Project Number: 248495

Project acronym: OptiBand

Project title: Optimization of Bandwidth for IPTV Video Streaming

Deliverable reference number: D2.5

Deliverable title: Final QoE research recommendations report

Due date of deliverable: M33

Actual submission date: 17/09/2012 (M34)

Start date of project: 1 January 2010 Duration: 33 months

Organisation name of lead contractor for this deliverable: FTW

Authors: Peter Fröhlich, Michal Ries, Raimund Schatz, Theresa Stürmer

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 D2.5 (“Final QoE research recommendations report”) is the final deliverable of the OptiBand QoE research activities (WP2). WP2 evaluated the QoE of the OptiBand solution in three iterations throughout the project.

Objectives of the document:

As specified in the OptiBand DoW, the main goal of the deliverable is to present and reflect on the QoE-related aspects of the OptiBand live test. The live test took place under the responsibility of WP8 by TIS at their premises in Torino, Italia (WP leader: TIS). The evaluation was focusing on the live integrated system that had been developed within WP7 (WP leader: HHI).

Furthermore, as announced in the second periodic progress report, the deliverable addresses the remaining WP2 objective of analysing the impact of socio-demographic variables on perceived quality of IPTV services.

Document structure

Section 1 outlines the aspects of the Live Test methodology that are related to the obtainment and interpretation of the QoE-related data. Section 2 presents the QoE Live Test results and provides conclusions interpreting these findings, also in comparison to the laboratory tests conducted earlier in the project.

Section 3 presents the analysis of socio-demographic variables in terms of their relatedness to rating tendencies, by pooling the data from 252 participants from OptiBand-related QoE studies.

Document history

The document had been produced and finalized on 31 August 2012 by FTW, in accordance to the delivery dates specified in the OptiBand DoW. It should be noted that not all data was available for writing this document, as not all necessary materials of WP8 have been delivered. Thus, not all necessary results and interpretations can be presented in this document.

However, on 04 September 2012, we were informed that the project was prolonged until 30 September 2012. Furthermore, on 06 September, some of the missing data (the subjective rating results for use cases 0 and 3 that had actually been gathered in the first week of July by TIS) had been sent from TIS to FTW. As a consequence, in the current version some adaptations have been made, and this document has been submitted to the commission by 17 September 2012.

Results summary

Relative quality: One of the main OptiBand objectives was to reduce the necessary bandwidth of an HD stream by more than 33% while preserving acceptable perceived quality. To this end, a reduction by 1.5 Mbps did not result in degradations of perceptual quality. When the available bandwidth was decreased by more than 33% per TV set, average quality values tended to be lower by about 0.3 MOS. We also found that when other TVs in the household were zapped with slow frequencies, there was no significant decline. Thus, we can conclude that with optimized tuning and zapping strategies, small relative impairments can be achieved with the OptiBand solutions.

Absolute quality: While the focus of the OptiBand QoE investigations is on assessing relative quality degradations, it is of course also important to look at the absolute achieved quality. A reduction by 1.5 Mbps using the OptiBand packet dropping solution resulted in a satisfactory absolute level of subjective quality (~4.0 MOS). The results for the more demanding use cases 1 and 2 are more differentiated, and they seem to be especially dependent on the underlying zapping scenario. We can conclude that absolute quality is satisfactory in slow zapping situations, as the quality threshold of 3.7 MOS is slightly surpassed in the respective test conditions (with an average MOS of 3.7). Also, disconnections and reconnections of other set top boxes do not seem to have a strongly disturbing impact.

An open point for the OptiBand solution is still how to deal with fast zapping conditions. Here, relatively low rating values of 3.1 MOS and an acceptance ratio of 0.63 have been obtained. The selected bandwidth levels for H.264 streaming services should have resulted in excellent streaming performance; however this was not the case in the presented test. In the report, we discuss feasible possibilities to improve performance under fast zapping conditions.

1. QoE Live Test: Outline of the Test Plan

This section provides an outline of the live test plan, in order to enable efficient comprehension of the analysed test results. The original concept is described in the 'live test plan', a document that had been prepared by FTW and TIS since summer 2011 the final test concept was accorded in June 2012). Actually the live test plan was supposed to be submitted to the commission as D8.1, but this has so far not yet been accomplished by the WP leader TIS. In order to facilitate an appropriate amount of reference information, we provide the accorded live test plan description in Annex I of this document (it will be removed, once D8.1 is submitted).

