/	0.8800
MOS^{II}_{av}	0.9987	/	/	0.1536	/	/	0.8915
MOS^{III}_{av}	0.6313	0.2144	0.0124	0.1184	/	/	0.9023
MOS^{IV}_{av}	0.5723	9.6508	0.2686	0.2244	-0.0171	-0.0940	0.9057

Table 5.4: Coefficients and correlation of "Video call" model.

expected that the AMR codec cannot achieve sufficient results for music content. For fast movement sequences we can clearly observe the mutually compensation effect. The MOS_{av} is significantly more influenced by audio quality for the lowest bit rate (56 kbps). It is caused by a loss of spatial information due to the compression, where higher audio quality can compensate the lower video quality. On the other hand MOS_{av} is not strongly influenced by audio quality for the higher bit rates (75, 105 kbps).

Therefore, it can be assumed that obtained results are significantly influenced by the sequence character. For instance, video calls contain lower amount of spatial and temporal information leading to the loss of the critical ability of the subjective judgment when the media has small video information contents. Therefore, it is necessary to propose one audiovisual model with different coefficients for Video call and Cinema trailer or Video clip. The mutual compensation property and synergy of component media have to be taken into account. The following model was investigated:

$$\begin{aligned} MOS_{av} = \quad & K_{cc} && +A_{cc}\cdot MOS_a && +V_{cc}\cdot MOS_v \\ & +AV_{cc}\cdot MOS_a\cdot MOS_v && +A'_{cc}\cdot MOS_a^2 && +V'_{cc}\cdot MOS_v^2, \end{aligned} \tag{5.6}$$

where K_{cc} is a constant, and A_{cc}, V_{cc}, AV_{cc}, A'_{cc}, V'_{cc} are weights of MOS_a and/or MOS_v (MOS video). Inputs of this model are the above described audio and video metrics. The focus was taken at the best trade-off between complexity and correlation with the subjective audiovisual quality. All these models are designed for speech (video call) and for non-speech (video clip/cinema trailer) scenarios. The coefficients are shown in Tables **5.4** and **5.5**. The highest Pearson correlations (4.22) in both cases are obtained by models MOS^{III}_{av} and MOS^{IV}_{av}. In **Figure 5.15** the correlation in each case is visualized. The correlations in the models MOS^{III}_{av} and MOS^{IV}_{av} are almost equal, but the model MOS^{IV}_{av} with the two terms of second degree for MOS_a and MOS_v is much more complex than model MOS^{III}_{av}. Therefore, a good trade-off between quality and complexity of the audiovisual quality is the model MOS^{III}_{av}. Although, the proposed models are a combination of two mono-modal models the cross-modal interaction is introduced by the product term in MOS^{II}_{av}, MOS^{III}_{av}, MOS^{IV}_{av}.

Model	K_{cc}	A_{cc}	V_{cc}	AV_{cc}	A'_{cc}	V'_{cc}	correlation
MOS^{I}_{av}	-1.5025	0.7380	0.7411	/	/	/	0.8879
MOS^{II}_{av}	0.9135	/	/	0.2329	/	/	0.8415
MOS^{III}_{av}	-0.9222	0.5691	0.5064	0.1697	/	/	0.9106
MOS^{IV}_{av}	-1.1895	0.5947	0.7126	0.0677	-0.0031	-0.0395	0.9117

Table 5.5: Coefficients and correlation of "Cinema trailer"/"Video clip" model.

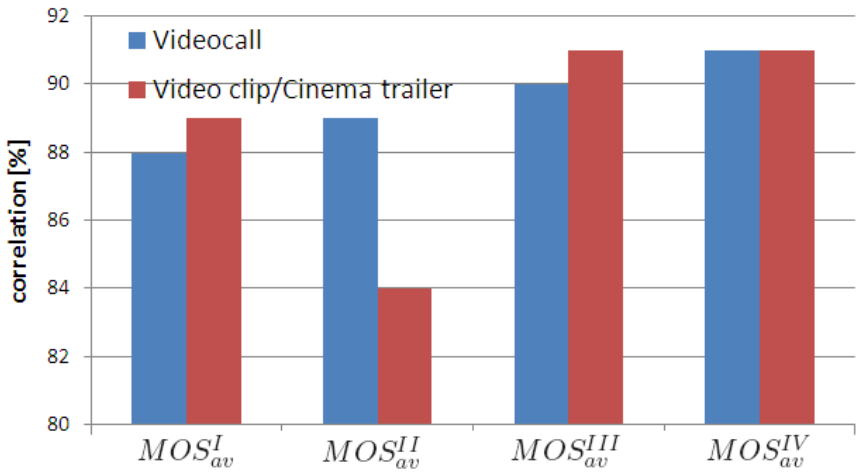

Figure 5.15: Pearson correlations of proposed models.

Self criticism

In this section the ANSI, AD and PESQ metrics were tuned for the mobile environment and extended to the proposed audiovisual model. The proposed audiovisual model has a few drawbacks:

- the model needs a reference video sequence,
- the model does not introduce new audio or video parameters,
- in former Section 4 it was shown that the model relies on the ANSI video quality parameters which do not perform very well in mobile scenarios.

The explanation for is that this was the initial work to quality estimation for mobile environments and reflects the actual state of the art and its applicability for mobile scenarios.

Chapter 6

Conclusions

THIS thesis is dedicated to the estimation of subjective quality for video streaming over wireless mobile networks. The topic covers a multitude of designs of estimation methods for subjective video quality. As an initial step, it was necessary to define the usage scenario and to design the setup for subjective assessments. Furthermore, it was necessary to investigate and develop methods for content segmentation in order to introduce content awareness in the quality estimation. Finally, reference-free estimation methods for the mobile scenario were investigated and proposed.

The proposed estimation models are focused at reference-free quality estimation due to their lower complexity and better applicability. The mobile video streaming scenario reflects an environment of usage, user equipment and typical types of video content. For this purpose mobile scenarios and a test methodology were investigated and defined in order to achieve the best emulation of the real world scenario, covering the most frequent content classes (CC). The scenario is characterized by the size and resolution of the user terminal screen, user mobility and network performance.

These experiences were exploited in a subjective assessment design. For the emulation of the "real word" scenario the ACR assessment method is the most suitable, because a user does not have access to original sequences. Furthermore, the assessments were performed for semantically different content classes. The tested sequences were encoded with encoding settings typical for wireless video streaming. The proposed methodology follows ITU-T recommendations except one point. After an initial video quality assessment it was observed that a systematic deviation appears between subjective evaluations performed at an LCD screen and at a screen of a mobile phone. Therefore, for further assessments, only cell phones or PDAs were used.

The so obtained results show the very important feature of subjective evaluations that the subjective video quality is **content dependent**. The maximal difference between two CCs encoded under equal conditions in a five grade MOS scale can be up to 3 MOS grades for QCIF resolution and 1.6 MOS grades for SIF resolution. Due to content dependent video quality, it is necessary to design features which allow temporal content segmentation and content classification of video streams. A temporal content segmentation was introduced in order to estimate the quality for a single sequence within a video stream. For this purpose an adaptive metric for scene change detection was developed. This metric is able to detect scene changes in all defined content types with both precision and recall higher than 97%.

The next important approach was to classify CCs for content specific metrics. For this purpose the

motion and colour sequence features were investigated, and content sensitive parameters were extracted. These parameters are then input for a content classification based on hypothesis testing. The proposed content classifier is a robust tool for content classification. Moreover, the hypothesis testing allows very fast extension of the number of CCs by adding a new content class.

Our proposals for quality estimation are trade-offs between applicability, processing demands and prediction accuracy. The aim was to estimate quality at the receiver with reasonable processing complexity. Furthermore, the proposed estimation methods demonstrate that it is possible to predict the video quality for the wireless video streaming scenario with reference-free video estimators. The relevance of the required parameters was considered according to the results of a multivariate analysis. This knowledge was successfully applied to the proposed direct reference-free quality and ANN for QCIF resolution. The direct reference-free quality metrics are dedicated to certain content classes. On the other hand, the ANN model is general for all considered content classes and the training performance is sufficient for video quality estimation. Therefore, a better fitting performance of direct reference-free quality metrics was expected.

Moreover, for SIF resolution estimators were proposed based on content adaptive motion parameters which are derived from MV features. Three reference-free estimation methods were proposed. The first method estimates video quality in two steps: the content class is estimated from the original video sequence at the sender side, and then the quality metric is calculated at the receiver with almost zero complexity. The second and the third estimation methods are suitable for stand alone estimation at the receiver side. The second, the ensemble based metric has a performance similar to the content class based metric. The content classification can be understood as a kind of pre-estimation in order to obtain a more homogeneous set of results within one content class, which allows for more accurate quality estimation. This effect has been achieved by introducing cross-validation in ensemble based metrics. Furthermore, the direct motion proposal has a slightly worse estimation performance but allows a completely reference-free estimation for all content classes. The performance of introduced video quality metrics shows a good agreement between estimated MOS and the evaluation set. Moreover, the proposed estimation methods for SIF resolution and proposed content classifier were submitted for patenting [122].

Within this thesis also the audiovisual quality was investigated. The aim was to estimate an audiovisual quality for mobile streaming services. The audiovisual quality assessments show that audio quality, video quality and sequence character are important factors to determine the overall subjective perceived quality. A mutual compensation property of audio and video can also clearly be seen from the obtained results. The proposed audiovisual metrics for speech and non-speech content show over 90% agreement with the test results.

In addition, also the WTP was investigated in the tested scenarios. The obtained results clearly show a linear dependence between the subjective measure MOS and WTP. The proposed metric exhibits more than 95% correlation with tests results. Moreover, this relation did not show dependence on codec or resolution.

