Technische Universität Wien

Faculty of Electrical Engineering and Information Technology

Institute of Telecommunications

Master of Science Thesis
Telecommunication Engineering

Quality Assessment of a 3D Mobile Video Service

by

José Ramón Monteverde Almela

Supervisors:

Dr. Michal Ries
Prof. Markus Rupp

Vienna, June 2012

Abstract

The recent development in technology of 3D displays for portable devices and in parallel deployment of Long Time Evolution (LTE) extend the portfolio of multimedia services and increase available bandwidth for user devices [1]. Moreover, new generation of smartphones with extensive processing power equipped with 3D displays [2] enhance quality of experience for subscribers of mobile services.

One of the key challenges is provisioning of acceptable video quality for mobile video applications and meeting processing and network requirements over limited resources. Therefore, our initial study is focused following three issues: assessment methodology, encoding settings and content influence.

This thesis presents subjective and objective quality assessments for 3D mobile videos. The assessments follow in general video quality assessment standards issued by standardization bodies. Furthermore, the test setup follows the state of the art technical settings and reflects the content popularity and variety.

Contents

List of figures

List of tables

List of equations

1. Introduction

In this section a context of the current situation of the 3D mobile media is presented. Some basics about the state-of-the art hardware for 3D mobile devices and the formats and codecs used for the representation of the 3D video are also shown. After that, an outline about the main subjective evaluation methods is given.

In the subsequent sections thesis the main goals and procedures followed will be expounded as well as analyzed and discussed in accordance with the structure below:

- Section 2 describes the test scenario where the assessment takes part.
- Section 3 shows the video materials used both in the user-based and automated assessment.
- Section 4 describes the methodology used in the user-based assessment.
- Section 5 specifies the methods proposed for the automated assessment.
- Section 6 shows the results obtained for both user-based and automated assessment.
- Section 7 explains the conclusions reached according to the results of the previous section.

1.1 Current situation of 3D mobile media

While emerging in areas such as 3D cinema and 3D television, 3D media has also been actively researched for its delivery to mobile devices [3]. The general concept of 3D media assumes that the content is to be viewed on big screens and simultaneously by multiple users.

Glasses-enabled stereoscopic display technologies have matured sufficiently to back the success of 3D cinema and have also been enabling the introduction of first generation 3DTV. Autostereoscopic displays have been developed as an alternative display technology offering glasses-free 3D experience for the next generation 3DTV.

The research challenge of achieving the symbiosis between 3D and mobile media is to adapt, modify, and advance the 3D video technology, originally targeted for large screen experience, for the small displays of handhelds.

The introduction of 3D media to handhelds is supported by the current trend of developing novel multicore processors as an effective way to reduce the power consumption while maintaining or increasing the performance [4].

Apart from the potential of the new LTE networks with the distribution of 3D video content for mobile devices as mentioned above, the research and development of mobile 3DTV based on Terrestrial Digital Multimedia Broadcasting (T-DMB) and Digital Video Broadcasting-Handheld (DVB-H) systems, are currently quite active in Korea and Europe [5]. The motivations for this research and development are based on the rapid development of mobile technologies, the speedy evolution of the mobile market for handheld devices and focused TV services, such as live sporting or art events, and consumer willingness to switch to new smartphones. The introduction of mobile 3DTV has been relatively easy due to available standards such as DVB-H and T-DMB. These standards do not require much change in terms of transmission infrastructure to develop backward compatible services.

1.2 Autostereoscopic displays [6]

As it has been pointed out in the previous section, autostereoscopic displays offer the possibility of 3D perception without the need for special glasses or any other device, what completely fits to mobile devices and stands out like the only technology suitable for them.

Comparing 2D to 3D video, the former lacks of four important aspects which make the difference:

- Stereo parallax: seeing a different image with each eye.
- Movement parallax: seeing different images when we move our heads.
- Accommodation: the eyes' lenses focus on the object of interest.
- Convergence: both eyes converge on the object of interest.

Autostereoscopic displays use one or more of these effects to create the 3D experience. According to the way of presenting the video, they can be ordered in three different groups:

- Two-view displays.
- Head-tracked displays (usually with just two views).
- Multiview displays (with three or more views).

1.2.1 Display types

Two-view displays

The horizontal resolution is divided into two sets of columns. One of the two visible images consists of every second column of pixels; the second image consists of the other columns. The two images are captured or generated so that one is appropriate for each of the viewer's eyes.

The viewer must be placed at a specific distance and position (angle) to the display, in order to be able to perceive the 3D image. This area is called “sweet spot”. Outside this area the user would see an incorrect, pseudoscopic image.

To avoid these limitations, different solutions have been purposed, namely head-tracking or multiview displays.

Head-tracked displays

This kind of displays knows the position of the viewer’s head. Consequently, for a two-view head-tracked display both views are directed to the sweet spot, avoiding pseudoscopy.

The system must be designed with minimum lag so that the user cannot notice the head tracking.

Multiview displays

The most important benefit of multiview displays is that if both eyes of the users are inside the viewing zone, they can perceive the 3D image. According to this, the display accommodates multiple viewers, each of them seeing the 3D video from their own pint of view.

One drawback is the difficulty of generating all the views simultaneously, as every view is displayed, whether there is a user in the direction of that particular view or not.

1.2.2 Display technologies

Three main technologies are used for making autostereoscopic displays:

- Spatial multiplex: the resolution of a display device is split between the multiple views.
- Multiprojector: a single projection display is used for each view.
- Time-sequential: a single very fast display device is used for all views.

Different solutions have been purposed for every class, but in this thesis the focus of attention will be for the most popular nowadays, specifically parallax barrier and lenticular lens. Examples of their design are depicted below, though only for the two-view case. Naturally, similar versions for more than two views are also possible.

Parallax barrier

Parallax barrier is essentially a mask with openings and closings that blocks the light from the screen in certain directions [3] as depicted in Figure 1. The intensity of the light rays passing through the filter changes as a function of the angle, as if the light is directionally projected. Each eye sees the display from different angle and thus sees only a fraction of all pixels, precisely those meant to convey the correct (left or right) view, otherwise combined in the rendered image.

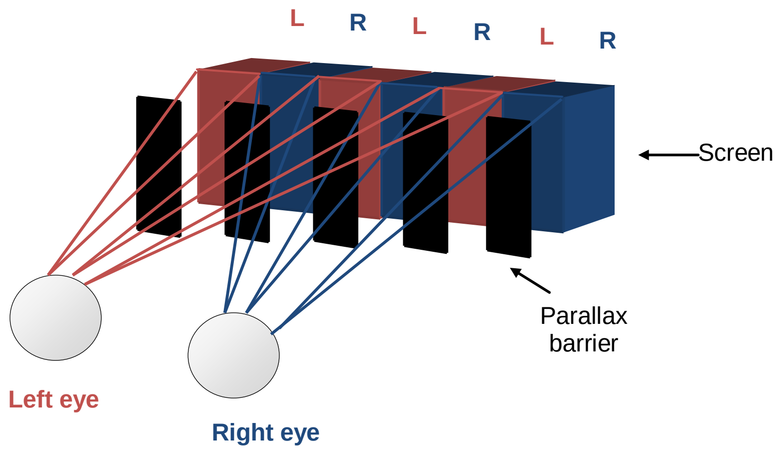

Figure 1: Parallax barrier system

As different subpixels are responsible for different-perspective images, the spatial resolution is decreased and the discrete structure of views becomes more visible. Parallax barriers block part of the light and thus decrease the overall brightness. In order to compensate for this limitation, one needs extra bright backlight, which would decrease the battery life if used in a portable device.

Lenticular lens

In this case, as depicted in Figure 2, instead of a barrier, an array of cylindrical lenslets is placed in front of the screen, directing the light from adjacent pixel columns to different viewing slots at the ideal viewing distance so that each of the viewer's eyes sees light from only every second pixel column.

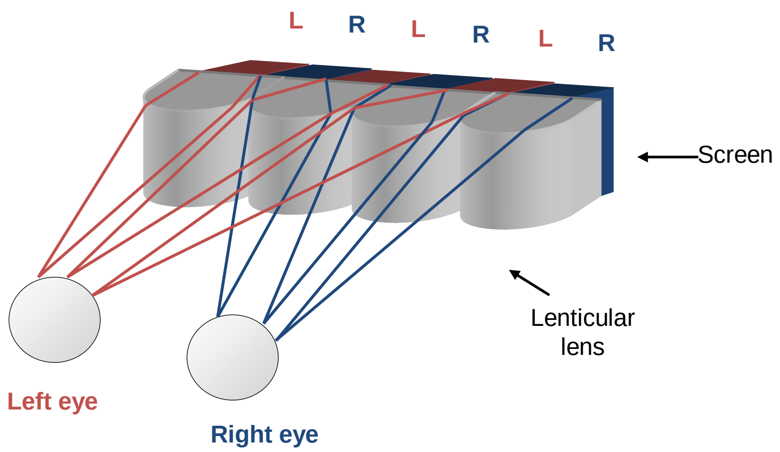

Figure 2: Lenticular lens system

1.3 3D video formats and codecs

In this section the most popular and used video formats for the 3D representation and their corresponding compression techniques, the main 3D coding standards, are commented. It is important not to confuse them; while a format is the way the video is stored in a file or transmitted over a stream to be interpreted by a computer, a coding standard is a compression a decompression technique for the file or stream.

As currently there is not a widely accepted standard for 3D video, a generic, flexible and efficient 3D video format that would serve a wide range of different 3D video systems is highly desirable in this context. Therefore MPEG is currently investigating such a new generic 3D video standard.

1.3.1 3D video formats

There are a lot of different 3D video formats available and under investigation. They include different types of data, mostly related to specific types of displays. This starts from classical two-view stereo video, extends to multiview video with more than two views, video plus depth, multiview video plus depth, and layered depth video.[7]

Conventional stereo video (CSV)[7]

It is the most well-known and simple type of 3D representation. The captured video signals (two or more) are meant to be directly displayed using a 3D display system, though there might also be some pre-processing.

If we compare CSV to other video formats, its algorithms are the least complex. It can be as simple as only separately encoding and decoding the different video signals. However, the video size is increased compared to 2D formats. This increase can be balanced by reducing the resolution (spatial and/or temporal), if necessary.

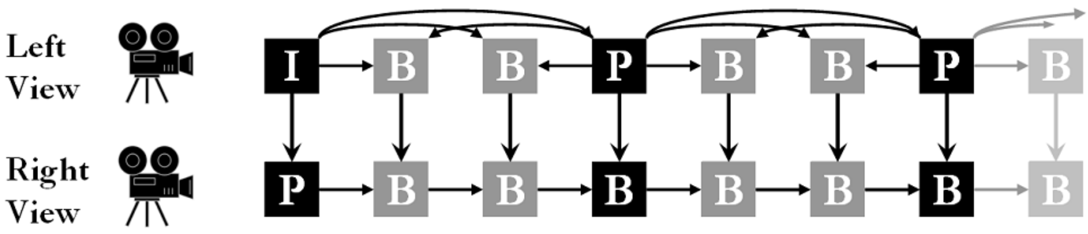

Figure 3: Stereo coding combined interview/temporal prediction [7]

As depicted in Figure 3 above, coding efficiency in CSV can be improved by combining temporal/interview prediction. This technique has been a standard for many years included in MPEG-2 Multiview Profile. It has been also added to H.264/AVC with the

Stereo SEI (Supplemental Enhancement Information) message, which will be described in the following section 1.3.2.

When the captured views are more than two, the standard used is Multiview Video Coding (MVC), accepted by MPEG-ITU, as an extension of H.264/AVC, though it can be applied to two views, too. (See section 1.3.2).