Section 1.1 introduces into the QoE research questions, section 1.2 then outlines the experimental design developed to answer the research questions, and section 1.5 provides a tangible impression of the implemented test prototype, especially with regard to its streaming parameters.

1.1 Research Questions

With regard to the scope of QoE, we addressed the following research questions in the live test:

Overall QoE: What is the QoE of the tested OptiBand solution under live network conditions?

Use Cases: How does QoE differ in the defined use cases, regarding number of TV sets used and available bandwidth (confer the live test document in Annex I)?

Zapping patterns: The pattern of TV set or set-top-box usage may have a strong impact on the QoE. Viewers are usually well susceptible to changes in the transmission quality, and especially transitions between different amounts of available bandwidth could be experienced as a nuisance.
A possible impact could as well be the **frequency** of such changes. For example, the viewing experience might be disturbed by other household inhabitants who are zapping between channels. To this end, we were interested to compare high and low frequencies of such transitions (i.e. fast vs. slow 'zapping patterns').

Furthermore, we wanted to investigate whether the **amplitude** of bandwidth variation has an impact on the QoE. Concretely, it could make a difference if other household inhabitants simply zap channels, or whether the whole set-top-box is being switched on and off.

Content classes: As in the previous studies, we were interested in understanding the influence of the content type on the QoE. As planned, the content types were Action, Soccer, and Documentary.

1.2 Sample

Overall, the sample comprised 30 subjects, with a balanced distribution of age and gender. Unfortunately, no further information on sample characteristics can be provided at this phase, as the necessary data has not been provided by TIS yet.

1.3 Experimental Design

In order to answer the research questions, the following factors have been defined: use cases, zapping patterns, and content classes. These will now be outlined further.

1.3.1 Use cases

The tested use cases were designed in order to emulate "real word conditions" and in parallel to reflect the research and technical criteria for the QoE evaluation of the OptiBand solution. Below, these use cases are outlined; a comprehensive description of the use cases is provided in the live test plan (see the version of June 2012 in Annex I):

Use case 0

From the OptiBand project perspective, this use case represents the state-of-the-art (SOTA) or baseline situation: only one TV-set is switched, the whole bandwidth is available and no packet-dropping needs to be applied. The available bandwidth is considered to be 15Mbps. In fact the entire system acts just like the state of the art networks, without any special operation being done at the PDD.

Use case 1

This is the most ambitious scenario for the OptiBand packet dropping solution, as it involves 3 TV sets in the household. It considers the technical requirements of ADSL2+ line and technical annex requirements. The available downstream bandwidth at the application layer is 15 Mbps.

Use case 2

In this use case, we assume that only one further TV-set is switched on (i.e., two TV-sets per household in total). The available bandwidth at the application layer is considered to be 10Mbps.

Use case 3

This use case reflects the often constrained current state of technology conditions for IPTV. Only one TV set per household is considered, and for video transmission a payload of 6.5 Mbps is dedicated. The OptiBand data dropping method is applied in this use case.

1.3.2 Zapping patterns

We specified three zapping patterns, in order to address the related research questions introduced above:

1. Fast

In order to understand the impact of the frequency of other household owners' zapping on the QoE of a video, we introduced fast zapping pattern and a slow zapping pattern for comparative analysis. For the 'Fast' zapping pattern, a 15-seconds interval was specified.

2. Slow

For the slow zapping pattern, intervals were defined for 30-45 seconds (see the exact implementation in section 1.3.4).

3. Disconnecting:

As mentioned in section 1.1, we also wanted to test the impact of the 'amplitude' of bandwidth transitions. When the other STBs are only zapping between channels and contents, there will not necessarily be a strong variation of bandwidth levels. However, when set-top boxes are switched off and on, the volume available of available bandwidth for the other set-top-boxes will vary more strongly. In order to also compare these two possibilities of bandwidth variation within the household, we added a further condition called "Disconnecting", in which the set-top-boxes of other household viewers were fully disconnected. By that condition, we could assure that the PDD switches from the lowest to the highest bandwidth level (compare a description of the bandwidth level management approach in 1.5.2). The zapping frequency was similar to the 'slow' condition, so that we had a possibility to compare strong with medium bandwidth changes.

1.3.3 Content classes

As before, the defined content classes were action, soccer, and documentary. The concrete implementation of video materials is described in section 1.5.1.

1.3.4 Experimental Conditions

For testing, representative combinations of the above factors needed to be selected. These will now be explained, grouped by the addressed use cases.