Appendix A

List of abbreviations

3G	3rd Generation
3GPP	3rd Generation Partnership Project
AAC	Advanced Audio Coding
AAC LC	AAC Low Complexity
AAC LTP	AAC Long Term Prediction
AC	Alternating Current
ACR	Absolute Category Rating
AD	Auditory Distance
AM	Acknowledged Mode
AMR	Adaptive Multirate Codec
ANN	Artificial Neural Network
ARQ	Automatic Repeat Request
AS	Application Server
AUC	Authentication Center
AVC	Advanced Video Coding
CABAC	Context Adaptive Binary Arithmetic Coding
CAVLC	Context Adaptive Variable Length Coding
CBR	Constant Bit Rate
CC	Content Class
CDF	Cumulative Distribution Function
CIF	Common Intermediate Format
CN	Core Network
CRC	Cyclic Redundancy Check
CS	Circuit Switched
DC	Direct Current
DCT	Discrete Cosine Transform
DCR	Degradation Category Rating
DL	Downlink
ECDF	Empirical Cumulative Distribution Function

FIFO	First-In/First-Out
FIR	Finite Impulse Response
FR	Frame Rate
GGSN	Gateway GPRS Support Node
GSM	Global System for Mobile Communications
HLR	Home Location Register
HSDPA	High Speed Downlink Packet Access
HSUPA	High Speed Uplink Packet Access
HTML	Hyper-Text Mark-up Language
HVS	Human Visual System
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IFD	Integrated Frequency Distance
IP	Internet Protocol
IRPROP+	Improved Resilient Propagation
ISO	International Organization for Standardization
ITU	International Telecommunication Union
JM	Joint Model
JSCC	Joint Source-Channel Coding
JVT	Joint Video Team
kNN	k-Nearest Neighbor
KS	Kolmogorov-Smirnov
LAN	Local Area Network
MAC	Medium Access Control
MB	MacroBlock
MMS	Multimedia Messaging Service
MOS	Mean Opinion Score
MPEG	Motion Picture Expert Group
MSC	Mobile Switching Center
MSE	Mean Square Error
MV	Motion Vector
NAL	Network Abstraction Layer
NALU	Network Abstraction Layer Unit
NRI	NAL Reference Identification
NTSC	National Television Systems Committee
OSI	Open System Interconnection
PAL	Phase Alternation by Line
PC	Pair Comparison
PCA	Principal Component Analysis
PDA	Personal Digital Assistant
PDF	Probability Density Function

PDP	Packet Data Protocol
PDU	Packet Data Unit
PESQ	Perceptual Evaluation of Speech Quality
PHY	PHYsical Layer
PPS	Picture Parameter Set
PQ	Perceived/Perceptual QoS
PS	Packet Switched
PSS	Packet-Switched Streaming
PSNR	Peak to Signal-to-Noise Ratio
QCIF	Quarter Common Intermediate Format
QoS	Quality of Service
QoE	Quality of Experience
QP	Quantization Parameter
QVGA	Quarter Video Graphics Array
RAB	Radio Access Bearer
RACH	Random Access Channel
RAN	Radio Access Network
RB	Radio Bearer
RD	Rate-Distortion
RGB	Red, Green, Blue
RNC	Radio Network Controller
RNS	Radio Network Subsystem
ROI	Region of Interest
RRC	Radio Resource Control
RS	Redundant Slices
RTCP	Real-Time Control Protocol
RTP	Real-Time Protocol
SAD	Sum of Absolute Differences
SAP	Service Access Point
SDP	Session Description Protocol
SDU	Service data Unit
SECAM	Séquentiel couleur à mémoire
SEI	Supplemental Enhancement Information
SGSN	Serving GPRS Support Node
SI	Spatial Information
SIF	Standard Interchange Format
SIM	Subscriber Identity Module
SIR	Signal to Interference Ratio
SM	Synchronization Mark
SMB	Submacroblock
SMIL	Synchronized Multimedia Integration Language

SPS	Sequence Parameter Set
SVC	Scalable Video Coding
TB	Transport Block
TCP	Transmission Control Protocol
TE	Terminal Equipment
TF	Transport Format
TI	Temporal Information
TPC	Transmitter Power Control
TTA	Telecommunication Technology Association
TTC	Telecommunication Technology Committee
UDP	User Datagram Protocol
UE	User Equipment
UEP	Unequal Error Protection
UL	Uplink
UMTS	Universal Mobile Telecommunications Network
URI	Universal Resource Identifier
UTRAN	UMTS Terrestrial Radio Access Network
VBR	Variable Bit Rate
VCL	Video Coding Layer
VGA	Video Graphics Array
VLC	Variable Length Code
VLR	Visitor Location Register
WAP	Wireless Application Protocol
WCDMA	Wideband Code Division Multiple Access
WLAN	Wireless LAN
WTP	Willingness To Pay

Appendix B

Description of ANSI T1.803 video parameters

THE standard ANSI T1.803 specifies a method for estimating the video performance of a one-way video transmission. The video performance estimator is defined for the end-to-end transmission quality. The encoder can utilize various compression methods. This estimation method is based on quality parameters that measure the perceptual effects of a wide range of impairments such as blurring, block distortion, unnatural motion, noise and error blocks. Each quality parameter is calculated through a quality feature, defined as a quantity of information associated with a spatial-temporal sub-region of a video stream.

B.1 Quality features

THE quality feature is defined as a quantity of information associated with, or extracted from, a spatial-temporal sub-region of a video stream (either original or processed). By comparing features extracted from the processed video with features extracted from the original video, a set of quality parameters is computed in order to detect perceptual changes in video quality. Initially, a perceptual filter is applied to the video stream to enhance some property of perceived video quality, such as edge information. After this perceptual filtering, features are extracted from Spatial-Temporal (S-T) sub-regions. Finally, a perceptual threshold is applied to the extracted features. Features calculation is performed in the following steps:

- Perceptual filtering.
- Video stream segmentation in S-T region.
- Feature extraction or summary statistics, from each S-T region (e.g. mean, standard deviation).
- Perceptual thresholding.

B.1.1 S-T regions

Each S-T region describes a block of pixels. S-T region sizes are described by the number of pixel horizontally and vertically and the duration of the region. One fifth of a second is a desirable temporal extend, due to ease of frame rate conversion. **Figure B.1** illustrates a S-T region of 8 horizontal pixels, 8 vertical pixels, 5 video frames, for a total of 320 pixels.

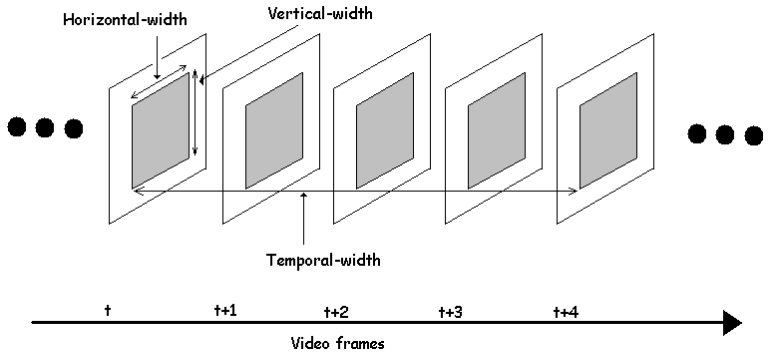

Figure B.1: Example spatial-temporal (S-T) region.

B.2 Features based on spatial gradients

THE features derived from spatial gradients can be used to characterize perceptual distortion of edges. The Y (luminance) components of the original and processed video streams are filtered using horizontal and vertical edge enhancement filters. Next, these filtered video streams are divided into spatial-temporal (S-T) regions from which features, or summary statistics, are extracted that quantify the spatial activity as a function of angular orientation. Then these features are clipped at the lower end to emulate perceptual thresholds. The edge enhancement filters, the S-T region size, and the perceptual thresholds were selected based on [18].

B.2.1 Edge enhancement filters

The original and precessed Y (luminance) video frames are first processed with horizontal and vertical edge enhancement filters. Two filters are applied separately, one to enhance the horizontal pixel difference while smoothing vertically and the other to enhance the vertical pixel difference while

smoothing horizontally.

$$H = \begin{bmatrix} -w_n & ... & -w_2 & -w_1 & 0 & w_1 & w_2 & ... & w_n \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ -w_n & ... & -w_2 & -w_1 & 0 & w_1 & w_2 & ... & w_n \\ -w_n & ... & -w_2 & -w_1 & 0 & w_1 & w_2 & ... & w_n \\ -w_n & ... & -w_2 & -w_1 & 0 & w_1 & w_2 & ... & w_n \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ -w_n & ... & -w_2 & -w_1 & 0 & w_1 & w_2 & ... & w_n \end{bmatrix}. \tag{B.1}$$

The two filters H for vertical edge enhancement and $V = H^T$ for horizontal edge enhancement with size 13×13 have the following filter weights:

$$w_x = k \cdot \left(\frac{x}{c}\right) \cdot e^{\left\{-\left(\frac{1}{2}\right)\left(\frac{x}{c}\right)^2\right\}},\ x = 0, 1, 2, ..., 6,$$

where x is the pixel displacement from the center of the filter, c is a constant that sets the width of the bandpass filter, and k is a normalization constant selected such that each filter would produce the same gain as a true Sobel filter [123]. For the optimal choice of the parameters see [18].

B.2.2 Description of features f_{SI13} and f_{VH13}

This section describes the extraction of two spatial activity features from the S-T region of the original and processed video streams. These features will be used to detect spatial impairments as blurring and blocking. The filter H enhances the spatial gradient in the horizontal direction while the filter $V = H^T$ enhances the spatial gradient in the vertical direction. The response at each pixel from the horizontal and vertical filters can be plotted on a two dimensional diagram with the horizontal filter response forming the abscissa value and the vertical filter response forming the ordinate value. For a given image pixel located at row i, column j, and time t, the horizontal and vertical filters will be denoted as $H(i, j, t)$ and $V(i, j, t)$, respectively. These responses can be converted into polar coordinates (R, θ) using the relationship:

$$R(i, j, t) = \sqrt{H(i, j, t)^2 + V(i, j, t)^2}$$

and

$$\theta(i, j, t) = \tan^{-1}\left[\frac{V(i, j, t)}{H(i, j, t)}\right].$$

The first feature is a measure of the overall spatial information (SI), denoted as f_{SI13}, since images were preprocessed using the 13×13 filter masks. This features is computed simply as the standard deviation over the S-T region of the $R(i, j, t)$ samples, and then clipped at the perceptual threshold [18].

$$f_{SI13} = \left\{ std_{i,j,t}[R(i, j, t)] \right\}\Big|_P : i, j, t \in \{S - T\ region\} \tag{B.2}$$

This feature is sensitive to changes in the overall amount of spatial activity within a given S-T region. For instance, localized blurring produces a reduction in the amount of spatial activity, whereas noise produces an increase.