Another efficient method for stereo view coding is the low-pass filtering of one of the images of the stereo pair as the perception of the overall quality is dominated by the higher quality image. This way, one view is downsampled to half or quarter resolution. Mixed resolution stereo video coding is based on this effect.

One of the problems of CSV is that depth perception cannot be adjusted to different display types and sizes, and also head motion parallax is not supported. Compared to other 3D video formats, its functionality is limited.

Video plus depth (V+D) [8]

This video format is made up of a 2D video signal and depth map. They are sent to the user and then, after rendering, a stereo pair is obtained. Each 2D image frame is supplemented with a grayscale (luminance) depth map which indicates if a specific pixel in the 2D image needs to be shown in front of the display (white) or behind the screen plane (black). The depth range indicates the maximum and minimum distance of the 3D point to the camera. Depth is usually codified with 8 bits. The depth map grayscale images can be included into the luminance channel of a video signal and the chrominance can bet set to a constant value. Therefore, the resulting standard signal can be processed by any state-of-the-art video codec. The depth data coding can reach just 10-20% of the necessary bitrate needed for colour video coding while still providing good quality. Nevertheless, for more complex depth maps, the necessary bitrate can be very similar to colour video coding.

An interesting functionality not present in CSV is that V+D is able to generate the stereo pair at the decoder, so the impression can be adjusted after the transmission and it also enables head motion parallax as well as more than two views.

The concept of V+D is highly interesting due to the backward compatibility and extended functionality. Moreover it is possible to use available video codecs. It is only necessary to specify high-level syntax that allows a decoder to interpret two incoming video streams correctly as color and depth. Additionally, information about depth range (maximum and minimum) needs to be transmitted.

On the other hand the advantages of V+D over CSV are paid by increased complexity for both sender side and receiver side. View synthesis has to be performed after decoding to generate the second view of the stereo pair. Before encoding the depth data have to be generated. This is usually done by depth/disparity estimation from a captured stereo pair.

Multiview video + depth (MVD) [8]

Multiview video plus depth is a 3D video standard developed by MPEG supporting free viewpoint video by including different output views or rendering them at the decoder. As its name reads, it combines multiview video with multiple depth maps.

Depth has to be estimated for the N views at the sender. N color with N depth videos have to be encoded and transmitted. At the receiver the data have to be decoded and the virtual views have to be rendered.

MVD can efficiently support multiview autostereoscopic displays.

This new standardization activity clearly targets high-quality, high-resolution, thus high-bitrate and high-complexity home user applications. General usage might be realized via a layered,scalable representation where a base layer (e.g. one color video and one depth map, perhaps at limited resolution) is accessible for low complexity devices without having to deal with the whole signal.

Layered depth video (LDV) [9]

Layered depth video is an alternative to MVD where instead of representing the scene with an array of depth pixels (pixel color with associated depth values), each position in the array may store several depth pixels, organised into layers (e.g. data from other viewing directions). It efficiently reduces the multi-view video bitrate, and it offers photo-realistic rendering, even with complex scene geometry.

LDV supports rendering of virtual views and therefore multiview autostereoscopic displays in a similar way as MVD does. LDV can be obtained from MDV and it is more efficient, as less data has to be transmitted. However, one associated drawback is that additional error-prone vision tasks are included which operate on partially unreliable depth data.

1.3.2 3D video coding standards

A variety of compression and coding algorithms are available for the different 3D video formats. Some of these standards and coding algorithms are standardized e.g. by MPEG, since standard formats and efficient compression are crucial for the success of 3D video applications [10]. This way, some of the most popular coding standards are presented, all of them based on MPEG-4 AVC / H.264 standards family.

H.264 Simulcast [11]

Simulcast uses stereo video format, consisting of two input images, one for the left and right view of the stereo pair. The codec used in this case is H.264/AVC. Both input images are encoded with it resulting in two encoded bit-streams or transport-streams

BS/TS. Once they are transmitted over the channel, both streams are decoded independently, obtaining then the two warped sequences for the stereo pair.

According to the H.264/MPEG-4-AVC standard, “H.264 Simulcast” is specified as the individual application of an H.264/AVC conforming coder to several video sequences in a generic way.

H.264 Supplemental Enhancement Information Message [11]

As Simulcast, H.264 Supplemental Enhancement Information (SEI) Message uses the stereo video format, interlacing line-by-line the two sequences into one. The codec applied to the interlaced sequence is H.264/AVC, consequently obtaining only one encoded bit-stream or transport-stream BS/TS. The sequence is transmitted over the channel, then decoded and interlaced, to obtain the stereo pair again. It enables the encoder to indicate to the decoder how to extract two distinct views of a video scene from a single decoded frame. The message also serves as a way to support stereo-view video in applications that require full compatibility with prior decoder designs.

SEI Message assists in processes related to decoding, display or other purposes. However, SEI message is not required for constructing the luma or chroma samples by the decoding process. Conforming decoders are not required to process this information for output order conformance to H.264/SVC.

SEI is basically metadata included in the video file itself, specifically the frame packing arrangement (FPA) field. According to the type of representation needed this field contains the 3D representation arrangement as specified in Table 1:

Value	Interpretation
0	checkerboard: pixels are alternatively from left and right views
1	column alternation: left and right views are interlaced by column
2	row alternation: left and right views are interlaced by row
3	side-by-side: left view is on the left, right view on the right
4	top-bottom : left view is on top, right view at the bottom
5	frame alternation: one view per frame (temporal interleaving)

Table 1: definition of frame packing arrangement type

H.264 Auxiliary Picture Syntax [11]

Auxiliary Picture Syntax uses video plus depth format, consisting of two inputs, one for the video sequence and the associated depth information for one of the two views of the stereo pair. Both of them are encoded with H.264/AVC simultaneously but independently, resulting in one encoded bit-stream or transport-stream BS/TS. After transmission, this stream is decoded, again simultaneously but independently for primary and auxiliary coded pictures, resulting in the distorted video sequence and the distorted depth sequence for one of the two views of a stereo pair.

This standard also enables extra supplemental information as SEI or VUI (Video Usability Information). One of these new types of data are auxiliary pictures, which are extra monochrome pictures sent along with the main video stream, that can be used for such purposes as alpha blend compositing.

MPEG-C part 3

MPEG-C Part 3 [8] is used as a container for the video plus depth format, consisting of the input video sequences and the associated depth information for one of the two views of a stereo pair. The codec used for MPEG-C Part 3 is H.264/AVC, which is applied to each of the two input sequences independently, resulting in two encoded bit--streams BS. For transmission these two bit-streams are interleaved frame-by-frame in a multiplexer, resulting in one MVC transport-stream TS, that may contain additional depth maps properties as auxiliary information. After transmission a demultiplexer separates this stream into the two individually coded streams. These two streams are decoded independently, resulting in the distorted video sequence and the distorted depth sequence for one of the two views of a stereo pair.

Supplemental information has been defined for this coding standard, such as AVSI (Auxiliary Video Supplemental Information) message, which characterizes the interpretation of an auxiliary video sequence that accompanies a primary video sequence. For instance, an AVSI can indicate that the auxiliary video represents depth map information, and can provide parameters for the proper interpretation of the auxiliary video as such depth information.

Multiview Video Coding (MPEG-4 MVC / H.264)

H.264 MVC uses the stereo video format, consisting of two input video sequences for the left and right view of the stereo pair. The codec used for H.264 Multiview Video Coding is H.264/MVC, which is applied to both sequences simultaneously for inter-view predictive coding, resulting in two dependent encoded bit-streams BS that may contain the camera parameters as auxiliary information. For transmission these two bit-streams are interleaved frame-by-frame in a multiplexer, resulting in one MVC transport-stream TS. After transmission this stream is decoded (and thereby demultiplexed), resulting in the distorted sequences of the stereo pair.

The specification of Multiview Video Coding in H.264/AVC [12] defines it as an extension to the family of H.264 standards. The basics of H.264/AVC for single view can be extrapolated to MVC, so that a current picture in the coding process can have temporal as well as inter-view reference pictures for motion-compensated prediction, but also includes a number of new techniques for improved coding efficiency, reduced decoding complexity, and new functionalities for multiview operations. MVC takes advantage of some of the interfaces and transport mechanisms introduced for the scalable video coding (SVC) extension of H.264/AVC. New requirements for 3D video related to interface, transport of the MVC bitstreams, and MVC decoder resource management lead to new features, that have been adopted for MVC, including marking of reference pictures, supporting for efficient view switching, structuring of the bitstream, signaling of view scalability supplemental enhancement information (SEI) and parallel decoding SEI.

1.4 Subjective evaluation

In subjective methods test users are set in a controlled test environment in order to evaluate the quality of a video. A way of doing this is providing the viewer with a distorted video for its evaluation [13].

Another way is providing a reference/original sequence so that the tester can determine the relative quality of the distorted video. Common to all procedures is the pooling of the votes into a mean opinion score (MOS) which provides a measure of subjective quality on the media in the given test set.

Regarding subjective quality assessment two main disadvantages stand out. The time taken to complete is extremely length, the assessment is expensive, tedious, and it is necessary to be very careful during the making in order to obtain meaningful results.

Recommendations from ITU-R for stereoscopic subjective assessment of television pictures [14] depend on whether we have a reference image or not. When a reference image is available, double-stimulus continuous quality-scale (DSCQS) or double-stimulus impairment scale (DSIS) are two of the most used methods. Examples include comparison of display systems, quality assessment of coding systems, and so on. When no reference is available, the categorical judgment method can be used (ACR - Absolute Category Rating), for example, to identify the merits of stereoscopic systems.

1.4.1 Double stimulus continuous quality scale (DSCQS)

In this method of subjective quality assessment, both the reference (original) and the test sequence are shown to the testers in sequence pairs twice alternating the order by a random process [15]. The viewers are not aware of which one is the either the reference or the test sequence. The duration of the sequences is reasonably short (8-10 s). The rating score ranges from “bad” to “excellent”, what can be equivalent to a 0 to 100 scale as it can be seen in Figure 4. The differences between the two ratings are further analyzed. This data is used to balance, the possible uncertainties caused by different users’ experiences and material content. This procedure is usually chosen when the reference and test sequences have similar quality; otherwise the subjects can easily notice the quality differences between both sequences.

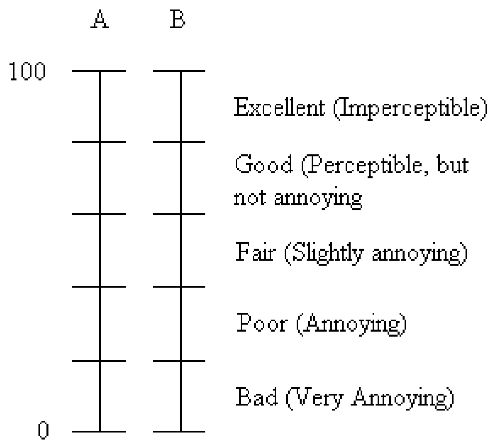

Figure 4: Video quality assessment scale used in subjective MOS tests [15]

1.4.2 Double stimulus impairment scale (DSIS)

In this method the reference-test sequence pair is shown just once and the reference is always presented firstly. The score rating of the impairments are in the same way as in DSCQS 0 to 100 scale or 1 to 5-point scale (from “very annoying” to “imperceptible”). This method is preferred when the impairments are noticeable, for example big degradations caused by encoding or transmission.

If our aim is to evaluate digital video systems over longer timescales, using quite short sequences can became a problem. These systems generate substantial quality variations that may not be uniformly distributed over the time. Due to the bursty and non- deterministic nature of encoded video transmission over packet networks, like the Internet, the design of DSCQS and DSIS is not suitable in this case. The users can perceive a significantly variation of the quality of the video over time in this kind of situation. The first problem is that if double stimulus (showing both the reference and test video sequences) is used for longer sequences, the time between comparable moments will be too lengthy to be rated accurately. Moreover, it is known that if the length of the sequence is increased from 10 to 30 s, memory is increased for more recent stimulus, so the recent parts of the sequence have a greater influence in the overall quality impression.

