Project Number: 248495
Project acronym: OptiBand
Project title: Optimization of Bandwidth for IPTV Video Streaming

Deliverable reference number: D2.1
Deliverable title: Criteria specification for the QoE research

Due date of deliverable: **M6**
Actual submission date: **30 June 2010**

Start date of project: 1 January 2010	Duration: 30 months
Organisation name of lead contractor for this deliverable: FTW Authors: Michal Ries, Peter Fröhlich, and Raimund Schatz	

Release no.	Date	Status	Distribution
V1	30 April 2010	FINAL DELIVERABLE	PU

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n 248495

Table of Contents

Table of figures

Glossary

Abbreviation / acronym	Description
ACR	Absolute category rating
ARQ	Automated repeat request
CC	Content class
DCR	Degradation category rating
HD	High definition
HVS	Human visual system
IPTV	Internet protocol television
KPI	Key performance indicators
MOS	Mean opinion score
MSE	Mean square error
PC	Pair comparison
PDA	Packet dropping algorithms
PSNR	Peak signal to noise ratio
QoE	Quality of experience
QoS	Quality of service
ROI	Region of interests
RTCP	Real-Time control protocol
SI	Spatial information
SSIM	Structural similarity
SVC	Scalable video coding
SOTA	State-of-the-art
TCP	Transmission control protocol
TI	Temporal information
UDP	User datagram protocol
VOD	Video on demand
VQM	Video quality metric

1. Executive Summary

Deliverable 2.1 presents the results of Task 2.1 (“Study and define criteria for QoE”). T2.1 is the first task of the OptiBand workpackage 2 (“Research of QoE metrics”), headed by Telecommunications Research Centre Vienna (FTW). The overall goal of WP2 is to ensure that the QoE of the end user is not negatively affected (or affected by a tolerable degree) by the reduction of the bandwidth by the content aware data drop algorithm. Quality of experience (QoE) is defined as the “overall acceptability of an application or service, as perceived subjectively by the end-user”.

The **goals of D2.1** specified for this deliverable are as follows:

1. To introduce into research problems and influence factors for the QoE of HDTV
2. To provide a detailed survey of QoE testing methods for HDTV
3. To specify a QoE test strategy for the OptiBand project
4. To define the QoE criteria, thresholds, and benchmarks

Relations to other deliverables and tasks:

- This deliverable is based on D1.1 (Functional Specification Document), in which the basic requirements for the OptiBand architecture and test setup have been specified.
- The results of the evaluation method specified in this deliverable will feed into algorithm development in WP3 (Network solution) and WP4 (Head-end solution).
- The (already substantially detailed) methods and criteria specification within the present deliverable is the basis for the QoE Research Plan developed subsequently within this workpackage (D2.2, due month 9).

Structure of D2.1:

D2.1 is structured according to the goals mentioned above:

- Section 2 introduces into the state-of-the-art of QoE testing.
  - Section 2.1 briefly introduces the motivation and general research of QoE
  - Section 2.2 provides an overview of the many influence factors contributing to the QoE of HDTV services, such as user characteristics, end-devices, and service types (-> addressing goal 1 of D2.1, see above).
  - Section 2.3 describes state-of-the-art QoE measurement methods (-> goal 3)
- Section 3 specifies the QoE test strategy within the OptiBand project (-> goal 3)). The main principles are testing in 3 iterations and combining user-based with automated QoE assessment methods.
- Section 4 specifies the quality criteria, thresholds, and benchmarks for interpreting the QoE research results (-> goal 4).

This deliverable has been approved by the OptiBand consortium partners in the 2nd OptiBand project meeting (29-30 June 2010).

2. QoE research for IPTV services: an introduction

After a brief introduction to QoE research (section 2.1), we describe the relevant factors influencing the perceived quality of video and TV services (section 2.2) and provide an overview of state-of-the-art QoE assessment methodologies.

2.1 Introduction to QoE

Already 15 years ago, the ITU-T has recognized the fact that any evaluation of service quality evaluation is incomplete without a strong user centric component. For example, ITU-T Rec. E.800 [1] provides a general overview of quality sub-dimensions contributing to the user perceived QoS (Quality of Service) of telecommunication services. The contribution of these sub-dimensions is quantitatively correlated with respective technical QoS parameters and accordingly the overall QoS results from a summation of those parameters. However, most subsequent research work has primarily focused on the technical aspects of this approach, thus lacking sufficient evidence on the user's quality perception [2].
In contrast to objective service quality, the subjective quality reflects the subjective perception of individual viewers. In general, its evaluation is performed by psycho-visual experiments and therefore is influenced by the following subjective and objective factors: video content, encoding parameters, usage scenario and network performance, etc. Moreover, the relation between objective parameters or QoS parameters [1] with subjective quality is not trivial.

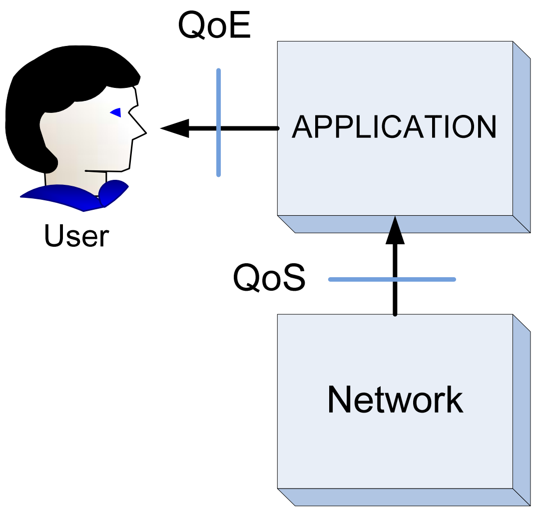

Figure 1: Relationship between user, application and network.

The subjective service and video quality belongs to a higher abstraction layer also called QoE. The QoE basically relates to a higher abstraction layer when compared to QoS [3], [4], [6], [7], [8], [9], and reflects the user's perceived experience of what is being presented by the application layer. The application layer acts as a user interface front-end that presents the overall result of the individual quality of services [6] (see Figure 1). It can be considered as a perceptual layer and an extension to the application layer defined in the OSI Model [7].

The different quality notions around Quality of Service (QoS) and Quality of Experience (QoE) are used and defined differently through the standardization bodies and the associated research domain. We know better how to measure QoS based on service performance metrics. On the other hand, trying to translate or map from the QoS domain to the more subjective domain of QoE is a significant challenge. QoE measurement functions can be obtained from QoS objective service parameters. QoE measurements provide the most reliable feedback on what levels of service quality customers actually perceived. The most suitable approach towards QoE measurement is the holistic approach that links together QoE and QOS [2], [11], [12]. This approach in general is highly suitable for telecommunication and multimedia services such as IPTV.

2.2 Influence factors on QoE

QoE in the context of telecommunications networks represents a purely subjective measure from the user's perspective of the overall value of the service provided. QoE cannot be taken simply as objective quality parameter of the service, but must also take into consideration every factor that contributes to overall user service perception.

In the following, the QoE influence factors will be introduced, along the taxonomy shown in Figure 2. Section 2.2.1 describes the role of the context surrounding a TV viewing situation, including personal characteristics and preferences, the user's TV viewing environment, cultural aspects, and the viewed content.

Section 2.2.2 will then highlight the strong impact of the expectations a user has towards a certain service. In this regard, the role of general TV viewing habits (such as Internet TV and user-generated video platforms), the personal usage history, as well as the image of a service or brand are analyzed.

Section 2.2.3 then elaborates on technology-related QoE influence factors, ranging from infrastructure issues (codec settings, end user devices, video delivery, packet dropping algorithms, service type).

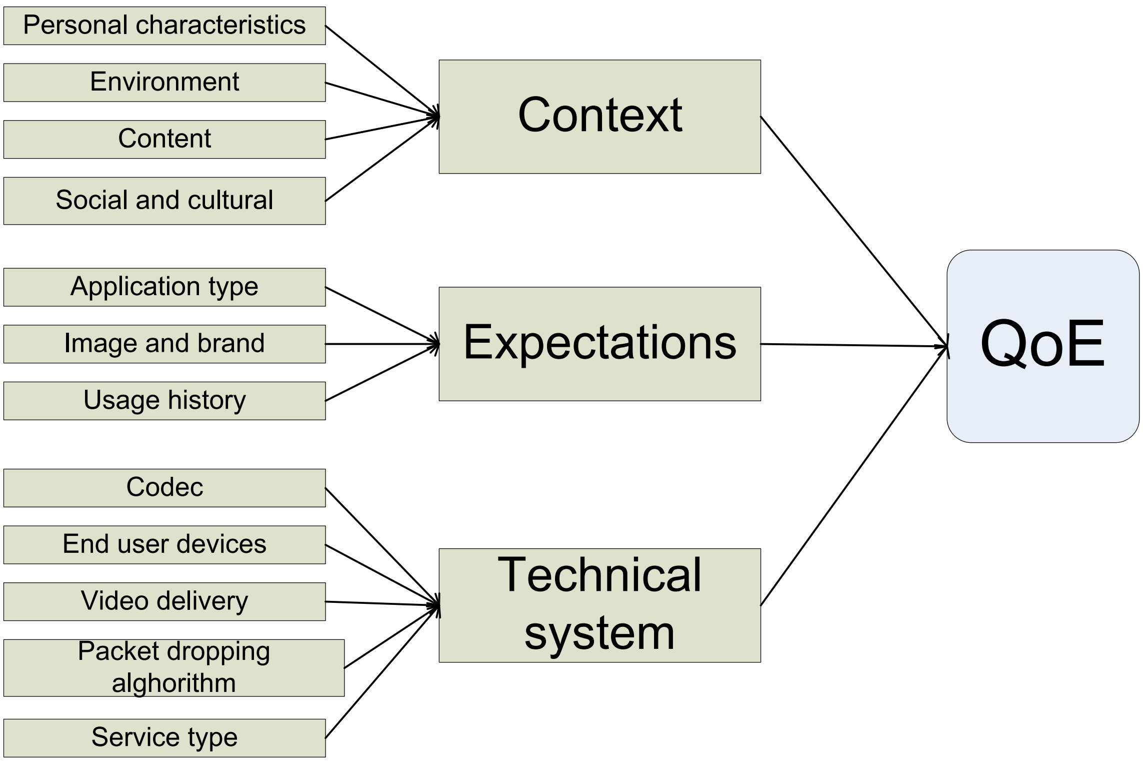

Figure 2: Influence factors of quality of experience.

2.2.1 Context

Each multimedia service with defined usage scenario is strongly determined by the context. Furthermore, the subjective perception of individual viewers is strongly influenced by the context. The context of multimedia services consists of three main factors: environment, content and technical system.

Personal characteristics

Different users have different viewing behaviours and quality judgments. Relevant characteristics in this regard can be static (e.g. age, gender, native language) as well as dynamic (e.g. motivation, emotional status) [58]. When designing novel services and algorithms for IPTV, it is therefore important to know what the distribution of user characteristic of the target users of IPTV services are.

TV viewing surveys show that there is a relatively even subscriber distribution in terms of **gender and age** (see Table 1). These statistics [15] also show that the younger generation (18 - 34) has a similar high share as the middle-aged and the senior population. This indicates that despite the high popularity of the Internet-based and user-generated video, TV services are still fighting their corner in this age group. Moreover, there is no statistically dominant viewer group in terms of gender and age in Europe [15].

	All	Cable TV	Satellite TV	IPTV	Pay Cable Channels (HBO, …)
Gender					
Male	49%	50%	52%	59%	54%
Female	51%	50%	48%	41%	46%
Age Range					
18-34	31%	32%	27%	32%	31%
35-54	38%	38%	39%	36%	40%
55+	31%	30%	34%	32%	29%
Average Age	45.1	44.9	46.4	44.9	44.5

Table 1: US subscribers of TV services [18]

US consumer research from 2009 [18] indicates a high share (44%) of households with higher **income** than $75,000. This reflects that IPTV currently only includes paid services and offer many additional pay services (e.g. VOD). The viewer distribution in Europe is very similar to the presented US statistics [15].