For use case 0 (1 TV, 15Mbps)

The experimental conditions for use case 0 were straight-forward, as they only involved one TV set and no zapping. The only performed variations were with regard to the content: per experimental condition, participants either watched action (CC1), soccer (CC2), or documentary (CC3).

Zapping time	15	30	45	60	75	90	105	120
STB1 (1st TV set)	TV set watched by test participants for CC1, CC2, CC3							

Table 1: Planned temporal profile for use case 0

For use case 1 (3 TVs, 15 Mbps)

Within this use case we evaluated all three zapping patterns and three contents. Thus, in total we had nine different experimental conditions defined by the combination of zapping patterns (fast, slow, slow/disconnecting) and content classes (action, soccer, documentary). These experimental conditions reflected common channel switching behaviours by users and the most favourite contents. With the 'disconnecting' zapping pattern, we also realized a high variety of available bandwidths.

The following tables provide an overview of the fast zapping pattern (Table 2), the slow zapping pattern (Table 3), and the slow pattern with disconnection (Table 4).

The fast zapping pattern models the situation that a person is watching a video with a set-top-box / TV set in a household. On the other two set-top-boxes (STB2 and STB3) are rapidly switched at an interval of 15 seconds. A 120 second session thus contained 8 transmissions (see Table 2).

Within this scenario, the zapping times between STB 2 and 3 are asynchronous (zapping is never performed exactly at the same moment), in order to avoid artefacts imposed by simultaneous channel switching. Simultaneous switching is not realistic for real usage conditions. STB1 either continuously presents action, soccer, or documentary (CC1, CC2, or CC3), but the same channel switching model of the other two set-top-boxes is applied for each content class.

Zapping time sec.	15	30	45	60	75	90	105	120
STB1 (1st TV set)	TV set under QoE test for CC1, CC2, CC3							
STB2 (2nd TV set)	CC1							
				CC2				
							CC3	
STB3 (3rd TV set)	CC1			CC1			CC1	
		CC2			CC2			CC2
			CC3			CC3		

Table 2: Fast zapping pattern for all CCs with active 3 STBs

As mentioned above, the slow zapping conditions within Use Case 1 were set to 45 seconds and the whole session contained 2 transmissions (see Table 3). From the perspective of a viewer of STB1, who views the same content for 120 seconds, the underlying bandwidth availability changes after 45, then after 45, and lasts 30 until the end. Again, participants watched soccer, action or documentary on STB 1, whereas the zapping pattern of STB 2 and STB 3 did not change.

Zapping time	15	30	45	60	75	90	105	120
STB1 (1st TV set)	TV set under QoE test for CC1, CC2, CC3							
STB2 (2nd TV set)	CC1							
							CC2	
STB3 (3rd TV set)								
	CC2							
				CC3				

Table 3: Slow zapping pattern for all CCs with active 3 active STBs.

As specified in section 1.3.2, the ‘Disconnection” pattern aimed at modelling a large variance of concurrent IPTV traffic within a household, varying between no concurrent set-top box (first 45 seconds), two concurrent set-top boxes both other STBs switched off in the following 45 seconds), and one concurrent set-top-box (last 30 seconds), see Table 4.

Zapping time	15	30	45	60	75	90	105	120
STB1 (1st TV set)	TV set under QoE test for CC1, CC2, CC3							
STB2 (2nd TV set)	CC1			X	X	X	X	X
				X	X	X	X	X
				X	X	X	X	X
STB3 (3rd TV set)				X	X			
	CC2			X	X			
				X	X	CC3		

Table 4: Slow zapping pattern for all CCs with connecting and disconnecting STB2 and STB3.

For use case 2 (2 TVs, 12 Mbps)

Also use case 2 (2TV sets in household with 12 Mbps) contained all three defined zapping patterns (fast, slow, and disconnecting) and three contents (action, soccer, documentary), resulting in nine possible combinations. The approach for specification of experimental conditions was similar as for use case 1.

The fast zapping pattern models the situation that a person is watching a video with a set-top-box / TV set in a household. The other set-top-boxes (STB2) are rapidly switched at an interval of15 seconds. A 120 second session thus contained 8 transmissions (see Table 2). The zapping times between STB one and two are asynchronous (zapping is never performed in the same moment). STB1 either presents action, soccer, or documentary (CC1, CC2, or CC3), but the same channel switching model of STB 2 is applied for each content class.

