OptiBand

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

Deliverable reference number: D2.4
Deliverable title: Intermediate QoE research recommendations report

Due date of deliverable: M22
Actual submission date: 31/10/2011

Start date of project: 1 January 2010 Duration: 30 months
Organisation name of lead contractor for this deliverable: FTW
Authors: Peter Fröhlich, Michal Ries, Matthias Fuchs, Theresa Stürmer, Kathrin Masuch & Raimund Schatz.

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

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

Table of Tables

Executive Summary

Deliverable D2.4 (“Intermediate QoE research report”) presents the second Quality of Experience (QoE) testing iteration within the OptiBand project T2.3. Its goals and methodology are based on D2.1 (“Criteria specification for the QoE research”) and D2.2 (“Detailed research plan”). The main goal of this QoE testing iteration was to test whether the bandwidth of HD IPTV can be reduced without perceivable quality losses.

Section 1 outlines the main aspects of interest, which are (1) the packet dropping scenario (1080i multirate and 1080p SVC), (2) the bandwidth reduction levels and (3) the content classes. Section 2 gives an overview of the video contents that had been used for testing, section 3 outlines the user-based assessment and section 4 the automated assessment methodology. Section 5 then provides the assessment and section 0 closes with some general conclusions and an outlook to further ongoing and future activities of WP2.

Relative degradations of quality of experience

The main result is that, at least with laboratory testing of offline encoded files, the OptiBand project has even exceeded its goal to reduce bandwidth by 33% while preserving acceptable perceived quality. The related results from user-based assessment are as follows:

- We did not find any significant perceptual quality degradations between the reference content and sequences with up to almost 40% bandwidth reduction.
- For even stronger bandwidth reductions of between 40% and almost 60%, the MOS ratings were only about 0.3 points lower than for the reference.
- It is worthwhile to note that this success applies for both coding scenarios selected for further investigation in the OptiBand project: 1080i multirate and 1080p SVC.
- The user-based acceptability testing results strengthen this picture. In the acceptance rating tests, no significant difference was found between the source content and any degraded sequence.

Also the automated assessment results indicate that the investigated OptiBand packet dropping approaches are suitable for reducing bandwidth while still preserving acceptable perceptual quality.

- All three metrics (SSIM, PSNR, and J.341) detected very small declines of bandwidth reduction on perceived quality. This especially applies for SSIM, where differences are below 1%.
- Also PSNR values hardly reveal any differences, except for the strong gaps, which were caused by identical frames and clipping PSNR-clipping in case of identical frames (this phenomenon is discussed in the report).
- If only the 1080i scenario is considered, i.e. the scenario that will be pursued in the OptiBand integrated prototype and live system, the J.341 metric correlated best with subjective quality ratings (0.97 for soccer, and 0.98 for action). However, for the estimated MOS lower absolute values were found than for the actual MOS from the user-based quality ratings.

Absolute quality

In this second iteration, the tested sequences overall attained higher absolute quality ratings than in the first iteration:

- For the 1080i multirate scenario, all tested sequences surpassed the MOS threshold of 3.7 was surpassed for bandwidth reductions of up to 50%.
- In the 1080p SVC scenario, although for the action content the tests were successfully passed, only a bandwidth reduction of 23% was possible. This effect can be well explained by the source contents that had to be used for testing in the 1080p SVC scenario.
- The acceptance ratings were in general very high: only 5-10% of the sequences were qualified inacceptable by the participants. The only exception here is soccer content in the 1080p SVC scenario, where already the state-of-the-art sequence was rated as inacceptable by 20% of the participants, which may be explained again by the quality of the reference source content that had to be used.

Further results and activities

- The results suggest that defined thresholds (MOS, SSIM, PSNR) should remain the same. For the estimated MOS of ITU J.341 we propose a threshold of 3.0, and for user-based acceptance ratings an acceptance ratio of 85%.
- As proposed in Annex 2 of D2.2, a study on the impact of content durations in QoE testing is being conducted and will be available as an Annex to D2.4. This study also includes results for on a third additional content type ‘Documentary”, which will also be featured during the Live Tests.

1. Focus of interest

In the following, the main variables that we focus on in the OptiBand QoE research are described, in order to understand their impact on QoE. The variety of test conditions is explained in the following, namely the packet dropping scenarios, the bandwidth levels, and the content classes. The variables defined below are the same as in the description of the overall method (see D2.2 [12]), and specifications of bandwidth levels and video materials are only marginally differing from the first test iteration (D2.3 [13]). For better readability, we include the whole description also in the present deliverable, even if we are risking some redundancy with those previous documents.

1.1 Packet dropping scenarios

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

To cover the most relevant implementation possibilities, we systematically evaluate packet dropping within different “encoding scenarios”. These had been defined in D2.2, Section 3.3, Table 3. In the first iteration, evaluations were made for the following two packet dropping scenarios

- 1080i (multirate video streaming)
- 1080p (SVC)

For the 1080i scenario, two partners, OPT and TVN, provided a set of test sequences each. For the 1080p scenario, HHI provided test sequences. The detailed specification of the packet dropping algorithms under evaluation is provided in D3.4.1 (IPTV data dropping algorithm prototype version 1 and associated documentation).

Scenario	**1080i multirate video streaming**	**1080p SVC**	
Partner	OPT + TVN	HHI	
Content	action movie / soccer	Soccer	action movie
Max. video BR	7 ÷ 8.5 Mbps	8.5 Mbps	7 ÷ 8.5 Mbps
Video output format	1080i	1080p	1080p
Frame rate	25 fps (50 fps interlaced)	50 fps	25 fps
Aspect ratio	16:9	16:9	16:9
GOP structure	IBBP	IB..BP	IB..BP
GOP length	32	8 (Intra frame every 48 frames)	8 (Intra frame every 24 frames)
Video mode	CBR	VBR	VBR

Table 1: Encoding parameters to be consistently applied to the test video sequences

1.2 Bandwidth levels

Due to its importance for the project, bandwidth is also varied systematically. The bandwidth is defined as average bandwidth including all necessary transport data for correct decoding of the video stream (e.g., video payload, signalization, and encapsulation format). For the OptiBand project (see the agreed video format and mode settings in Table 5), the following bandwidth levels have been specified in D2.2:

Bandwidth level	**Relative bandwidth reduction, as compared to the SOTA**	**Bandwidth (Mbps)**		
		Min	Mean	Max
7. SOTA	0%	7.5	7.8	8.5
6.	10%	6.9	7.0	7.1
5.	20%	6.1	6.2	6.3
4.	**33% - OptiBand project target**	5.1	5.2	5.3
3.	40%	4.6	4.7	4.8
2.	50%	3.8	3.9	4.0
1.	55%	3.4	3.5	3.6

Table 2: Bandwidth level conditions and operation points specified for the OptiBand QoE evaluation (from D2.2)

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

Table 3 and Table 4 show the actual properties of the delivered video files for the 1080p SVC and 1080i multirate scenarios, respectively. Please note that especially in the SVC scenario the bandwidth levels were derived from the spatial, temporal and quality scalability features of the specific MGS encoded test sequences (please refer to D4.1). Seven SVC operation points, corresponding to the bandwidth levels defined above, were provided which cover more than the required range of bandwidth levels. Due to the SVC encoding structure, the operation points could not exactly match each of the originally specified bandwidth levels.

Bandwidth level	Operation point	Bandwidth (MBbps)			Relative bandwidth reduction, as compared to the SOTA (derived from minimum bandwidth)
		Min	Mean	Max	
7. SOTA	0	8.2	8.3	8.5	0%
6.	1	6.2	7.1	8.1	15%
5.	2	5.2	6.4	7.4	23%
4.	3	4.6	5.9	7.1	30%
3.	4	4.0	4.5	4.9	46%
2.	8	3.6	4.2	4.8	50%
1.	13	2.8	3.5	4.3	58%

Table 3: Actual test file properties as delivered for the 1080p SVC scenario

Bandwidth level	Bandwidth (MBbps)			Relative bandwidth reduction, as compared to the SOTA (derived from minimum bandwidth)
	Min	Mean	Max	
7. SOTA	7.5	7.8	8.5	0%
6.	6.9	7.0	7.1	8%
5.	6.1	6.2	6.3	19%
4.	5.1	5.2	5.3	32%
3.	4.6	4.7	4.8	39%
2.	3.8	3.9	4.0	49%
1.	3.4	3.5	3.6	55%

Table 4: Actual test file properties as delivered for the 1080i multirate scenario

1.3 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 HD IPTV consumption situation. To support this motivation, Telecom Italia performed own market research and the available popularity ratings were investigated (see D2.1 [11], section 2.2.1). The results of this survey reflect the sport and film genre dominance within the available HD channels in Europe.

Therefore, we selected the following content classes for focused investigation in OptiBand QoE research (WP2):

- Action movie (CC1)
- Soccer (CC2)

For these two content classes also HD resolution provides the most significant benefit from the point of view of perceived quality. Action movie and soccer are classes with different levels of detail, complexity of structures, and movements. The content classes soccer and action movie are the most sensitive sequences.

Figure 1: Snapshots of Soccer (left) and Action movie (right) sequences.

Note that in the upcoming extra study on content durations as well as in the Live Test (WP8) a third content type will be introduced: ‘documentary’.

2. 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:

- 2 coding/packet-dropping scenarios (multirate and SVC approaches)
- 7 bandwidth levels (H.264 SOTA reference sequence + 6 degraded sequences)
- 2 content classes (action and soccer)

As stated in section 1.1, for the 1080i scenario test sequences one set of test sequences was delivered for each of the two approaches: one by OPT and another one by TVN. For the 1080p scenario the necessary sequences were provided by HHI. This means that three sets of video materials were available.

For each content class, representative video clips were selected. The rationale behind choosing to use action and soccer content is that they are representative for typical HD IPTV consumption situations in Europe, not the usually used test video sequences that are either unnatural (e.g., a riverbed or police boat) or coming from a US-American context (e.g. American football or basketball matches).

Of course, the video sources used for QoE testing should remain consistent throughout the three OptiBand QoE testing iterations, in order to attain perfect comparability of results between iterations. However, as noted in D2.3 (section 6.3), we decided to relax this principle and to search for full HD soccer content sources. The hitherto used content had been up-scaled from 720p to 1080p and its limitations in quality were potentially responsible for comparatively low absolute subjective quality results.

We were successful in finding highly suitable HD soccer source content, but it could only be used by the partners working with 1080i format. Under these conditions, we decided to exploit the availability of these new soccer contents in the second iteration evaluation only for the 1080i multirate scenario. This means that the soccer contents from the first iteration, which provided limited quality potentially responsible for the low subjective quality results, had to be kept for the 1080p SVC scenario, due to the lack of available content.

For the resulting three video source material sets (action, new soccer and old soccer) we generated three 10-seconds sequences each. The spatio-temporal characteristics (as defined in ITU-T Recommendation P.910) for these sequences are shown in (see an introduction into spatial and temporal information in D2.1, section 3.1.3). Figure 2 (next page) provides an overview of the resulting 84 combinations.