1.4.3 Single stimulus absolute category rating (ACR)

The single stimulus absolute category rating method is a category judgment where the test sequences are presented one at a time and are rated independently on a category scale [16]. (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 time pattern for the stimulus presentation can be illustrated by Figure 5. If a constant voting time is used (e.g., several viewers run simultaneously from a tape), then the voting time should be less than or equal to 10 s. The presentation time may be reduced or increased according to the content of the test material.

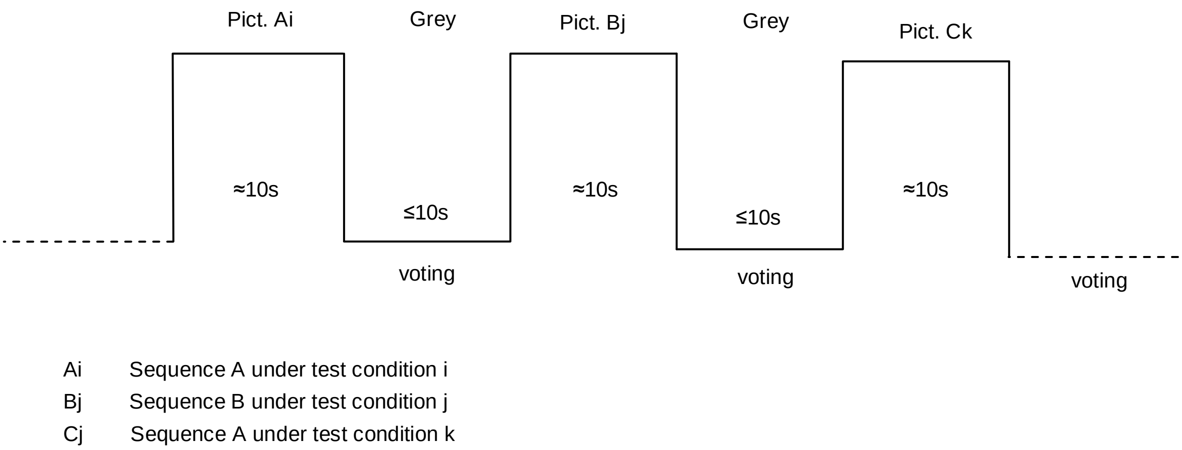

Figure 5: Stimulus presentation in the ACR method [16]

The five-level scale (from “bad” to “excellent”) should be used in this method too. (Figure 4)

For the ACR method, the necessary number of replications is obtained by repeating the same test conditions at different points of time in the test.

2. Test scenario

In the following, the main aspects of interest are described, in order to understand their impact on QoE. The variety of test conditions is explained in the following, namely the source videos, the bandwidth levels, and the content classes. These factors constitute the basis for creating the test sequences that will be used in our test (see section 3).

Both user-based assessment, which test takes place in a real world condition scenario (see section 4), and automated assessment (see section 5), will reflect the influence of these aspects on the QoE.

2.1 Source videos

One of the main influence factors of QoE is the extent to which bandwidth reduction related to video sequences is performed, but also the various settings and selected parameters are very important.

For our scenarios, four source videos were provided (one per every type of content), which we will use, in order to meet processing, network and quality requirements, with the following parameters:

Content	Action	Basketball	Cartoon	Football
Bandwidth levels	4	4	4	4
Resolution	640x480 pixels	640 x 480 pixels	640 x 480 pixels	640 x 480 pixels
Frame rate	15 fps	15 fps	15 fps	15 fps
Aspect ratio	4:3	4:3	4:3	4:3
Display aspect ratio	16:9	16:9	16:9	16:9
GOP structure	3 (Hierarchical B frames)	3 (Hierarchical B frames)	3 (Hierarchical B frames)	3 (Hierarchical B frames)
Intra period	[15,250]	[15,250]	[15,250]	[15,250]
Reframes	4	4	4	4
Search range	16	16	16	16
Video mode	CBR	CBR	CBR	CBR

Table 2: Source videos coding parameters

The H.264 High profile at 2.2 level was used for encoding of all video sequences at 15 fps. This frame rate allows considerable bandwidth reduction suitable for mobile video, while showing good performance with different levels of motion. [17]

Moreover, the Stereo SEI Message coding, with frame packing arrangement side-by-side (left-right type 3) [11] was set, as it is the only arrangement supported by our test device. The 3D format used is CSV (Conventional Stereo Video).

Additional description about this standard was made in section 1.3.2, though it is worth mentioning that it is becoming very popular and commonly found on many of the new 3D capable devices.

Furthermore, the state of the art settings were considered for encoding setting and recent media consumer statistics and quality investigations of mobile video services were considered for our test scenario definition.

2.2 Bandwidth levels

Due to its importance for the project, bandwidth is varied systematically. The bandwidth is defined as average video bit rate including only video payload.

Bandwidth levels	Bandwidth [kbps]	Bits per pixel
4.	1150 Kbps	0.250
3.	850 Kbps	0.184
2.	550 Kbps	0.120
1.	250 Kbps	0.0540

Table 3: Bandwidth levels

With these four different bandwidth levels (see Table 3), we aimed at investigating the QoE according to systematically progressing bandwidth reduction levels. The rationale for choosing these bandwidth levels was to have a wide range of them and reach a low level where the degradation is perceptible.

2.3 Content classes

The most important rationale for content class definition was to use genres that viewers could be realistically expected to watch in a typical 3D mobile video service consumption situation. According to recent marketing surveys [18] [19], action, sport and cartoon are the most popular contents for 3d video. Therefore, we selected the following content classes (CCs):

- Action movie (CC1)
- Basketball (CC2)
- Cartoon (CC3)
- Football (CC4)

From a subjective point of view every content class has different character of spatial complexity, 3D depth and amount of movements.

Screenshots of the sequences chosen for every content class are depicted in Figure 6.

Figure 6: From left to right and up to down: action, basketball, cartoon and soccer sequences

3. Video materials

In order to enable a structured comparison of the main variables of interest described in the previous section, videos were produced that varied according to the following properties:

- 4 content classes (action, basketball, cartoon and soccer)
- 4 bandwidth levels (four H.264 encoded sequences at different bandwidth levels (see Table 3))

Figure 7 provides an overview of the resulting 32 combinations.

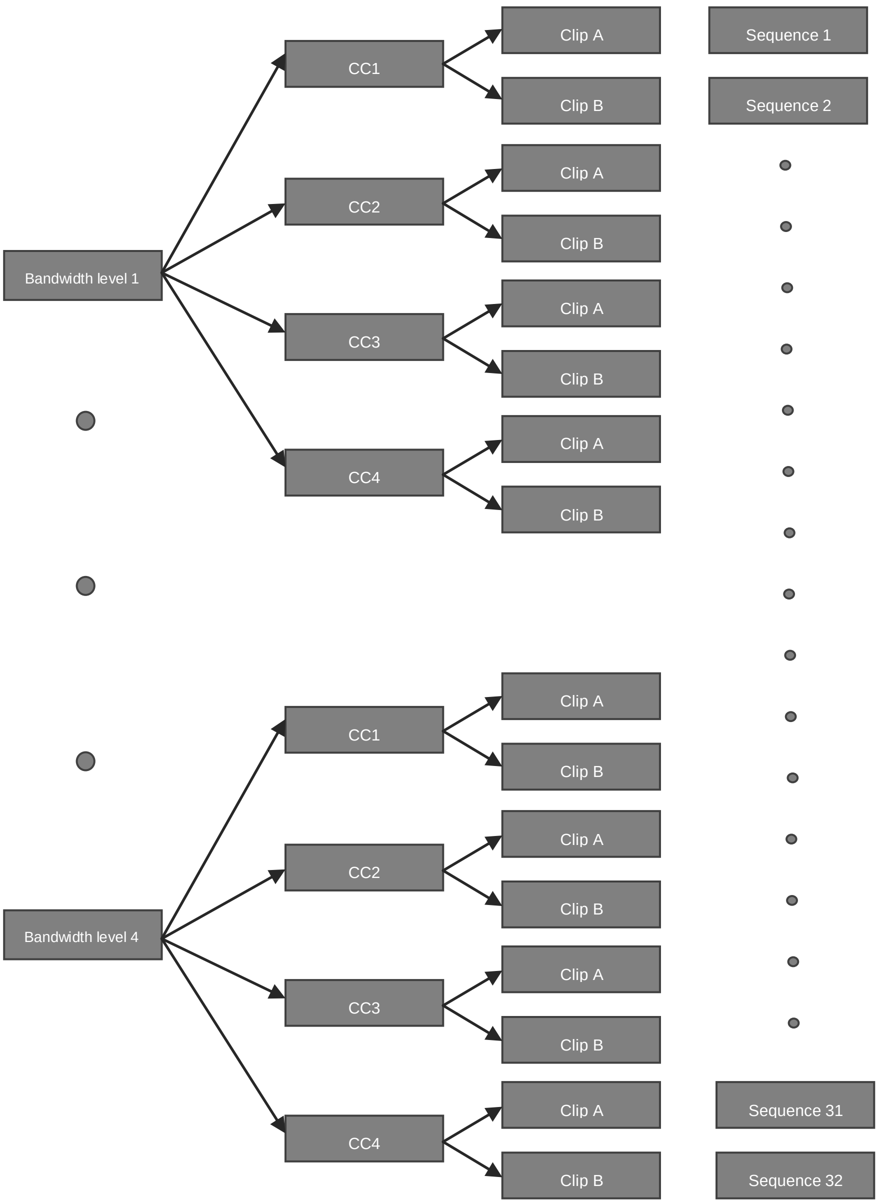

Figure 7: Video contents used in user-based QoE assessment

For the tests we selected sequences of twenty-second duration and VGA resolution, which is not consistent with valid ITU-T P.910 recommendation. The reason for duration extension of test sequences is that 3D viewing needs much more attention in compare to 2D viewing [20]. Therefore, the recognition of 3D scene takes more time. On other hand ITU alone introduced study item [21], [22] for methodologies of subjective and objective assessment of picture and sound quality of 3D multimedia services.

4. User-based assessment method

This section describes the methodology for user-based assessment of 3D mobile video services. Section 4.1 describes the test user sample. Section 4.2 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 4.3 describes the user-based QoE assessment: the quality rating test. Section 4.4 then outlines further data gathering measures and section 4.5 concludes with a tabular overview of the overall test procedure and durations.

4.1 Test user sample

25 subjects took part in the study. The main characteristics of the analyzed sample summarized in Table 4.

Age	26.32 years (± 2.15, min: 18, max: 40)
Gender	3 female, 22 male
Professional Status	11 employed, 2 unemployed, 12 students
Highest educational degree	Undergraduate (19), Graduate (5), Doctoral (1)
Kind of videos watched in cell phone	Offline (11), Online (11), None (3)
Minutes a day watching videos in cell phone	9.56 minutes (± 6.43, min:0, max: 50, median: 0)
Ever watched 3D video	Yes (19), No (6)
Experience in image processing	Yes (6), No (19)

Table 4: Test user sample characteristics

We can notice the profile selected for this test is suitable for potential users of this new mobile 3D technology. Young people (~26) who have already experienced some 3D service, and regularly use their phone to watch videos.