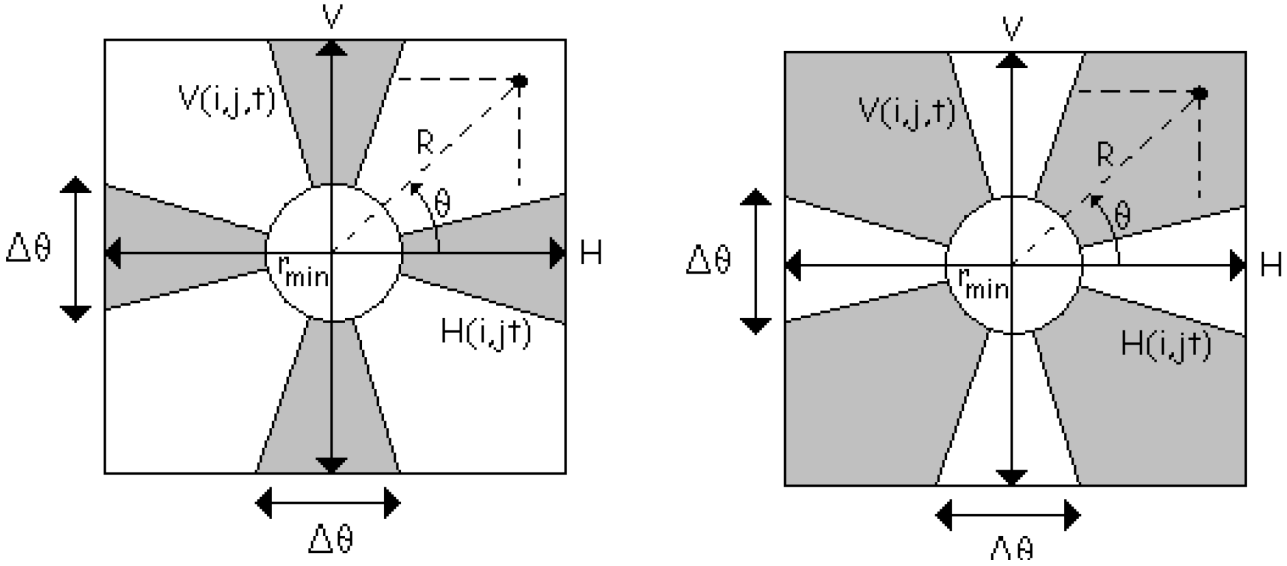

Figure B.2: Division of horizontal and vertical spatial activity into HV (left) and $\overline{HV}$ (right) distribution.

The second feature, f_{HV13}, is sensitive to changes in angular distribution, or orientation, of spatial activity. Complementary images are computed with the shaded spatial gradient distributions shown in **Figure B.2**. The image with horizontal gradient, denotes as HV, contain the $R(i,j,t)$ pixels that are horizontal or vertical edges (pixel that are diagonal edges are zeroed). The image with the diagonal gradients, denoted as $\overline{HV}$, contains the $R(i,j,t)$ pixels that is diagonal edges (pixel that are horizontal or vertical edges are zeroed). A gradient magnitude $R(i,j,t)$ less then r_{min} [18] is zeroed in both images to assure accurate θ computation. Pixel in HV and $\overline{HV}$ can be represented mathematically as:

$$HV(i,j,t) = \begin{cases} R(i,j,t) & \text{if } R(i,j,t) \geq r_{min} \text{ and } m\frac{\pi}{2} - \Delta\theta < \theta(i,j,t) < m\frac{\pi}{2} + \Delta\theta \\ & (m = 0,1,2,3) \\ 0 & \textit{otherwise} \end{cases} \tag{B.3}$$

and

$$\overline{HV}(i,j,t) = \begin{cases} R(i,j,t) & \text{if } R(i,j,t) \geq r_{min} \text{ and } m\frac{\pi}{2} + \Delta\theta < \theta(i,j,t) < (m+1)\frac{\pi}{2} - \Delta\theta \\ & (m = 0,1,2,3) \\ 0 & \textit{otherwise} \end{cases} \tag{B.4}$$

where: $i,j,t \in \{\text{S-T region}\}$.

The feature f_{HV13} for one S-T region is then given by the ratio of the mean of HV to the mean of $\overline{HV}$, where these resultant means are clipped at their perceptual threshold P, namely:

$$f_{HV13} = \frac{\left\{ mean[HV(i,j,t)] \right\}\Big|_P}{\left\{ mean[\overline{HV}(i,j,t)] \right\}\Big|_P}.$$

The f_{HV13} feature is sensitive to changes in the angular distribution of spatial activity within a given S-T region. For example, if horizontal and vertical edges suffer of more blurring than diagonal edges, f_{HV13} of the processed video will be less than f_{HV13} of the original video. On the other hand, if erroneous horizontal or vertical edges are introduced, say in the form of blocking or tilling distortion, then f_{HV13} of the processed video will be greater than f_{HV13} of the original video. The f_{HV13} feature thus provides a simple means to include variations in the sensitivity of the human visual system with respect to angular orientation.

B.3 Video parameters

THE video parameters that measure distortion in video quality due to gains and losses in the feature values are first calculated for each S-T region by comparing the original feature value, $f_o(s,t)$, with the corresponding processed feature values, $f_p(s,t)$. Several functional relationships are used to emulate the visual masking of impairments for each S-T region. Next, an error-pooling function [18] across space and time emulates how humans deduce subjective quality ratings. Error pooling across space will be referred to as spatial collapsing, and error pooling across time will be referred as temporal collapsing. Sequential application of the spatial and temporal collapsing function [18] to the stream of S-T quality parameters produces quality parameters for the entire video clip, witch is nominally 5 to 10 seconds in duration. The final time-collapsed parameter values may be scaled and clipped to account for a non-linear relationship between the parameter value and the perceived quality and to further reduce the parameter sensitivity.
The video parameters are calculated in the following steps:

- Compare original feature values with processed feature values.
- Perform spatial collapsing.
- Perform temporal collapsing.
- Perform nonlinear scaling and/or clipping (optional).

B.3.1 Comparison functions

The perceptual impairment at each S-T region is calculated using functions that model visual masking of the spatial and temporal impairments. This section presents the masking functions that are used by the various parameters to produce quality parameters as a function of space and time.

B.3.2 Error ratio and logarithmic ratio

Loss and gain are normally examined separately, since they produce fundamentally different effects on quality perception (e.g., loss of spatial activity due to blurring and gain of spatial activity due to noise or blocking). Of the many comparison function that have been evaluated, two forms have consistently produced the best correlation to subjective ratings. The distortion measures are calculated for each S-T region by comparing the original feature values, $f_o(s,t)$, with the corresponding processed feature values $f_p(s,t)$:

$$ratio\_loss(s,t) = np\left\{\frac{f_p(s,t) - f_o(s,t)}{f_o(s,t)}\right\},$$

$$ratio\_gain(s,t) = pp\left\{\frac{f_p(s,t) - f_o(s,t)}{f_o(s,t)}\right\},$$

$$log\_loss(s,t) = np\left\{\log_{10}\left[\frac{f_p(s,t)}{f_o(s,t)}\right]\right\},$$

$$log\_gain(s,t) = pp\left\{\log_{10}\left[\frac{f_p(s,t)}{f_o(s,t)}\right]\right\},$$

where pp is the positive part operator (i.e, negative values are replaced with zero), and np is the negative part operator (i.e, positive values are replaced with zero).
The visual masking function implies that impairment perception is inversely proportional to the amount of localized spatial or temporal activity that is present. In other words, spatial impairments become less visible as the spatial activity increases (i.e. spatial masking), and temporal impairments become less visible as the temporal activity increases (i.e. temporal masking). While the logarithmic and ratio comparison function behave very similarly, the logarithmic function tends to be slightly more advantageous for gains, while the ratio function tends to be slightly more advantageous for losses. The logarithm function has a larger range, and this is useful when the processed feature values greatly exceed the original feature values.

B.3.3 Spatial collapsing functions

The parameters from the S-T region from three-dimensional matrices spanning one temporal axis and two spatial dimensions (i.e. horizontal and vertical placement of the S-T region). Next, impairments from the S-T regions with the same time index t are pooled using a spatial collapsing function. Spatial collapsing yields a time history of parameter values. This time history of parameter values, denoted generically as $p(t)$, must then be temporally collapsed using a temporal collapsing function.

B.3.4 Temporal collapsing function

The parameter time history results $p(t)$ output from the spatial collapsing function is next pooled using a temporal collapsing function to produce an objective parameter O for the video clip, which is nominally of 4 to 10 seconds length. Viewers seem to use several temporal collapsing function when subjectively rating video clips that are approximately 10 seconds length. The variable entitled *mean* over the time is indicative of the average quality that is observed during the time period. The 90%

and 10% percentiles over the time are indicative of the worst transcient quality that is observed for gains and losses, respectively.

B.3.5 SI and HV video parameters

Initially, the *ratioloss* function on the f_{HV13} feature and on the f_{SI113} are computed, where the threshold P in (B.2) is 12. These two parameters (matrices) are denoted as *hvloss* and *siloss*, respectively. Then the *loggain* function on the (f_{HV13} feature and on the f_{SI13}) where the threshold P in (B.2.2) is 8, is calculated. These two parameters (matrices) are denoted as *hvgain** and *sigain**, respectively. The *hvloss*$*$, *siloss*$*$, *hvgain*$*$ and *sigain*$*$ (see Appendix A of [120]) matrices give both spatial information (given by the rows) and temporal information (given by the columns). Now, impairments from the S-T region with the same time index t are pooled using a spatial collapsing function. Spatial collapsing yields a time history of parameters values.

The spatial function [18] is set below 5% and is applied on *hvloss*$*$ and *siloss*$*$. For each temporal index t (i.e. for each column of the matrices), this function sorts the parameter values from low to high. Then it computes the average of all the parameter values that are less than or equal to the 5% threshold level. For loss parameters, it produces a parameter that is indicative of the worst quality over space.

The spatial function is set above 95% and is instead applied on *hvgain*$*$. For each temporal index t (i.e. for each column of the matrix), this function sorts the parameter values from low to high. Then it computes the average of all the parameter values that are greater than or equal to the 95% threshold level. For gain parameters, it produces a parameter that is indicative of the worst quality over space. Finally, the mean is computed over all the parameter values. It produces a parameter that is indicative of the average quality over space and parameters have become row vectors (parameters denoted with $**$). Now all the elements of each vector are pooled using a temporal collapsing function to calculate single parameter for the video sequence.