Scenario	Content class	Clip	Spatial information	Temporal information
1080p	CC1 (action)	Clip A	46,1989	57,7716
1080p	CC1 (action)	Clip B	51,5644	69,4213
1080p	CC1 (action)	Clip C	59,0294	61,5609
1080p	CC2 (soccer old)	Clip A	128,2129	32,9140
1080p	CC2 (soccer old)	Clip B	110,5319	59,3168
1080p	CC2 (soccer old)	Clip C	112,7072	27,9660
1080i	CC1 (action)	Clip A	46,1989	57,7716
1080i	CC1 (action)	Clip B	51,5644	69,4213
1080i	CC1 (action)	Clip C	59,0294	61,5609
1080i	CC2 (soccer)	Clip A	92.5785	46.9622
1080i	CC2 (soccer)	Clip B	99.4217	42.0293
1080i	CC2 (soccer)	Clip C	100.8267	42.8192

Table 5: Spatio-temporal characteristics of the test video sequences (following procedure from ITU Rec. P.910)

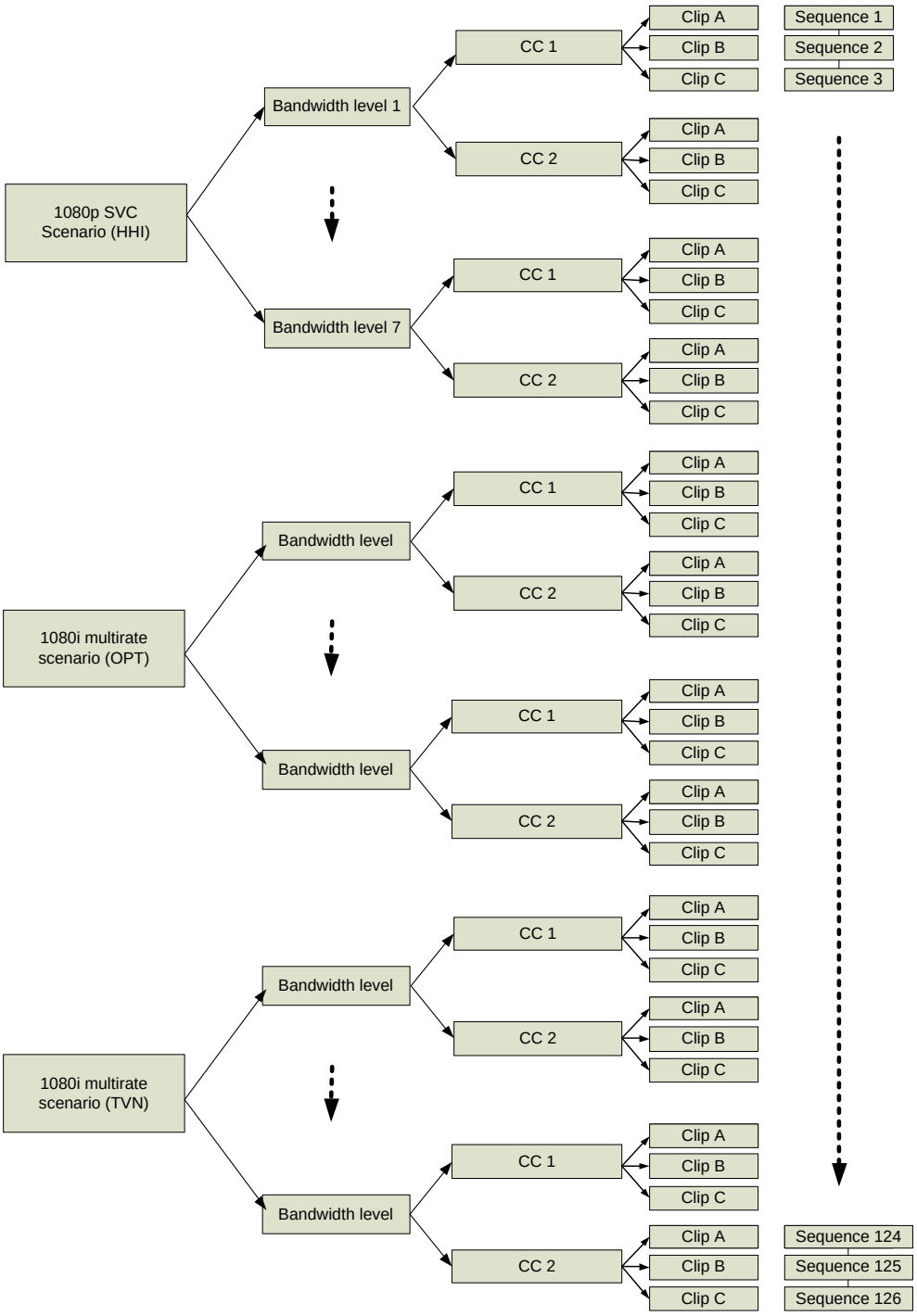


Figure 2: Video contents used in user-based QoE assessment (iteration 2)

3. User-based assessment method

This section describes the methodology for user-based assessment of full HD IPTV video services. Section 3.1 describes the test user sample. Section 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 specifies the varied test conditions for the factors of interest for systematic research (packet dropping scenario, bandwidth levels, content class).

Sections 3.4 and 3.5 describe the two user-based QoE assessment types that come into use in the OptiBand project: the quality rating test, and the acceptance threshold test. Section 3.6 then outlines further data gathering measures and section 3.7 concludes with a tabular overview of the overall test procedure and durations. Of course, the test method follows the overall OptiBand QoE research plan (D2.2) and thus is highly consistent to the method of the first iteration test (D2.3). For better readability, we include the whole test description also in the present deliverable, even if we are risking some redundancy with those previous documents.

3.1 Test user sample

54 subjects took part in the study. They had been recruited with the help of FTW's test persons database and announcements at a public internet job portal. An incentive in form of a 25 EUR gift voucher per participant was handed out. The main characteristics of the analysed sample as follows:

Age	34.22 years (± 12.2, min: 20, max: 65)
Gender	30 female, 24 male
Professional Status	28 employees, 1 unemployed, 16 students, 3 retiree, 6 employed students
Highest educational degree	Compulsory school (1), Vocational training (2), vocational secondary school (9), secondary school (13), university (23)
Number of TV sets in the participant's household	0 (3), 1 (33), 2 (8), 3 (5), more than 3 (5)
TV reception	DVB-T (16), cable digital (9), cable analogue (24), satellite (14), IPTV (17)
Hours a day watching TV	1.6 hours (±1.2, min:0, max: 5.0, median: 1.5)

The reason for including about twice as many test users than specified in D2.2 was that we wanted to control for the effects of participants' rating tendencies in studies, in which only high quality sequences are investigated (which is the case in the present study). To accomplish this, half of the participants were confronted only with the (high-quality) OptiBand sequences, and the other half viewed additional low quality sequences (confer the Annex for more details on the related results).

3.2 Overall test conditions

3.2.1 Room conditions

Our study was tailored to emulate home TV watching conditions. We followed general viewing conditions for subjective assessments in home environment defined by ITU-R BT.500-11 [1]. The viewing conditions defined therein are fulfilled by the i:lab, FTW's facility for perceived quality and usability testing. It is equipped with user testing equipment of TV services and is compliant with ITU-R BT.500-11 (see Figure 3). The viewing distance is approximately 3 m.

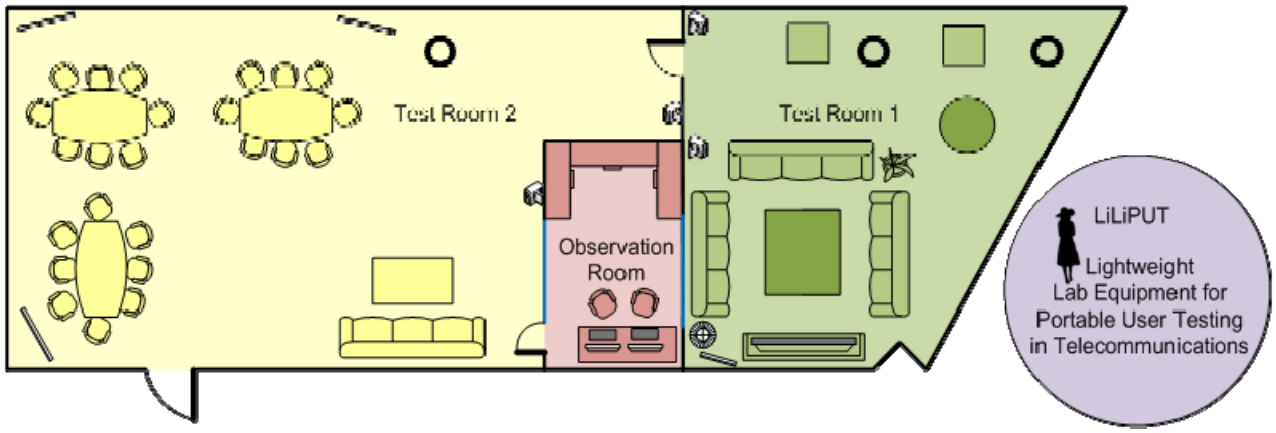


Figure 3: i:lab: Tests were conducted in test room 2

3.2.2 Service type

The technology setting was based on an emulation of a full HD IPTV service in the first two steps. The full HD IPTV service was emulated, because the performance of packet dropping algorithms should be evaluated before its implementation. The emulation allowed us to play the sequences exactly in the same way as the proposed system solutions of the OptiBand project. The fact of emulating the sequences had no impact on perception of video sequences, also because the test subjects were not aware of the emulation.

3.2.3 End-user device

The display for subjective testing was chosen according to the recent trends on TV flat panels market (confer D2.1). Thus, the parameters for a highly representative TV set were defined as follows:

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

We used the Sony KDL-40EX500, which meets all these requirements. It has the following technical specifications:

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

3.3 Varied test conditions

The varied test conditions corresponded to the main variables of interest in the OptiBand project specified in section 3.3:

Factor	**Factor steps**
Packet dropping scenario	1. 1080i (multirate video streaming) 2. 1080p (SVC)
Bandwidth level	7 Bandwidth levels (1 SOTA and 6 degraded sequences)
Content class (CC)	1. Action film (CC1) 2. Soccer (CC2)

Note that we slightly varied the type of content presentation across test sessions: one half of test users exclusively included the above-specified OptiBand video sequences, and the other half additionally contained sequences with deliberately low quality (300 kbit/s, 500 kbit/s, and 800 kbit/s (these low-quality sequences were prepared without using the OptiBand packet dropping algorithms). By this means we wanted to control for possible effects on participants' rating caused by the presence or absence of low quality content in the overall sample of video sequences.

3.4 Acceptance threshold test

While typical quality rating tests provide fine-grained mean opinion scores (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 not yet been defined. To address this gap, we explicitly aimed at identifying this acceptance threshold in dedicated test runs. The prospect would then be to enable a more valid interpretation of the quality rating test results described below.

Warm-up sequences

At the beginning of each test session, a trial run was conducted with 10 video sequences that were similar to tested video sequences, covering a wide range of bandwidth levels.

Scenario

The users were asked to imagine that they are trying out for the first time a new HDTV service from a well known operator, and that they will be paying a monthly fee for that. They should picture themselves sitting in their living room and watch programs via this new cable service.

Acceptance threshold rating procedure

After each sequence, the subject was asked a binary question: "Is this sequence acceptable or inacceptable for you?" Subjects were told that "inacceptable" would mean that they would get angry or even call the provider's complaint hotline.

Electronic questionnaire

The Telecommander is an electronic questionnaire that has been developed at FTW. The system automatically presents video sequences to test participants, handles the completion of electronic questionnaires, manages data collection and provides tools for efficient statistical analysis activities.