This profile is partially consistent with the literature of mobile TV and 3D, which offers different insights into the users and usage motivations [23]. Based on a mobile TV field trial, a typical user is described as a well-educated male aged between 23 and 35 with a yearly income of €20,000-30,000. The main motivations for usage are killing time while waiting or staying up-to-date with daily news while on the move. Another reason for usage is the novelty of the system and the desire to belong to the group of first users as motivating factors. Owning and sharing of content is also valued by the users

Studies of mobile 3DTV reveal another aspect of the user. The related literature describes a repertoire of aspects for creating an additional entertaining experience with

3D. Presence as the feeling of being there, engagement, naturalness, and enhanced realism describe the 3D experience and motivate users to watch 3D content. The negative aspect of the 3D experience is physical discomfort or simulator sickness. The reasons for simulator sicknesses are not fully understood, though a variety of different kinds of eye-related symptoms have been enumerated. This is a disadvantage for 3D and it is known that the user's enthusiasm for new technology, the learning benefits of viewing 3D and excitement about 3D content decreases significantly with the increase in such symptoms.

4.2 Overall test conditions

4.2.1 Room conditions

Our study was conducted in a real world condition. The optimal viewing distance and angles for perceiving the 3D video are described in Figure 8, as given by the manufacturer [24].

The most usual distance (d) between the eyes and 3d display was approximately 40 cm for the tested mobile device.

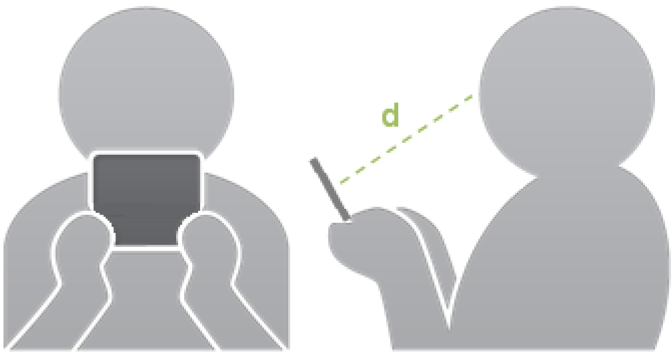

Figure 8: Optimal conditions for 3D perception

4.2.2 End-user device

The device for subjective testing chosen was the recently introduced HTC Evo 3D with the following technical specifications[25]:

- Display technology: LCD 3D display using "parallax barrier"
- Screen size: 4.3''
- Resolution: 960 x 540 pixels
- Aspect ratio: 16:9
- Player: mVideoPlayer 2.7.0
- Decoder type: H.264

Figure 9: Test user device (HTC Evo 3D)

4.3 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 judgment where test sequences are presented one at a time and are rated independently on a category scale (see section 1.4.3). After each presentation the subjects are asked to evaluate the quality of the sequence shown.

Warm-up sequences

At the beginning of each test round a trial run is presented with three sequences. Sequences are similar to tested video sequences, covering a wide range of bandwidth levels.

Scenario

In order to emulate real world conditions the test subjects were asked to imagine that they are trying out for the first time a new 3D mobile video service from a well-known mobile service provider, and that they will be paying a monthly fee for that. They should picture themselves using this new service.

Rating procedure

After each sequence, the user is asked to rate the subjective quality, according to an ordinal 5-point scale ranging from ‘bad’ to ‘excellent’. The voting time was defined as 10 seconds.

Questionnaire

A specifically designed questionnaire for this test is given. (See appendix A)

Allocation of test sequences

The allocation of the test video sequences to test subjects was as follows: Each user was exposed to one set of sequences. This set included videos for the specified four content classes. Test sequences were presented in random order, with the exception that neither the same clip nor the same bandwidth level appeared in succession.

Repetition of quality rating test

The quality rating test was conducted twice, in order to check consistency of obtained results. Between the rating tests, there was a 30 min break.

4.4 General inquiry

At the end of the test, further data is gathered from the test persons, namely demographic data and data on general user behavior. This data is showed in section 4.1.

4.5 Overview of test procedure

In the following, an overview of the parts within each test session is provided.

Part	Duration (min.)
Welcome, briefing	3
Quality rating test I	18
Break	30
Quality rating test II	18
Debriefing, general inquiry	5
	~ 75

Table 5: Overview of test phases and test duration

The increase of the length of the sequences to 20 s. each, made the test longer, consequently with some users complaining about some eye strain. (See section 6.1.3).

5. Automated assessment method

This section specifies the automated QoE methods used to complement user-based assessment. The evaluated video material was the same as in the user-based assessment. (See section 3.)

Automated evaluation of the investigated video sequences was performed with the video quality measurement and estimation tool MSU Video Quality Measurement Tool [26]. As described in the following, the features of this tool used for the automated evaluation are SSIM and PSNR computation. Both metrics are very common in 2D video assessment but they can also be used for the evaluation of our 3D service.

The possibility of predicting 3D video quality using 2D quality metrics in packet loss conditions for both left-and-right (our case) and color-and-depth stereoscopic videos has been addressed in [27], which could accelerate the further development of consumer products and introduction of new 3D video services in time. The results of this study show that the output from 2D objective metrics can be mapped, so that it correlates strongly with both the overall viewer perception of image quality and depth. This implies that, while subjective test results remain as the best and precise judgment of 3D video quality, and developing a single quality metric for 3D video still remains to be a big challenge due to the number of perceptual attributes associated with 3D perception and their relationship with the human visual system, the use of objective quality assessment metrics is an acceptable compromise for the 3D video research community.

5.1 Structural Similarity Index (SSIM index)

The structural similarity (SSIM) index 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 system diagram [28] of the proposed quality assessment system is shown in Figure 10. Suppose *'x'* and *'y'* are two nonnegative image signals, which have been aligned with each other. The system separates the task of similarity measurement into three comparisons: luminance, contrast and structure, where structural information is defined in an image as those attributes that represent the structure of objects in the scene, independent of the average luminance and contrast.

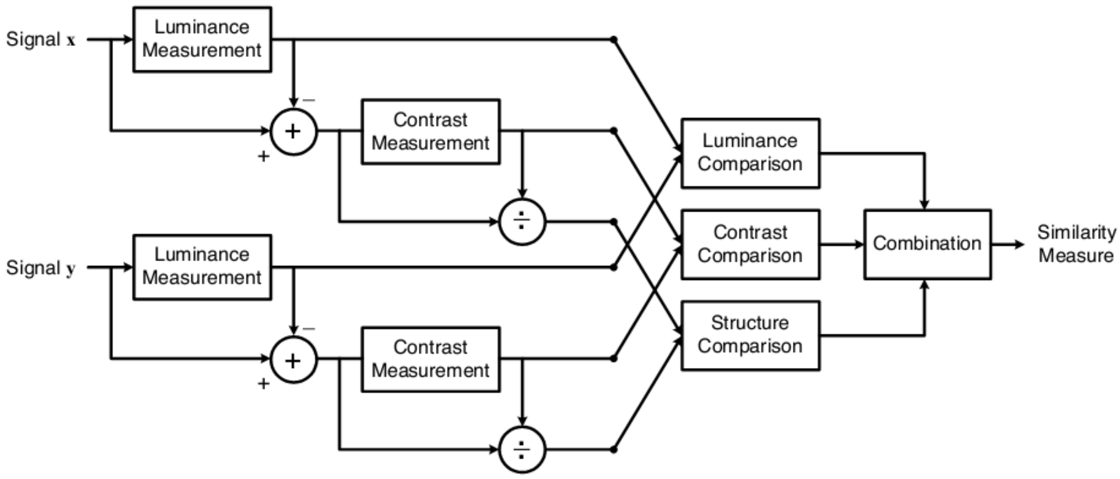

Figure 10: Diagram of the structural similarity (SSIM) measurement system

The specific form of the SSIM index is:

$$SSIM\ (x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Equation 1: SSIM definition

In our SSIM definition C_1 and C_2 are calculated using the following equations:

$$C_1 = 0.01 * 0.01 * video1Max * video2Max$$

$$C_2 = 0.03 * 0.03 * video1Max * video2Max$$

Equation 2: Constants used in the SSIM definition

,

where *video1Max* is the maximum value of a given color component for the first video, *video2Max* is the maximum value of the same color component for the second video.

The SSIM index was calculated between the H.264 reference sequence (that was also downsampled to VGA@15 fps) and the H.264 sequences for the four different bandwidths. SSIM Index is based on measuring of three components (luminance similarity, contrast similarity and structural similarity) and combining them into result value.

The SSIM used is a precise version which uses Gauss blur [29]. In practice, one usually only requires a single overall quality measure of the entire image, therefore, the mean of the SSIM index map was computed to evaluate the overall frame quality. SSIM values were calculated for single frames and average SSIM over a twenty-second length time window.

5.2 Peak Signal-to-Noise Ratio (PSNR)

The basic definition of PSNR is [29]:

$$PSNR(x,y) = 10 \cdot log_{10} \frac{MaxErr^2 \cdot w \cdot h}{\sum_{i=0,j=0}^{w,h} (x_{i,j} - y_{i,j})^2}$$

Equation 3: PSNR definition

,

where *MaxErr* - maximum possible absolute value of colour components difference, w - video width, h - video height.

Generally, this metric is equivalent to Mean Square Error, but it is more convenient to use because of logarithmic scale. It has the same disadvantages as the MSE metric.

The PSNR (peak signal-to-noise ratio) was calculated as a difference measure between the H.264 reference sequence and H.264 sequences with the four different bandwidths. The calculation was performed for the gray scale (or luma component) colour space. This procedure reduced processing complexity by a factor of three and eliminated erroneous PSNR calculation due to different colour spaces of original and encoded sequences (which is typically a calculation error). Finally, PSNR values are calculated for single frames and average PSNR over a twenty-second length time window.

6. Results

In the following, the results of the user-based and automated assessment are presented. As a way of comparing both assessments and the validity of the objective metrics selected, the correlations between subjective and objective results are presented.

Furthermore, statistical analysis of the standard deviation is performed to reflect the user's diversity.

6.1 User-based assessment

In the following, the user-based assessment results for all the sequences (4 content classes x 4 bandwidth levels x 2 clips) are presented. Please refer to sections 2 and 3 for the specification of test sequences and to section 4 for the assessment methodology. The figures present the quality ratings and for each content class (the values for the two clips per content class were averaged).

6.1.1 MOS (Mean Opinion Scores)

The obtained MOS data was scanned for unreliable and inconsistent results [30]. Votes from one viewer to a certain sequence that differ two or more MOS grades from the first to the second run were considered unreliable and therefore rejected. In total, 1.5% of the results were rejected. This correction had negligible effect on the test global mean score. The 95% confidence intervals [16] were as well computed, assuming the votes follow a normal distribution.

When looking at the mean values and their confidence intervals within the four content classes in Figure 11, there are some very noteworthy relative differences.

The differences among the MOS values of the content types are very significant. MOS for every bandwidth level always received the lowest values for basketball, and the highest for cartoon. The order is, for the four levels, from lowest to highest: basketball, football, action, and cartoon. For the bandwidth levels 1 and 2 this trend is outstanding, as a stepped shape can be noticed. As the bandwidth is increased, the MOS values equalize and form two groups: sports content (basketball and football) and movies (action and cartoon).

The testers did not perceive a big difference between levels 4 (1150 Kbps) and 3 (850 Kbps). The evaluations are very similar. This bandwidth reduction only led to a maximum degradation of 0.27 MOS points for soccer (from 3.28 to 3.01), 0.17 MOS points for basketball (from 3.12 to 2.93), 0.15 MOS points for action (from 4.12 to 3.97) and just 0.02 MOS points for cartoon (from 4.14 to 4.16). In this last case the evaluation is surprisingly even higher for a smaller bandwidth.