For the *siloss*$**$ row vector, the temporal collapsing function produces a parameter that is indicative of the worst quality over time. The temporal function is set to 10%. It sorts the time history of the parameter values from low to high and selects the 10% threshold level. This is the final *siloss* parameter.

The temporal function *mean* is instead applied on the *sigain*$**$, *hvloss*$**$ and *hvgain*$**$ row vectors. The three mean parameters are indicative of the average quality over time. The *hvgain* parameter is so defined completely.

Finally, a clipping function is calculated for products of a temporal collapsing function on *sigain*$**$ and *hvloss*$**$ parameters in order to reduce the sensitivity of the parameters to small impairments. A clipping function is mathematically represented as:

$$clip_T(p) = \begin{cases} \max(p, T) - T; & p \geq 0 \\ \min(p, T) - T; & p < 0 \end{cases} \tag{B.5}$$

where p is the clipped parameter and T the threshold. The threshold T is 0.004 for the *sigain*$**$ parameter and 0.06 for the squared *hvloss*$**$. After the application of this clipping function, also the *sigain* and *hvloss* parameters are completely defined.

Appendix C

Description of audio quality parameters

C.1 Model for perceptual evaluation of speech quality

THE perceptual model for the Perceptual Evaluation of Speech Quality (PESQ) [113] defined by ITU is designed to calculate a distance between the original and degraded speech signal (PESQ score). The PESQ score is usually mapped to a MOS scale, a single number in the range of -0.5 to 4.5, although for most cases the output range will be between 1.0 and 4.5, the typical range of MOS values found in an ACR listening quality experiment.
In PESQ the time signals are mapped to the time-frequency domain using a short-term Fast Fourier Transform (FFT) with a Hann window [125] of size 32 ms. Adjacent frames are overlapped by 50%. The absolute hearing threshold $P_0(f)$ is interpolated to get the values at the center of the Bark bands [126] that are used. These values are stored in an array and are used in Zwicker's loudness formula [126]. The Hertz scale is converted to the Bark frequency scale:

$$b = 6 \cdot \sinh^{-1}\left(\frac{f}{600}\right). \tag{C.1}$$

There is a constant gain following the FFT for time-frequency analysis. This constant is computed from a sine wave of a frequency of 1000 Hz with an amplitude at 29.54 (40 dB SPL - Sound Pressure Level) transformed to the frequency domain using the windowed FFT over 32 ms. The (discrete) frequency axis is then converted to a modified Bark scale by binning of the FFT bands. The peak amplitude of the spectrum binned to the Bark frequency scale (called the "pitch power density") must then be 10 000 (40 dB SPL). The latter is enforced by a post multiplication with a constant, the power scaling factor S_p.
The same 40 dB SPL reference tone is used to calibrate the psychoacoustic (Sone) loudness scale. After binning to the modified Bark scale, the intensity axis is warped to a loudness scale using Zwicker's law, based on the absolute hearing threshold. The integral of the loudness density over the Bark frequency scale, using a calibration tone at 1000 Hz and 40 dB SPL, must then yield a value of 1 Sone. The latter is enforced by a post-multiplication with a constant, the loudness scaling factor S_l.

C.1.1 Description of PESQ algorithm

Because many of the steps in PESQ are quite algorithmically complex, a description is not easily expressed in mathematical formulas. This description is textual in nature and the reader is referred to the C source code [127] for a detailed description. **Figure C.1** shows the core of the perceptual model. For each of the blocks a high level description is given.

The part depicted on **Figure C.1** of PESQ algorithm is explained in following sections.

C.1.2 IRS filtering

It is assumed that the listening tests were carried out using an IRS receive or a modified IRS receive characteristic in the handset. A perceptual model of the human evaluation of speech quality must take account of this to model the signals that the subjects actually heard. Therefore IRS-like receive filtered versions of the original speech signal and degraded speech signal are computed.
In PESQ this is implemented by a FFT over the length of the file, filtering in the frequency domain with a piecewise linear response similar to the (unmodified) IRS receive characteristic (ITU T P.830), followed by an inverse FFT over the length of the speech file. This results in the filtered versions $X_{IRSS}(t)$ and $Y_{IRSS}(t)$ [127] of the scaled input and output signals $X_S(t)$ and $Y_S(t)$ [127]. A single IRS-like receive filter is used within PESQ irrespective of whether the real subjective experiment used IRS or modified IRS filtering. The reason for this approach was that in most cases the exact filtering is unknown, and that even when it is known the coupling of the handset to the ear is not known. It is therefore a requirement that the objective method be relatively insensitive to the filtering of the handset.
The IRS filtered signals are used both in the time alignment procedure and the perceptual model.

C.1.3 Computation of the active speech time interval

If the original and degraded speech file starts or ends with large silent intervals, this could influence the computation of certain average distortion values over the files. Therefore, an estimate is made of the silent parts at the beginning and end of these files. The sum of five successive absolute sample values must exceed 500 from the beginning and end of the original speech file in order for that position to be considered as the start or end of the active interval. The interval between this start and end is defined as the active speech time interval. In order to save computation cycles and/or storage size, some computations can be restricted to the active interval.

C.1.4 Short-term Fast Fourier Transform

The human ear performs a time-frequency transformation. In PESQ this is simulated by a short-term Fast Fourier Transform (FFT) with a window size of 32 ms. The overlap between successive time windows (frames) is 50%. The power spectrum - the sum of the squared real and squared imaginary parts of the complex FFT components - are stored in separate real valued arrays for the original and degraded speach signals. Phase information within a single Hann window is discarded in PESQ and all calculations are based on only the power representations $PX_{WIRSS}(f)_n$ and $PY_{WIRSS}(f)_n$ [127]. The start points of the windows in the degraded signal are shifted over the delay. The time axis of

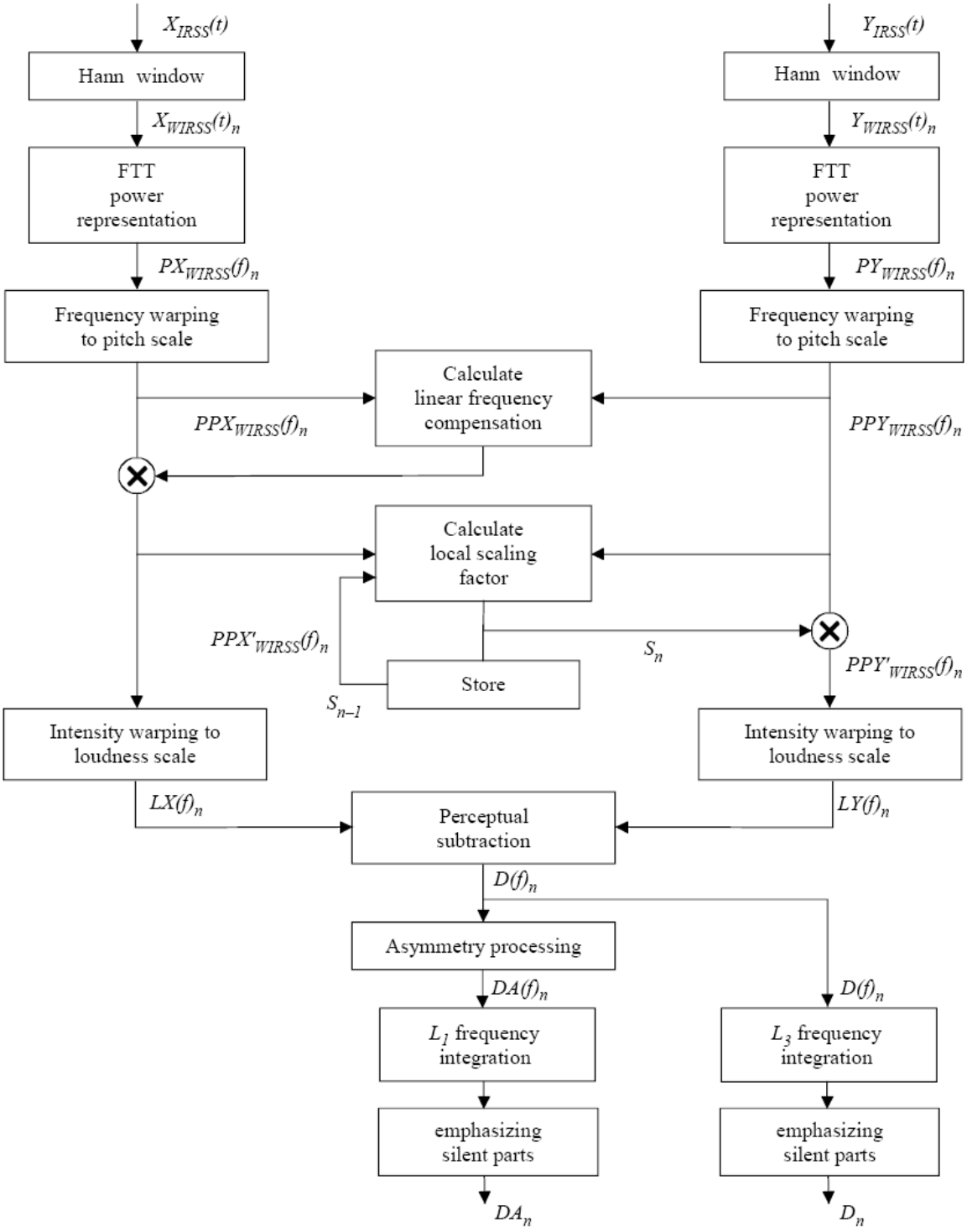

Figure C.1: Overview of perceptual model.

the original speech signal is left as is. If the delay increases, parts of the degraded signal are omitted from the processing, while for decreases in the delay parts are repeated.

C.1.5 Pitch power densities

The Bark scale reflects that at low frequencies, the human hearing system has a finer frequency resolution than at high frequencies. This is implemented by binning FFT bands and summing the corresponding powers of the FFT bands with a normalization of the summed parts. The warping function that maps the frequency scale in Hertz to the pitch scale in Bark scale is introduced. The resulting signals are known as the pitch power densities $PPX_{WIRSS}(f)_n$ and $PPY_{WIRSS}(f)_n$ [127].