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 for each of the 3 packet dropping scenarios. Each of these 3 sets included videos for the specified two content classes. The rationale for selecting one of the three clips was based on a randomization algorithm, in order to ensure that no content class / bandwidth combination was presented twice to the users. For example, if a test subject was viewing ClipA of CC1 / bandwidth level 6 in the 1080p SVC scenario, she was provided with ClipB for the same combination in 1080i multirate scenario. Test sequences were presented in random order, with the exception that the same clip did not appear in succession.

3.5 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.

In order to achieve discriminative results, even if the perceived quality may have a limited variance (in case of very good performance of the tested packet dropping approach), we apply ACR with hidden SOTA (source) sequences. For example, instead of exposing a subject to one original sequence and 6 degraded sequences, the original sequence will be presented 3 times.

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

The same basic scenario was used as in the acceptance threshold test described above.

Rating procedure

After each sequence, the user is asked to rate the subjective quality, according to an ordinal 9-point scale ranging from ‘bad’ to ‘excellent’ (ITU-T P.910). The voting time was defined as 10 seconds.

Electronic questionnaire

Also here, the Telecommander was used.

Allocation of test sequences

The same principles as in the quality rating test were applied for allocating test sequences.

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 5 min break.

3.6 General inquiry

At the end of the test, further data is gathered from the test persons, namely demographic data and data on general user behaviour.

Demographic data

At the end of the test, the test participants were asked about basic demographic data relevant for the test.

General user behaviour (QGUB)

The participants were provided with the QGUB questionnaire developed at FTW, which collects data about previous usage of telecommunications services, with a focus on IPTV services.

3.7 Overview of test procedure

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

Part	**Duration (min)**
Welcome, briefing	3
Warm-up	5
Acceptance threshold test	20
Break	5
Quality rating test I	20
Break	5
Quality rating test II	20
Debriefing, general inquiry	10
	~ 90

Table 6: Overview of test phases and test duration

4. Automated assessment method

This section specifies the automated QoE methods used in the OptiBand project to complement user-based assessment. The evaluated video material was the same as in the user-based assessment (see section 3.3). Automated evaluation of the investigated video sequences was performed with a video quality measurement and estimation tool, which had been custom-developed by FTW for the OptiBand project. As is described in the following, the main features of this tool are PSNR and SSIM computation, as well as the added ITU-T recommendation J.341.

4.1 Structural Similarity Metrics (SSIM)

The SSIM index (Structural Similarity Metrics, compare D2.1, section 2.3.2) was calculated between the H.264 reference sequence and the H.264 sequences with packet dropping. The SSIM index was computed using three image measurement comparisons: luminance, contrast and structure. Each of these measures was 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 were calculated within the local window. Because SSIM index maps often exhibit undesirable “blocking” artefacts, each window was filtered with a Gaussian weighting function (11×11 pixels). 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 ten second length time window.

4.2 Peak Signal-to-Noise Ratio (PSNR)

The PSNR (peak signal-to-noise ratio) was calculated as a difference measure between the H.264 SOTA reference sequence and H.264 sequences with packet dropping. 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). Error-free frames (identical investigated and reference sequence) were clipped up to a maximum value of 111.30 dB for 1080 video mode. We used clipping in order to avoid infinite PSNR values resulting in a zero MSE. This PSNR clipping value corresponds to one error in one colour in one frame [4]. Finally, PSNR values are calculated for single frames and average PSNR over a 10-second length time window.

4.3 ITU-T-REC-J.341

As outlined in D2.1 and D2.2, since January 2011 a new standard recommendation is available on the “Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference” (ITU-T-REC-J.341, [7]). J.341 is a full-reference assessment method suitable for full HD resolution, where reference sequence should normally be the unique original. It is available as commercial software by the company Swissqual, and it is called VQuad-HD (confer www.swissqual.com). The project has decided to include J.341 as a third automated assessment technique, for comparative evaluation. In this test iteration we evaluate under which conditions J.341 can provide meaningful results.

5. Results

In the following, the results of the user-based and automated assessment are presented. Section 5.1 focuses on the 1080i multirate scenario video materials and section 5.2 on the 1080p SVC scenario video materials.

Statistical analysis of overall differences was performed by means of ANOVA tests, and pairwise comparisons were performed by means of post-hoc tests, applying Bonferroni corrections. Differences devised as statistically significant have a level of significance of 0.05.

5.1 1080i multirate scenario

In the following, the automated and user-based evaluation results for the 1080i multirate scenario are shown. For better readability and generalizability, an overall view on the results of both partners contributing to this scenario is presented. Please confer Annex I and II for separate results of OPT and TVN.

5.1.1 User-based assessment

In the following, the user-based assessment results for the 1080i multi-rate video sequences (2 content classes x 7 bandwidth levels x 3 clips) are presented. Please refer to section 1 for the specification of test sequences and to section 3 for the assessment methodology. The figures present the quality and acceptance threshold values for each content class (the values for the three clips per content class were averaged).

Quality ratings

When looking at the mean values and their confidence intervals within the two content classes in Figure 4, there are very few noteworthy relative differences. With regard to action contents, only bandwidth levels 1 and 2 received a significantly lower MOS rating than the SOTA reference sequence. For soccer, the only pairwise difference was between the SOTA reference sequence and bandwidth level 1.

This means that quality degradations could only be detected for bandwidth reductions of 39% (action) and 49% (soccer). However, this difference was relatively small: 55% bandwidth reductions of bandwidth level 7 only led to a MOS degradation of 0.3 MOS points for action (from 4.2 MOS to 3.9 MOS) and 0.1 MOS points for soccer (from 3.9 MOS with the SOTA reference to 2.8). The 95% confidence intervals shown in Figure 4 were very small, so the results can be trusted well. A factor contributing to this is certainly the relatively high number of 54 subjects included in the study.

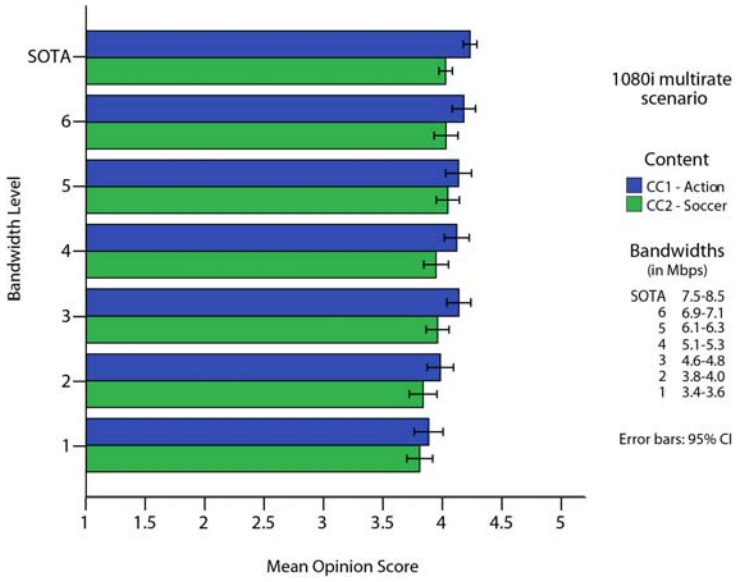


Figure 4: Quality rating results for CC1 and CC2 in the 1080i multirate scenario (error bars represent 95% confidence intervals)

The quality ratings yielded good absolute results: all bandwidth levels were above the defined threshold of 3.7. Interestingly the mean ratings did not exceed a MOS of 4.2, even for the state of the art sequence. Furthermore, there were small differences between the MOS values of the content types: action was rated about 0.1 – 0.3 MOS higher than soccer. Differences were significant for bandwidth levels 3, 4, 6, and 7. It may be unexpected that we found very small differences between the bandwidth levels, as usual result plots show a visible decline of quality with reduced bandwidth. However, comments from test participants as well as experts from our team indicate that quality of basically all video files was regarded as very good and that differences between the files were hardly perceivable. In rating situations characterized by the presence of homogeneous high-quality stimuli, rating artefacts may be possible.

Most notably, mean quality rating results may be lower when only good sequences are presented than if also low-quality 'distractor' sequences are among the list of test stimuli. As noted in section 3.3, we controlled for this possible rating bias by slightly varying the type of content presentation across test sessions: one half of test users exclusively included the above-specified OptiBand video sequences, and the other half additionally contained sequences with deliberately low quality (300 kbit/s, 500 kbit/s, and 800 kbit/s). More detailed considerations of the marginal differences between the results for these two sample subsets are provided in Annex III. In the main analysis shown in the present report, we pooled both sub-samples to achieve 'independence' from effects of these two possible content presentation variants.

In this context, in order to measure more fine-grained differences between the reference sequence and the degraded sequences, one could have in principle used a method like Degraded Category Rating (DCR, [1]). However, with DCR, information about absolute quality would have been lost, which appeared much more important in the practical context of the project than identifying overly detailed quality differences. Furthermore, and even more importantly, no comparability with the results of the first test iterations would have been possible. Please confer section 0 for a further discussion on the user-based testing results.

Acceptance test

When looking at the absolute acceptance rating results in Figure 5, it can be seen that acceptance was in general very high. Usually in fewer than 5% of the sequences, viewers rated the quality as inacceptable, even if bandwidth was diminished by 50%.

Most importantly, relative decreases of acceptance with lower bandwidths could not be verified. We did not find any statistically significant difference, even between bandwidth levels 1 and 7 (SOTA). There were as well no observable differences between the two content classes; no statistically significant differences were found.

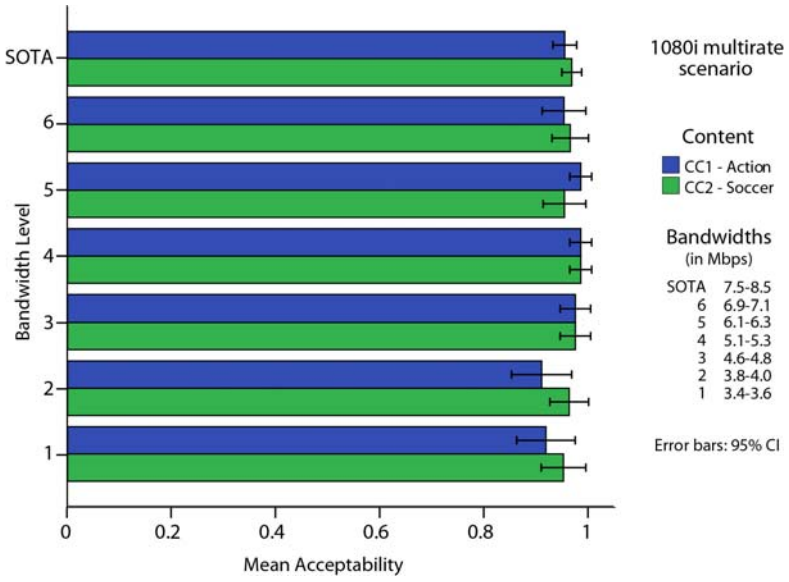


Figure 5: Acceptance rating results for CC1 and CC2 in the 1080i multirate scenario (error bars represent 95% confidence intervals)

5.1.2 Automated assessment

The PSNR, SSIM, and J:341 analyses were performed for the 1080i video sequences (2 content classes x 6 degraded bandwidth levels x 3 clips), 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 J.341 values for each content class (averages of the three clips per content class).

Structural similarity metrics (SSIM)

SSIM results are all above the defined threshold of 0.9. There are relative degradations due to the bandwidth reduction, but these are very small (from 0.945 to 0.92).

Differences between action and soccer contents were negligible: SSIM values of action sequences were about 0.01 higher than soccer sequences.