Gender	
Male	59%
Female	41%
Age Range	
18-34	32%
35-54	36%
55+	32%
Average Age	44.9
Annual household incomes	
Less than $25,000	12%
$25,000 - $49,999	20%
$50,000 - $74,999	23%
$75,000 +	44%
Average Income	$77,641

Table 2: US IPTV consumer research statistics from 2009 [18]

Consumer statistics show that people living in the countries of northern Europe have a lower **TV watching time** than those living in the east or south [15]. The viewing times in Europe range from approximately 140 minutes a day in Austria, Switzerland or Germany) to approximately 240 minutes in Southern or Eastern European countries (e.g. Italy and Hungary).

Given the above considerations, it is important to specify a representative and balanced set of subjects for QoE tests. The ITU-R BT.500-11 recommendation [46] provides recommendations for sample size and composition. The test subjects should not be directly involved in video or picture quality evaluation as part of their work and should not be experienced assessors. Furthermore, they should have normal visual acuity. And, given the above data, subjects should watch TV approximately as often as the average population.

Environment

TV services are still mostly consumed in a living room environment: 93 % of living rooms are equipped with a TV set [17]. The main environmental characteristics within a living room TV viewing setting are screen size and viewing distance. Purchase statistics show that consumers prefer flat panel displays with screen sizes above 32“. The optimal viewing distance for a screen size of 32“ is about 3 meters, which will in most cases be only realistic in the (usually larger) living room.

The opposite of living room environment is usage on small terminals while on the move. The mobile environment is determined by the flexibility of portable devices supporting rich multimedia services. The portable devices allow consuming TV services in various environments, which makes it hard to define.

Due to mobility and flexibility requirements, screen sizes of portable devices are small, usually ranging between 2,5“ and 5”. The viewing distance for watching TV on mobile devices is usually 20 - 30 cm [40].

Content

The most important aspects to be considered during content selection for QoE tests are perceptual impact and popularity. These will be introduced in the following.

Perceptual impact

Visual perception of video sequences is influenced by their spatial, colour, and temporal characteristics. Spatial, colour, and temporal characteristics can be used to characterize the content type of a sequence [39], [40]. Selection of these features describes a content class (CC) as a cluster of video sequences with similar motion, colour and spatial characteristics. Figure 3 shows that there are substantial differences in the temporal and spatial domain of popular content types. Where the spatial information (SI) [53] measurement reflects the complexity (amount of edges) of still pictures and the temporal information (TI) [53] measurement is based upon the motion difference feature. TI is calculated for every two consecutive frames as difference of the luminance pixel values. TI is than computed as a maximum over time of the standard deviation over space.

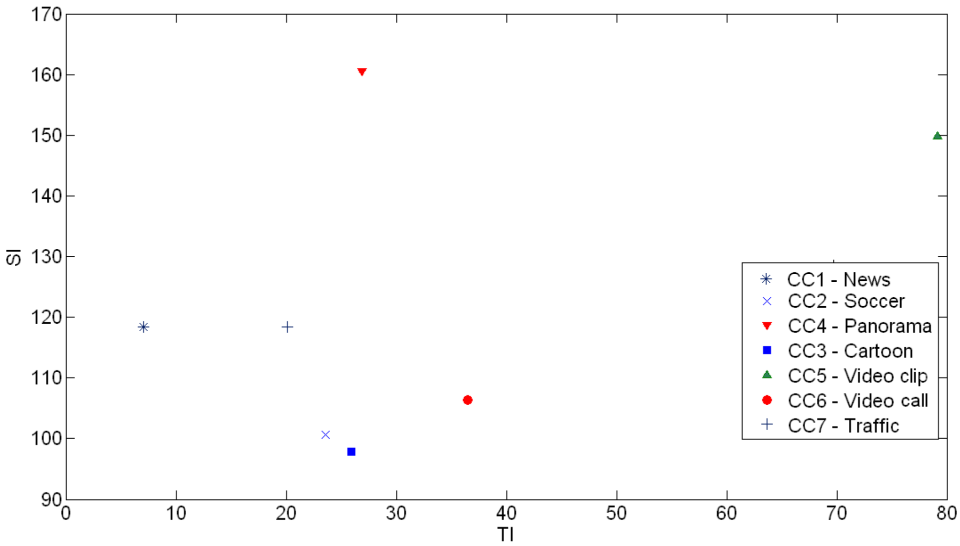

Figure 3: Spatial and temporal information (SI, TI) of popular content types.

Figure 4 shows the results from a study on subjective judgments of video sequences depending on the video content and the type of degradation [41]. The results indicate that, apart from the lowest bit-rate, movies (magenta square) and news interviews (green star) obtain the best rating in terms of quality, while movie trailers (red star) and video clips (cyan triangle) yield equal ratings. The ranking of the soccer video improves when the bit-rate increases. Soccer videos appear to be really sensitive to the bit-rate and require high bit-rate encoding for getting an acceptable quality. The ranking of the contents for the lowest bit-rate is clearly different from the ranking of the contents from medium and high bit-rates. This [41] and other work [40] calls for a concept for description of content classes according to their different impacts on user perception.

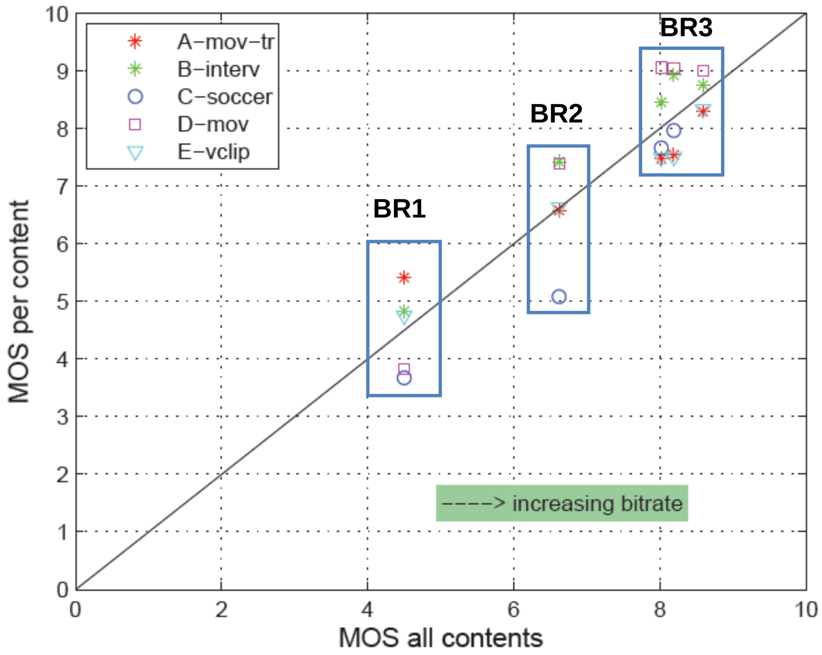

Figure 4: Perceived quality depending on the bit-rate and on the content [41].

Considering these perceptual aspects, the following content classes can be differentiated:

- Soccer: This CC contains wide angle camera sequences with uniform camera movement (panning). The camera is tracking small rapid moving objects (ball, players) on the uniformly coloured (typically green) background.
- Video clip or action movie (movie thriller): 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. This class concentrates more content class due to very similar perception of the video contents (see Figure 4).
- News: 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.
- Video call (interview): Typical example if this CC is 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 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 or interview.
- Cartoon: In this CC 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.
- Panorama: Global motion sequences taken with a wide angle panning camera. The camera movement is uniform and in a single direction.

Content popularity

At the moment, consistent statistics of video content popularity in Europe [55] are not available, due to the high variety of video services and cultural differences within Europe. Generally, independent market research indicates highest popularity for sport and TV series programmes. Sport is the driving force behind the program schedules of both general and special channels [15]. The sport content with especially high impact is soccer. For example, in Spain, the final of Euro 2008 had an 80.9% market share with 14.5 million viewers. Moreover, marketing research of TI shows dominating popularity of sport content followed by movies (movie genre not specified).

The European audiovisual observatory [15] (see Figure 5) confirms the popularity HD of sport and film channels in Europe. Thus, these two content types can be regarded as highly important for driving HD IPTV services.

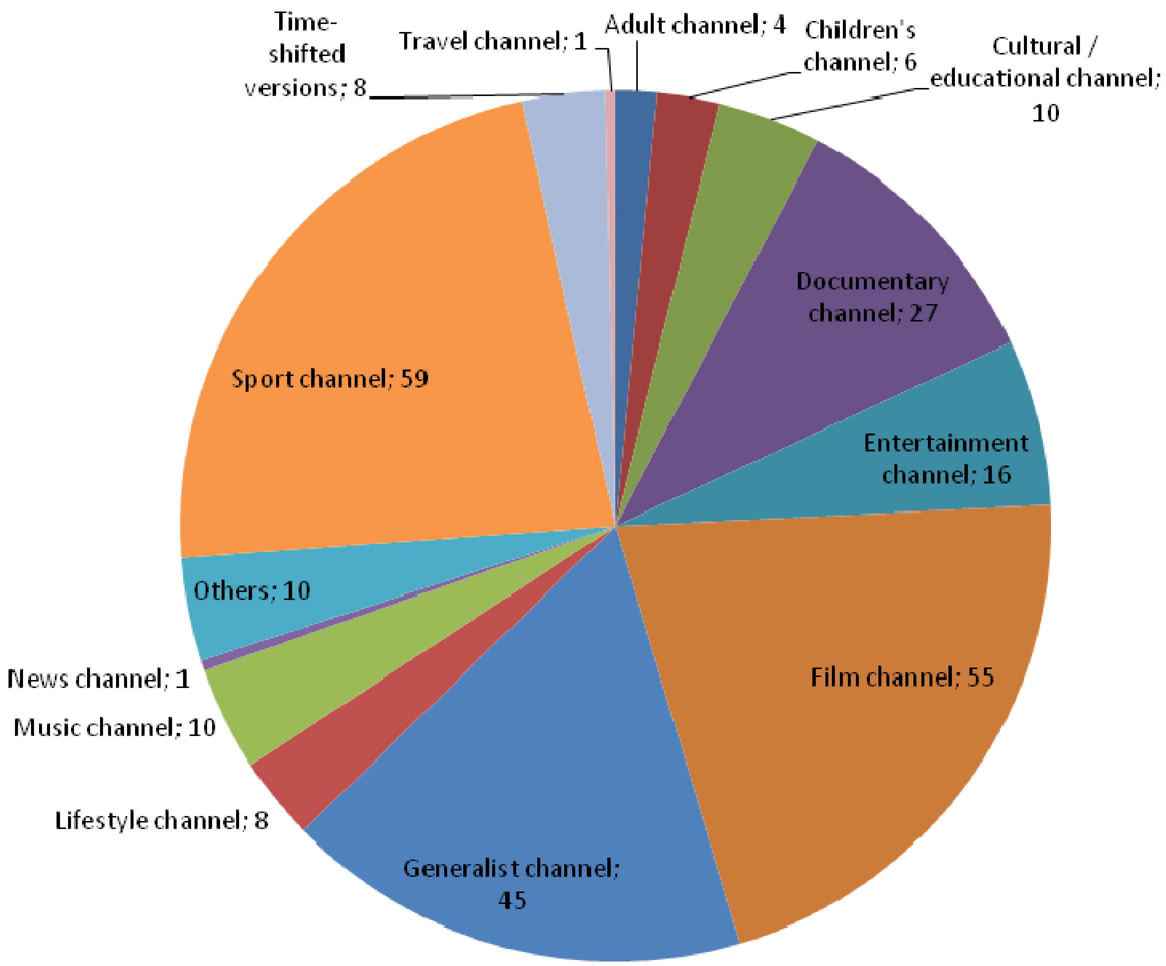

Figure 5: HD channels available in Europe by genre [15]

Social and cultural context

The strong social component of television has been subject of research since long. Lull [61], for example, argues that TV is essentially a social medium, with its context-providing function being more important than the actual content transmitted. Television shows are designed to function as social happenings, not only during their public airing but also before and after their broadcast. The most important social context factors of TV viewing in this regard are clearly individual vs. group usage, and the group composition characteristics. Another important component for QoE is the cultural and interregional TV consumption differences [60].