C.1.6 Partial compensation of the original pitch power density for transfer function equalization

To deal with filtering in the system under test, the power spectrum of the original and degraded pitch power densities are averaged over the time. This average is calculated over speech active frames only using time-frequency cells whose power is more than 1000 times above the absolute hearing threshold. Per modified Bark bin, a partial compensation factor is calculated from the ratio of the degraded spectrum to the original spectrum. The maximum compensation is never more than 20 dB. The pitch power density $PPX_{WIRSS}(f)_n$ of the original speech signal of each frame n is then multiplied with this partial compensation factor to equalize the original to the degraded signal. This results in an inversely filtered original pitch power density $PPX'_{WIRSS}(f)_n$ [127].
This partial compensation is used because severe filtering can be disturbing to the listener. The compensation is carried out on the original signal because the degraded signal is the one that is judged by the subjects.

C.1.7 Partial compensation of the distorted pitch power density for time varying gain variations between distorted and original signal

Short-term gain variations are partially compensated by processing the pitch power densities frame by frame. For the original and the degraded pitch power densities, the sum in each frame n of all values that exceed the absolute hearing threshold is computed. The ratio of the power in the original and the degraded files is calculated. A first order low-pass filter (along the time axis) is applied to this ratio. The distorted pitch power density in each frame n is resulting in the partially gain compensated pitch power density $PPY'_{WIRSS}(f)_n$ [127] for degraded signals.

C.1.8 Calculation of the loudness densities

After partial compensation for filtering and short-term gain variations, the original and degraded pitch power densities are transformed to a Sone loudness scale using Zwicker's law [121].

$$LX(f)_n = S_l \cdot \left(\frac{P_o(f)}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5 \cdot \frac{PPX'_{WIRSS}(f)_n}{P_o(f)}\right)^{\gamma} - 1\right] \tag{C.2}$$

with $P_0(f)$ the absolute threshold and S_l the loudness scaling factor. Above 4 Bark, the Zwicker power is 0.23, the value given in the literature [121]. Below 4 Bark, the Zwicker power is increased slightly to account for the so-called recruitment effect. The resulting two-dimensional arrays $LX(f)_n$ and $LY(f)_n$ [127] are called loudness densities.

C.1.9 Calculation of the disturbance density

The signed difference between the distorted and original speech signal loudness density is computed. When this difference is positive, components such as noise have been added. When this difference is negative, components have been omitted from the original signal. This difference array is called the raw disturbance density.
The minimum of the original and degraded loudness density is computed for each time frequency cell. These minimums are multiplied by 0.25. The corresponding two-dimensional array is called the mask array. The following rules are applied in each time-frequency cell:

- If the raw disturbance density is positive and larger than the mask value, the mask value is subtracted from the raw disturbance.
- If the raw disturbance density lies in interval $[-|mask\ value|, |mask\ value|]$, the disturbance density is set to zero.
- If the raw disturbance density is $< -|mask\ value|$, the mask value is added to the raw disturbance density.

The impact is that the raw disturbance densities are pulled toward zero. This represents a dead zone before an actual time frequency cell is perceived as distorted. This models the process of small differences being inaudible in the presence of loud signals (masking) in each time-frequency cell. The result is a disturbance density as a function of time and frequency, $D(f)_n$ [127].

C.1.10 Cell-wise multiplication with an asymmetry factor

The asymmetry effect is caused by the fact that when a codec distorts the input signal it will in general be very difficult to introduce a new time-frequency component that integrates with the input signal, and the resulting output signal will thus be decomposed into two different percepts, the input signal and the distortion, leading to clearly audible distortion [124]. When the codec leaves out a time frequency component, the resulting output signal cannot be decomposed in the same way and the distortion is less accurate. This effect is modeled by calculating an asymmetrical disturbance density $DA(f)_n$ [127] per frame by multiplication of the disturbance density $D(f)_n$ [127] with an asymmetry factor. This asymmetry factor equals the ratio of the distorted and original pitch power densities raised to the power of 1.2. If the asymmetry factor is less than 3, it is set to zero. If it exceeds 12, it is clipped at that value. Thus, only those time frequency cells remain, as non zero values, for which the degraded pitch power density exceeded the original pitch power density.

C.1.11 Aggregation of the disturbance densities over frequency and emphasis on soft parts of the original

The disturbance density $D(f)_n$ and asymmetrical disturbance density $DA(f)_n$ are integrated (summed) along the frequency axis using two different L_p norms and a weighting on soft frames (having low loudness):

$$D_n = M_n \sqrt[3]{\sum_{f=1,...Number\ of\ Bark\ bands} (|D(f)_n| \cdot W_f)^3} \tag{C.3}$$

$$DA_n = M_n \sum_{f=1,...Number\ of\ Bark\ bands} (|DA(f)_n| \cdot W_f) \tag{C.4}$$

with multiplication factor M_n defined in [127], resulting in an emphasis of the disturbances that occur during silences in the original speech fragment, and W_f a series of constants proportional to the width of the modified Bark bins. These aggregated values, D_n and DA_n, are called frame disturbances.

C.1.12 Integrated frequency distance parameter

Integrated frequency distance (IFD) is a non reference parameter with a good hyperbolic correlation with the subjective audio quality. The values of the $PPX'_{WIRSS}(f)_n$ matrix and of the $PPY'_{WIRSS}(f)_n$ matrix are aggregated over time with a mean function (see Appendix A of [120] and [127]). Every bark frequency is weighted with the PESQ perceptual model values [113]. Therefore the distance between the original and compressed values in frequency domain, for one bark frequency value, is also adequately weighted. We can see in **Figure C.2** that the distance between original and degraded signals in Bark scale may be a good and simple indicator of the audio quality degradation. IFD is the integrated difference between the non original audio signal and the degraded one.

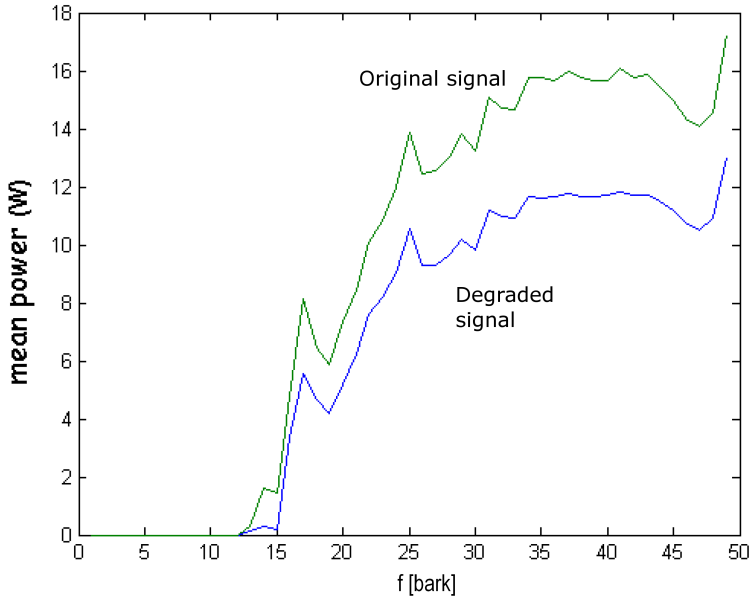

Figure C.2: Distance between original and degraded signals in Bark scale.

$$IFD = \sum_{f=1,...Number\ of\ Bark\ bands} (mean\ power(f)_{original_n} - mean\ power(f)_{degraded_n}). \tag{C.5}$$

Where $mean\ power(f)_{original_n}$ (see Appendix A of [120]) is for every bark frequency the mean over time of the original audio signal (the mean of the $PPY'(f)_n$ rows); and the $mean\ power(f)_{degraded_n}$

(see Appendix A of [120]) is for every bark frequency the mean over time of the compressed audio signal (the mean of the $PPX'(f)_n$ rows).

C.2 Auditory distance

THE Auditory Distance (AD) [115] is calculated in following steps the *perceptual transformation* and the *distance measure* (see **Figure C.3**).

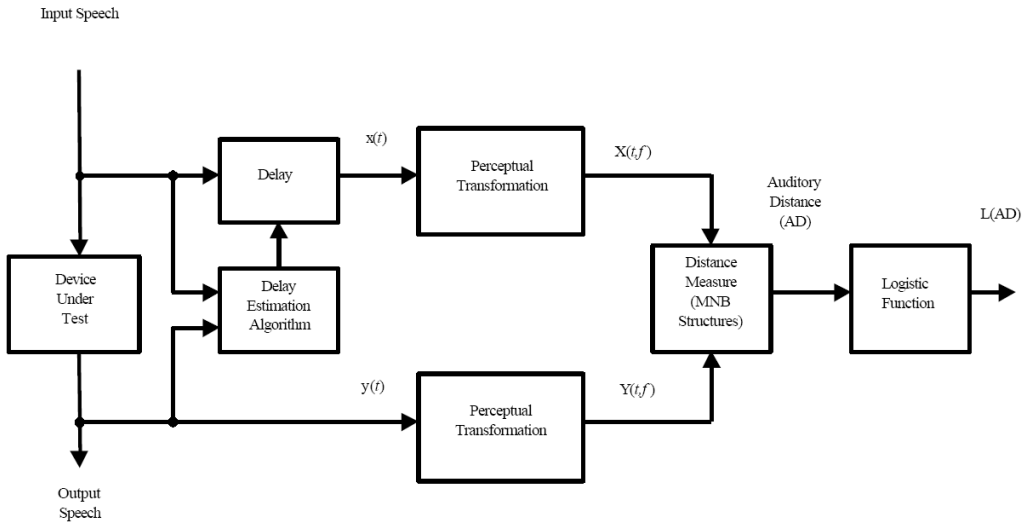

Figure C.3: High-level block diagram of the objective estimation approach of AD.

C.2.1 Perceptual Transformations

This perceptual transformation is applied to frequency domain representations of the speech signals. Speech signals are segmented into frames, multiplied by a Hamming window, and then transformed to the frequency domain using an FFT. The nonuniform frequency resolution is treated by the use of a *psychoacoustic frequency scale*: a Bark frequency scale. The Hertz scale is converted to the Bark frequency scale (C.1).

C.2.2 Distance Measures

The measure of the *distance* between the two perceptual transformed signals happens through a hierarchical structure of Measuring Normalizing Blocks (MNB).