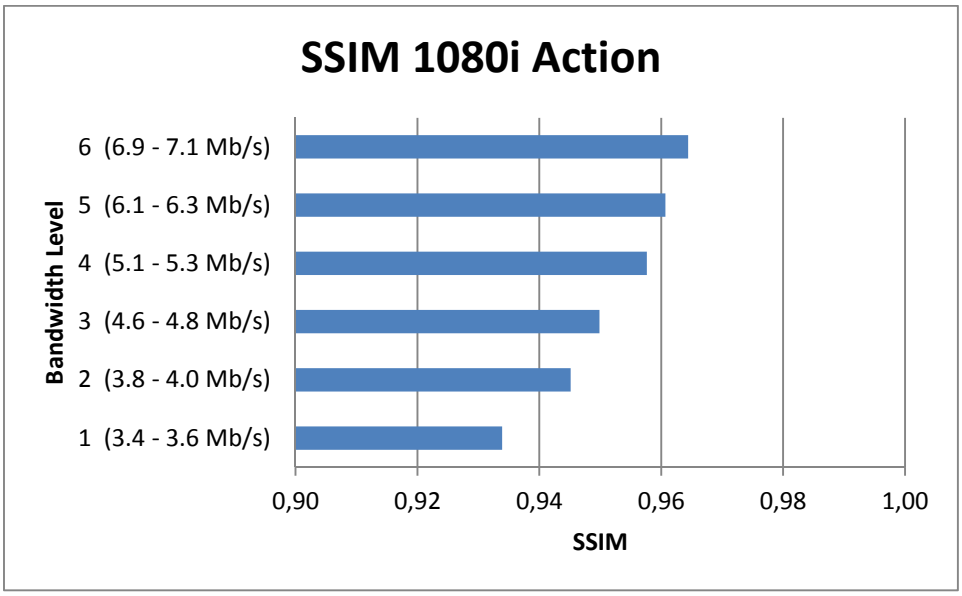


Figure 6: SSIM Results for CC1 (action) in the 1080i multirate scenario

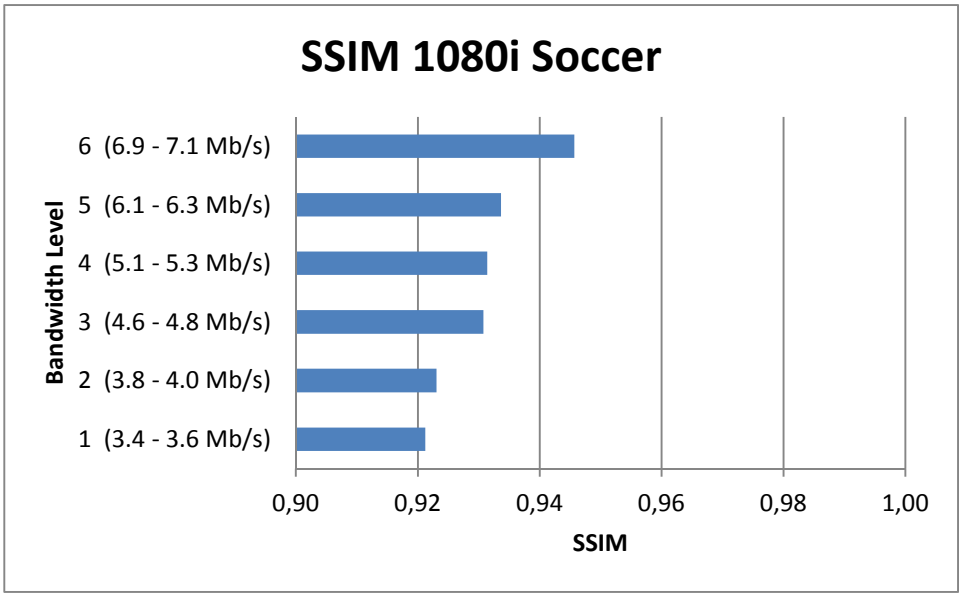


Figure 7: SSIM Results for CC2 (soccer) in the 1080i multirate scenario

Peak Signal-to-Noise Ratio (PSNR)

PSNR results were all above the threshold of 27 / 30 dB for whole sequence / single frames. This means that relative differences between the sequences with bandwidth reduction and the SOTA reference sequence were rather small. Furthermore, hardly any differences between content types were observed. Regarding the differences between the different bandwidth levels, Figure 17 and Figure 18 show a significant decline of PSNR values between bandwidth level 5 and 6 (from ~96 dB to ~37 dB), however almost no difference between the remaining bandwidth levels.

As already explained in D2.3, the unusually high PSNR value of bandwidth level 6 can be explained by the chosen PSNR evaluation method. For bandwidth level 6 many identical frames from the un-degraded sequence may have been used by the packet dropping algorithm, which may have resulted in many frames with high PSNR values (as specified in D2.2, the specific point of our method was that the reference source was not the unique original but the state-of-the-art H.264 sequences, and that error-free frames were clipped up to a maximum value of 111.30 dB).

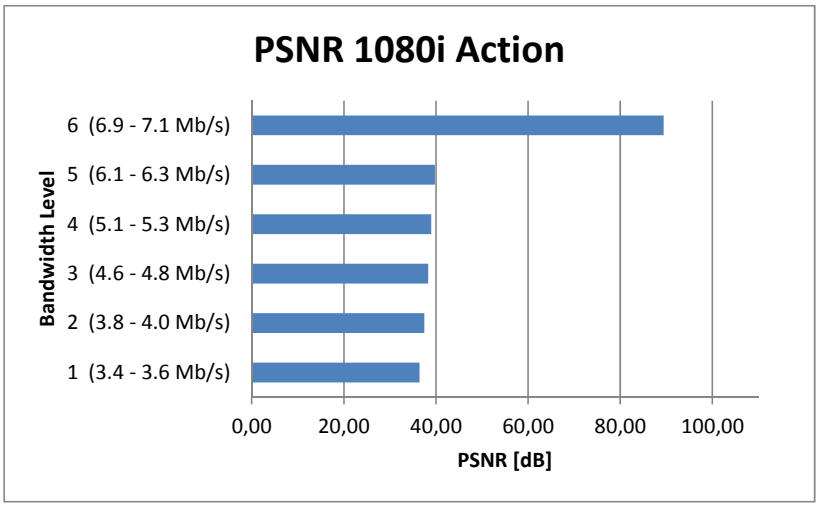


Figure 8: PSNR results for CC1 (action) in the 1080i multirate scenario)

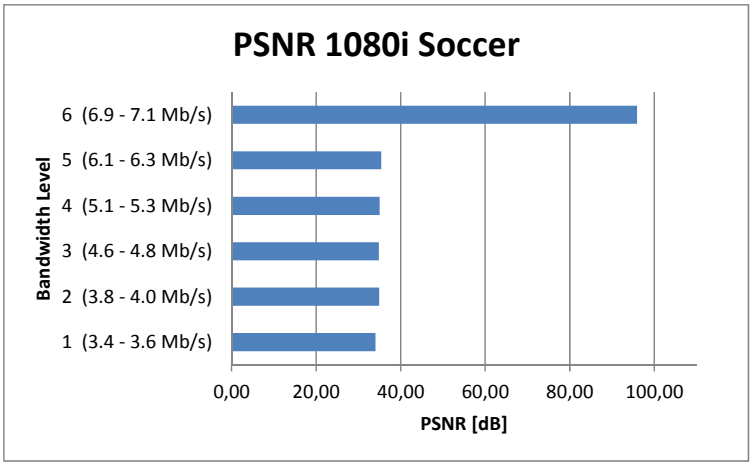


Figure 9: PSNR results for CC2 (soccer) in the 1080i multirate scenario)

ITU-T-REC-J.341

As highlighted in D2.2 and section 4.3, J:341 results are based on perceptual modelling and mapped to an estimated MOS scale - they are NOT derived by subjective testing. The presented MOS values therefore of course need to be interpreted cautiously.

The estimated MOS scores start with 4.0 for bandwidth level 6 and decrease to about 2.7. This decline of 1.3 MOS is interesting, because in user-based assessment the difference was only about 0.3 MOS.

J.341 was added to the compendium of automated assessment only since this second iteration, and so far we had not pre-specified or validated a threshold value for this method in D2.2. If one considers also a 3.7 score applicable as a threshold for the estimated J.341-based MOS, then a reduction of about 35% bandwidth for action (level 4) and of 42% for soccer would still yield results above this value.

The overall estimated MOS is slightly higher for soccer sequences than for the action sequences. This is not fully consistent with the results from user-based assessment, where ratings for action sequences were slightly higher for action than for soccer sequences (see section 5.1.1). Also, while in user-based assessment very little relative degradations are observable throughout the tested content types, automated assessment reveals relatively strong declines for action, but not for soccer. Please refer to further discussion of J.341 performance in section 0.

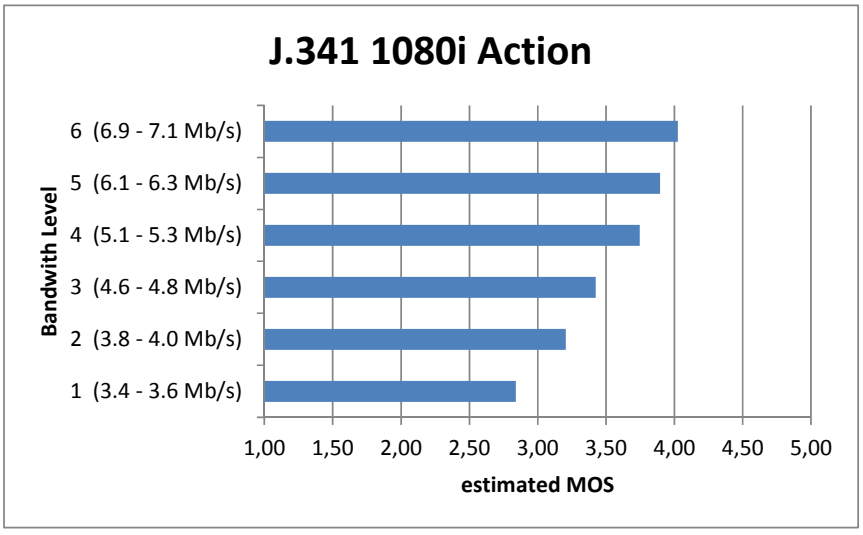


Figure 10: J.341 results for CC1 (action film) in the 1080i multirate scenario)

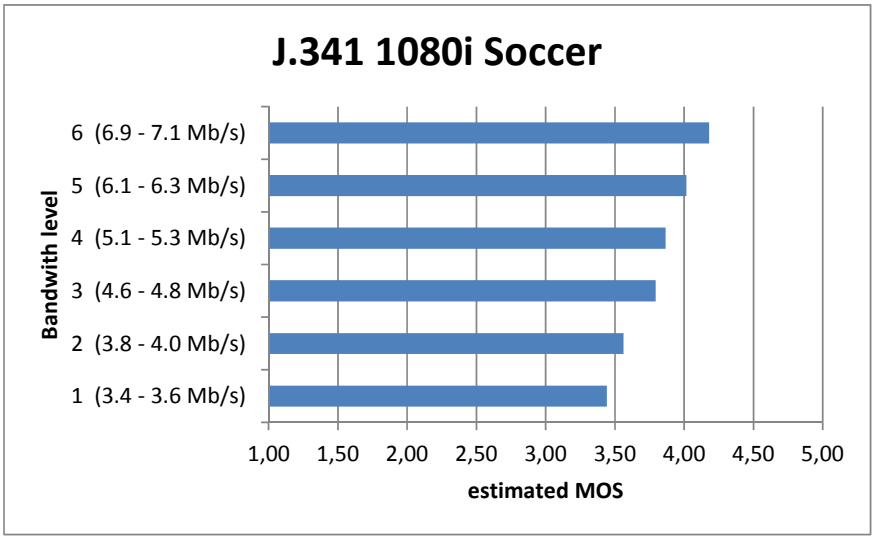


Figure 11: J.341 results for CC2 (soccer) in the 1080i multirate scenario)

5.1.3 Correlation of user-based and automated testing results

Figure 12 shows the correlations of automated assessment metrics results (SSIM, PSNR, and J.341) with user-based quality rating results (MOS).