Both the influence of TV watchers group composition as well as interregional differences on QoE is largely unknown, and future research should investigate this. However, as the OptiBand project does not have a scientific focus on such social aspects, these factors are controlled in the user studies, by only investigating individual subjects, and by including users from two typical HDTV markets in Europe: Austria and Italy (D1.1).

2.2.2 Expectations

Since "QoE is how the user feels about how an application or service was delivered, relative to their requirements," [42] this experience is strongly influenced by the user's expectations. This section describes how expectations towards a specific service type or brand can shape subjective quality, and how important a person's individual usage history can be.

Application type

Recent studies show that consumers of TV services expect higher quality and more control above the video content and TV viewing [15], [16]. These expectations partly reflect the digital broadcasting services (terrestrial and satellite) with increase of resolution to HD. On the other hand, these services provide minimum control of viewing and content to viewer. The emerging IPTV and VOD reflect these increasing user expectations. IP-based technology offers superior HD quality and full viewing control.

Another fast growing service is Internet TV and video. The expected quality of user-generated video content is often considerably lower than traditional video services, involving social status and relationships, as well as the capabilities of capturing devices (e.g. often mobile phone cameras), network connection, as well as display devices.

When planning user studies, it is important to know about the influence of different quality expectations on the perceived quality. For example, the same video sequence can be judged more favourably when watchers think it is a YouTube video than when they think it is from a HDTV service. Consequently, the introductory briefing about the expected service needs to be carefully designed to avoid bias.

Image and brand

User expectations are also strongly coupled to the image of a certain service or brand. For example, people expect more quality from a premium service than from a free-of-charge offer.

Thus, also the information on the image and brand needs to be taken into account in the briefing of a test user.

Usage history

The subjective assessment of a service by a watcher is highly dependent on his or her previous experience. For example, a long-year subscriber of a premium TV service will be more critical than a user who is used to watching TV with a room antenna.

When conducting quality of experience user tests, it is therefore important to take the previous experience into account. The most valid strategy here is to take account of current usage patterns and trends from market research studies (see also D1.1, section 3.3.3).

2.2.3 Technical system

Of course, there are several important technology factors influencing the QoE. This part describes the most important components of digital IPTV services: codec settings, end-devices, video delivery, streaming optimization, and service type.

Codec

Uncompressed video requires large amounts of bandwidth and storage space. Moreover, the end-user cost in transmission networks is typically proportional to available bandwidth and transmitted data volumes. Therefore, the videos transmitted over any networks are compressed with very effective and lossy compression algorithms. The provided video services are capable of supporting more than one video and audio codecs depending on the head-end equipment and set-top box.

The following video codecs are widely supported for video services:

- **H.262** (or: MPEG-2 Video) [19] is a digital video compression and encoding standard developed in 1994. H.262 is rather old, but still frequently applied in DVB broadcasting services, due to low deployment and hardware cost. On the other hand, encoding performance of this codec has been significantly outperformed by the newer standard H.264.
- **H.263** [20] – is also a rather old encoding standard designed for low complexity and low rate applications. Usually this codec is used in mobile video applications.
- **H.264** (MPEG-4 AVC or MPEG-4 Part 10) [21] is the newest commercial coding standard. This coding standard is a successor of two coding standards (H. – 262, 263) and an extension of MPEG-4. This emerging coding standard is frequently used in the whole spectrum of video applications, such as DVB, internet video services, and mobile video services. Its scalable extensions aka SVC are built on an H.264/MPEG-4 AVC compliant, i.e., backwards-compatible base layer corresponding to a minimum quality, frame rate, and resolution. SVC allows adding one or more enhancement

layer(s). Sub-bitstreams including one, several or all enhancement layers represent the same video at increased quality and/or increased resolution and/or increased frame rate.

- **SMPTE** 421M (the standardized version of Windows Media™ 9) [22]: had initially been developed as a proprietary video format by Microsoft before it was released as a formal SMPTE standard video format in 2006. It is today a supported wide spread standard for HD DVDs, Blu-ray Discs, Windows Media Video 9, and Microsoft's Silverlight framework.

The best trade-off between encoding efficiency and perceptual quality are offered by the H.264 and SMPTE 421M [3] coding standards. Decoding of full HD H.264 video streams is supported by many end-user devices and video applications. . These technological advantages and also its growing market penetration ([23], see Figure 6) make H.264 the most suitable coding standard for HD IPTV services. .

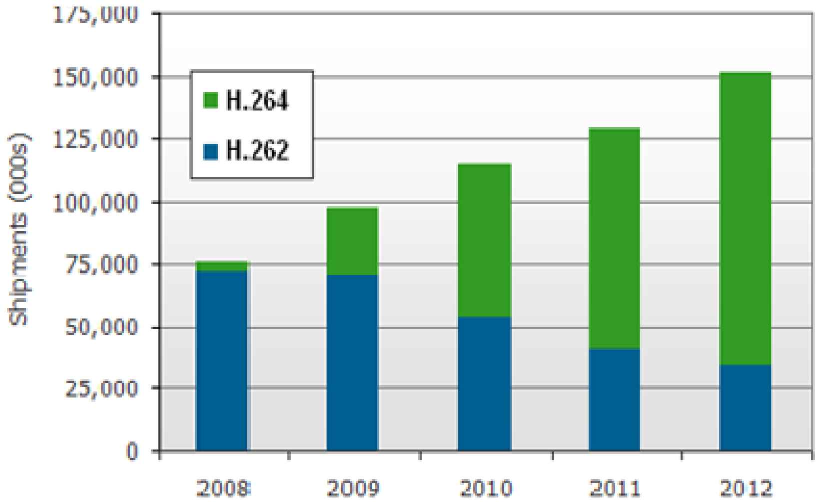

Figure 6: Shipments of digital TVs by decoder type [23].

End user devices

In the recent decade, the consumer electronic industry has provided a large scale of products for multimedia consumption. New generation end-user devices substantially vary in size, processing power, mobility, usage, and other respects.

Generally, three classes of end-user devices can be distinguished: mobile video devices (e.g. cell phones, iPods), personal computers, and (HD) TV sets. Nowadays, the consumption of video services on mobile video devices and personal computers is rapidly growing, due to technological progress and availability of new video services.

- **Mobile video devices**, with rich multimedia services. Typical representatives of these devices are video-capable cell phones, iPods, and palmtops. The video applications for these devices are limited in processing power, access network performance and video resolution. These limitations allow provisioning of video services only at lower bit and frame rates. On the other hand, perceptual optimisation and efficiency of the newest video encoding standard H.264 allow provisioning of video services in excellent quality.
- **Personal computers**, with increasingly popular streaming services and multimedia capabilities, have become a primary end-user device class for video applications. Due to high processing power, data storage capability and increasing bandwidth availability, personal computers support almost all types of video applications.
- **HD - TV sets** are the successor of traditional television systems (with 480 or 576 lines). Since 2009, the market is dominated by sets providing 1080p inputs, HDMI interfaces, and full HD resolution. 1080p resolution screens are available in all display technologies, including plasma, DLP front and rear projection and LCD front projection. Recent market research shows the domination of LCD technology in TV panels ([24], see Figure 7). Higher resolution and new projection technologies allow for increasing the display size. Today, 50% of the users have 32” or higher display sizes, a trend also corroborated by internal marketing research of TIS. According to current reports of TV flat panels sales, 40” plus are dominating [26].

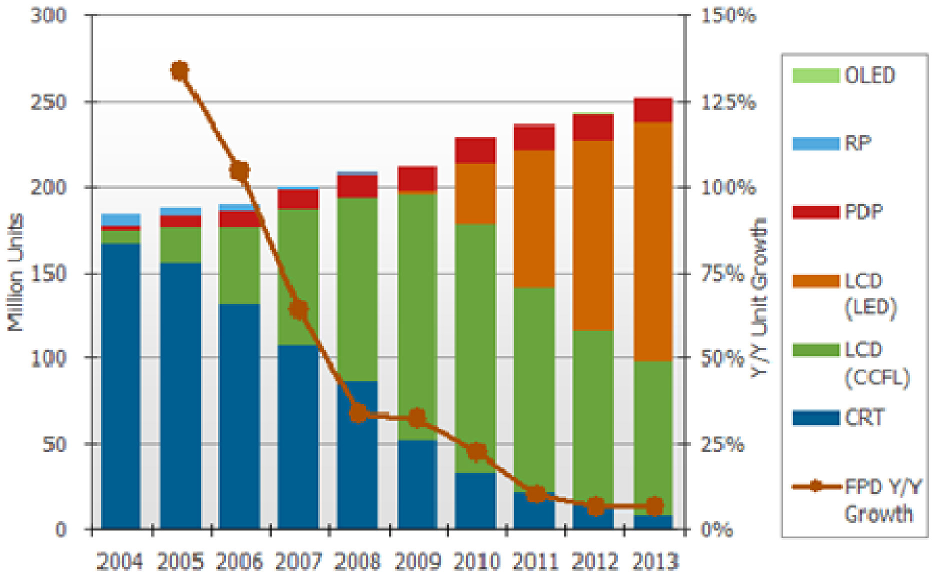

Figure 7: Worldwide TV Market by Technology [26].

The actual trend shows that end user devices are going to be more efficient, powerful and available. The increase of processing power and efficiency allows for playing video content in superior quality on all types of devices in the very close future. This aspect removes the differences between the multimedia products on one side but necessary bring to attention the real product usage on the other side. Good example of this trend is a cell phone equipped with digital TV receiver.

Video delivery

Video delivery refers to the ability of an application to play synchronized (audio and video) media streams in a continuous way while those streams are being transmitted to the client over a packet data network. The usual way of transport of video streaming over IP packet networks is based on the 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. In contrast to the Transmission Control Protocol (TCP), UDP does not provide any Automated Repeat reQuest (ARQ) mechanism to perform retransmissions. 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 for most of the IP based video services, which represent IPTV, VOD and Internet TV.

Packet dropping algorithm

One of the main tasks of video streaming is to address the problem of network congestion, when the bit rate of the transmitting video is greater than the current available network bandwidth. The packet dropping algorithm goal is to save bandwidth and ensure that the frames which have been selected to be sent can entirely arrive at the decoder before their playback deadlines. Furthermore, the packets which were selected to be dropped shouldn't impair the correct video decoding, meaning that the decoding of the sent packets shouldn't be dependent on the decoding of the dropped packets.

Of course, packet dropping algorithms should not significantly influence the perception of HD IPTV services. Commonly known perceptual impairments related to packet dropping are the extension of the video buffering verifier delay (usual for Smoothing MPEG packet solutions), video artefacts caused by dropping B-frames (usual for selective frame dropping solutions) and encoding artefacts (usual for multi rate streaming and scalable video coding). It is crucial for the OptiBand QoE experience research to further investigate characteristic artefacts of the specific packet dropping solutions developed during the project, and to assess their impact on the perceived quality of HD IPTV services.

Service type

The most frequently used video services are TV broadcasting, Video on Demand and Video sharing services. All of these services can be based on an IP platform. Of course, TV broadcasting services can be provided via terrestrial, cable or satellite radio transmission. On the other hand, the IP based technology provides almost an unlimited option in all aspects for provisioning of all types of video services.

TV broadcasting is the distribution of video content which transmit programs to an audience. The audience may be the general public or a relatively large sub-audience, such as children or young adults. This service has very limited options for bidirectional interaction with users or between the users. Majority of broadcasting services are dedicated for static user but in recent years started pilot services for mobile users (e. g. DVB-H).

Video on demand (VOD) are systems which allow users to select and watch video content on demand. Television VOD systems either stream content through a set-top box, allowing viewing in real time, or download it to a device such as a personal computer, PVR or portable media player for viewing at any time. The majority of cable- and telco-based television providers offer both VOD streaming, including pay-per-view and free content, whereby a user buys or selects a movie or television program and it begins to play on the television set almost instantaneously, or downloading to a PVR rented from the provider, or downloaded onto a personal computer, for viewing in the future. Internet television, using the Internet, is an increasingly popular form of video on demand.