A time measuring normalizing block (TMNB) is shown in **Figure C.4** and a frequency measuring normalizing block (FMNB) is given in **Figure C.5**. Each of these blocks take the perceptually transformed input and output signals $X(f,t)$ and $Y(f,t)$, respectively as inputs, and returns a set of measurements.

The TMNB integrates the original and the degraded signals over some frequency scale, then measures differences between the so integrated degraded and original signals, and normalizes the output signal at multiple times.

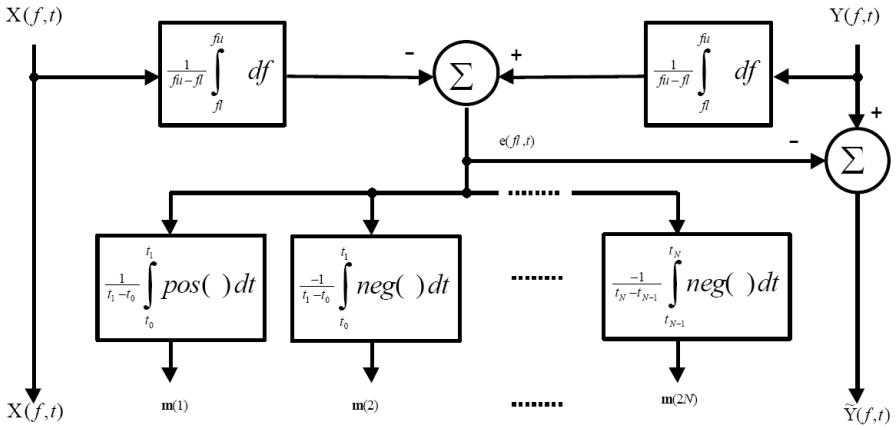

Figure C.4: Time measuring normalizing block (TMNB).

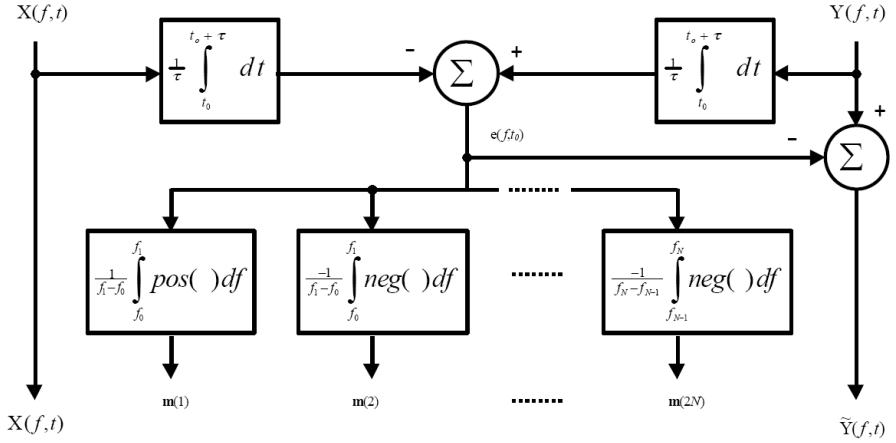

Figure C.5: Frequency measuring normalizing block (FMNB).

An FMNB integrates over some time scale the original and the degraded signals, then measures differences between the so integrated degraded and original signals, and normalizes the output signal at multiple frequencies. There are two MNB structures that offer relatively low complexity and high performance as estimators of the perceived speech quality. These structure are shown in **Figures C.6** and **C.7**.

Both MNB structures start with an FMNB that is applied to the input and output signals at the longest available time scale. Four measurements are extracted and stored in the measurement vector m. These measurements cover the lower and upper band edges of telephone band speech (from 0 to 500 and from 3000 to 3500 Hz). In MNB structure 1, a TMNB is then applied to the input and output signals at the largest frequency scale (approximately 15 Bark). Finally, a residual measurement is made.

We can loosely describe the action of these MNB structures as a dynamic decomposition of a codec output signal. This decomposition proceeds in a space that is defined partly by human hearing and judgment (through the MNB structure) and partly by the codec input signal. The parameters of this dynamic decomposition are combined linearly to form a measure of the perceptual distance between two speech signals. Each measure **m** taken at different frequencies, is weighed with different coeffi-

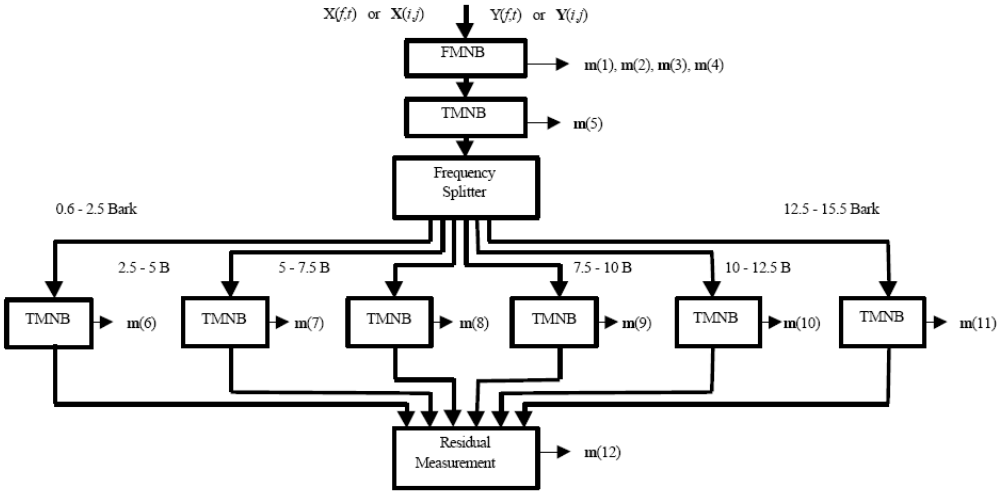

Figure C.6: MNB structure 1.

cients, because the human ears do not perceive all the frequencies in the same way. The MNB structure 1 consists of $\mathbf{m} = (m(1), m(2), ..., m(12))$ and MNB structure 2 consists of $\mathbf{m} = (m(1), m(2), ..., m(11))$, for example see **Figures C.6** and **C.7** . The value that results from this linear combination is called ***auditory distance***:

$$AD' = \mathbf{w}^T \cdot \mathbf{m} \tag{C.6}$$

where $\mathbf{w}$ is a length 12 (MNB structure 1) or 11 (MNB structure 2) vector of weights. In practice, AD values are nonnegative. When the input and output signals are identical, all measurements are zero and AD is zero. As the input and output signals move apart perceptually, AD increases. Finally, the AD' was mapped to a finite range (0,1):

$$AD = \frac{1}{1 - e^{AD' - \max(AD')}}. \tag{C.7}$$

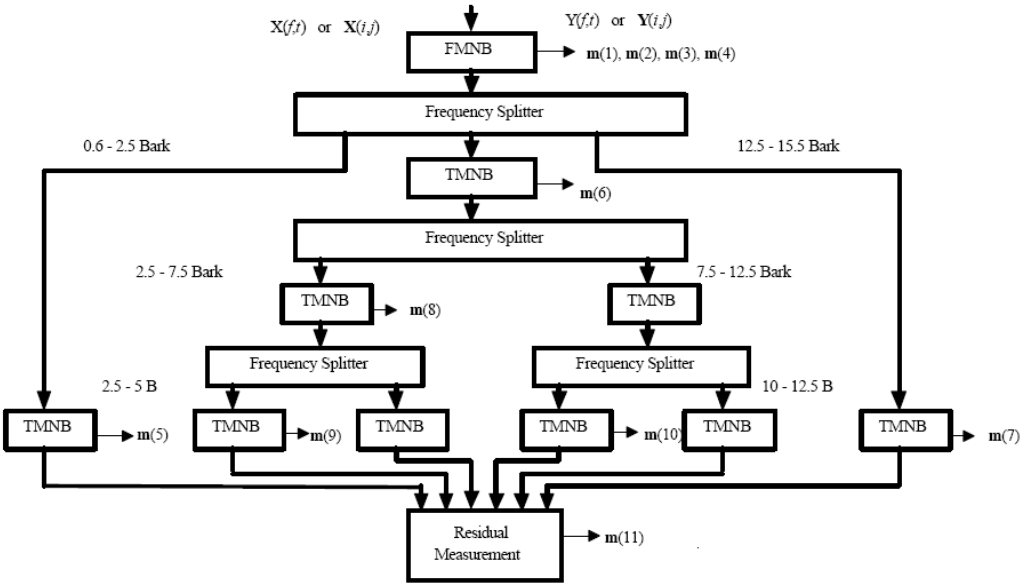

Figure C.7: MNB structure 2.

Bibliography

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

[2] 3GPP TS 26.234 V6.11.0:"End-to-end transparent streaming service; Protocols and codecs," Jun. 2007.

[3] ITU-T Recommendation H.264 (03/05): "Advanced video coding for generic audiovisual services" — ISO/IEC 14496-10:2005: "Information technology - Coding of audio-visual objects - Part 10: Advanced Video Coding," International Telecommunication Union, 2005.

[4] M. Ries, O. Nemethova, M. Rupp, "Reference-Free Video Quality Metric for Mobile Streaming Applications," in Proc. of the DSPCS 05 & WITSP 05, pp. 98-103, Sunshine Coast, Australia, Dec. 2005.

[5] O. Nemethova, M. Ries, E. Siffel, M. Rupp, "Quality Assessment for H.264 Coded Low-Rate and low-Resolution Video Sequences," in Proc. of Conf. on Internet and Inf. Technologies (CIIT), St. Thomas, US Virgin Islands, pp. 136-140, Nov. 2004.

[6] H. Koumaras, A. Kourtis, D. Martakos, "Evaluation of Video Quality Based on Objectively Estimated Metric," in Journal of Communications and Networking, Korean Institute of Communications Sciences (KICS), vol. 7, no.3, Sep. 2005.

[7] C. John, "Effect of content on perceived video quality," Univ. of Colorado, Interdisciplinary Telecommunications Program: TLEN 5380 Video Technology, Aug. 2006.

[8] Yevgeni Koucheryavy, Giovanni Giambene, Dirk Staehle, Francisco Barcelo-Arroyo, Torsten Braun, Vasilios Siris," Traffic and QoS Management in Wireless Multimedia Networks," Springer US, 2008.

[9] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Sep. 1999.

[10] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia application," International Telecommunication Union, 1998.