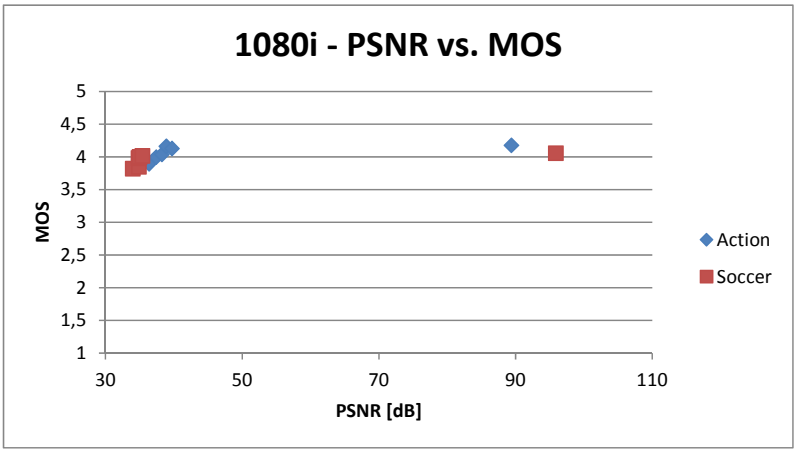
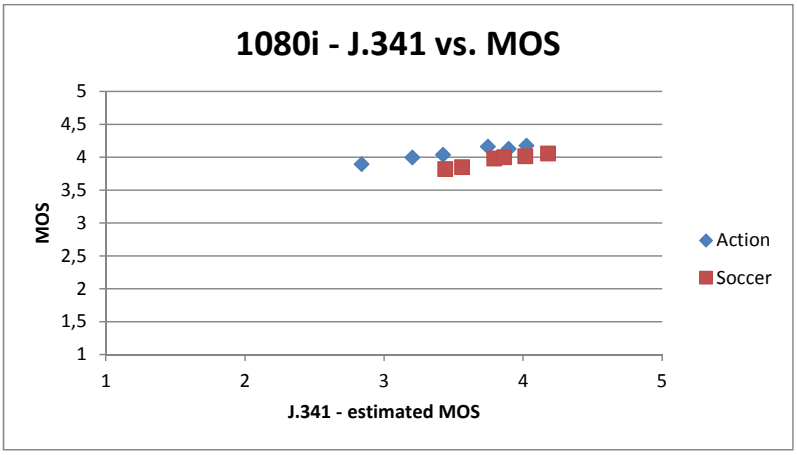
Correlation of MOS with …	Scatterplot	Pearson correlation
SSIM		Action: r = 0.97 Soccer: r = 0.85
PSNR		Action: r = 0.54 Soccer: r = 0.54
J.341		Action: r = 0.98 Soccer: r = 0.97

Figure 12: Correlations of automated assessment metrics (SSIM, PSNR; and J.341) results with user-based assessment results (MOS) for the 1080i multirate scenario (action and soccer contents separately)

The correlations of the user-based assessment (MOS) against SSIM and J.341 were quite high, ranging between 0.85 and 0.97. J.341 had the highest correlations: 0.98 and 0.97, respectively. PSNR correlation was lower, due to the strong gap between bandwidth level 6 and the other bandwidth levels (see the explanation for that in section 5.1.2). Differences between content types were only manifested in the SSIM analysis: soccer contents (r=0.85) were less correlated than action contents (r=0.97).

However, it is important to note that the power of these correlations appears very limited. Our obtained MOS values are all very high, due to the good quality of our source contents (~ 4 MOS). Almost no significant difference between the bandwidth levels was found. Our results show that overall correlation of user-based and automated assessment scales is given, but, in order to investigate the predictive power of automated measurment metrics, further correlation studies involving sequences with more strongly differing content quality may be needed.

5.2 1080p SVC scenario

5.2.1 User-based assessment

In the following, the user-based assessment results for the 1080p SVC video sequences (2 content classes x 7 bandwidth levels x 3 clips) are presented. For the specification of test sequences, please compare section 1 and for the assessment methodology section 3. The figures present the quality and acceptance rating values for each content class (the values for the three clips per content class were averaged).

Quality rating

Hardly any relative MOS difference between the bandwidth levels was found. An analysis of variance test (ANOVA) only revealed an overall significant difference for action, not for soccer. For action, only one pairwise comparison between bandwidth levels within each content type was significant: the highest bandwidth (bandwidth level 7, SOTA) vs. the lowest (bandwidth level 1). This means that bandwidth reduction was highly effective, even at a rate of up to 50% for action and 60% for soccer.

Overall, action contents receive very high ratings, with mostly values of around 4.0 MOS, and soccer contents (including the SOTA) are on or slightly below the specified 3.7 MOS threshold. Differences between the content types are significant, except for the lowest bandwidth level 1. The reason for this discrepancy is that the soccer sources used for this scenario were of limited quality (please confer section 2 for more information on the chosen source contents, and section 0 for a discussion of the role of source content quality in perceptual quality measurements).

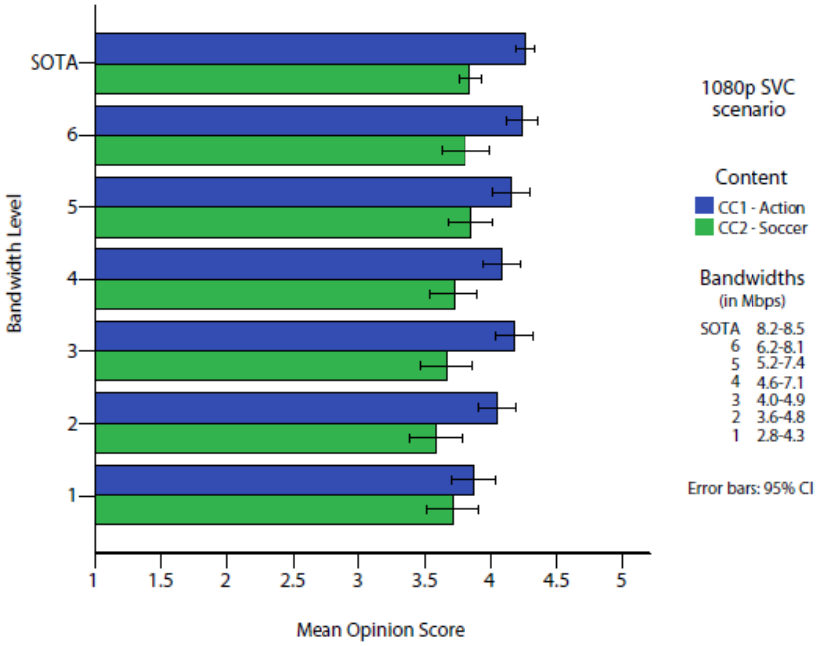


Figure 13: Quality rating results for CC1 and CC2 in the 1080p SVC scenario (error bars represent 95% confidence intervals)

Acceptance test

The acceptance mean rating values and confidence intervals in Figure 14 show that bandwidth levels do not differ significantly in terms of perceptual quality, they all have similar acceptance rates. This applies both for action and soccer. It thus appears that even strong bandwidth reductions did not have any effect on the perceived quality.

There are some differences between the content types with regard to absolute quality: action contents were considered unacceptable only in about 5% of the situations, but for soccer contents this applied for 20% of the situations. Due to relatively large confidence intervals of the soccer-related results, however, this difference was only significant for the SOTA sequence (!) and the three lowest bandwidth levels. This

difference may again be ascribed to the situation that for the 1080p SVC scenario we had to use the soccer content sources of limited quality , i.e. the original content was 720p and was up-scaled to 1080p.

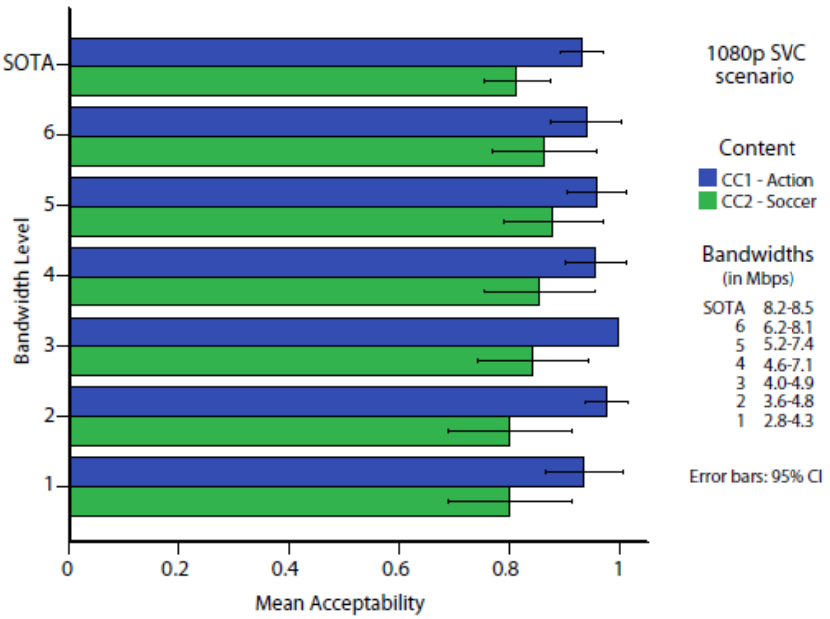


Figure 14: Acceptance rating results for CC1 and CC2 in the 1080p SVC scenario (error bars represent 95% confidence intervals)

5.2.2 Automated assessment

Also for the 1080p SVC video sequences, PSNR, SSIM, and J.341 analyses were performed (2 content classes x 6 degraded bandwidth levels x 3 clips), taking the respective state of the art videos (SOTA) as a reference. The analysis followed the procedure outlined in section 4.

The figures below present the SSIM, PSNR and J.341 values for each content class (averages of the three clips per content class).

Structural similarity metrics (SSIM)

SSIM automated assessment results are above the specified SSIM threshold, which had been raised after the first iteration from 0.8 to 0.9 . The figures above basically show that relative quality differences to the reference sequences are rather small, with SSIM values reaching from almost 0.96 to almost 0.999.