Video sharing refers to web services where users can distribute theirs videos. Some services may charge, but the large majority of them offer free services. Many services have options for private sharing and other publication options. Among video sharing services are the most dominant user generated video sharing websites. User generated sites mostly offer free services whereby users can upload video clips and share it with the masses. Since 2008 the most dominant web sharing services support full HD formats. The most dominant global web pages for video sharing are [56]:

- YouTube with 64M videos and 1200M Views Watched Per Day,
- MetaCafe with 38M videos and 17M Views Watched Per Day,
- Dailymotion with 10M videos and 60M Views Watched Per Day,

Actual development in provisioning of video services show domination of IP based video delivery platforms in all service types. The only exception is the analogue TV service which is going to be turned off within the next two years all over the developed world.

2.3 QoE measurement methods for video

This section provides a comprehensive overview of QoE measurement methods. The presentation is divided into user-based assessment and automated assessment (i.e., which does not require user involvement). We conclude with a description of QoE measurement methods for IPTV services.

2.3.1 User based assessment

When conducting QoE studies with users, it is important to ensure reproducible and representative conditions for viewing and interacting with video services. In this regard, ITU-R Recommendations [46], [47], [53], [54] propose several methodologies:

- **Absolute Category Rating (ACR)** method is a category judgement where test sequences are presented one at a time and are rated independently on a category scale. (This method is also called Single Stimulus Method.) The method specifies that after each presentation the subjects are asked to evaluate the quality of the sequence shown.
- **The Degradation Category Rating (DCR)** implies that the test sequences are presented in pairs: the first stimulus presented in each pair is always the source reference, while the second stimulus is the same source presented through one of the systems under test. (This method is also called the Double Stimulus Impairment Scale method.)
- The method of **Pair Comparisons** implies that the test sequences are presented in pairs, consisting of the same sequence being presented first through one system under test and then through another system. The systems under tests (A, B, C, etc.) are generally combined in all the possible n (n–1) combinations AB, BA, CA, etc. Thus, all the pairs of sequences should be displayed in both possible orders (e.g. AB, BA). After each pair a judgement is made which element in a pair is preferred in the context of the test scenario.
- **Threshold definition methods** are in general for identification of subjective tolerance, acceptance, or for better/worse quality level. The judgement is based on simple binary question related to quality level.

The difference between such methods is in utilizing explicit references and methods that do not use any explicit reference. The non-reference methods 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 DCR method [48] has long been a key method for the assessment of television pictures whose typical quality represents extremely high quality levels of video telephony and videoconferencing. The specific comments of the DCR scale (imperceptible/perceptible) are valuable when the viewers' 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.

The threshold methods are very suitable for setting of tolerance and acceptance thresholds. This method is suitable only for services with well defined usage and service scenarios.

When a large number of items are 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.

Standards for subjective assessment methodology (incl. environment, etc)

- **ITU-R Recommendation BT.500 [46]:**

  Recommends subjective assessment methods to establish the performance of television systems using measurements that anticipate more directly the reactions of those who might view the systems tested. In this regard, it is understood that it may not be possible to fully characterize system performance by objective means; consequently, it is necessary to supplement objective measurements with subjective measurements described in this recommendation.

- **ITU-T Recommendation P.930 [47]:**

  This Recommendation describes the principles of an adjustable video reference system that can be used to generate the reference conditions necessary to characterize the subjective picture quality of video produced by compressed digital video systems. A Reference Impairment System for Video (RISV) can be utilized to simulate the impairments resulting from the compression of video sequences, independent of compression scheme. The subjective evaluation methods are described in Recommendation P.910.

- **ITU-T Recommendation P.910 [53]**

  This Recommendation describes non-interactive subjective assessment methods for evaluating the one-way overall video quality for multimedia applications such as videoconferencing, storage and retrieval applications, tele-medical applications, etc.

- **ITU-R Recommendation BS.1534-1[54]**

  Recommend methods for the subjective assessment of intermediate quality level of audio coding systems.

2.3.2 Automated assessment

Automated assessment allows fast computation or estimation of video quality. Automated assessment can be divided into methods that estimate the subjective quality and methods that measure the objective quality of video sequences. Based on these differentiations, these methods will now be described.

Estimation of subjective video quality

Estimation of subjective quality reduces the time consumption of user based assessment and at the same time broadens the possibilities of the quality estimation deployment. Estimating subjective video quality relies on one of following two approaches. The first approach is based on modelling of human visual system. The second approach is linking together a set of objective parameters with subjective quality.

Metrics based on models of the Human Visual System (HVS)

Metrics based on models of the Human Visual System (HVS) were proposed in recent years [10], [30], [31], [32], [33], [34]. 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. The HVS metrics in close future will reflect the recent neuroscience findings and psychophysical experiments [50], [51], [52] which established that there is interaction across the visual channels and that such interactions are important for visual masking [13]. The main disadvantage of these HVS models is their high computational complexity.

Metrics based on a set of objective parameters

Metrics based on a set of objective parameters [14], [35], [39], [40], [43] 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.

- **Reference-based metrics** [35], [36]: 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.

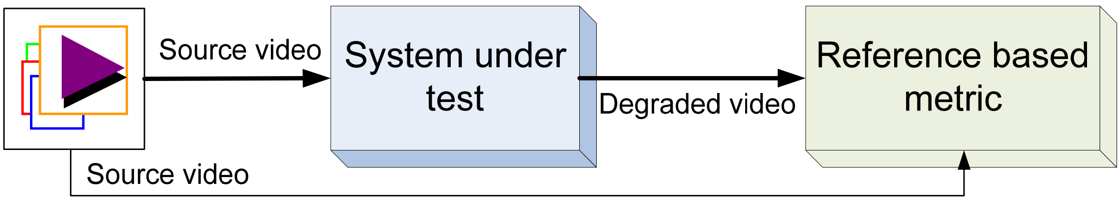

Figure 8: Reference-based metrics

The best known reference-based metric is Video Quality Metric (VQM) which also became an ANSI T1.801.03 standard [35]. The standard VQM 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. 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 sub-regions. Finally, a perceptual threshold is applied to the extracted features. This method is widely use for quality estimation of diverse video services. Recently performance evaluations on HD video stream were published [28] which show a good correlation between subjective and estimated MOS results.

Structural Similarity (SSIM) index [13] is a method for measuring the similarity between two images. The SSIM index can be viewed as a quality measure of one of the images being compared provided the other image is regarded as of perfect quality. The SSIM index is designed for image quality measures, based on the assumption that the human visual system is highly adapted to extract structural information from the viewing field. It follows that a measure of structural information change can provide a good approximation to perceived image distortion.

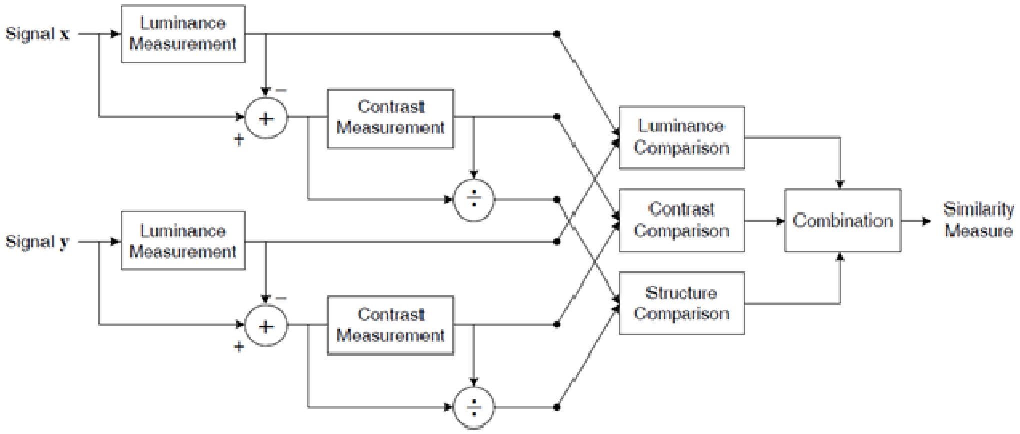

Figure 9: Diagram of the structural similarity (SSIM) measurement system [13].

- **Semi-reference-based metrics**: measurements obtained by computing a set of parameters on the degraded video sequence and comparing them with the same parameters computed on the reference picture [30], [31], [34], [40]. 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.

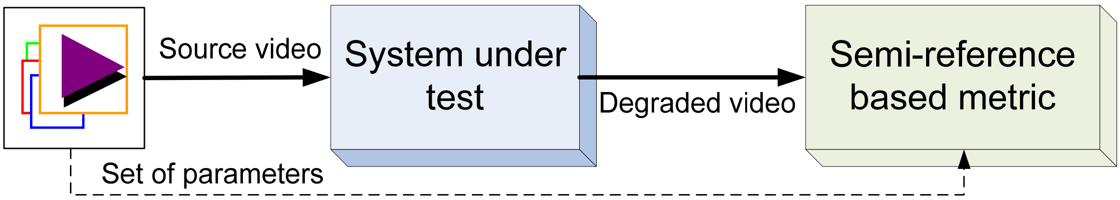

Figure 10: Semi-reference-based metrics

- **Reference-free metrics** [14], [38], [39], [48], [49]: they do not require any knowledge of the original video source. The video quality estimation is based on reference-free measures calculated from distorted video. Moreover, the reference-free measures reduce complexity and at the same time broaden the possibilities of the quality prediction deployment. 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.

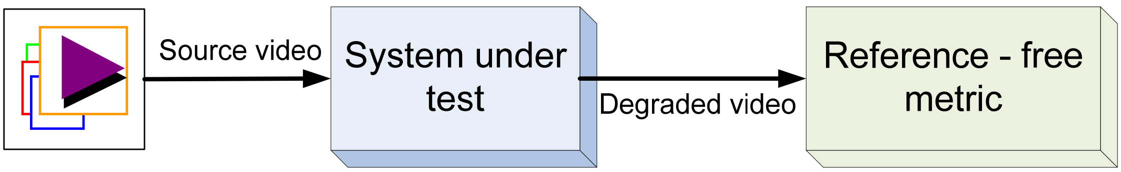

Figure 11: Reference-free metrics

In case of a well-defined usage and service scenario, the video quality estimation can be based on set of key performance indicators (KPI) [25]. These KPIs are objective parameters that have a strong correlation with QoE measures. The set of KPIs usually consists of video, application and network parameters, which are highly specific to the system under investigation.

Objective video quality measurement

The objective video quality represents the physical meaning of quality distortion. This measurement was introduced for performance evaluation of video codec and video transmission system.

The quality distortion measures are Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR). Both of them poorly reflect the subjective video quality [13], [14], [27], [28], [29]. Nevertheless, the PSNR is still widely used as reference method for comparing performance of video coding algorithms.

MSE and PSNR

The simplest and most widely used full-reference quality metric is the mean squared error (MSE), computed by averaging the squared intensity differences of distorted and reference image pixels, along with the related quantity of PSNR. These are appealing because they are simple to calculate, have clear physical meanings, and are mathematically convenient in the context of optimization. PSNR is an objective video quality measure for the ratio between the maximum possible power of a signal and the power of corrupting pixels that affects the fidelity of its representation. PSNR is usually expressed in terms of the logarithmic decibel scale. The PSNR is most commonly used as a measure of quality of reconstruction of lossy compression codecs (e.g., for image compression). The signal in this case is the original data, and the noise is the error introduced by compression. PSNR is most easily defined via the mean squared error (MSE) which for two $m \times n$ monochrome images I and K where one of the images is considered an erroneous approximation of the other is defined as:

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

Where the PSNR is defined as:

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

Here, MAX_I is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255.

2.3.3 Measurement methods for Full HD Television

The most recent and relevant investigations have been performed by Stephen Wolf and Margaret H. Pinson in National Telecommunications and Information Administration [28] and at the University of Zagreb [29]. They investigated performance of H.264 in compression system that uses progressive and interlaced HDTV formats. It is often claimed that progressive video is compressed more efficiently than interlaced video. Therefore, their evaluation was performed on resolution formats 1080i and 1080p using different compression ratios achieved by H.264. compression algorithm.

The subjective assessment was performed according ITU-R Recommendation BT.500 [46] and video quality was estimated according following metrics:

- PSNR,
- SSIM index [13],
- VQM [35].