[11] ITU-R Recommendation BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," International Telecommunication Union, Geneva, Switzerland, 2003.

[12] ITU-T Recommendation P.930, "Principles of a reference impairment system for video," International Telecommunication Union, 1996.

[13] I.E.G. Richardson, C.S. Kannangara, "Fast subjective video quality measurements with user feedback," IEE Electronic Letters, Vol. 40, Number 13, pp. 799-800. Jun. 2004.

[14] A. W. Rix, A. Bourret, and M. P. Hollier, "Models of Human Perception," in Journal of BT Tech., vol. 17, no. 1, pp. 24-34, Jan. 1999.

[15] S. Winkler, F. Dufaux, "Video Quality Evaluation for Mobile Applications," in Proc. of SPIE Conference on Visual Communications and Image Processing, Lugano, Switzerland, vol. 5150, pp. 593-603, Jul. 2003.

[16] S. Winkler, Digital Video Quality, JohnWiley & Sons, Chichester, 2005.

[17] E.P. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, F. Moschetti, "Low bit rate quality assessment based on perceptual characteristics," in Proc. of Int. Conf. on Image Processing , vol. 3, pp. 182-192, Sep. 2003.

[18] ANSI T1.801.03, "American National Standard for Telecommunications - Digital transport of one-way video signals. Parameters for objective performance assessment," American National Standars Institute, 2003.

[19] M.H. Pinson, S. Wolf, "A new standardized method for objectively measuring video quality," in Journal of IEEE Transactions on broadcasting, Vol. 50, Issue: 3, pp. 312-322, Sep. 2004.

[20] T. M. Kusuma, H. J. Zepernick, M. Caldera, "On the Development of a Reduced-Reference Perceptual Image Quality Metric," in Proc. of the 2005 Systems Communications (ICW05), pp. 178-184, Montreal, Canada, Aug. 2005.

[21] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A No-Reference Perceptual Blur Metric," in Proc. of IEEE Int. Conf. on Image Processing, pp. 57-60, Sep. 2002.

[22] IETF STD 0007: "Transmission Control Protocol," Postel J., Sep. 1981.

[23] IETF STD 0006: "User Datagram Protocol," Postel J., Aug. 1980.

[24] IETF RFC 1889: "RTP: A Transport Protocol for Real-Time Applications," Schulzrinne H. et al., Jan. 1996.

[25] IETF RFC 3267: " RTP payload format and file storage format for the Adaptive Multi-Rate (AMR) Adaptive Multi-Rate Wideband (AMR-WB) audio codecs," Mar. 2002.

[26] IETF RFC 3016: "RTP Payload Format for MPEG-4 Audio/Visual Streams," Kikuchi Y. et al., Nov. 2000.

[27] IETF RFC 2429: "RTP Payload Format for the 1998 Version of ITU-T Rec. H.263 Video (H.263+)," Bormann C. et al., Oct. 1998.

[28] 3GPP TS 26.071 V.4.0.0: "Mandatory Speech Codec speech processing functions; AMR Speech Codec; General description," 2001.

[29] ITU-T Recommendation G.722.2, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)," International Telecommunication Union, 2002.

[30] ISO/IEC 14496-3:2001, "Information technology – Coding of audio-visual objects – Part 3: Audio".

[31] ITU-T Recommendation H.263: "Video coding for low bit rate communication," International Telecommunication Union, May 1996.

[32] ISO/IEC 14496-2:2001, "Information technology – Coding of audio-visual objects – Part 2: Visual".

[33] ISO/IEC 14496-2:2001/Amd 2:2002, "Streaming video profile".

[34] 3GPP TS 26.244 V6.7.0: "Transparent end-to-end packet switched streaming service (PSS); 3GPP file format (3GP)," Jun. 2007.

[35] S. Wenger, "H.264/AVC over IP," in Journal of IEEE Trans. on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 645–657, Jul. 2003.

[36] S. Branguolo, N. Tizon, B.P. Popescu, B. Lehembre, "Video Transmission over UMTS Networks Using UDP/IP," in Proc. of 14th European Signal Processing Conference (EUSIPCO), Florence, Italy, Sep. 2006.

[37] S. Wenger et al., "RTP Payload Format for H.264 Video," IETF RFC 3984, Feb. 2005.

[38] 3GPP TS 26.233 V.6.0.0: "End-to-end transparent streaming service; General description," 2004.

[39] ITU-T H.263, Series H: Audiovisual and multimedia systems, Infrastructure of audiovisual services — coding of moving video, "Video coding for low bit rate communication," International Telecommunication Union, Jan. 2005.

[40] I.E.G. Richardson, "H.264 and MPEG-4 Video Compression, Video Coding for the Next-Generation Multimedia", John Wiley & Sons Ltd., Mar. 2005.

[41] ITU-R BT.601-6, Series BT: Broadcast service (television), "Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios," International Telecommunication Union, Jan. 2007.

[42] IETF RFC 2326: "Real Time Streaming Protocol (RTSP)", Schulzrinne H., Rao A. and Lanphier R., Apr. 1998.

[43] ITU-T, "Recommendation E.800: QoS Terms and Definitions related to Quality of Service and Network Performance including dependability," International Telecommunication Union, Aug. 1994.

[44] M. Siller and J. Woods, "Improving quality experience for multimedia services by QoS arbitration on a QoE framework," in Proc. of the 13th Packed Video Workshop 2003, Nantes, France, 2003.

[45] A. S. Patrick, et al., "A QoE Sensitive Architecture for Advanced Collaborative Environments," in Proc. of the First International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks, 2004.

[46] T. M. O' Neil, "Quality of Experience and Quality of Service for IP Video Conferencing", Polycom, Whitepaper, 2002.

[47] J. Goodchild, "Integrating data, voice and video - Part II," IP Video Implementation and planning guide, United States Telecom Association, 2005.

[48]] B. Bauer, A. S. Patrick, "A Human Factors, Extension to the Seven-Layer OSI Reference Model," Jan. 2004, Available: http://www.andrewpatrick.ca/OSI/10layer.html.

[49] C. Zetzsche, G. Hauske, "Multiple channel model for the prediction of subjective image quality," in Proc. SPIE, vol. 1077, pp. 209 - 216, 1989.

[50] S. Westen, R. L. Lagendijk, J. Biemond, "Perceptual image quality based on a multiple channel HVS model," in Proc. of International Conference on Acoustics, Speech and Signal Processing, vol. 4 of ICASSP, pp. 2351 - 2354, 1995.

[51] P. Teo, D. Heeger, "Perceptual image distortion," in Proc. of SID International Symposium, pp.2 09 - 212, Jun. 1994.

[52] D. Heeger, "Normalization of cell responses in cat visual cortex," in Journal of Visual Neuroscience, vol. 9, pp. 181 - 197, 1992.

[53] A. Bonds, "Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex," in Journal of Visual Neuroscience, vol. 2, pp. 41 - 55, 1989.

[54] G. Hauske, T. Stockhammer, R. Hofmaier, "Subjective Image Quality of Low-rate and Low-Resolution Video Sequences" , in Proc. of International Workshop on Mobile Multimedia Communication, Munich, Germany, Oct. 5-8, 2003.

[55] Z. Wang, L. Lu, and A. C. Bovik, "Video Quality Assessment Based on Structural Distortion Measurement," in Journal of Signal Processing: Image Communication, special issue on Objective Video Quality Metrics, vol. 19, no. 2, pp. 121-132, Feb. 2004

[56] O. Nemethova, M. Ries, A. Dantcheva, S. Fikar, M. Rupp, "Test Equipment of Time-Variant Subjective Perceptual Video Quality in Mobile Terminals," in Proc. of Int. Conf. on Human Computer Interaction (HCI), pp. 1–6, Phoenix, USA, Nov. 2005.

[57] F. Kozamernik, P. Sunna, E. Wyckens, D. I. Pettersen, "Subjective quality of internet video codecs - Phase 2 evaluations using SAMVIQ," EBU TECHNICAL REVIEW, Jan. 2005.

[58] JVT - Test and Video Group, "Report of The Formal Verification Tests on AVC (ISO/IEC 14496-10 — ITU-T Rec. H.264) ," Waikoloa, Dec. 2003.

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

[60] M. Ries, O. Nemethova, B. Badic, M. Rupp, "Assessment of H.264 Coded Panorama Sequences," in Proc. of the First International Conference on Multimedia Services and Access Networks, Orlando, Florida, Jun. 2005.

[61] J. Gozdecki, A. Jajszczyk, R. Stankiewicz: "Quality of service terminology in IP networks," IEEE Communications Magazine, pp. 153-159, Mar. 2003.

[62] A. Watson, M. A. Sasse: "Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications," In ACM Multimedia, pp. 55-60, ACM, Oct. 1998.

[63] M. Ries, O. Nemethova, M. Rupp, "Motion Based Reference-Free Quality Estimation for H.264/AVC Video Streaming," in Proc. of IEEE Int. Symp. on Wireless Pervasive Computing (ISWPC), San Juan, Puerto Rico, US, Feb. 2007.

[64] M. Ries, O. Nemethova, C. Crespi, M. Rupp, "Content Based Video Quality Estimation for H.264/AVC Video Streaming," in Proc. of IEEE Wireless Communications and Networking Conf. (WCNC), pp. 2668–2673, Hong Kong, Mar. 2007.

[65] Adima Manoli, "User Assessment for Negotiating the Quality of Service for Streaming Media Applications," in Proc. of the 19th International Symposium on Human Factors in Telecommunication, Berlin, Germany, Dec. 2003.

[66] K.Yamori, H.Ito, Y.Tanaka, "Optimum Pricing Methods for Multiple Guaranteed Bandwidth Service," in Proc. of the 2005 Networking and Electronic Commerce Research Conference, Riva Del Garda, Italy, pp. 349–355, Oct. 2005.

[67] M. Ries, O. Nemethova, M. Rupp, "On the Willingness to Pay in Relation to Delivered Quality of Mobile Video Streaming," in Proc. of IEEE ICCE 2008, Las Vegas, USA, pp. 195 - 196, 2008.

[68] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-Reference Perceptual Quality Assessment of JPEG Compressed Images," in Proc. of IEEE Int. Conf. on Image Processing, pp. 477-480, Sep. 2002.