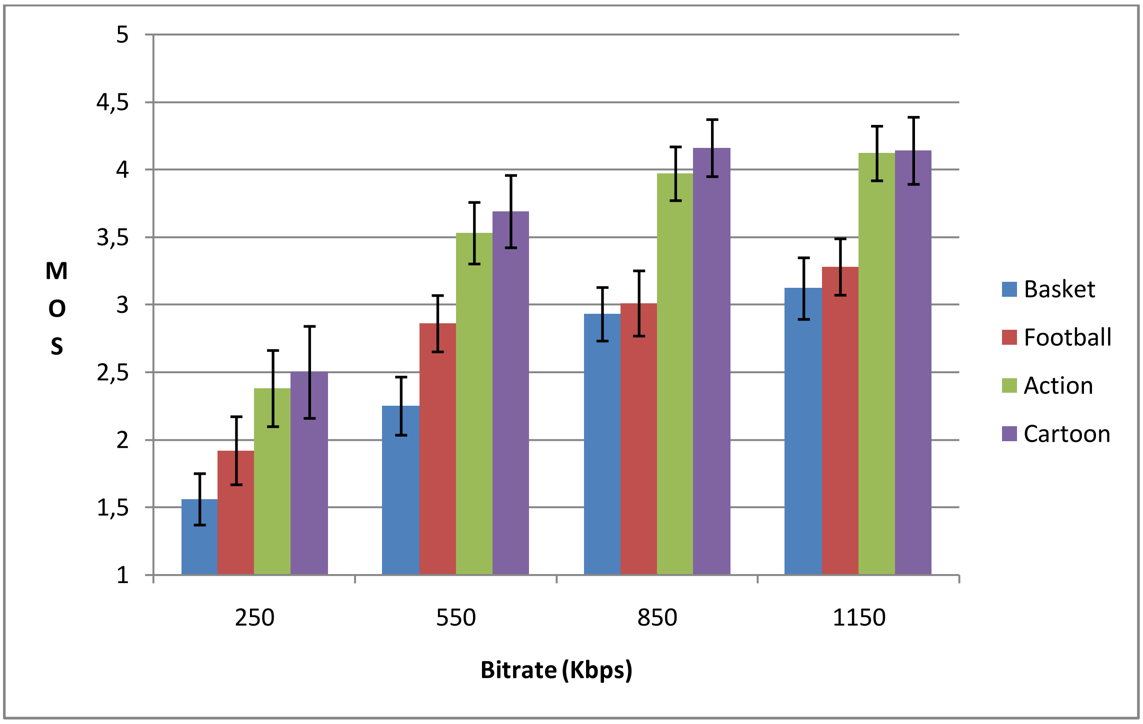

Figure 11: Quality rating results for CC1, CC2, CC3 and CC4 for every bandwidth level

Content class / Bandwidth (Kbps)	Basket (MOS ± CI)	Soccer (MOS ± CI)	Action (MOS ± CI)	Cartoon (MOS ± CI)
250	1,56 ± 0,19	1,91 ± 0,25	2,38 ± 0,28	2,51 ± 0,34
550	2,25 ± 0,22	2,86 ± 0,21	3,53 ± 0,23	3,70 ± 0,27
850	2,94 ± 0,20	3,01 ± 0,24	3,98± 0,20	4,16 ± 0,21
1150	3,13 ± 0,23	3,28 ± 0,21	4,12 ± 0,20	4,15 ± 0,25

Table 6: MOS confidence interval for every CC and bandwidth combination

The distribution of the 95% confidence intervals for the MOS can be used as a quality indicator of the collected data. The average size of the 95% confidence intervals (Table 6) is 0.23 on the 1 - 5 MOS scale. This indicates a good agreement between observers. They are especially higher for low bitrates of the highest scored content classes (cartoon and action). In this case the standard deviation is higher because the MOS differs from user to user, and so the confidence intervals.

Interestingly the mean ratings did not exceed a MOS of 4.16, even for the level 4 bandwidth (1150 Kbps). This result could be astonishing on first sight, but it is in line with other experience reports from QoE research (see section 7 for further discussion of this phenomenon).

Furthermore, we can see a similar representation of this results in the figure below in Figure 12, but in this case, comparing MOS to Bits per pixel (bits/(pixel*frame)) instead of bandwidth levels. The trend is the same as for the former case, but here we can evaluate better the dependence of the ratings with a more objective measure of the density of bits in every pixel of every frame.

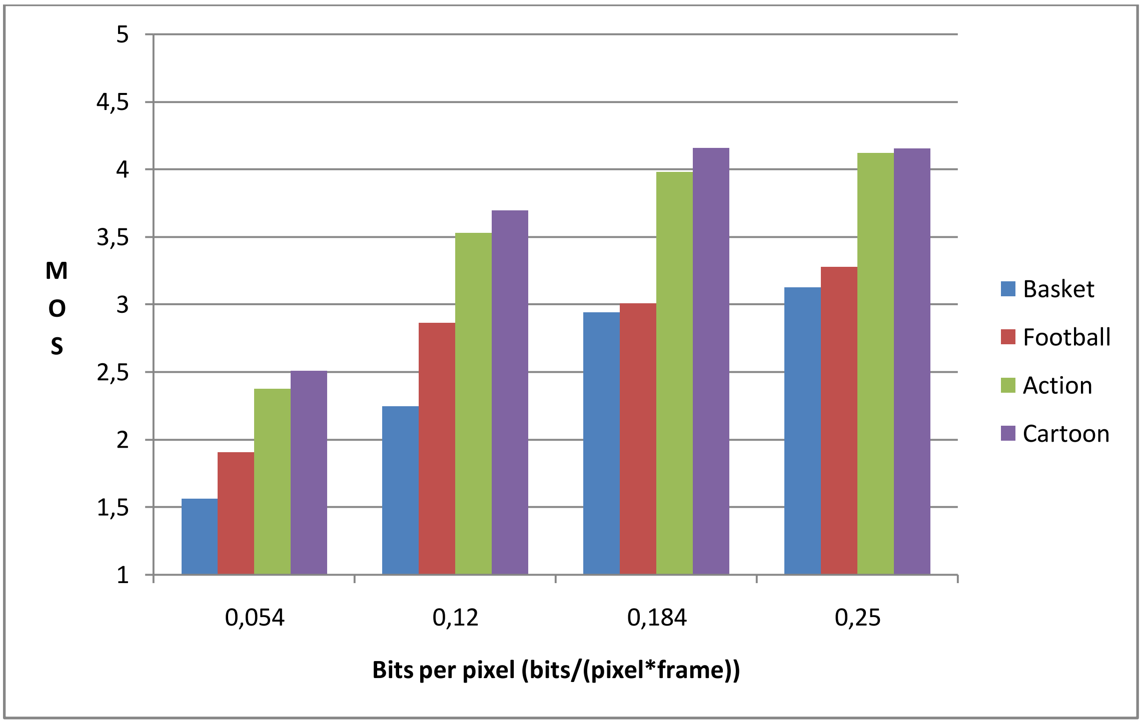

Figure 12: Quality rating results compared to "bits per pixel"

6.1.2 SOS (Standard deviation of Opinion Scores)

When performing Absolute Category Rating (ACR) we must consider possible deviation caused by the testers, such as memory effects, not knowing how to rate absolutely the quality, etc. Averaging the results (as MOS does) can erase some of this deviation, but SOS helps to describe the diversity of the users' ratings. [31]

Figure 13 shows the SOS in dependence of the MOS in our test. Each "experimental" point represents the (MOS, SOS) for every video sequence showed to the testers. The "theoretical" line represents a fitting function to evaluate the diversity of the users' ratings. This function is called the SOS Hypothesis and is defined as follows:

$$SOS(x)^2 = -ax^2 + 6ax - 5a = a\,(-x^2 + 6x - 5) \qquad \text{or}$$

$$SOS(x) = \sqrt{a(-x^2 + 6x - 5)}$$

Equation 4: SOS Hypothesis definition

, where '*x*' corresponds to MOS.

The equation is obtained by deriving the bounds of SOS, according to the 5-point scale defined. The SOS parameter "a" is obtained by minimizing the least squared errors between the measurement data and the fitting function. (Non-linear least squares with initial value a=-0.2)

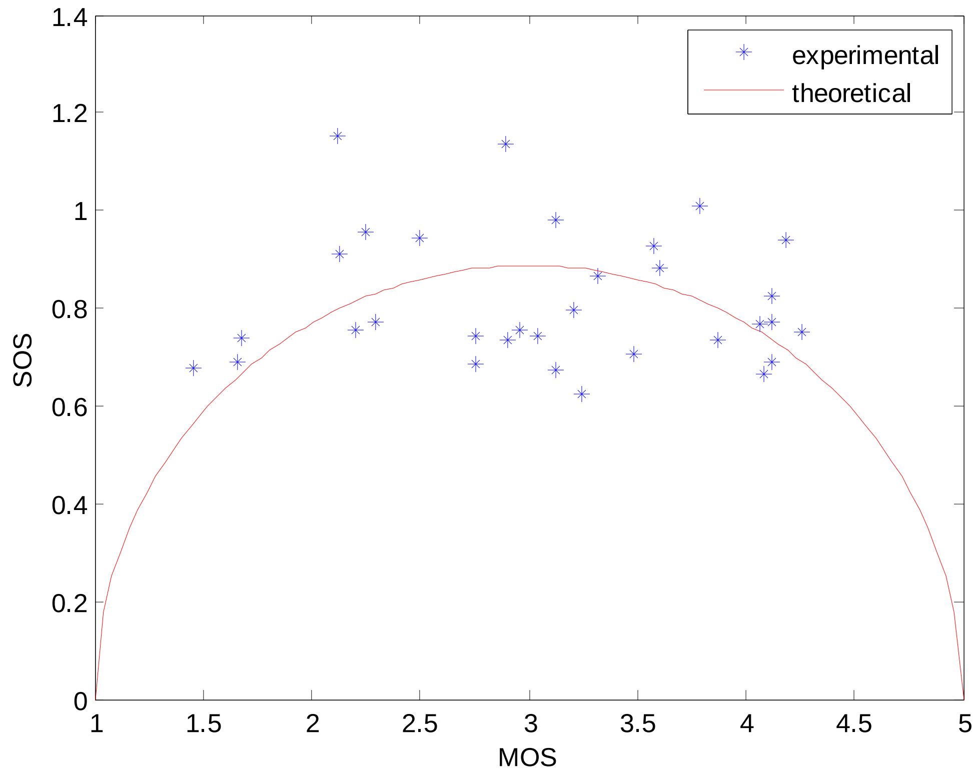

Figure 13: Diversity of the quality rating results for every sequence

.

As we can observe in Figure 13, experimental points are located close to the theoretical SOS fitting function. Here, the range around function represents the mean squared error (MSE) between the measurements and the fitting, in this case 0.0201. At the extremes, SOS obviously has to be zero, since a MOS of 1 and 5 can only be reached if all users rate 1 and 5, respectively.

The parameter "*a*" obtained in this case has a value of 0.197 which results into a mid-range user diversity. This value is consistent with video streaming applications, according to several other experiments which got similar results. [31]

We can conclude that, taking the figure and the parameter "*a*" in consideration, the SOS Hypothesis cannot be rejected for this experiment and the results are consistent. Therefore, the experimental results reflect the theoretical results for video streaming services.

6.1.3 User impressions

In this section some general impressions of the users during the test and in the final debriefing are collected.

Content classes

In general, users define basketball videos as the ones with the worst quality and in second place football sequences. Particularly they comment no heads nor faces are distinguished easily when the quality becomes lower, in this case they perceive the sequences as too "pixelled". This is possibly due to the impossibility represent so many individuals with a good quality as they appear in a team-sport match. Concerning football, the quality is rated better in close-ups, e.g. in a goal celebration.

The experiences with action and cartoon videos are much better in general, according to the comments, though cartoon is slightly higher ranked due to less perception of the distortion.