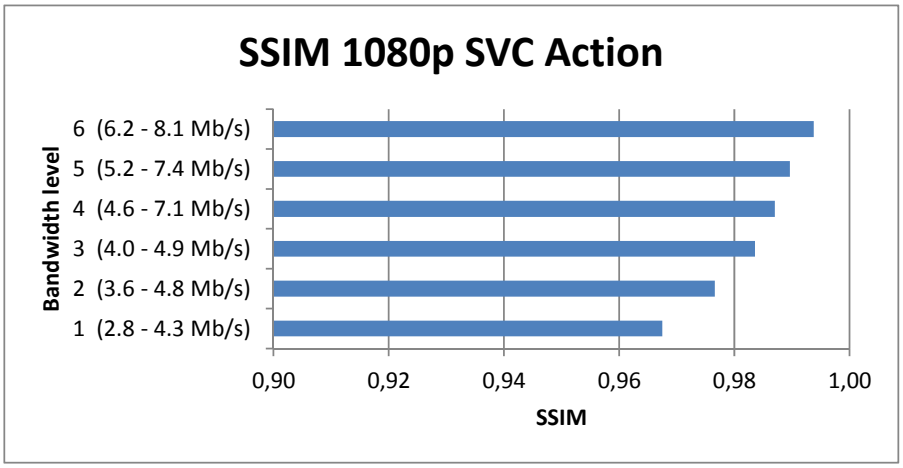


Figure 15: SSIM Results for CC1 (action) in the 1080p SVC scenario

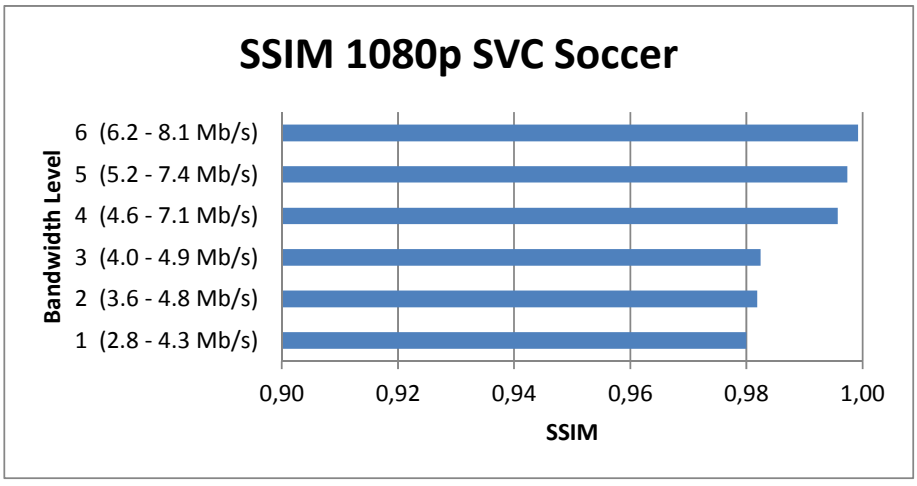


Figure 16: SSIM Results for CC2 (soccer) in the 1080p SVC scenario

Peak Signal-to-Noise Ratio (PSNR)

PSNR values are also above the specified threshold of 27 / 30 dB for whole sequence / single frames. Similarly to the 1080i multirate scenario results, we found a significant decline of PSNR values from over 100 dB to about 40dB, here it was between bandwidth level 3 and 4 (Figure 17 and Figure 18). As already explained in section 5.1.2, the unusually high PSNR values can again be explained by the chosen PSNR evaluation method. For bandwidth level 6, many identical frames from the un-degraded sequence may have been used by the packet dropping algorithm, which may have resulted in many frames with high PSNR values. In such cases, error-free frames were clipped up to a maximum value of 111.30 dB (see D2.2 for a full description of the approach).

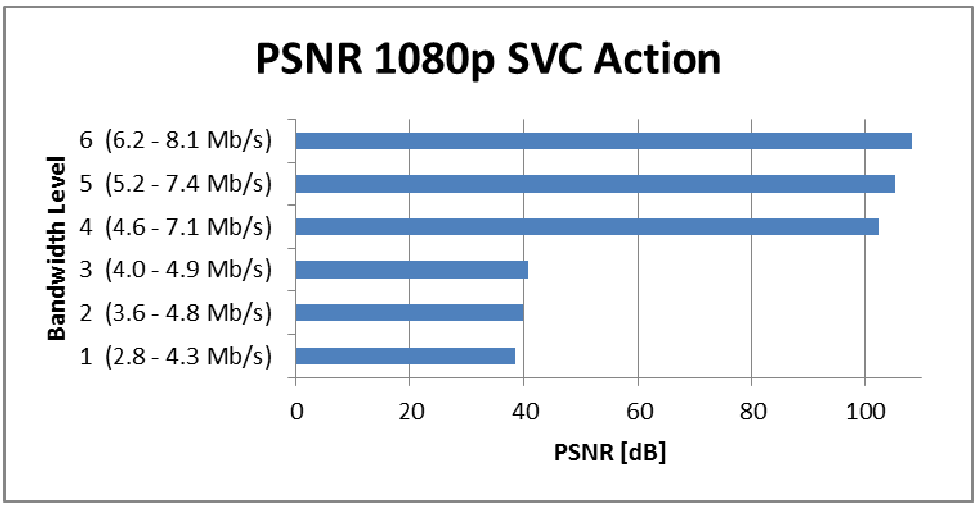


Figure 17: PSNR results for CC1 (action) in the 1080p SVC scenario)

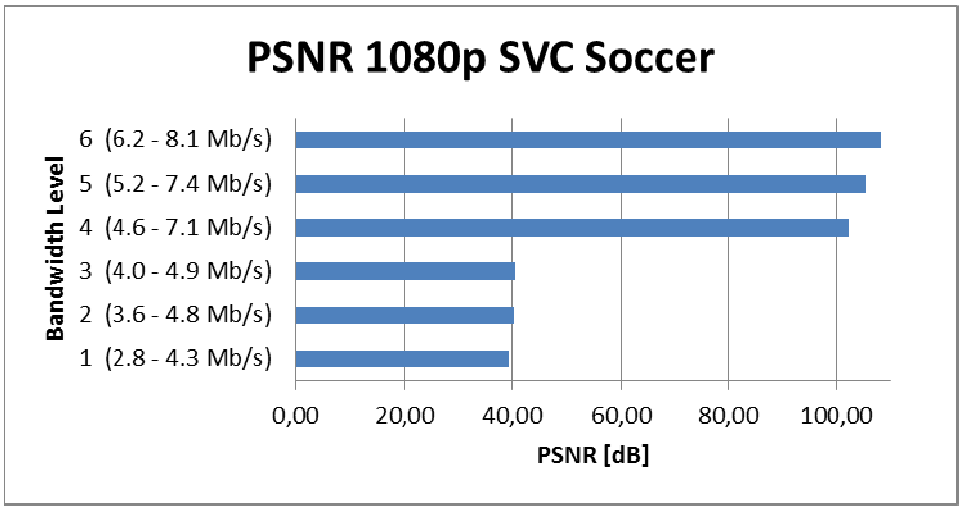


Figure 18: PSNR results for CC2 (soccer) in the 1080p SVC scenario)

ITU-T-REC-J.341

In analogy to the discussion of the 1080i scenario, the MOS values shown in Figure 19 and Figure 20 are estimated from automated assessment, not collected in user-based assessment. Therefore, careful interpretation is necessary.

In general, when assuming the MOS threshold of 3.7 that we had specified for subjective test results, the bandwidth levels, we might say that even strong bandwidth reductions of up to 50% still maintain acceptable quality.

Interestingly, J.341 soccer sequences results are even slightly better than the action sequences results, which is in contradiction to the low MOS from the user-based assessment (see section 5.2). This may indicate that while the quality of the source sequence was lower, relative degradations (as measured by the reference-based metric J.341) were rather small.

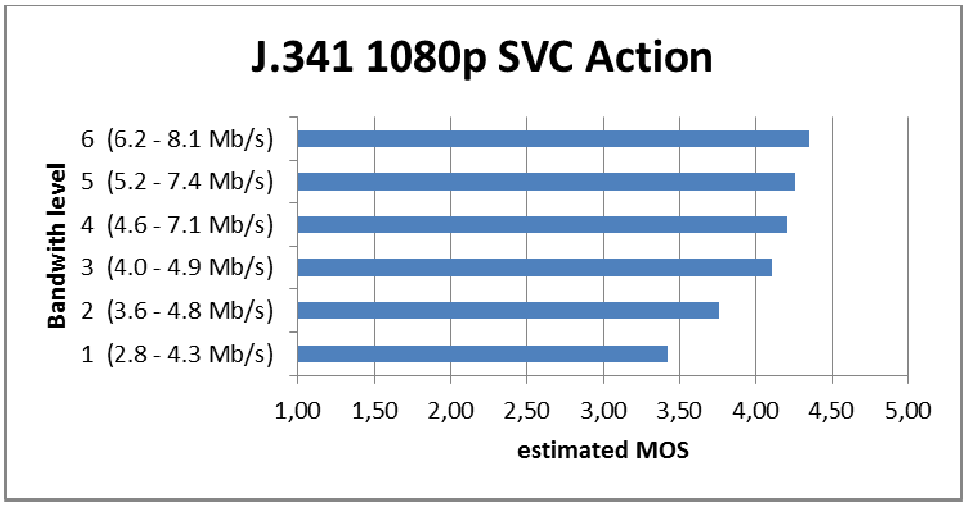


Figure 19: J.341 results for CC2 (action) in the 1080p SVC scenario)

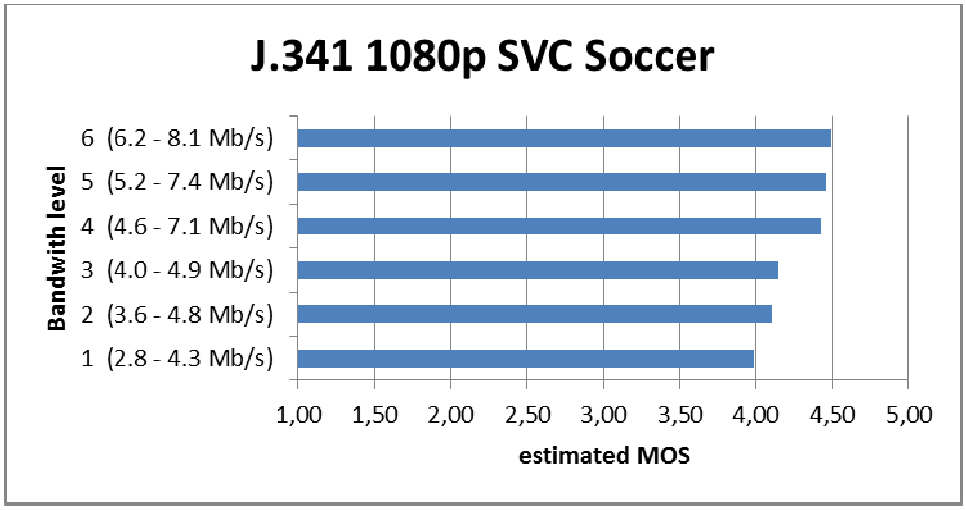


Figure 20: J.341 results for CC2 (soccer) in the 1080p SVC scenario)