Subjective results of different HDTV formats in these studies show that progressive scanning should be considered rather than interlaced for all future HDTV emissions (1080 lines). For the same type of video sequence, PSNR and VQM measures yielded reasonable good correlation with subjective testing. On the contrary, SSIM index and VQM the provide better correlations in case of video sequence combined from more different content types, which is more realistic scenario. These results allow us to conclude that SSIM and VQM are the most suitable metrics for estimation of full HD video quality.

3. Definition of QoE Test Strategy

3.1 Goals and Focus of interest

Deliverable 2.1 follows the actual state of the art in the field of QoE measurement for HD IPTV services. The links between SOTA and proposed methodology are shown in Table 3. Furthermore, Table 3 shows our focus on the different variables and deliverable targets.

Influence factor		Treatment within OptiBand	SOTA described in section	Methodology described in section	To be specified by
Context	Personal characteristics	Control	2.2.1	3.3.1	D1.1
	Environment	Constant	2.2.1	3.3.2	D1.1 / FTW
	Content	Vary	2.2.1	3.3.3	D1.1 / FTW
	Social and Cultural	Constant/control	2.2.1	3.3.2	FTW
Expectations	Application type	Control	2.2.2	3.3.2	FTW
	Image and brand	Constant	2.2.2	3.3.2	FTW
	Usage history	Control	2.2.2	3.3.2	FTW
Technical system	Codec	Constant	2.2.3	3.3.2	Agreed: H.264
	End user devices	Constant	2.2.3	3.3.2	According to standard
	Video delivery	Constant	2.2.3	3.3.2	D1.1
	Packet dropping algorithm	Vary (main independent variable)	2.2.3	3.3.3	According to the partners' algorithm proposal
	Service type	Constant	2.2.3	3.3.2	D1.1/FTW

Table 3: Goals and focus of interest.

Column 2 indicates how the different influence factors are treated within the OptiBand QoE test design. There are the following options:

- Vary: The influence factors that are of main interest within this project are systematically varied, i.e. experimental conditions are dedicated to representative characteristics of this factor (e.g., certain packet dropping algorithms developed by the OptiBand partners)
- Control: Influence factors, such as age and gender, which are expected to have a mediating impact on the QoE, will be controlled. For example, in the sample specification, important variables like age and gender will be distributed according to the target population. This will help to achieve representative results, and by splitting the sample, possible effects on QoE can be retraced.
- Constant: Some influence factors are kept constant, either due to the focus of the project (e.g., architecture and codec settings), or due to standard recommendations (testing environment).

3.2 Overall methodological approach

Three iterations

The QoE will be evaluated within three iteration steps (see Figure 12). Iterative design and evaluation is one of the most fundamental principles of user-centred development. While the first evaluation iterations are formative, i.e. for improving the prototype solution, the last iteration provides a summative evaluation, i.e. assessing the solutions against the specified criteria and benchmarks.

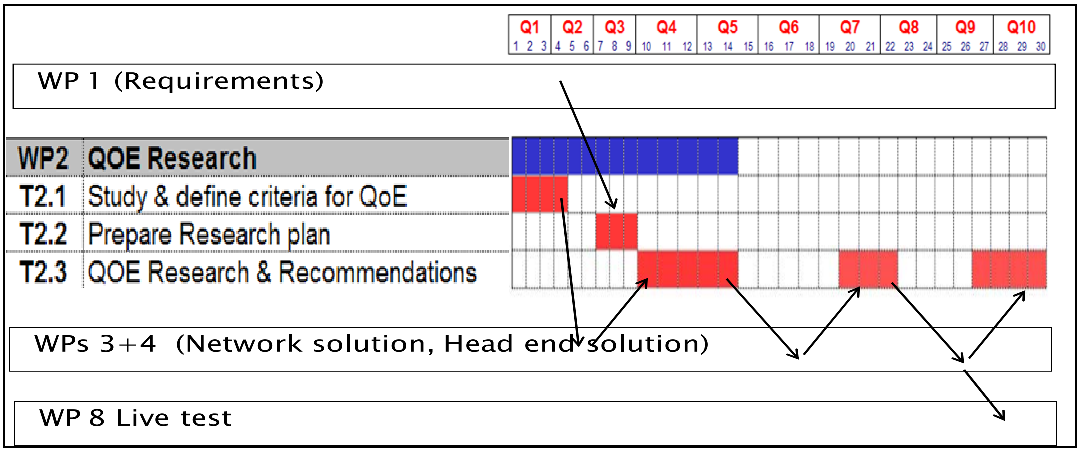

Figure 12: Timeline from OptiBand Annex I: 3 test iterations

In accordance with Annex I, the evaluation steps are as follows:

- **Step 1 - evaluation of simulation** (months 10-12)**:** In order to select suitable packet dropping strategies in WP3, the perceptual quality of a certain video sequence, based on an early version of the algorithm (the video sequences will be produced in seven bandwidth levels) will be compared to the original reference video sequence, which is the reference video clip before the application of the data drop algorithm. Based on the results, recommendations will be made, to guide the further algorithm development process in T3.3.
- **Step 2 - evaluation of prototype** (months 20-22)**:** In a subsequent user study, the selected algorithm will be evaluated again (the video sequences will be produced in seven bandwidth levels). Recommendations for adapting the packet dropping algorithm will be given to guide the prototypes implementation for the network and head end solution (T3.4).
- **Step 3 - final evaluation** (months 27-30): This test follows the same method of steps 1 and 2, in order to assess the final achievements of the OptiBand algorithm development. These tests run in parallel with the live test, which is defined in WP8. The procedure of step 3 will support and run in co-ordination with the live test.

Combination of user-based and automated assessment

In each of the three iterations, we will conduct both user-based and automated assessment. A combined assessment allows a more appropriate analysis of overall quality and improvement potentials, because the mutual correlations between subjective evaluations and video quality measures can be analyzed.

To allow for comparative progress assessment, the methods applied will not change throughout the three iterations (the only exception is that the user-based acceptance threshold test will only be conducted in iteration 1).

In the following, the methods used for combined and automated assessment will be specified in detail.

3.3 User-based assessment

This section describes the methodology for user-based assessment of full HD IPTV video services. Section 3.3.1 specifies the test persons' profile. Section 3.3.2 then describes the overall conditions, which are invariant throughout the QoE tests (room conditions, technology settings and service type, as well as the end-user device and decoder). Section 3.3.3 specifies the varied test conditions for the factors of interest for systematic research (packet dropping algorithms, bandwidth levels, content class, and reference sources), as well as an overview of video materials that are needed to test the necessary factor combinations.
Section 3.3.4 and 3.3.5 describe the two user-based assessment types that come into use in the OptiBand project: the quality rating test, and the acceptance threshold test.

3.3.1 Test persons profile

The recommendable number of subjects in a viewing test ranges from 15 to 40 [46], [53]. Fifteen is the absolute minimum for statistical reasons [46], while there is rarely any point in going beyond 40. The test subjects should not be directly involved in video or picture quality evaluation as part of their work and should not be experienced assessors. Nevertheless, in the training phase of our experiment we will perform the evaluation with small groups of experts (3-5) or other critical subjects in order to obtain indicative results.

The test subject will be briefly screened for normal visual acuity. The chosen group can range in different age, gender, education and experience with image processing. Two runs should be taken. In order to avoid a learning effect it is important to make a break of half an hour between the first and the second run.

Furthermore the test participants should be questioned about average viewing time. The average viewing time of the Austrian test sample should be approximately 2,5 hours per day and the average viewing time of the Italian test sample should be approximately 4 hours.

A more detailed test persons profile will be provided in D2.2.

3.3.2 Overall test conditions

Room conditions

Our test laboratory emulates the home conditions. Therefore, we follow general viewing conditions for subjective assessments in home environment defined by ITU-R BT.500-11[46] (see Table 4).

Ratio of luminance of inactive screen to peak luminance:	≤ 0.02
Display brightness and contrast:	see Recommendations ITU-R BT.818 and ITU-R BT.815
Monitor resolution:	1920 x 1080 pixels
Peak luminance:	> 200 cd/m2
Environmental illuminance on the screen (Incident light from the environment falling on the screen, should be measured perpendicularly to the screen):	< 200 lux

Table 4: Viewing conditions for subjective assessments in home environment.

The viewing distance and the screen sizes are to be selected in order to satisfy the Figure 13 below [46].

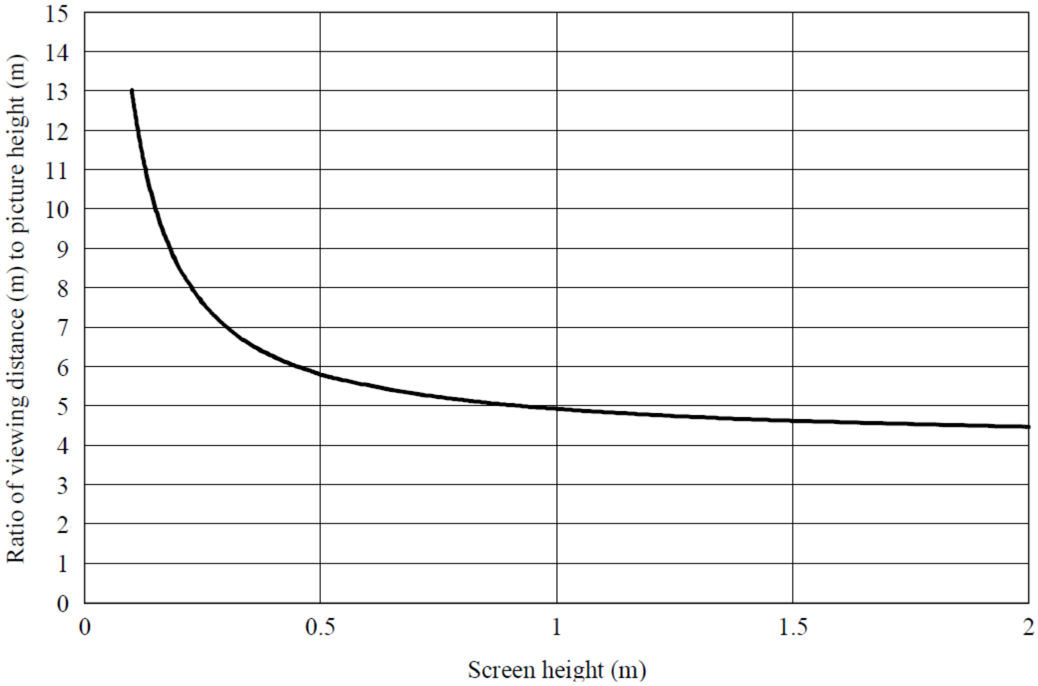

Figure 13 Function of viewing distance and the screen sizes [46]

FTWs iLab fulfils the above defined viewing conditions for perceived quality and usability testing. Our iLab (see Figure 14) is equipped with user testing equipment of TV services and is compliant with ITU-R BT.500-11 [46]. The viewing distance is approximately 4 m.

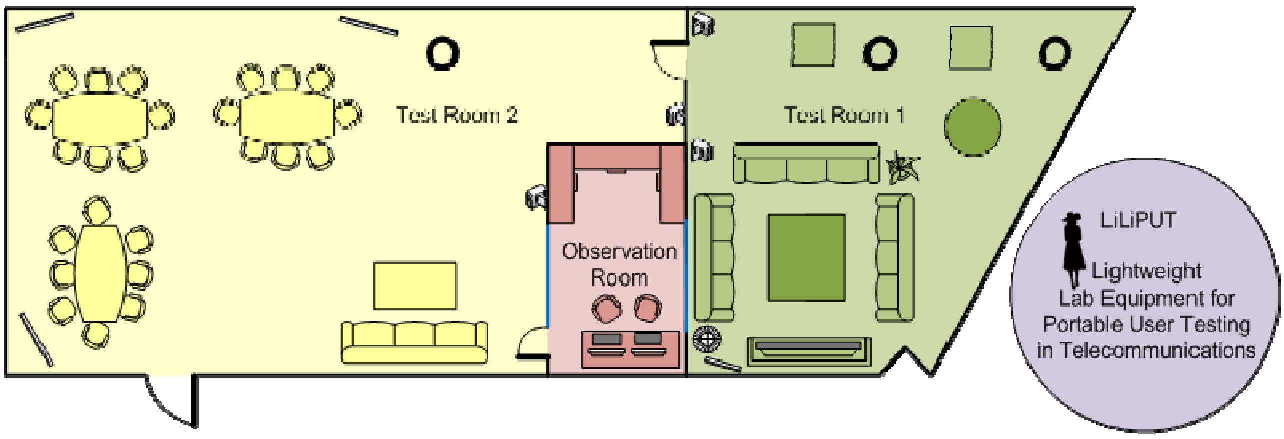

Figure 14: iLab

Service type

The technology setting is based on an emulation of full HD IPTV service in the first two steps. The full HD IPTV service is emulated because the performance of packet dropping algorithms should be evaluated before its implementation. The emulation allows us to play the sequences exactly in the same way as the proposed system solutions of the OptiBand project. This solution has no impact on perception of video sequences. Moreover, the test subjects are not aware of emulation.