It is worth mentioning that, although users were asked to evaluate the overall quality of the sequences and not focusing in a specific parameter, cartoon and action videos have some specific shots were the 3D depth effect is intensified on purpose. This fact is impressive for user with a first or little contact with 3D technologies and can lead them to a false perception of better quality of these two content classes. This effect is not present in the sport (basketball and football) sequences.

Test length

Test length is pointed out by the users in its entirety as too long (~ 75 min.). The fatigue is identified due to repetition of the same videos. Some eye strain is experience by the users possibly because of the simulator sickness (see section 4.1).

Quality ratings

Some users commented their reticence towards rating a sequence with a 5.0 reserving it for 'even better' sequences that might come along later on (see an explanation of this phenomenon in section 7.1). This is the reason why in the second iteration the amount of "5.0's" is larger.

6.2 Automated assessment

The PSNR and SSIM analyses were performed, taking the respective state of the art videos as a reference. The analysis followed the procedure outlined in section 4.

The figures below present the SSIM, PSNR and values for each content class (averages of the two clips per content class).

6.2.1 Structural similarity metrics (SSIM)

As well as in the subjective evaluation of MOS, the differences between the SSIM values of the content types are significant. SSIM calculations for every bandwidth level always gave as a result the lowest values for basketball, and the highest for cartoon as shown in Figure 14. The order is, for the four levels, from lowest to highest: basketball, football, action, and cartoon.

SSIM values are very similar for levels 4 (1150 Kbps) and 3 (850 Kbps). The difference for every content class is close to 0.01. This bandwidth reduction only led to a degradation of about 0.01 SSIM points for every content class. Between levels 3 and 2 (550 Kbps) the degradation is about 0.02. The big difference is shown in level 1 (250 Kbps). For instance, if we compare it to level 2 the degradation is 0.04 SSIM points for cartoon, 0.05 for action and football, and 0.09 for basket.

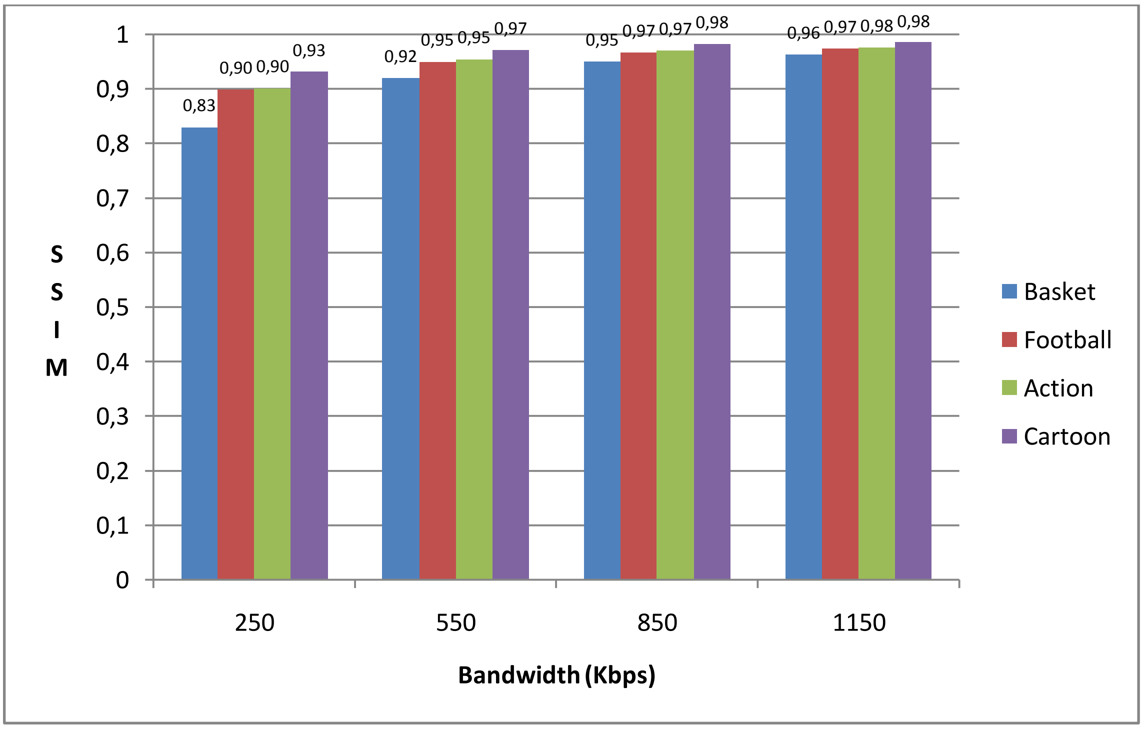

Figure 14: SSIM results for the four bandwidth levels and content classes

6.2.2 Peak Signal-to-Noise Ratio (PSNR)

PSNR results were all above 29 dB for whole sequence. This means that relative differences between the sequences with bandwidth reduction and the reference sequence were rather small.

As well as in the subjective evaluation of MOS and the objective SSIM, the differences between the PSNR values of the content types are significant. SSIM calculations for every bandwidth level always gave as a result the lowest values for basketball, and the highest for cartoon as shown in Figure 15. The order is, for the four levels, from lowest to highest: basketball, football, action, and cartoon.

In this case an almost constant-stepped shape is also shown, but for every bandwidth level and with the PSNR values increasing with it.

For example, the bandwidth reduction from level 4 to level 3 only led to a maximum degradation of 2.5 dB for basketball (from 35.5 dB to 33 dB) , 2.6 dB for football (from 39.9 dB to 37.3 dB) , 2.8 dB for action (from 42.2 dB to 39.4 dB) and 2.6 dB for cartoon (from 43 dB to 40.4 dB).

The degradation for every content class from one level to the previous ranges from 1.4 dB to 3.1 dB thus we can perceive there are not big differences, just a constant lowering trend.

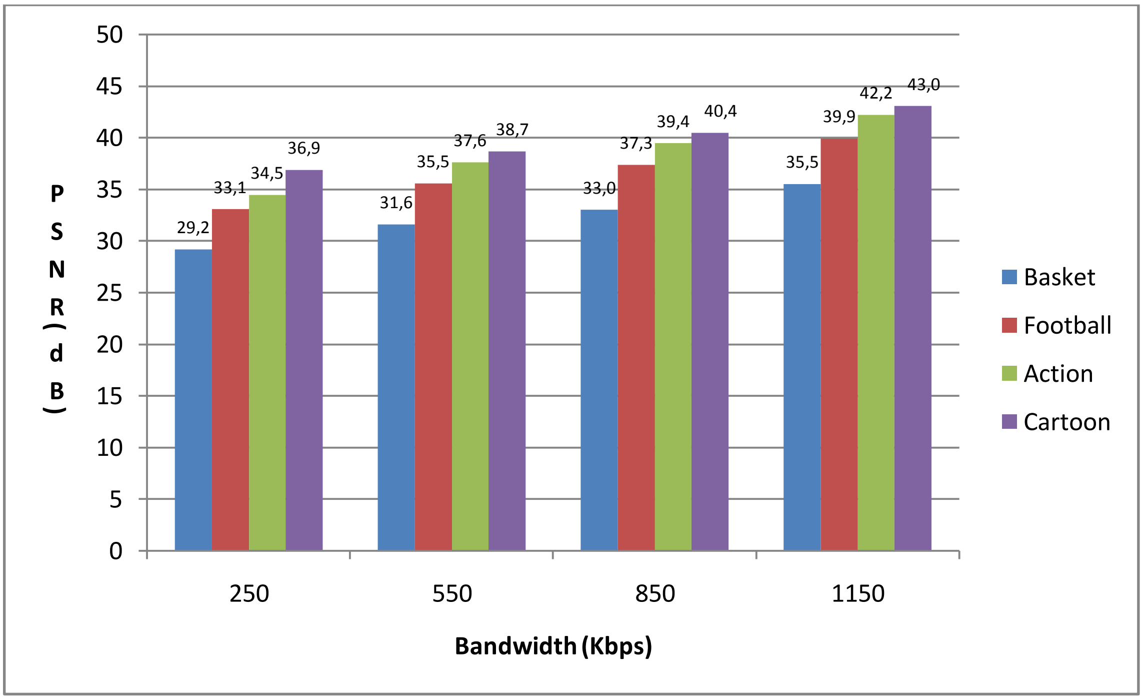

Figure 15: PSNR results for the four bandwidth levels and content classes

6.3 Correlation of user-based and automated testing results

Figure 16 and Figure 17 show the correlations of automated assessment metrics results (SSIM and PSNR) with user-based quality rating results (MOS).

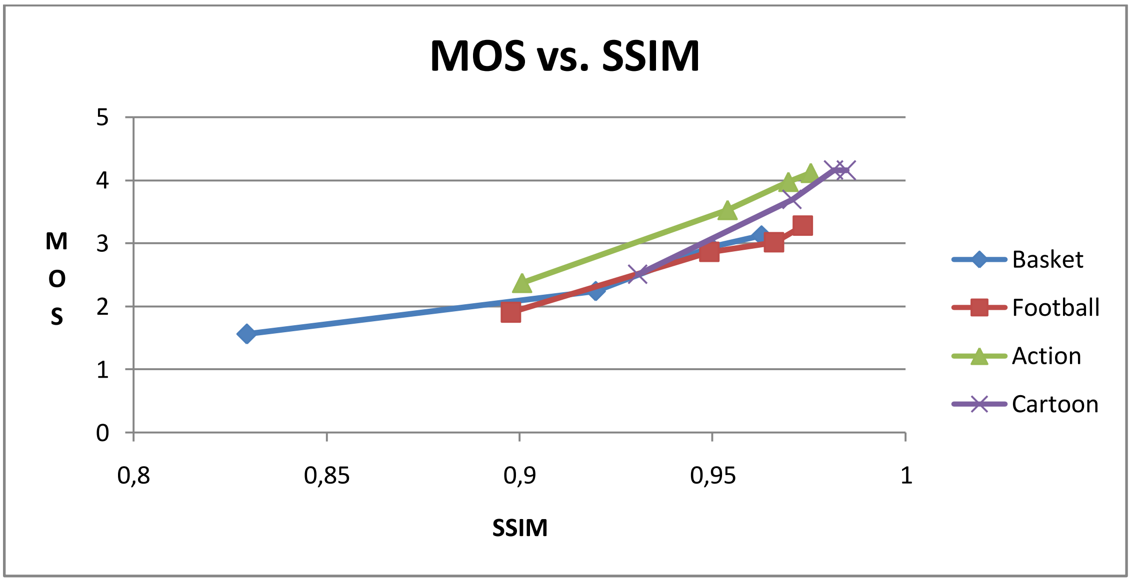

Figure 16: Correlations of the automated assessment metric SSIM results with user-based assessment results (MOS) for the four content classes

.

Content class	Basket	Football	Action	Cartoon	Overall
Pearson coefficient (r)	0,96695398	0,99347874	0,99849887	0,99699549	0,89330846

Table 7: Pearson correlation coefficient for SSIM and MOS for every CC

The correlations of the user-based assessment (MOS) versus SSIM and PSNR were quite high, ranging between 0.967 and 0.998. SSIM had the highest correlations: 0.998 and 0.997, respectively (Table 7). PSNR correlation (Table 8) was lower for every content class but for basketball. In this case it was a slightly higher, with 0.9670 for SSIM and 0.9695 PSNR. Little differences between content types were manifested in both analyses: basketball contents (r<0.97) were less correlated than the rest of the contents (r>0.98).

Concerning the overall correlations also we obtain high values (0.893 for SSIM and 0.966 for PSNR), but we get a better adjustment to the user results with PSNR.

Therefore it is important to note that due to the wide range of quality of our sequences (from 1.56 to 4.16 MOS) the correlation showed in the tables and figures above reveal the predictive power of automated measurement metrics.