Zapping time	15	30	45	60	75	90	105	120
STB1 (1st TV set)	TV set under QoE test for CC1, CC2, CC3							
STB2 (2nd TV set)	CC1			CC1			CC1	
		CC2			CC2			CC2
			CC3			CC3		

Table 5: Fast zapping pattern for all CCs with active 2 STBs

As mentioned above, the **slow zapping** conditions within Use Case 2 were set to 45 seconds and the whole session contained 2 transmissions (see Table 3). While the viewer of TV/STB1 watches the same content for 120 seconds, the underlying bandwidth availability changes after 45, then after 45, and lasts 30 until the end. Again, participants watched soccer, action or documentary on STB 1, whereas the zapping pattern of STB 2 was consistent.

Zapping time	15	30	45	60	75	90	105	120
STB1 (1st TV set)	TV set under QoE test for CC1, CC2, CC3							
STB2 (2nd TV set)	CC1							
				CC2				
							CC3	

Table 6: Slow zapping pattern for all CCs with active 2 STBs

In the ‘Disconnection” scenario, the concurrent set-top box was switched off after 45 seconds, and then switched on again after another 30 seconds, see Table 7.

Zapping time	15	30	45	60	75	90	105	120
1st TV set	TV set under QoE test for CC1, CC2, CC3							
2nd TV set	CC1			X	X			
				X	X			
				X	X	CC3		

Table 7: Slow zapping pattern for all CCs with connecting and disconnecting STB2.

For use case 3 (1 TV, 7Mbps)

Similarly as in use case 0, also here no zapping was made, that is, participants continuously watched either action (CC1), soccer (CC2), or documentary (CC3).

Zapping time	15	30	45	60	75	90	105	120
STB1 (1st TV set)	TV set watched by test participants for CC1, CC2, CC3							

Table 8: Planned temporal profile for use case 3

1.4 Procedure and Measures

Consistent to the first and second testing iteration, we collected the following measures:

- Quality rating scores (five-point MOS scale)
- Acceptance scores (binary decision between 0 and 1).

Furthermore, in order to assess prototype performance and to better interpret the subjective ratings we looked at a representative sample of live streams (see next section).

The viewing duration was defined as 120 seconds, due to findings of our scientific studies within the OptiBand project [6].

1.5 Test Prototype

1.5.1 Video materials

In the following, we provide an introduction of the video materials that have been provided for system integration and live testing.

These have not been the same ones as in iterations 1 and 2. The reason for changing the video contents was that we needed longer sequences of 2 minutes, which could be realized with the previous ones. Furthermore, the first two iterations had not included the documentary content class.

In order to accomplish this goal we dedicatedly created several video sequences fulfilling our requirements in autumn 2011. Note that the videos had been already introduced into the integrated prototype and have been demonstrated during the second periodic project review in Torino.

Also, before the live test, these files had already been used in other QoE tests, namely the OptiBand content durations study [6] and a Crowdsourcing study in co-operation with the OptiBand project [7]. Within these studies, we did not identify significant differences of the ratings between the three content classes.

During the integration phase, to make the files useable for the OptiBand integrated demonstrator system, TIS transcoded these files to a format suitable for the integrated video streaming technology. It is not known whether an interim subjective evaluation had been conducted or initiated by TIS, thus it is not fully clear what impact this transcoding activity had on the video quality.

1.5.2 Bandwidth levels

The functioning of the PDD to switch the suitable level according to available bandwidth has been defined in D 4.2. The involved look-up table with its key parameters, as defined following the La Coruna consortium meeting at October 2011, is shown in Table 9.

Level	Mbps	Soccer (MOS)	Action (MOS)	Documentary (MOS)
L	8.00	4.1	4.3	4.2
L1	6.50	4.02	4.14	4.08
L2	4.95	3.98	4.07	4.025

Table 9: Look-up table for bandwidth selection and targeted MOS for the three content classes

Bandwidth level selection is autonomously managed by the PDD algorithm in order to maximise the available bandwidth. The following section will provide indications on how bandwidth was actually distributed amongst the available set-top-boxes.

1.5.3 Video stream characteristics

We analysed sample video streams that had been captured during the Live tests. The goal of this analysis was to derive an understanding of how consistently the system performed with regard to the defined algorithm (see previous section) and the experimental conditions (see 1.3.4). For that purpose, TIS provided us with 3 sample stream capture datasets (created with the Wireshark software package).