Finally, the third step of subjective assessment is performed on functional prototypes. The complete OptiBand architecture will be evaluated in the third step.

End-user device

The display for subjective testing is chosen according to the recent trends on TV flat panels market (see section 2.2.3). This makes the most representative TV set with following parameters:

- Technology: LCD
- Screen size: 40” – 42”
- Resolution: Full HD (1080 x 1920)
- Decoder type: H.264

A device, which meets these requirements, is the Sony KDL-40EX500. It has the following technical specification:

- Display technology: LCD
- Screen size: 40”
- Resolution: Full HD (1080 x 1920)
- Aspect ratio: 16:9
- HDMI input: 4
- Decoder type: H.264

Codec

One of the main components of QoE is the compression of video source materials and the various settings and parameters selected. In the source video materials in our test setup are encoded with H.264 encoder. The encoder setting (see Table 5) is tuned for full HD IPTV service according the ITU-T - Recommendation G.1080 [3] and experiences of Telecom Italia. However, some other settings may be used (as will be specified in D4.1), if more suitable for some packet dropping algorithms.

Parameter	**Setting**
Format Profiles	H.264 High H.264 Scalable Baseline H.264 Scalable High
Level	4.1 (4.2 for 1080p optional)
Video bitrate	7 ÷ 8.5 Mbps (14 Mbps for 1080p)
Video format	1080i, as specified in D1.1
Frame rate	25 fps
Aspect ratio	16:9
GOP length	32
GOP structure	IBBBP
Video mode	CBR

Table 5: Encoding parameters applied to the test video sequences

3.3.3 Varied test conditions

Unlike the stable conditions mentioned above, the main factors of interest will be systematically varied, in order to understand their impact on QoE. These variations of test conditions are explained in the following, notably the type of reference sources, the packet dropping algorithms, the bandwidth levels, and the content classes.

Packet dropping algorithms

There will be three methods of packet dropping tested within each test iteration. Each of the three algorithm development partners will be responsible for developing one of these. The exact specification of the packet dropping algorithms tested will be provided in D3.1.

Bandwidth levels

Due to its importance for the project, bandwidth will also be varied systematically. The bandwidth is defined as average bandwidth including all necessary transport data for correct decoding of the video stream (i.e. video payload, signalization, encapsulation format ....). There will be seven bandwidth levels (see Table 6).

The precise bandwidth level specification is highly dependent on the specified video format and mode. For the OptiBand project (see the agreed video format and mode settings in Table 5), the following bandwidth levels have been specified:

Bandwidth level	Relative bandwidth reduction, as compared to the state-of-the art (SOTA):	Bandwidth level (MBbps)		
		min	centre	max
SOTA	0%	7,5	7,8	8,5
7.	10%	6,9	7,0	7,1
6.	20%	6,1	6,2	6,3
5.	**33% - OptiBand project target**	5,1	5,2	5,3
4.	40%	4,6	4,7	4,8
3.	45%	4,2	4,3	4,4
2.	50%	3,8	3,9	4,0
1.	55%	3,4	3,5	3,6

Table 6: Bandwidth level conditions specified for the OptiBand QoE evaluation

With these seven different test conditions, we will be able to investigate the QoE according to systematically progressing bandwidth reduction levels. The rationale for choosing these bandwidth levels is to enable the assessment whether the project goal of 33% bandwidth reduction can be partly achieved, fully achieved, or even exceeded.

Content class

In order to select the most representative content classes (CCs) Telecom Italia performed own market research and the available popularity ratings were investigated (see section 2.2.1). Therefore, were selected for test setup following CCs:

- Soccer
- Action movie

Figure 15: Snapshots of Soccer (left) and Action movie (right) sequences.

These two sequences represent CCs with different amount of details and complexity of structures and movements. Moreover, Figure 4 shows the subjective judgments of the video sequences depending on the video content and the type of degradation. Furthermore, it can be seen that CCs Soccer, Action movie, and Video clip are the most sensitive sequences [41].
Each of three packet dropping algorithms (proposed by three partners in the project) is evaluated with seven different bandwidth levels on two CCs. Therefore, each CC will be tested approximately under 21 different settings (3 packet dropping algorithms x 7 bandwidth levels). Together, 42 test sequences (21 different settings x 2 CC) will be evaluated. Given this setup, extending the number of CCs tested would lead to a combinatorial explosion of test conditions.
The selection of these two content classes reflects the sport and film genre dominance within the available HD channels in Europe (see Figure 5). For these two content classes also HD resolution provides the most significant benefit from the point of view of perceived quality.

Reference sources

As agreed in the project consortium, packet dropping will be applied to H.264 encoded video streams. In order to compare the results, two sorts of reference material can be thought of, in principle:

a. Unique original
b. H.264 encoded

In the OptiBand QoE evaluations (user-based and automated) we will only take H.264 encoded materials as a reference, not the unique original. This is because we primarily need to compare the new packet dropping algorithm in comparison with the current state of the art, which is H.264. Today's HD TV users are not any more exposed to unique original video material, therefore the inclusion would not add ecological validity.

Furthermore, as will be described in section 3.4, H.264 and H.264 plus packet dropping will also be used in the automated assessment, because the results can be correlated with the user-based assessment, and also because a more transparent comparison of packet dropping algorithms with the H.264 is possible.

Required video materials

Figure 16 provides an overview of the 42 needed video sequences for each of the three test iterations within the project. This number of video sequences results from the necessary combinations between the abovementioned factors:

- 3 packet dropping algorithms (PDA), i.e., from the three algorithm development partners, respectively (specified in D3.1).
- 7 bandwidth levels (cf. the specification above)
- 2 content classes (specified in section 3.3.3)

Additionally, there will be 2 reference videos (1 per content class).

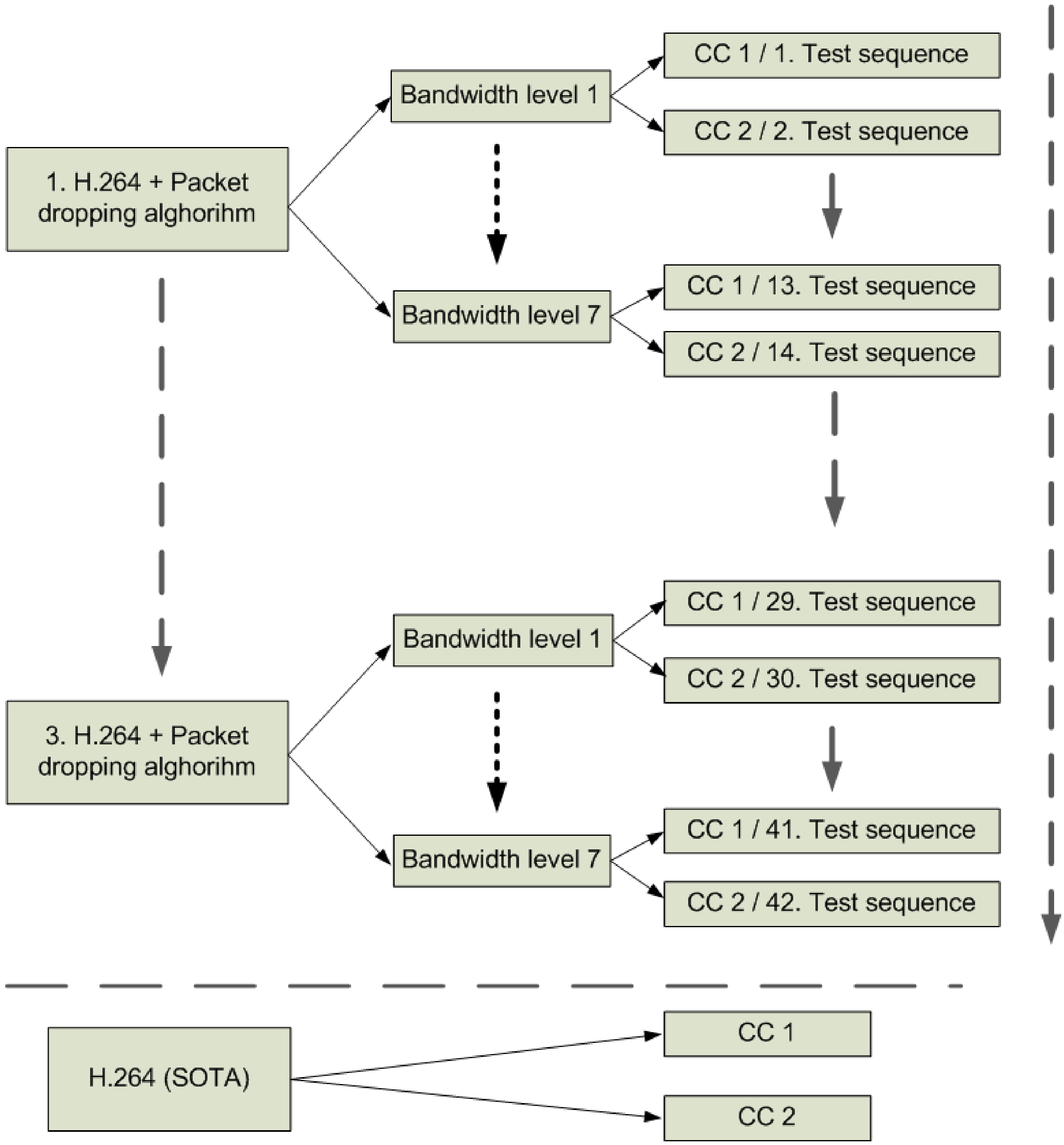

Figure 16: Video materials needed for each test iteration

In sections 3.3.4 and 3.3.5, the procedure for the two test types conducted in the user-based assessment is described: quality rating test and acceptance threshold test.

3.3.4 Quality rating test

In this section, the quality rating test is described. The general approach follows the Absolute Category Rating (ACR) method. The ACR method is a category judgement where test sequences are presented one at a time and are rated independently on a category scale. After each presentation the subjects are asked to evaluate the quality of the sequence shown.

The methodology description is structured into the briefing, experimental and debriefing phases.

Briefing phase

- The description of the test scenario and test method is given in written form to the test subjects. In addition the five grade opinion scale is explained.
- TeleCOMMANDER voting application is demonstrated (see Figure 17:).

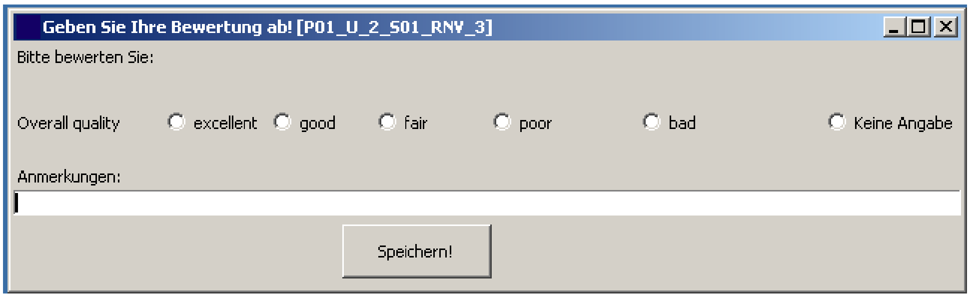

Figure 17: TeleCOMMANDER

- Furthermore, the test participant is asked for basic statistical data, such as the following:
    - Age
    - Gender
    - Type of used TV services in the living room (e.g. IPTV, Satellite, DVB-T, Internet TV ...). For this question, multiple answers are accepted.