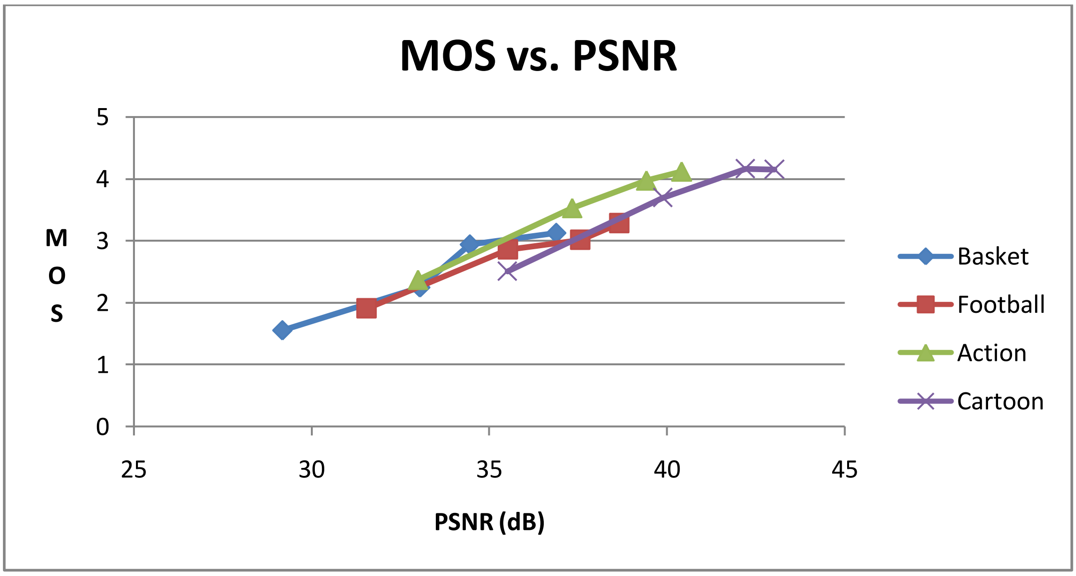

Figure 17: Correlations of the automated assessment metric PSNR results with user-based assessment results (MOS) for the four content classes

Content class	Basket	Football	Action	Cartoon	Overall
Pearson coefficient (r)	0,96953597	0,98437798	0,99649385	0,9884331	0,96540147

Table 8: Pearson correlation coefficient for PSNR and MOS for every CC

6.4 Correlation of user-based assessment and bits per pixel

In Figure 18 the correlation of MOS and bits per pixel for every content class is presented.

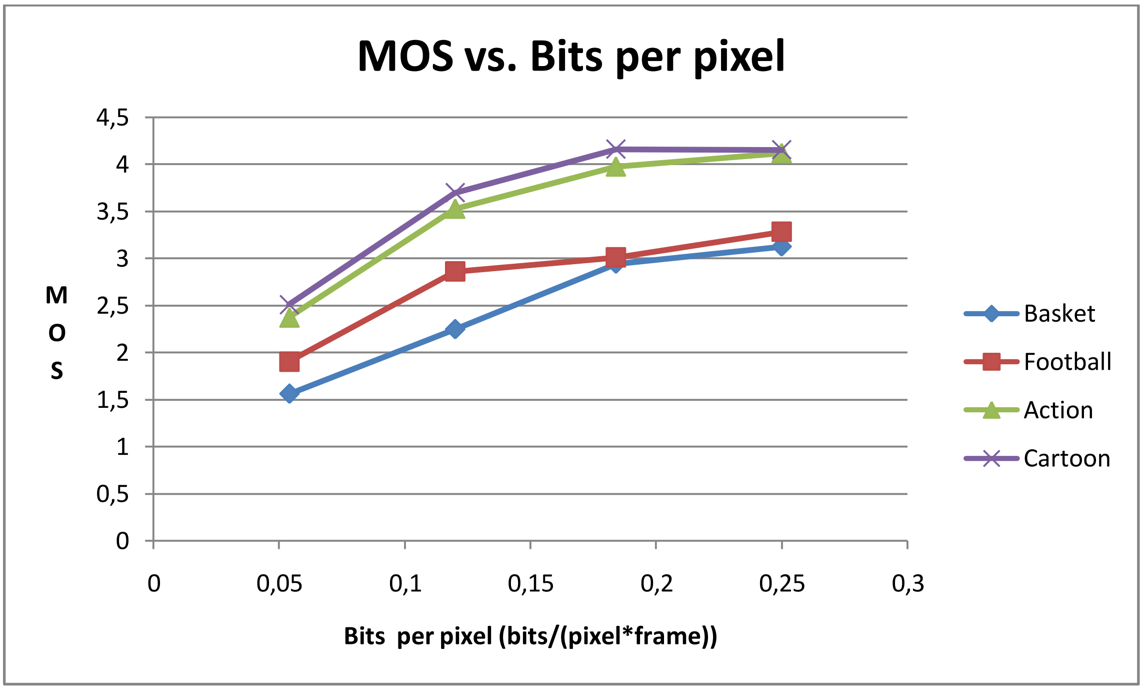

Figure 18: Correlations of bits per pixel and user-based assessment results (MOS) for the four content classes

Content class	Basket	Football	Action	Cartoon	Overall
Pearson coefficient (r)	0,97365292	0,92357999	0,92736185	0,89554453	0,727979

Table 9: Pearson correlation coefficient for bits per pixel and MOS for every CC

If we have a look at the values of the Pearson correlation coefficient in Table 9, we can notice a reduction compared to the results obtained with SSIM and PSNR for every content class but for basketball, with a similar value. Furthermore, the trend is also different, with basketball obtaining the higher value, then football, action and cartoon in descending order. But the real difference is obtained when calculating the overall Pearson correlation coefficient. In this case we get r=0,73, which indicates a weaker correlation than for MOS vs. SSIM (r=0,89)(see Table 8) and MOS vs. PSNR (r=0,97) (see Table 7).

This lower value of the Pearson correlation coefficient shows that the subjective perception of quality is not necessary fully dependant on the bit resolution.

7. Conclusions

In this thesis the quality assessment of a 3D mobile video service is expounded, both subjective and objective quality, also known as user-based and automated assessment. The possible solutions, in terms of existing methods, are considerable, though standards about this topic are still being developed, especially concerning automated assessment.

As an initial step, it was necessary to define the usage scenario and to design the setup for subjective assessments. Then reference-free estimation methods for the mobile scenario were investigated and proposed.

The proposed estimation models for the user-based assessment are focused at reference-free quality estimation (using the Absolute Category Rating method, see section 4.3). The 3D mobile video service scenario reflects an environment of usage, user equipment and typical types of video content. For this purpose mobile scenarios and a test methodology were investigated and defined in order to achieve the best emulation of the real world scenario, covering the most frequent content classes (CCs).

In addition, the estimation models for the automated assessment are based on metrics widely used in 2D conditions, as it has been validated its predicting potential of 3D video quality (see section 5) This fact is related to the current situation, where developing a single quality metric for 3D video still remains to be a big challenge.

In the following, we summarize and interpret the main results from the user-based assessment (section 7.1) and the automated assessment (section 7.2) as well as the correlation between both and with bits per pixel (section 7.3).

7.1 User-based assessment

According to our results we did not find any significant perceptual quality degradations between bandwidth levels 4 (1150 Kbps) and 3 (850 Kbps) (see Figure 11 and Table 6). The difference for every content class is lower than 0.2 MOS points for about 25% bandwidth reduction. Even so, we must remark our results for this absolute category rating are very content dependant. In this context we observe very acceptable results for the content classes cartoon and action (≈ 4 MOS points) for both levels mentioned above. We can also include football in this group (≈ 3.6 MOS points), but for the rest bandwidth levels for this three content classes and basketball for every bandwidth level, the MOS points reached are very disappointing according to quality from a subjective point of view.

We would like to stress that our obtained MOS scores can be regarded as considerably valid, because we achieved small confidence intervals, which suggests a high consensus among heterogeneous participant backgrounds and viewing experiences from 2D to 3D videos.

Quality rating scores

As noted in the previous sections, mean subjective ratings in the quality test only seldom surpassed the 4.0 mark, even for the highest bandwidth level. An important factor contributing to the limitation to a maximum mean rating score of 4.16 is probably the rating paradigm itself [32]. For example, also in previous research it has become evident that mean quality ratings actually never reach the maximum of 5.0 MOS, even when the unique original files are presented (subjects hesitate to indicate that a certain video presentation has "perfect" quality).

Absolute judgments are generally difficult for humans, and especially in the quickly evolving domain of home multimedia people may be unsure about which sequence should be qualified as excellent. Actually, many implicit participant motivations during rating studies remain to be of comparative nature, even if subjects are repeatedly instructed to make absolute judgments. For instance, while attempting to maintain consistency throughout their ratings, subjects often may reserve a '5.0' for 'even better' sequences that might come along later on in a test session. Even the situation of having to provide many successive ratings may be an implicit suggestion to participants to vary their judgments, so as to avoid giving the same rating values over and over again.

7.2 Automated assessment

The results obtained for SSIM index evaluation are similar to the user-based for high bandwidth levels. In this case, the evaluations for levels 4 (1150 Kbps) and 3 (850 Kbps) do not indicate any significant quality degradations, as the difference, with a 25% bandwidth reduction is 1% or even less (see Figure 14). But this trend continues if we compare levels 3 (850 Kbps) and 2 (550 Kbps), where the gap is just 2 or 3%, that means a reduction of 3-4% of overall quality with about 50% of the bandwidth. In the comparison of levels 2 (550 Kbps) and 1 (250 Kbps) we detect a bigger decrease, specially for the basketball content class. We conclude that this metric predicts an acceptable quality for a reduction of almost 50% of the bandwidth for all the content classes.

In the PSNR evaluation the results show a big difference among the content classes (see Figure 15). The reduction of 2 or 3 dB is constant for every bandwidth level and content class, and the difference between cartoon and basketball is for every level about 7 dB. Even so, the calculated PSNR reaches more than 30dB for every content class and bandwidth level combination (but for basketball@250Kbps with 29.2 dB), what points out an acceptable quality.

7.3 Correlations

The correlations between automated and user-based assessment are significantly high for MOS vs. SSIM (ranging from r=0.967 to r=0.998) (see Figure 16 and Table 7), and also for MOS vs. PSNR (ranging from r=0.970 to r=0.996) (see Figure 17 and Table 8) concerning the four different content. According to overall correlation the highest

Pearson coefficient is 0.965 for MOS vs. SSIM (see Table 8), meanwhile for MOS vs. SSIM we obtain r=0.893 (see Table 7).

The results show strong correlation between subjective and objective assessment and, as remarked in section 5, both metrics reveal a good predicting performance in our scenario of quality assessment of a 3D mobile video service.

In the case of the correlation between the user-based assessment and bits per pixel (bits/pixel*frame), the results obtained are quite different as they are lower for almost every content class, but specially important is the fact that the overall Pearson correlation coefficient only takes a value of 0.73, rather less than for MOS vs. SSIM and MOS vs. PSNR. In this way, we can deduce that the bit resolution is not a decisive factor in the subjective evaluation of the stereoscopic sequences in our 3D mobile video service scenario.