5.2.3 Correlation of user-based and automated assessment results

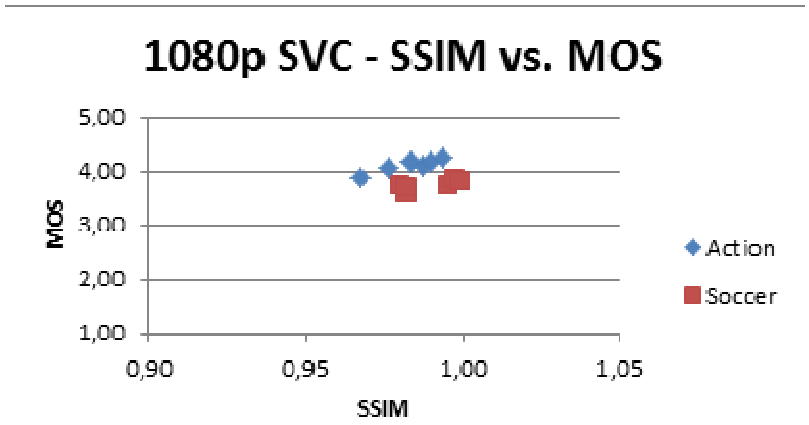
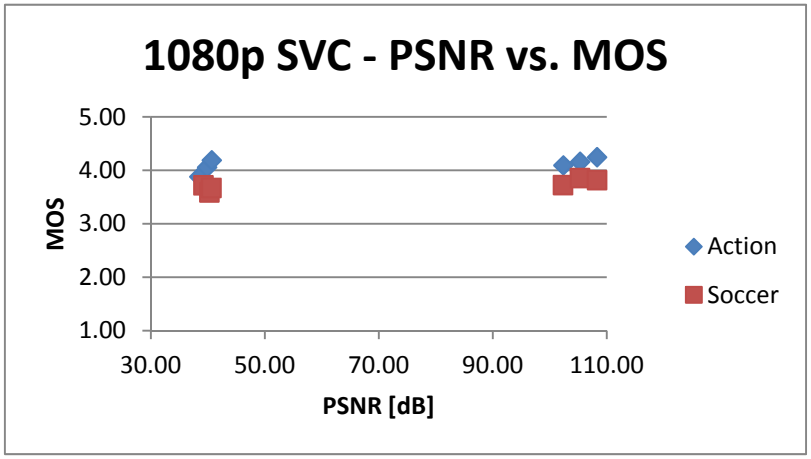
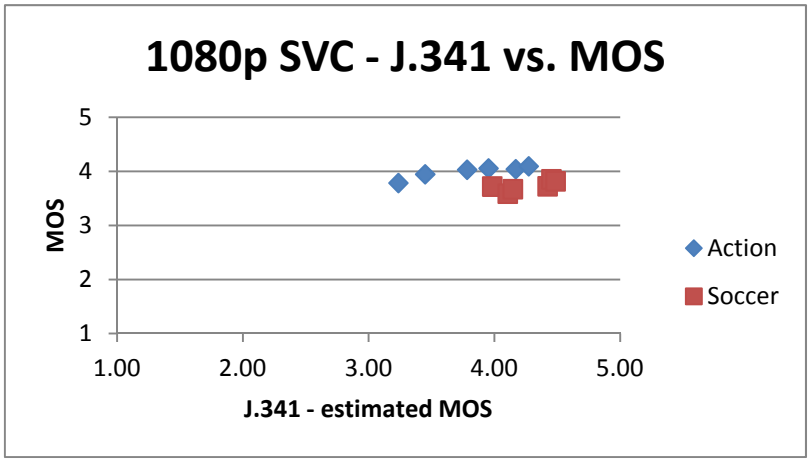
Correlation of MOS with …	Scatterplot	Pearson correlation
SSIM		Action: r = 0.78 Soccer: r= 0.79
PSNR		Action: r = 0.56 Soccer: r = 0.80
J.341		Action: r = 0.93 Soccer: r = 0.72

Figure 21: Correlations of automated assessment metrics (SSIM, PSNR; and J.341) results with user-based assessment results (MOS) for the 1080p SVC scenario (action and soccer contents differentiated by colour)

Correlations for SSIM vs. MOS are moderate, at r~0.8, for both content classes. J.341 results correlate highest with our obtained MOS values, but more strongly for action (r=0.92) than for soccer contents (r=0.72). Correlations of PSNR and MOS were low for action contents (r=0.56), obviously due to the clipping of error-free frames within our applied PSNR calculation method (compare section 5.2.2).

6. Conclusions

In the following, we summarize and interpret the main results from the user-based assessment (section 6.1) and the automated assessment (section 6.2). Based on these considerations, we provide an outlook to ongoing and upcoming project activities.

6.1 User-based assessment

6.1.1 Main user-based assessment results

Relative impairments

One of the most indicative results from our quality ratings is that we did not find any significant perceptual quality degradations between the reference content and sequences with up to almost 40% bandwidth reduction. For even stronger bandwidth reductions of between 40% and almost 60%, the MOS ratings were only about 0.3 points lower than for the reference.

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 analogue TV to IPTV. Even with these small confidence intervals, almost no significant difference between any of the tested bandwidth levels could be found.

This appears to be a very impressive improvement in comparison to the first testing iteration, where a strong progressive decline of perceived quality from the source content to the successively compressed contents was observed. It appears that, at least with laboratory testing of offline encoded files, the OptiBand project has even exceeded its goal to reduce bandwidth by 33% while preserving acceptable perceived quality. It is worthwhile to note that this success applies for both coding scenarios selected for further investigation in the OptiBand project: 1080p SVC and 1080i multirate. The acceptability testing results confirm this picture. In the acceptance test blocks, which constituted separate units within the test procedure and preceded the MOS-based quality testing blocks, no significant difference was found between the source content and any degraded sequence.

Absolute quality

In this second iteration, the tested sequences overall attained higher absolute quality ratings than in the first iteration. With action sequences, the MOS threshold of 3.7 was surpassed for all bandwidth levels (up to 55% reduction). Soccer was expected to be more challenging, also for the state-of-the-art sequences. Results for 1080i soccer were nevertheless quite satisfactory: the threshold could be attained even for bandwidth reductions of 50%. As the 1080i scenario is the basis for the integrated prototype and for the live tests, these results indicate that the project is well prepared for these activities.

In the 1080p SVC scenario, although for the action content the tests were successfully passed, a bandwidth reduction of 23% (level 6) was possible, but the bandwidth level 5 with a reduction of 30% did not fully reach the absolute quality threshold. This effect can be well explained by the quality of source contents that had to be used for testing in the 1080p SVC scenario (see a further discussion on this in section 0).

Despite the improvements with regard to absolute quality threshold attainment, it has to be noted that the averaged MOS ratings hardly exceeded the 4.0 mark. This is especially interesting, as test users as well as OptiBand project experts often commented that the overall perceptual quality of the video sequences was exceptionally good. This phenomenon will be further discussed in the following section.

The acceptance ratings were in general very positive: only 5-10% of the sequences were qualified inacceptable by the participants. The only exception here is soccer content in the 1080p SVC scenario, where already the state-of-the-art sequence was rated as inacceptable by 20% of the participants, which may be a result of the selection of the original content, which had to be up-sampled from 720p to 1080p.

6.1.2 Interpretation

Source content quality

A major recommendation based on our experiences from the first test iteration was to find better soccer source contents, as this was expected to raise overall absolute perceptual quality to HD level (see D2.3, section 6.3). As noted in section 2, we were successful in obtaining and preparing such new test contents for the second test iteration. These new soccer sequences should enable a fully trustable interpretation of absolute user-based assessment results, without doubts induced by limited content source quality. In fact, the absolute overall quality has slightly increased in the 1080i scenario, raising most of the investigated bandwidth levels beyond our pre-specified MOS threshold. However, unfortunately the new sequences could not be used for the 1080p SVC scenario, as no way of transcoding them to 1080p was available. This is probably the main reason for our finding that packet dropping in the 1080p SVC scenario did not fully reach the absolute quality threshold.

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 SOTA reference. This is interesting because experts in the pre-test phase regarded the sequences as superb, an impression which was also supported during the tests by many of the participants' qualitative comments. Subjects said that the sequences looked good to them and that they did not perceive notable differences. Furthermore, the acceptance results appear to indicate that quality was highly satisfactory. As noted above, with the acquisition of HD soccer content, limitations of video source content cannot any more be suspected to have caused this effect.

An important factor contributing to the limitation to a maximum mean rating score of 4.2 is probably the rating paradigm itself. 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; cf. Tominaga et al [3]). Also in previous video quality research (such as in mobile TV) a maximum MOS of 4 has been found to be a very common result for the best quality levels [8].

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.

Quality thresholds

In our test report of the first QoE testing iteration (D2.3), a recommendation was given to adopt a "relative" MOS threshold instead of an absolute one. The reason for this was that, given the weak state-of-the-art, it appeared to be difficult to name "the right" absolute quality threshold. Furthermore, the pre-specified value of 3.7 was difficult to reach in the first iteration, which raised discussions on relaxing this absolute threshold by reducing the targeted MOS value to, say 3.5. Such discussions actually divert attention from the actually more important aspect of relative MOS tradeoffs.

If we apply the suggested formula for a relative MOS (MOS-threshold = MOSoriginal * 0.85), it is evident that even the sequences with the highest bandwidth reduction pass this criterion. As also the absolute threshold was attained in most of the cases, we do not suggest to remove or lower the absolute threshold of 3.7, as it turns out to be a realistic and challenging "stretch goal" for HD IPTV services.

Until now, no threshold for acceptance testing has been defined. This type of testing may become more frequently used in practical contexts, due to its practical handling and due to its direct relatedness to the main question of service acceptability. Therefore, first considerations on setting thresholds also for acceptability may be helpful. While the definition of a minimum acceptance threshold cannot easily be drawn from previous empirical research, our experiences from the first two testing iterations in the OptiBand project suggest a threshold value of 85% for the acceptance ratio.

6.2 Automated assessment

Below we summarize the main results of automated testing, which are based on SSIM, PSNR, and the newly added ITU-T recommendation J.341. In section 6.2.2, we then interpret the suitability of these metrics to validly assess the QoE in an HD IPTV context, by comparing them to the user-based testing results.

As stated already in D2.3, the scope of the research on automated assessment within WP2 is not on on-line QoE metrics for live IPTV quality monitoring scenarios, but on the formative and summative quality assessment of the packet dropping algorithms developed in the project.[1] This also implies that the main source of insight within this work package stems from the subjective testing results. Research on automated testing is mostly "curiosity-driven", i.e. we are interested in investigating the general interrelation of currently used objective quality metrics and the obtained subjective results.

6.2.1 Main automated assessment results

This section presents the main results for automated assessment, which is followed by an interpretation in section 6.2.2.

Relative impairments

Also the automated assessment results indicate that the investigated OptiBand packet dropping approaches are suitable for reducing bandwidth while still preserving acceptable perceptual quality. All three metrics detected very small declines of quality along reduced bandwidth reduction. This especially applies for SSIM, where differences are below 1%. Also PSNR values hardly reveal any differences, except for the strong gaps, which were caused by identical frames and clipping PSNR-clipping in case of identical frames (see a summarizing explanation of this aspect in the next section).

The correlations between automated and user-based assessment are considerably high for J.341-MOS (ranging from 0.72 to 0.98), and SSIM-MOS (ranging from 0.78 to 0.97). PSNR correlated less with MOS (between 0.5 and 0.8), due to the mentioned reasons above. If only the 1080i scenario is considered, i.e. the scenario that will be pursued in the OptiBand integrated prototype and live system, the J.341 metric performed most consistently to the subjective quality ratings (0.97 for soccer, and 0.98 for action). The good prediction performance of this metric corroborates the results obtained in a recent competition by the video quality experts group (VQEG) [6], which preceded its standardization.

Attainment of quality thresholds

The defined quality thresholds for SSIM (0.9) and PSNR (27 / 30 dB for whole sequence / single frames) have been reached by all bandwidth levels. For the newly introduced J.341 metric, no threshold had been defined so far (see a discussion about worthwhile uses and quality thresholds for this metric in the next section).

6.2.2 Interpretation

In the following, we provide explanations of the gained automated assessment results. For better readability and comprehensiveness, we risk redundancy to previous deliverables by repeating important aspects inherent to the general methodology that has already been brought up (especially in D2.3).