Experimental phase

- The subjective assessment consist of two test rounds:
    - Two test rounds are performed (with reverse order of video materials), in order to check consistency of obtained results. The duration of one round is approximately 45 minutes.
    - Between the test rounds is a 30 minutes break.
- Trial sequences:
    - At the beginning of each test round a trial run is presented with three sequences
    - Subjective quality of these trial sequences varies substantially
    - Sequences are similar to tested video sequences

Figure 18: the time pattern for the presentation.

- After each sequence, the user is asked to rate the subjective quality:
    - Rating scale: 1-bad, 2-poor, 3-fair, 4-good, 5-excellent
    - Voting time: 10 seconds
    - This is done with the TeleCOMMANDER, an electronic questionnaire that has been developed at FTW (relevant technical features of TeleCOMMANDER will be presented in D2.2).
- The tested sequences:
    - The sequences are approximately of 10 seconds length, in order to keep scene content integrity.
    - There will be three clips per algorithm/content class/bandwidth combination. Thus, a user is not presented with the same sequence during the 42 test conditions, but with a different sequence for each of the three algorithms (the detailed test design will presented in D2.2). Combinations of clip, algorithm and bandwidth are varied throughout the test users.

- o Furthermore, the sequences represent different bandwidth levels, in order to exploit the full MOS scale.
- o Test sequences are presented according to the following rules:
  - two clips of the same CC must not appear in succession
  - two clips of the same impairment level must not appear in succession
- o Content (see previous section 3.3.3):
  - Soccer
  - Action movie
- At the subjective assessment 26 test subjects participate (see section 3.3.1).

Debriefing phase

At the end of the quality assessment, the test participants are interviewed regarding their overall impressions. The concrete interview guidelines will be presented in D2.2.

3.3.5 Acceptance threshold test

While the above described quality rating tests provide fine-grained MOS results, it is not possible to validly answer the most fundamental question of QoE with this form of test: “Is the quality acceptable, and when does it get unacceptable?” Unlike in other domains, such as VoIP telephony or mobile TV, this acceptance threshold has yet been defined. To address this gap, we will explicitly identify this acceptance threshold in dedicated test runs. This will then enable a more valid interpretation of the quality rating test results described above.

In the following, we describe the test procedure of this acceptance threshold test.

Briefing phase

The briefing phase is similar to the quality rating test.

Experimental phase

The same video material compositions and successions as in the quality rating study are used.

There are two blocks with two test rounds each, in which the participants first watch the clip and then provide a voting via the TeleCOMMANDER interface. In the first block, the first test round is to find out the acceptance threshold, which consists of a binary choice, i.e. whether the quality is acceptable or not. In the second round, the users provide a rating on a MOS scale, in the same way as described in section 3.3.4.

The second block is a repetition of the first block, with reversed order of video materials.

Debriefing phase

Similarly to the quality rating study, the acceptance threshold study will include a short final interview, in which the general impressions regarding the test are gathered. The detailed interview guidelines will be defined in D2.2.

3.4 Automated assessment

This section specifies the automated QoE methods used in the OptiBand project to complement user-based assessment. The evaluated video material will be the same as in the user-based assessment (see section 3.3.3).

Automated evaluation of the investigated video sequences is performed with a video quality measurement and estimation tool, which has been custom-developed by FTW for the OptiBand project. As will be described in the following, the main features of this tool are PSNR and SSIM index calculation and several results visualisations.

3.4.1 Estimation of subjective video quality (SSIM)

The SSIM index (see Section 2.3.2) is calculated between H.264 encoded and H.264 plus packet dropping (see Section 3.3.3). The SSIM index is computed using three image measurement comparisons: luminance, contrast and structure. Each of these measures is calculated over the 8×8 local square window moved pixel-by-pixel over the entire image. At each step, the local statistics and SSIM index are calculated within the local window. Because resulting SSIM index map often exhibits undesirable "blocking" artifacts, each window is filtered with a Gaussian weighting function (11×11 pixels). In practice, one usually requires a single overall quality measure of the entire image, so mean of SSIM index map is computed to evaluate the overall frame quality. SSIM values are calculated for single frames and average SSIM over time window.

3.4.2 Objective video quality measurement (PSNR)

The PSNR (see Section 2.3.2) is calculated as a difference measure between H.264 encoded and H.264 plus packet dropping (see Section 3.3.3).The calculation is performed for the gray scale (or luma component) colour space. This procedure reduces processing complexity by a factor of three and eliminates erroneous PSNR calculation due to different colour spaces of original and encoded sequences (which is typically a calculation error). Error-free frames (identical investigated and reference sequence), are clipped up to a maximum value of 111.30 dB for 1080 video mode and 107.78dB for 720 video mode. Such clipping has been used, because we need to avoid infinite PSNR values resulting for zero MSE. This PSNR clipping value corresponds to one error in one colour in one frame [59]. Finally, PSNR values are calculated for single frames and average PSNR over a specified time window.

3.4.3 Temporal visualization

The PSNR and SSIM index values are calculated for single frames of the investigated sequences. This allows us further visualization of PSNR and SSIM index results in the temporal domain (see Figure 19), comparable to empirical histograms. Results visualization allows quick visual performance inspection of a certain packet dropping algorithm in an investigated video sequence.

The example depicted in Figure 19 shows the PSNR and SSIM results of a distorted video along the time axis. The high values at 111,298 dB and SSIM index at 1 represent the clipped values for the error free video frames. The lower PSNR and SSIM refer distorted video frames.

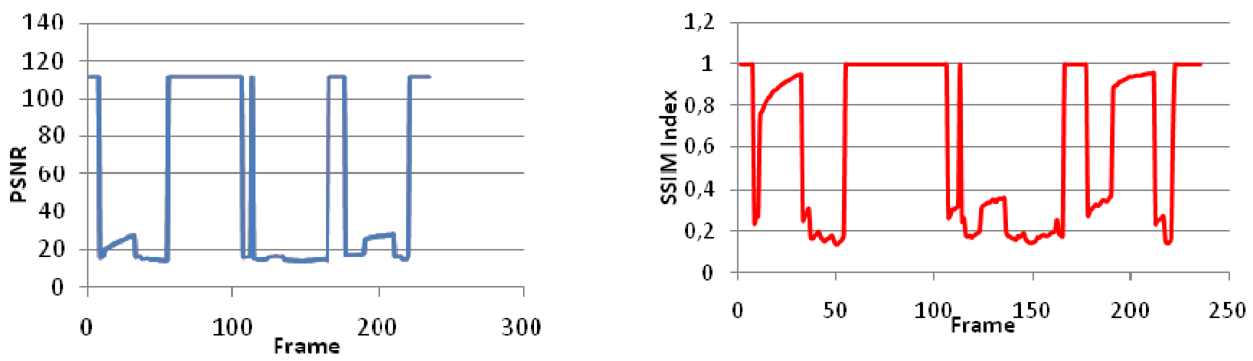

Figure 19: PSNR and SSIM index in temporal domain. The related video can be found at: http://userver.ftw.at/~froehlich/optiband/riverbed_upravene.rar

3.4.4 Visualization of results distribution

The PSNR and SSIM values for single frames of investigated sequences can be visualized like empirical probability density functions (PDF) in histogram format (see Figure 20 and Figure 21). This visualization allows us to see the distribution of impaired and error free frames. Furthermore, such a visualization allows a quick estimation (and calculation) of the ratio of frames below and above a given threshold.

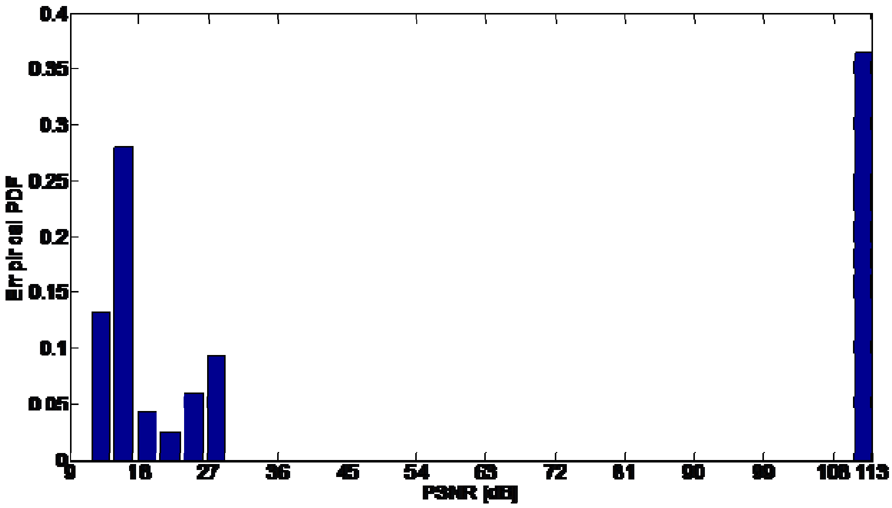

Figure 20: The empirical PDF for PSNR

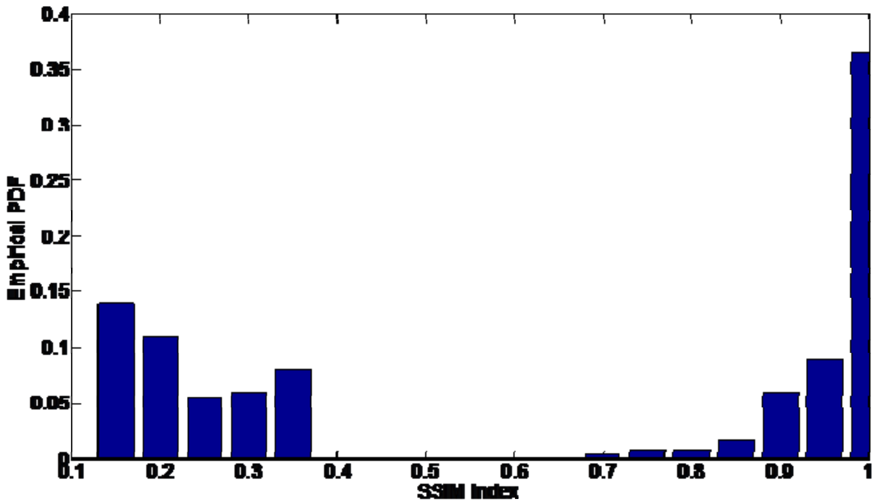

Figure 21: The empirical PDF for SSIM index

3.4.5 Mapping of KPI in the live test

The obtained data from user based and automated assessment allow us to define KPIs for estimation of acceptance thresholds. The KPI will be selected according to applicability for live measurement and correlation with user based assessment. The set of KPI will consist of encoding, network and packet dropping parameters.

3.5 Summary

Iteration step	System type tested	User based assessment	Automated assessment
1	Emulation (WP2)	Quality ratings + Acceptance threshold test	PSNR, SSIM index computation
2	Emulation (WP2)	Quality ratings + Acceptance threshold test	PSNR, SSIM index computation
3	Emulation (WP2)	Quality ratings + Acceptance threshold test	PSNR, SSIM index computation
	Live system (WP8)	Test procedure defined according to D8.1	Planned: threshold estimation based on KPI (finally specified within D8.1)

Table 7: Overview of assessment methods for each iteration step.

4. Specification of acceptance thresholds, quality criteria and benchmarks

This section describes the principles on how to interpret the results of the QoE research performed within the OptiBand project. Consistent specification of criteria, minimum quality thresholds, and benchmarking procedures are necessary to enable valid and meaningful interpretation of results throughout the three development iterations.

In section 4.1, the strategy for deriving the criteria for QoE evaluation is described, and section 4.2 specifies the quantitative minimum threshold values for each of these criteria. Section 4.3 describes the approach for benchmarking the investigated solutions, which have surpassed the minimum thresholds, to provide recommendations for the optimal packet dropping approach.

4.1 QoE criteria specification

Criteria for QoE testing are derived by the user-based and automated assessment methods, as described in the previous section. The main principles for this are summarized in the following.

4.1.1 User-based assessment

The specification of the quality threshold is based on acceptance threshold, as derived in the acceptance threshold test (compare section 3.3.5).