8. Bibliography

[1] Irmer, R.; Mayer, H.-p.; Weber, A.; Braun, V.; Schmidt, M.; Ohm, M.; Ahr, N.; Zoch, A.; Jandura, C.; Marsch, P.; Fettweis, G.; , "Multisite field trial for LTE and advanced concepts," *Communications Magazine, IEEE* , vol.47, no.2, pp.92-98, February 2009 doi: 10.1109/MCOM.2009.4785385
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4785385&isnumber=4785366

[2] Thompson, C.W.; , "Next-Generation Virtual Worlds: Architecture, Status, and Directions," *Internet Computing, IEEE* , vol.15, no.1, pp.60-65, Jan.-Feb. 2011 doi: 10.1109/MIC.2011.15
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5676166&isnumber=5676153

[3] Gotchev, A.; Akar, G.B.; Capin, T.; Strohmeier, D.; Boev, A.; , "Three-Dimensional Media for Mobile Devices," *Proceedings of the IEEE* , vol.99, no.4, pp.708-741, April 2011 doi: 10.1109/JPROC.2010.2103290
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5725138&isnumber=5733920

[4] Chen, Y.-K.; Chakrabarti, C.; Bhattacharyya, S; Bougard,B; , "Signal processing on platforms with multiple cores - Part 1: Overview and methodologies, " *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 24–25, Nov. 2009.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5230800

[5] Hur, N.; Hyun Lee; Gwang Soon Lee; Sang Jin Lee; Gotchev, A.; Sang-Il Park; , "3DTV Broadcasting and Distribution Systems," *Broadcasting, IEEE Transactions on* , vol.57, no.2, pp.395-407, June 2011 doi: 10.1109/TBC.2011.2114710
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5741868&isnumber=5772155

[6] Dodgson, N.A.; , "Autostereoscopic 3D Displays," *Computer*_, vol.38, no.8, pp. 31- 36, Aug. 2005
doi: 10.1109/MC.2005.252
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1492263&isnumber=32069

[7] Smolic, A.; Mueller, K.; Merkle, P.; Kauff, P.; Wiegand, T.; , "An overview of available and emerging 3D video formats and depth enhanced stereo as efficient generic solution," *Picture Coding Symposium, 2009. PCS 2009* , vol., no., pp.1-4, 6-8 May 2009 doi: 10.1109/PCS.2009.5167358
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5167358&isnumber=5167340

[8] ISO/IEC JTC1/SC29/WG11, "ISO/IEC CD 23002-3: Representation of auxiliary video and supplemental information", Doc. N8259, Klagenfurt, Austria, July 2007.

[9] Muller, K.; Smolic, A.; Dix, K.; Kauff, P.; Wiegand, T.; , "Reliability-based generation and view synthesis in layered depth video," *Multimedia Signal Processing, 2008 IEEE 10th Workshop on* , vol., no.,pp.34-39,8-10 Oct.2008 doi:10.1109/MMSP.2008.4665045

URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4665045&isnumber=4665023

[10]Merkle, P.; Brust, H., Dix, K; Wang, Y; Smolic, A; , “Adaptation and optimization of coding algorithms for mobile 3DTV”, *MOBILE3DTV*, Project no. 216503 D2.2 URL: http://sp.cs.tut.fi/mobile3dtv/results/tech/D2.2_Mobile3DTV_v1.0.pdf

[11]ITU-T Recommendation H.264, “Advanced video coding for generic audiovisual services”, Jun. 2011.

[12]ISO/IEC JTC1/SC29/WG11, “Text of ISO/IEC 14496-10:200X/FDAM 1 Multiview Video Coding”, Doc. N9978, Hannover, Germany, July 2008.

[13]Engelke, U.; Zepernick, H.-J.; , "Perceptual-based Quality Metrics for Image and Video Services: A Survey," *Next Generation Internet Networks, 3rd EuroNGI Conference on* , vol., no., pp.190-197, 21-23 May 2007 doi: 10.1109/NGI.2007.371215
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4231838&isnumber=4231801

[14]ITU-R, Recommendation BT.1438, “Subjective assessment of stereoscopic television pictures”, 2000

[15]Mira D.;, "A Survey on Network QoS Needs of Advanced Internet Applications, " University College London. November 2002.
URL: http://havnet.stanford.edu/pdfs/b.Measuring%20the%20quality%20of%20video%202002.pdf

[16]ITU-T, Rec. P.910, "Subjective video quality assessment methods for multimedia applications," Sept. 1999.

[17]Yadavalli, G.; Masry, M.; Hemami, S.S.; , "Frame rate preferences in low bit rate video," *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on* , vol.1, no., pp. I- 441-4 vol.1, 14-17 Sept. 2003 doi: 10.1109/ICIP.2003.1246993
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1246993&isnumber=27937

[18]http://www.roxio.com/enu/company/3d/survey.html

[19]http://blog.nielsen.com/nielsenwire/consumer/purchase-intent-for-3dtv-varies-around-the-globe/

[20]Quan Huynh-Thu; Barkowsky, M.; Le Callet, P.; , "The Importance of Visual Attention in Improving the 3D-TV Viewing Experience: Overview and New Perspectives," *Broadcasting, IEEE Transactions on* , vol.57, no.2, pp.421-431, June 2011 doi: 10.1109/TBC.2011.2128250
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5752215&isnumber=5772155

[21]ITU-R Study Group 6, "Digital three-dimensional (3D) TV broadcasting," Question ITU-R 128/6, 2008.

[22]ITU-T Study Group 9, "Objective and subjective methods for evaluating perceptual audiovisual quality in multimedia services within the terms of Study Group 9," Question 12/9, 2009.

[23]Jumisko-Pyykkö, S., Weitzel, M., Strohmeier, D., "Designing for user experience - What to expect from mobile 3D television and video", *1st Int. Conf. Des. Interactive User Experiences TV Video*, Mountain View, CA, Oct. 2008, pp. 183–192. URL:http://sp.cs.tut.fi/mobile3dtv/results/conference/5_FINAL_JumiskoPyykko_et_al_DesigningForUserExperience_UXTV_Oct08.pdf

[24]http://www.htcdev.com/devcenter/s3d

[25]http://www.htc.com/www/smartphones/htc-evo-3d/#specs

[26]http://compression.ru/video/quality_measure/video_measurement_tool_en.html

[27]Yasakethu, S.L.P.; Hewage, C.; Fernando, W.; Kondoz, A.; , "Quality analysis for 3D video using 2D video quality models," *Consumer Electronics, IEEE Transactions on* , vol.54, no.4, pp.1969-1976, November 2008 doi: 10.1109/TCE.2008.4711260
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4711260&isnumber=4711185

[28]Zhou Wang; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P.; , "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on* , vol.13, no.4, pp.600-612, April 2004 doi: 10.1109/TIP.2003.819861
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1284395&isnumber=28667

[29]http://compression.ru/video/quality_measure/info_en.html#start

[30]Ries, M.; Nemethova, O.; Rupp, M. ; , "Video Quality Estimation for Mobile H.264/AVC Video Streaming," *Journal of communications*, vol. 3, no1, January 2008
URL: http://www.ojs.academypublisher.com/index.php/jcm/article/download/03014150/79

[31]Hobfeld, T.; Schatz, R.; Egger, S.; , "SOS: The MOS is not enough!," *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop* on , vol., no., pp.131-136, 7-9 Sept. 2011
doi:10.1109/QoMEX.2011.6065690
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6065690&isnumber=6065682

[32]Fröhlich, P.; Ries, M; Fuchs, M; Stürmer, T; Masuch, K; Schatz, R.; , "Optimization of Bandwidth for IPTV Video Streaming" Optiband D2.4, Project number: 248495, Jan. 2011.
URL: http://www.optiband-project.eu/content/files/OptiBand_D2_4_FTW_31-10-2011_FINAL.pdf

Appendix A. Questionnaire

In this section, a copy of the questionnaire used for the evaluation by the test users is attached.

As it was commented in section 4.4, first of all some demographic data and general user behavior is collected. Name was only used to associate a test number with the user, in case of needing them for repeating the evaluation because of an error. This situation did not happen and no evaluation had to be repeated.

After the general data, the next sheets are for the evaluation itself, with the two evaluations as described in section 4.5.

Test Number: Round: ☐ 1st ☐ 2nd Date:

- **Name:** ______________________________

- **Age:** ______________________________

- **Sex:**

  ☐ female ☐ male

- **Nationality:** ______________________________

- **Highest educational degree:**

  ☐ undergraduate ☐ graduate ☐ doctoral

- **Professional status:**

  ☐ employed ☐ unemployed ☐ student ☐ retired

- **What kind of videos do you watch on your mobile phone?**

  ☐ online videos ☐ offline videos (films, TV series,…)

- **Have you ever watched 3D video?**

  ☐ Yes ☐ No

- **How long do you spend watching videos on your phone a day?**

  ____

- **Have you got some experience in image processing (films, photos, etc)?**

  ☐ Yes ☐ No

Test Number: **Round:** ☐ 1st ☐ 2nd **Date:**

TU WIEN

Clip number	**Evaluation**				
1.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
2.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
3.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
4.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
5.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
6.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
7.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
8.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number: Round: ☐ 1st ☐ 2nd Date:

Clip number	Evaluation				
9.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
10.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
11.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
12.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
13.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
14.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
15.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
16.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number: Round: ☐ 1st ☐ 2nd Date:

Clip number	Evaluation				
17.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
18.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
19.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
20.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
21.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
22.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
23.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
24.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number:	Round: ☐ 1st ☐ 2nd	Date:

Clip number	Evaluation				
25.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
26.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
27.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
28.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
29.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
30.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
31.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
32.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number: Round: ☐ 1st ☐ 2nd Date:

Clip number	Evaluation				
33.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
34.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
35.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number:	Round: ☐ 1st ☐ 2nd	Date:

Clip number	Evaluation				
1.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
2.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
3.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
4.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
5.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
6.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
7.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
8.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number: **Round:** ☐ 1st ☐ 2nd **Date:**

TU WIEN

Clip number	**Evaluation**				
9.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
10.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
11.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
12.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
13.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
14.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
15.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
16.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number: **Round:** ☐ 1st ☐ 2nd **Date:**

TU WIEN

Clip number	**Evaluation**				
17.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
18.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
19.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
20.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
21.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
22.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
23.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
24.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Test Number:	Round: ☐ 1st ☐ 2nd	Date:

Clip number	Evaluation				
25.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
26.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
27.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
28.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
29.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
30.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
31.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent
32.	1 Bad	2 Poor	3 Fair	4 Good	5 Excellent

Clip number	Evaluation				
33.	1	2	3	4	5
	Bad	Poor	Fair	Good	Excellent
34.	1	2	3	4	5
	Bad	Poor	Fair	Good	Excellent
35.	1	2	3	4	5
	Bad	Poor	Fair	Good	Excellent

Appendix B. List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
3DTV	Three-Dimensional Television
ACR	Absolute Category Rating
AVC	Advanced Video Coding
AVSI	Auxiliary Video Supplemental Information
BS	Bit-Stream
CBR	Constant Bit Rate
CC	Content Class
CI	Confidence Interval
CSV	Conventional Stereo Video
DSCQS	Double-Stimulus Continuous Quality-Scale
DSIS	Double-Stimulus Impairment Scale
DVB-H	Digital Video Broadcasting-Handheld
FPA	Frame Packing Arrangement
ITU	International Telecommunication Union
ITU-R	International Telecommunication Union - Radio
ITU-T	International Telecommunication Union - Telecommunications
LCD	Liquid Crystal Display
LDV	Layered Depth Video
LTE	Long Term Evolution
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MSE	Mean Squared Error
MVC	Multiview Video Coding
MVD	Multiview Video plus Depth

PSNR	Peak Signal-to-Noise Ratio
QoE	Quality of Experience
SEI	Supplemental Enhancement Information
SOS	Standard deviation of Opinion Scores
SSIM	Structural Similarity
SVC	Scalable Video Coding
T-DMB	Terrestrial Digital Multimedia Broadcasting
TS	Transport-Stream
V+D	Video plus Depth
VGA	Video Graphics Array
VUI	Video Usability Information