Using the state-of-the-art H.264 video stream as reference source

Consistently to the first iteration, we did *not* use the unique original as our reference sequence, but a sequence encoded with state-of-the-art H.264 (compare D2.2). The motivation behind this was to exclude the influence of the encoder performance (each project partner uses his own H.264/AVC implementation; e.g. offline vs. real-time encoding) and to consider only the performance of the PDA.

[1] For a discussion of feasible QoE metrics for the live system please refer to T4.2. The responsible partner for that task is HHI.

Naturally, using as a reference the state-of-the-art H.264 video stream instead of the unique original resulted in much smaller measured degradations of the compressed video sequences. As a consequence, the values of SSIM and especially PSNR were comparatively high and thus surpassing the set thresholds.

Performance of full-reference methods like PSNR, SSIM and J.341

It is generally well known that subjective tests have a higher normative power than objective quality metrics (see e.g. [6] for recent results). This has very obvious reasons: full-reference quality metrics like PSNR, SSIM and J.341 rely on the comparison of non-degraded with degraded pictures, that is, they primarily assess visual fidelity. Thus, by their very nature, they do not match human quality perception very well, because in normal usage situations (i.e. living-room watching), the viewer only sees the degraded video and cannot compare it with the “original”.

Reference-based assessment methods of course rely very much on the properties of the reference sequence, and they should represent “perfect” quality. If the reference sequence is clearly less than perfect, as it was the case with the old soccer sequences that had to be used for the 1080p scenario, the test result may be distorted. As a consequence, the scores need to be interpreted with caution, and its values should not be misunderstood as absolute quality indicators.

Performance of PSNR and SSIM in HD IPTV environments

SSIM and PSNR both have very high values for basically all bandwidth levels. As explained already in D2.3 this result is not so unusual for today’s HD IPTV environment, which is characterized by a higher display resolution and better performance of new codecs (such as new deblocking filters in H.264). Here, SSIM and PSNR may have limitations with regard to the discrimination between quality impairments at high bandwidth levels. Especially SSIM results are mostly found at a very high range of 0.95 – 1, which makes meaningful comparisons of reference sequences with impaired sequences difficult (even if SSIM-MOS correlations are satisfactory).

Also other studies show that SSIM and PSNR quality values are generally very high, which makes it difficult to set meaningful threshold values. For example, a study by Monjas [5] on HD video quality evaluation showed that the SSIM index was in a range of 0.94 to 1). Another publication with similar findings comes from Kulikov et al. [14], where also generally high SSIM values greater than 0.92 were reported for bandwidths above 1Mbps. In a similar vein, Lopez et al [15] measured the PSNR of HDTV video sequences and also here found very good results for all tested bandwidth levels, which only differed by 1.3 dB within a range of 10MBits (5 vs 15 MBit/s).

In the first iteration the very good SSIM and PSNR scores and almost non-existent differences between bandwidth levels stood in discrepancy to the subjective results, which showed comparatively strong declines with reduced bandwidth. However, in this iteration, both automated and user-based testing do not indicate strong quality differences between bandwidth levels. In this sense, low discriminance of SSIM and PSNR are not so critical, as they are consistent to MOS results.

PSNR clipping

Similarly to the first iteration, we found mean PSNR values of more than 100 dB for the higher bandwidth levels and then a strong decline to 30-40 dB at the lower bandwidth levels (compare sections 5.1.2 and 5.2.2). For the highest bandwidth levels, many identical frames from the un-degraded sequence have been used by the packet dropping algorithm, which has resulted in many frames with high PSNR values.

The underlying motivation is that the SVC approach does not reduce quality of each frame: certain frames (according to bandwidth level) are identical with the original (in our case the state-of-the-art H.264 encoded sequence, as explained above). Therefore, in case of an identical reference of the state-of-the-art sequence with the degraded frame it was necessary to introduce “PSNR clipping” (as specified in D2.2, error-free frames were clipped up to a maximum value of 111.30 dB). Each clipping of PSNR values represents one error in one pixel of a frame.

In comparison to this abrupt development of PSNR results, the MOS rating results were declining very steadily. This means that PSNR (at least in the way we were using it, i.e. using the H.264 state-of-the-art sequences as a reference) is not well capable of predicting video QoE in an SVC scenario with identical frames. This is also reflected in the PSNR-MOS correlations, which were mostly lower than for J.341-MOS and SSIM-MOS.

Suitability of ITU-T J.341

As noted above, the recently introduced J.341 method highly correlates with the user-based quality rating results, especially for the 1080i multirate scenario that the project will be focusing on in the integration and live test phase. Thus, if objective metrics are needed for automated quality assessments, J.341 may have the best prediction power for this scenario. In case of automated quality assessment is necessary for operative uses of OptiBand technology, such as when deploying or adapting an IPTV service, J.341 could be recommendable.

However, predictions with J.341 results need to be made cautiously. First, quality scores of the various bandwidth levels differed more strongly in the predicted J.341 MOS than in the actual MOS from the user tests (maximum difference of 1.5 vs. 0.3, refer to Figure 12). Second, while the lowest MOS obtained in the 1080i quality rating tests was about 3.6, the lowest estimated MOS in the J.341 assessment was 2.5. Thus, at least in cases comparable to our situation, one may also be satisfied with a threshold for an estimated MOS of 3.0, instead of the 3.7 for the quality rating tests.

6.3 Outlook

In the following, we provide an outline of ongoing and future activities related to WP2.

Quality thresholds

One important goal for this second iteration was to validate whether the threshold of 3.7 MOS of the user-based quality rating should be changed for the further course of the project and beyond. It is clear that setting this threshold could not be based on systematic tests but had to rely on expert statements and needed to be transferred from related domains, such as standard definition TV or Mobile TV. And indeed, in the first iteration it looked as if the threshold value may have been set too high, as even the highest bandwidth levels did not reach this threshold in every case.

However, in the second iteration, the threshold was surpassed also for strongly reduced bandwidths. With this confirmation, we decided to keep with the threshold of 3.7 also for interpreting the upcoming live test. In general, one should keep in mind that the most important criterion for OptiBand QoE testing should be the effectiveness of *relative* bandwidth reduction by packet dropping with regard to quality trade-offs.

For acceptance tests, we had until now not defined a quality threshold. Also here, a decision is generally rather normative and depending from the service context, but based on scarce literature [2] and our experiences from the first two iterations of the OptiBand project, we would propose a minimum acceptance ratio of 85%.

Extra study on comparing content durations

In addition to the initially planned QoE research within the OptiBand project, we are conducting a study on the impact of content durations in QoE testing. First, the study helps to fine-tune the methodology of both the OptiBand live test and the final laboratory tests in 2012, such that we gain highly ecologically valid results that are sensitive to the content duration. We expect to gain an improved understanding for the definition of optimal clip lengths in user-based QoE testing, which shall mirror a realistic TV watching experience, while still maintaining manageable and controlled laboratory conditions.

With this study, we hope to provide first guidance on optimal content durations in experiments focusing on true quality of EXPERIENCE tests (as opposed to the hitherto restriction to pure video quality on the pixel level). We expect that our results will be interesting for the research community, as we know about increased awareness of so far missing guidance with regard to the aspect of content duration in holistic QoE studies. The results of this study will be delivered as an additional Annex to this document by January 2012.

Increasing the number of evaluated content classes

In the hitherto evaluation, we have focussed on the two most popular and challenging contents for HD IPTV: action film and soccer. By testing with these most representative and demanding content classes, general conclusions can be drawn on whether the OptiBand packet dropping approach is generally feasible and acceptable for IPTV usage settings.

In order to broaden the scope of content types, we are adding the class of documentary films, which are showing natural phenomena in high quality. Upcoming results from the extra study on content durations and

from the live test will include experiences with this third content class. Similarly as action and soccer, such popular and challenging contents are also highly suitable for the evaluation of the efficiency of OptiBand packet dropping approaches in an HD IPTV setting.

With the given goals and research approach within the OptiBand project, it is not feasible to add more content types into our quality perception studies. This is because our rigorous study design has many factorial combinations (bandwidth levels, different packet dropping approaches, several content clips per content type), which limits the number of content classes to be included in a test session of acceptable duration. Of course, for real-life operation, quality parameters also need to be defined for all possible other content classes covered in a given IPTV service (please confer to WP4 for topics related to QoE management in the live system). If such data is needed at a later stage during the deployment of an OptiBand-enabled IPTV service, we propose to use the ITU-T J.341 metrics (applying the threshold proposed in section 6.2.2).

Live test (WP8)

WP2 is strongly co-ordinated with live test planning, and there will also be intense involvement during conduction and data analysis. A preliminary plan for the live test has been submitted to the commission in June 2010 [10]. Care was taken to make the test consistent to the laboratory tests of WP2. Especially, the same action and soccer contents classes are being used to enable comparison of results. Furthermore, as stated above, documentary as a further content type will be included. We will also use consistent measures as in the laboratory tests, most importantly the acceptance rating. The main difference will be that sequences will be longer (up to 5 minutes length), and that video playback will dynamically change due to TV set usage profiles within the household. The definite plan for this research activity will be specified in D8.1.

7. Bibliography

[1] ITU-R Recommendation BT.500, ”Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, 2002.
[2] M. Ries, O. Nemethova, M. Rupp: "On the Willingness to Pay in Relation to Delivered Quality of Mobile Video Streaming"; in: "International Conference on Consumer Electronics 2008", IEEE Conference Proceedings, Las Vegas, USA, 2008, ISBN: 1-4244-1459-8, 195 - 196.
[3] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi. Performance comparisons of subjective quality assessment methods for mobile video. Proc. QoMex Workshop 2010.
[4] 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.
[5] A. Monjas, Performance evaluation of quality metrics for video quality assessment, Diploma thesis at Technical University Berlin, 2010
[6] Video quality experts group (VQEG). Report on the validation of video quality models for High Definition Video Content. Version 2.0, June 30, 2010.
[7] ITU-T Recommendation J.341 “Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference”, 2011.
[8] M. Ries, “Video Quality Estimation for Mobile Video Streaming”, PhD Thesis, Technical University of Vienna, 2008.
[9] E. Dumic, S. Grgic, M. Grgic, “Comparison of HDTV formats using objective video quality measures,” in Journal of Multimedia Tools and Applications archive, Vol. 49 Issue 3, September 2010, Kluwer Academic Publishers Hingham, MA, USA, doi>10.1007/s11042-009-0441-2.
[10] Lazzara, S., Franco, C., Fröhlich, P., and Ries, M., „Review of the use cases for the live tests”. Internal OptiBand project document, 2011.
[11] Ries, M., Fröhlich, P., and Schatz, R., D2.1 - Criteria specification for the QoE research. Available at the OptiBand project homepage: www.optiband-project.eu”, 2011.
[12] Fröhlich, P., Ries, S., Egger, S., D2.2 - Detailed research plan. Internal OptiBand project document, 2011.
[13] Fröhlich, P., Ries, M., Schatz, R., Egger, S., and Holzleitner, I., “D2.3 - Initial QoE research recommendations report. Available at the OptiBand project homepage: www.optiband-project.eu”, 2011.
[14] Kulikov, D. Sixth MPEG-4 AVC/H.264 Video Codecs Comparison. Available at: http://compression.graphicon.ru/video/codec_comparison/h264_2010/index.html#Sequences
[15] Lopez, J.P.; Diaz, M.; Jimenez, D.; Menendez, J.M., “Tiling effect in quality assessment for High-definition digital television”, IEEE International symposium on consumer electronics, 2008.