The results of acceptance thresholds provide links between acceptable video quality and the mean opinion score (MOS) and also with results of automated assessment (see below). The results of the acceptance threshold test allow us to define acceptable video quality for subjective and objective video quality measures.

Furthermore, for more fine-grained quality assessment, such as benchmarking, the mean-opinion score (MOS) results from the video quality rating test (see section 3.3.4) will be used.

4.1.2 Automated assessment

The methods for estimating acceptance threshold2 should be defined in more ways (see section 3.4), according to the applicability:

1. Mapping of the objective video quality parameter **PSNR** on the acceptance threshold provides a direct link between physical and subjective measures. The PSNR is used for acceptance threshold estimation and packet dropping algorithm benchmarking (see Section 3.4.2).
2. **The SSIM Index** provides the most accurate **method** for estimating the acceptance threshold [29]. The SSIM index is used for acceptance threshold estimation and packet dropping algorithm benchmarking (see section 3.4.1).
3. **Mapping of KPIs** (network, application and packet dropping algorithm parameters) to acceptance threshold. KPIs will only be used for acceptance threshold estimation (see Section 3.4.5).

4.2 Minimum QoE thresholds

In the following, the minimum QoE thresholds are specified for the user-based, and then for the automated assessment criteria.

4.2.1 User-based assessment:

The minimum acceptance threshold for current mobile video services is defined as 3,7 MOS [57]. Therefore, in OptiBand we specify 3,7 MOS as a threshold (as an initial value) for the reference video sequence as well as for the video sequence with the implemented packet dropping algorithm. The definite threshold for HD IPTV services will be set according to the results of the acceptance threshold test (see section 3.3.5.).

4.2.2 Video quality estimation:

- Objective video quality measures:
  The benchmark measure for objective video quality is PSNR. The initial PSNR acceptance threshold value for a single frame in our specific testing scenario has been defined as 27 dB. Furthermore, the PSNR value is set for averaged frames over a certain time window at 30 dB. The duration of the time window will be set according to the design of packet dropping algorithms in range between 7 and 30 seconds. This initial value was set based on a preliminary video quality test with a small group of video and image processing experts.
  The value will be determined more accurately after the 1st and 2nd iteration of the QoE evaluation (see Figure 12).
- Subjective video quality measures:
  The benchmark for subjective video quality is the SSIM index. The initial value for the acceptance threshold expressed in objective video quality measure is **SSIM = 0,8**. The initial value was set according to a preliminary video quality test with small group of experts on video and image processing.
  This value will be determined more accurately after the 1st and the 2nd iteration step of QoE evaluation (see Figure 12).
- Key performance indicators:
  The selection of KPI parameters with most significant impact on QoE (e.g., bitrate, packet error rate) will be done after the first iteration of QoE evaluation. After the 2nd iteration the KPI values for acceptance threshold will be settled. Finally, the fine tuning of KPI values will be done after the third iteration step of QoE evaluation.
  Please note that for selection of KPIs it is necessary to obtain video sequences with applied packed dropping algorithms in different quality levels with full setting description of packed dropping algorithm.

4.3 QoE Benchmarking

This section explains the principles for benchmarking.

4.3.1 Prioritization of criteria

For benchmarking, the criteria are prioritized as follows:

1. User-based assessment
   1.1. MOS
2. Automated assessment
   2.1. SSIM
   2.2. PSNR

4.3.2 Ranking procedure

The proposed algorithms will be ranked according to the following procedure:

1. Achieved 33 % bandwidth reduction in average with keeping minimal QoE thresholds (see section 4.2).

Algorithm	33 % bandwidth reduction in average with keeping minimal QoE thresholds [yes/no]
Algorithm 1	

Algorithm 2	
Algorithm 3	

For example, 33% bandwidth reduction reflects 5 Mbps reduction of the bandwidth on 15 Mbps HD video stream after introduction of the packet dropping algorithms.

2. In the second step only packet dropping algorithms are evaluated, which fulfil the condition described in the previous point (33 % bandwidth reduction in average with keeping minimal QoE thresholds). The best performance of proposed packet dropping algorithms for each tested encoding setting will be presented as follows:

   a. User based assessment:

Rank	Algorithm	CC1[MOS]	CC2 [MOS]	Average [MOS]
1.				
2.				
3.				

   b. Automated assessment:
      i. SSIM index results:

Rank	Algorithm	CC1[SSIM]	CC2 [SSIM]	Average [SSIM]
1.				
2.				
3.				

      ii. PSNR results:

Rank	Algorithm	CC1[PSNR]	CC2 [PSNR]	Average [PSNR]
1.				
2.				
3.				

5. Bibliography

[1] ITU-T, ”Recommendation E.800: QoS Terms and Definitions related to Quality of Service and Network Performance including dependability,” International Telecommunication Union, Aug. 1994.
[2] P. Reichl, J. Fabini, and C. Happenhofer, M.and Egger, “From QoS to QoX: A Charging Perspective,” in Proceedings of the 18th ITC Specialist Seminar on Quality of Experience. Blekinge: Blekinge Insitute of Technology, pp. 35 –44, May 2008.
[3] ITU-T, ”Recommendation G.1080: Quality of experience requirements for IPTV services,” International Telecommunication Union, Dec. 2008.
[4] ITU-T, ”Recommendation G.1081: Multimedia Quality of Service and performance – Generic and user-related aspects, ”International Telecommunication Union, Oct. 2008.
[5] Y. Koucheryavy, G. Giambene, D. Staehle, F. Barcelo-Arroyo, T. Braun, V. Siris, ”Traffic and QoS Management in Wireless Multimedia Networks,” Springer US, 2008.
[6] 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.
[7] 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.
[8] T. M. O’ Neil, ”Quality of Experience and Quality of Service for IP Video Conferencing”, Polycom, Whitepaper, 2002.
[9] J. Goodchild, ”Integrating data, voice and video - Part II, IP Video Implementation and planning guide, ” United States Telecom Association, 2005.
[10] 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.
[11] T. Deryckere, W. Joseph, L. Martens, L. De Marez, K. De Moor, “A software tool to relate technical performance to user experience in a mobile context,” in Proc. WOMWOM, Newport Beach, USA, Oct. 2008.
[12] S. Egger, M. Ries, P. Reichl, “Quality-of-Experience Beyond MOS: Experiences with a Holistic User Test Methodology for Interactive Video Services,“ in Proc. of 21st ITC Specialist Seminar, Miyazaki, Japan, Mar. 2010.
[13] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, Apr. 2004.
[14] 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.
[15] European audiovisual observatory, “Yearbook 2009 - Film, television and video in Europe,” European Commission, 2010.
[16] Olswang, “Olswang Convergence Consumer Survey 2007,” Nov. 2007.
[17] Parks Associates Industry Reports, “FEATURED IN: Converge! Network Digest/Blueprint: Telco Triple Play,” 2003, available at: http://www.parksassociates.com/press/articles/2003/converge_mg.htm.
[18] BIGresearch, “Pay TV Industry Demographic Profile,” BIGresearch CIA, Oct. 2009, available at: http://www.valassis.com/1024/resources/TELE_demo.aspx.
[19] Recommendation ITU-T H.262 (2000) | ISO/IEC 13818-2:2000, Information technology – Generic coding of moving pictures and associated audio information: Video.
[20] 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.
[21] Recommendation ITU-T H.264 (2005), Advanced video coding for generic audiovisual services.
[22] SMPTE 421M (2006), Television – VC-1 Compressed Video Bitstream Format and Decoding Process.
[23] Displaysearch, “TV Makers Looking to New Designs and Features to Break out of Commodity Trap,” Oct. 2009, available at: http://www.displaysearch.com/cps/rde/xchg/displaysearch/hs.xsl/091230_tv_makers_looking_to_new_designs_and_features_to_break_out_of_commodity_trap.asp
[24] Displaysearch, “LCD TVs to Exceed 180 Million Units in 2010”, Jan. 2009, available at: http://www.displaysearch.com/cps/rde/xchg/displaysearch/hs.xsl/01_27_09_displaysearch_sees_flat_panel_display_growth_slowing_through_2015.asp

[25] ITU-T Recommendation J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference,” International Telecommunication Union, Geneva, Switzerland, 2008.
[26] Displaysearch, “Quarterly Global TV Shipment and Forecast Report”, May 2010, available at: http://www.displaysearch.com/cps/rde/xchg/displaysearch/hs.xsl/quarterly_global_tv_shipment_and_forecast_report.asp
[27] 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.
[28] S. Wolf, M. Pinson, “Application of the NTIA general video quality metric (VQM) to HDTV quality monitoring,” in Proc. of The Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, AZ, USA, January 2007.
[29] E. Dumic, S. Grgic, M. Grgic, “Comparison of HDTV formats using objective video quality measures,” Journal of Multimedia Tools and Applications, Springer Netherlands, Jan. 2010.
[30] 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.
[31] 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.
[32] 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.
[33] C. Zetzsche, G. Hauske, ”Multiple channel model for the prediction of subjective image quality,” in Proc. SPIE, vol. 1077, pp. 209 - 216, 1989.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] 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.
[41] A. Raake, M.N. Garcia, S.Moeller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann, “T-V-MODEL: Parameter-based prediction of IPTV quality,” in ICASSP08, in process, 2008.
[42] B. Corrie, H. Wong, T. Zimmerman, S. Marsh, A.S. Patrick, J. Singer, B. Emond, S. Noël, “Towardsquality of experience in advanced collaborative environments,” in Proc. of 3rd Annual Workshop on Advanced Collaborative Environments, Seattle, USA, Jun., 2003.
[43] 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.
[44] 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.
[45] M. Ries, J. Kubanek, M. Rupp, ”Video Quality Estimation forMobile Streaming Video Quality Estimation for Mobile Streaming,” in Proc. of Measurement of Speech and Audio Quality Networks, Prag, Czech Republic, 2006.
[46] ITU-R Recommendation BT.500, ”Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, 2002.
[47] ITU-T Recommendation P.930, ”Principles of a reference impairment system for video,” International Telecommunication Union, 1996.

[48] 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.
[49] D. De Vera, P. Rodríguez-Bocca, G. Rubino, “QoE Monitoring Platform for Video Delivery Networks,” in LECTURE NOTES IN COMPUTER SCIENCE, NUMB 4786, pp. 131-142, 2007.
[50] P. Teo, D. Heeger, ”Perceptual image distortion,” in Proc. of SID International Symposium, pp.2 09 - 212, Jun. 1994.
[51] D. Heeger, ”Normalization of cell responses in cat visual cortex,” in Journal of Visual Neuroscience, vol. 9, pp. 181 - 197, 1992.
[52] 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.
[53] ITU-T Recommendation P.910, ”Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, Sep. 1999.
[54] ITU-R Recommendation BS.1534-1, ”Method for the subjective assessment of intermediate quality level of coding systems,” International Telecommunication Union, Geneva, Switzerland, 2003.
[55] List of most-watched television broadcasts, available at: http://en.wikipedia.org/wiki/List_of_most-watched_television_broadcasts.
[56] Comparison of video services, available at: http://en.wikipedia.org/wiki/Comparison_of_video_services#cite_note-15
[57] M. Ries, M. Rupp, "Technical Recommendation S. 10/07: Source Encoding of High Definition Mobile TV Services," Vienna University of Technology, 2007.
[58] Möller, S., Engelbrecht, K.-P., Kühnel, C:, Wechsung, I., and Weiss, B. (2009). A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction. QoMEX 2009, First International Workshop on Quality of Multimedia Experience.
[59] G. Giambene, G. Samuele, P. Cristina, M. Ries, A. Sali, "Traffic Management in HSDPA via GEO Satellite," Space Communications, Volume 21, pp. 21 – 61, 2007.
[60] K. De Moor, I. Ketyko, W. Joseph, T. Deryckere, L. De Marez, L. Martens, G. Verleye, ”Proposed Framework for Evaluating Quality of Experience in a Mobile, Testbed-oriented Living Lab Setting,” Mobile Netw Appl. DOI 10.1007/s11036-010-0223-0, 2010.
[61] J. Lull, “Family communication patterns and the social uses of television,” Communication Research, 1980.