[69] S. Saha and R. Vemuri, "An Analysis on the Effect of Image Features on Lossy Coding Performance," IEEE Signal Processing Letter, vol. 7, no. 5, pp. 104-107, May 2000.

[70] Alan Hanjalic, "A Content-Based Analysis of Digital Video," Springer US, 2004.

[71] A. Kankanhalli, H. J. Zhang, S. W. Smoliar, "Automatic Partitioning of Full-Motion Video," Springer-Verlag New York, USA, 1993.

[72] Q. LI, J. Chung-Mong, Lee and Wei Xiong, "A Video Information Management System," Springer Netherlands, 2004.

[73] J. R. Hampapur, A. and T. Weymouth, "Digital Video Segmentation", in Proc. of Second ACM international conference on Multimedia, San Francisco, USA, 1994.

[74] C. Garcia, R. Ronfard, G. Tziritas, P. Bouthemy, "Scene Segmentation and Image Feature Extraction for Video Indexing and Retrieval," Springer Berlin / Heidelberg, 2004.

[75] K. Mai, R. Zabih, J. Miller, "A Feature-Based Algorithm for Detecting and Classifying Scene Breaks," in Proc. of Third ACM international Conference on Multimedia, San Francisco, USA, 1995.

[76] C.L.Huang, B.Y.Liao, "A Robust Scene-Change Detection Method for Video Segmentation," in Journal of IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 12, pp. 1281-1288, Dec. 2001.

[77] S.Youm, W. Kim, "Dynamic Threshold Method For Scene Change Detection," in Proc. of ICME, vol. 2, pp. 337-340, 2003.

[78] X.Wang, and Z.Weng, "Scene Abrupt Change Detection," in Proc. of Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 880-883, 2000.

[79] K.W.Sze, K.M.Lam, G.Qiu, "Scene Cut Detection Using The Colored Pattern Appearance Model," in Proc. of ICIP, vol. 2, pp. 1017-1020, 2003.

[80] P. K.Sahoo and S.Soltani, A. K.C.Wong and Y.C.Chen, "A survey of thresholding techniques," in Proc. of CVGIP, vol. 41, pp. 233-260, 1988.

[81] H.C.Liu, G.Zick, "Automatic Determination of Scene Changes in MPEG Compressed Video," ISCAS, vol. 1, pp. 764-767, 1995.

[82] H.Li, G.Liu, Z.Zhang, Y. Li, "Adaptive Scene-Detection Algorithm for VBR Video Stream," in Journal of IEEE Transactions on Multimedia, vol. 6, no. 4, pp. 624-633, Aug. 2004.

[83] A. Dimou, O. Nemethova, M. Rupp, "Scene Change Detection for H.264 Using Dynamic Threshold Techniques," in Proc. of 5th EURASIP Conf. on Speech and Image Processing, Smolenice, Slovakia, Jun. 2005.

[84] A.G. Dantcheva, "Video Quality Evaluation," Diploma thesis Vienna University of Technology, Apr. 2007.

[85] VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," 2000, available at http://www.vqeg.org/.

[86] Woo Young CHOI, Rae-Hong PARK, "Motion vector coding with conditional transmission," in Journal of Signal Processing, Vol. 18, No. 3, Nov. 1989.

[87] Maurizio Pulu, "On Using Raw MPEG Motion Vectors To Determine Global Camera Motion", Aug. 1997.

[88] Sung-Hee Lee, Ohjae Kwon, Young-Wook Sohn, Jeong-Woo Kang and Rae-Hong Park, "Motion Vector Correction Based on the Pattern-like Image Analysis," in Proc. of ICCE, LV, USA, Jun. 2003.

[89] C. Crespi de Arriba, "Subjective Video Quality Evaluation and Estimation for H.264 codec and QVGA Resolution Sequences," Diploma thesis Vienna University of Technology, Aug. 2007.

[90] W. J. Krzanowski, "Principles of Multivariate Analysis," Clarendon press, Oxford, 1988.

[91] K. R. Gabriel, "The biplot graphic display of matrices with application to principal component analysis," Biometrika 58, pp. 453-467, 1971.

[92] K. Bosch, "Statistik-Taschenbuch," Oldenbourg Wissensch. Vlg, Munich, 1998.

[93] M. Ries, O. Nemethova, M. Rupp, "Video Quality Estimation for Mobile H.264/AVC Video Streaming," in Journal of Communications, Vol. 3, pp. 41 - 50, 2008.

[94] M. Ries, O. Nemethova, M. Rupp,"Performance evaluation of mobile video quality estimators," invited paper, in Proc. of 15th European Signal Processing Conference (EUSIPCO), Poznan, Polen, Sep. 2007.

[95] B.V. Dasarathy, B.V. Sheela, "Composite classifier system design: Concepts and methodology," in Proc. of the IEEE, vol. 67, no. 5, pp. 708-713, 1979.

[96] L.I. Kuncheva, "Combining Pattern Classifiers, Methods and Algorithms," New York, Wiley Interscience, 2005.

[97] A. Krogh, J. Vedelsby, "Neural Network Ensembles, Cross Validation and Active Learning," Advances in Neural Information Processing Systems 7, MIT Press, 1995.

[98] Hastie, Tibshirani, Friedman, "The Elements of Statistical Learning," Springer, 2001.

[99] C. Igel, M. Hsken, "Improving the Rprop learning algorithm," in Proc. of the 2nd Int. Symp. on Neural Computation, pp. 115-121, Berlin, ICSC Academic Press, 2000.

[100] M. Ries, R. Puglia, T. Tebaldi, O. Nemethova, M. Rupp,"Audivisual Quality Estimation for Mobile Streaming Services," in Proc. of 2nd Int. Symp. on Wireless Communications (ISWCS), Siena, Italy, pp. 173–177, Sep. 2005.

[101] C. Bishop, "Neural Networks for Pattern Recognition," Oxford University Press, 1995.

[102] K. Hornik, M. Stinchcombe, H. White, "Multilayer feedforward networks are universal approximators," in Journal of Neural Networks, Vol. 2, pp. 359-366, 1989.

[103] F. Girosi, M. Jones, T. Poggio, "Regularization Theory and Neural Networks Architectures," Neural Computation, Vol. 7, pp. 219-269, 1995

[104] G. Rubino, M. Varela, "A new approach for the prediction of end-to-end performance of multimedia streams," in Proc. of International Conference on Quantitative Evaluation of Systems (QEST'04), IEEE CS Press, University of Twente, Enschede, The Netherlands, Sep. 2004.

[105] A. Weigend, "On overfitting and the effective number of hidden units," in Proc. of the 1993 Connectionist Models Summer School, pp. 335-342, 1993.

[106] S. Geman, E. Bienenstock, R. Doursat, "Neural Networks and the Bias/Variance Dilemma," Neural Computation, Vol. 4, pp. 1-58, 1992.

[107] F. D. Foresee, M. T. Hagan, "Gauss-Newton approximation to Bayesian regularization," in Proc. of the 1997 International Joint Conference on Neural Networks, pp. 1930-1935, 1997.

[108] R. M. Neal, "Bayesian Learning for Neural Networks," New York: Springer-Verlag, 1996.

[109] J. M. Bernardo, A. F. M. Smith, "Bayesian Theory," New York: John Wiley, 1994.

[110] Hagan, M. T., and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," IEEE Transactions on Neural Networks, Vol. 5, no. 6, pp. 989-993, 1994.

[111] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink, D. L. Alkon, "Accelerating the convergence of the backpropagation method," Biological Cybernetics, Vol. 59, pp. 257-263, 1988.

[112] M. Ries, J. Kubanek, M. Rupp, "Video Quality Estimation for Mobile Streaming Video Quality Estimation for Mobile Streaming," in Proc. of Measurement of Speech and Audio Quality Networks, Prag, Czech Republic, 2006.

[113] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs," International Telecommunication Union, 2001.

[114] ITU-R Recommendation BS.1387-1, "Method for objective measurements of perceived audio quality," Geneva, Switzerland, 1998-2001.

[115] S. Voran, "Objective Estimation of Perceived Speech Quality. Part I: Development of the Measuring Normalizing Block Technique," in Journal of IEEE transactions on speech and audio processing, vol. 7, no. 4, Jul. 1999.

[116] A. Rimell, M. Hollier, "The significance of cross-modal interaction in audiovisual quality perception. Advanced Communications Research," BT Laboratories, Ipswich, IP5 3RE, UK, 1999.

[117] C. Jones, D.J. Atkinson, "Development of Opinion-Based Audiovisual Quality Models for Desktop Video-Teleconferencing," in Proc. of 6th IEEE International Workshop on Quality of Service, Napa, California, May 18-20, 1998.

[118] S. Tasaka, Y. Ishibashi, "Mutually Compensatory Property of Multimedia QoS," in Journal of IEEE Transactions, 2002.

[119] R. Puglia, "Influence of Audio and Video Quality on subjective Audiovisual Quality - H.263 and Adaptive Multi Rate (AMR) coding," Diploma thesis Vienna University of Technology, Apr. 2005.

[120] T. Tebaldi, "Influence of Audio and Video Quality on subjective Audiovisual Quality - MPEG-4 and AAC coding," Diploma thesis Vienna University of Technology, Apr. 2005.

[121] E. Zwicker, R. Feldtkeller, "Das Ohr als Nachrichtenempfänger," S. Hirzel Verlag, Stuttgart, 1967.

[122] M. Ries, O. Nemethova, M. Rupp, "Reference-free video quality estimator for video streaming," submitted for Austrian patent 22.12.2006, No. PCT/AT2006/000539.

[123] A. K. Jain: "Fundamentals of digital image processing," Englewood Clifs, NJ: Prentice-Hall Inc., pp. 348-357, 1989.

[124] J.G. BEERENDS, "Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality, Speech Quality Assessment," Workshop papers, pp. 1-9, Bochum, Germany, Nov. 1994.

[125] R. B. Blackman, John Tukey, "Particular Pairs of Windows," published in "The Measurement of Power Spectra, From the Point of View of Communications Engineering", pp. 98-99, New York: Dover, 1959.

[126] E. Zwicker, "Subdivision of the audible frequency range into critical bands," in Journal of the Acoustical Society of America, 33, Feb., 1961.

[127] PESQ Algorithm - C code, "Perceptual Evaluation of Speech Quality (PESQ) ITU-T Recommendation P.862," Version 1.2 - 2, Aug. 2002.