In the following, we provide for each use case typical temporal profiles, as well as summary tables of related key indicators

- Average video payload during the 120 second long test session (the tests session starts in second 40).
- Variance of video payload during the 120 second long test session
- Maximum video payload during the 120 second long test session.

Naturally, the following analysis can provide considerable support for the interpretation of the QoE results presented in section 2.

Use case 0 (1 TV, 15Mbps)

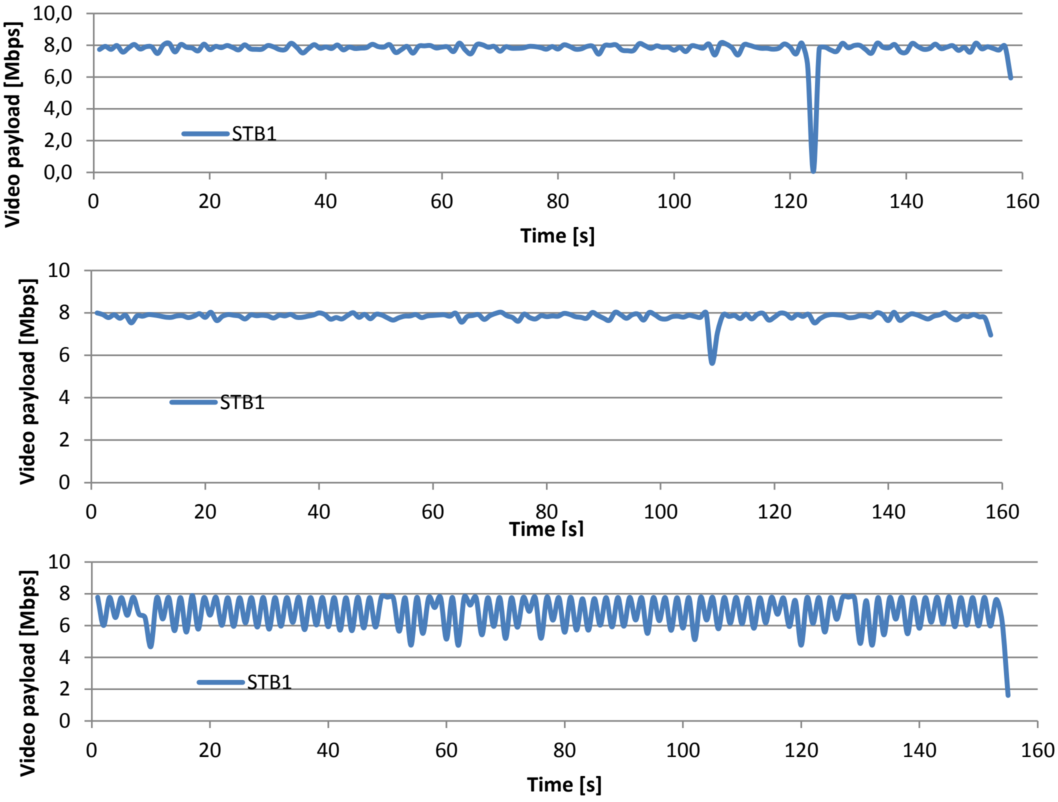

Figure 1: Temporal payload profile for use case 0 video stream at STB1.
Top: action, middle: soccer, bottom: documentary

To be consistent with the planned experimental conditions, the video streams should be constant and should not show a high variance in the temporal profile. As can be seen in Figure 1, this has been achieved by the test prototype to a high extent. Outliers such as the one in the action video stream could be assumed to not have a strong negative impact on quality perception. However, we see a notable difference between the high temporal variance between the documentary and the low variance of the action and soccer streams. As is also shown in Table 10, the variance of 1.19 is notably higher than for action (0.44) and soccer (0.05). This is quite surprising, due to near-to-optimal laboratory transmission conditions, which could expected to have very low noise levels within the DSL frequency band.

Content class	Action	Documentary	Soccer
Average video payload [Mbps]	7,77	6,79	7,82
Variance video payload [Mbps]	0,44	1,19	0,05
Maximum video payload [Mbps]	8,13	7,89	8,03

Table 10: Video stream statistics for use case 0

Also it is interesting that the average payload of documentary is about 1 Mbps lower than action and soccer. This could indicate inefficient scheduling of constant bitrate algorithm of the real-time encoder. The maximum payload is very similar among the content types, which is consistent with our testing requirements.

Use case 1 (3 TVs, 15Mbps)

The **fast zapping** stream captures for particular STBs are shown below (see Figure 2, test session starts in second 40).

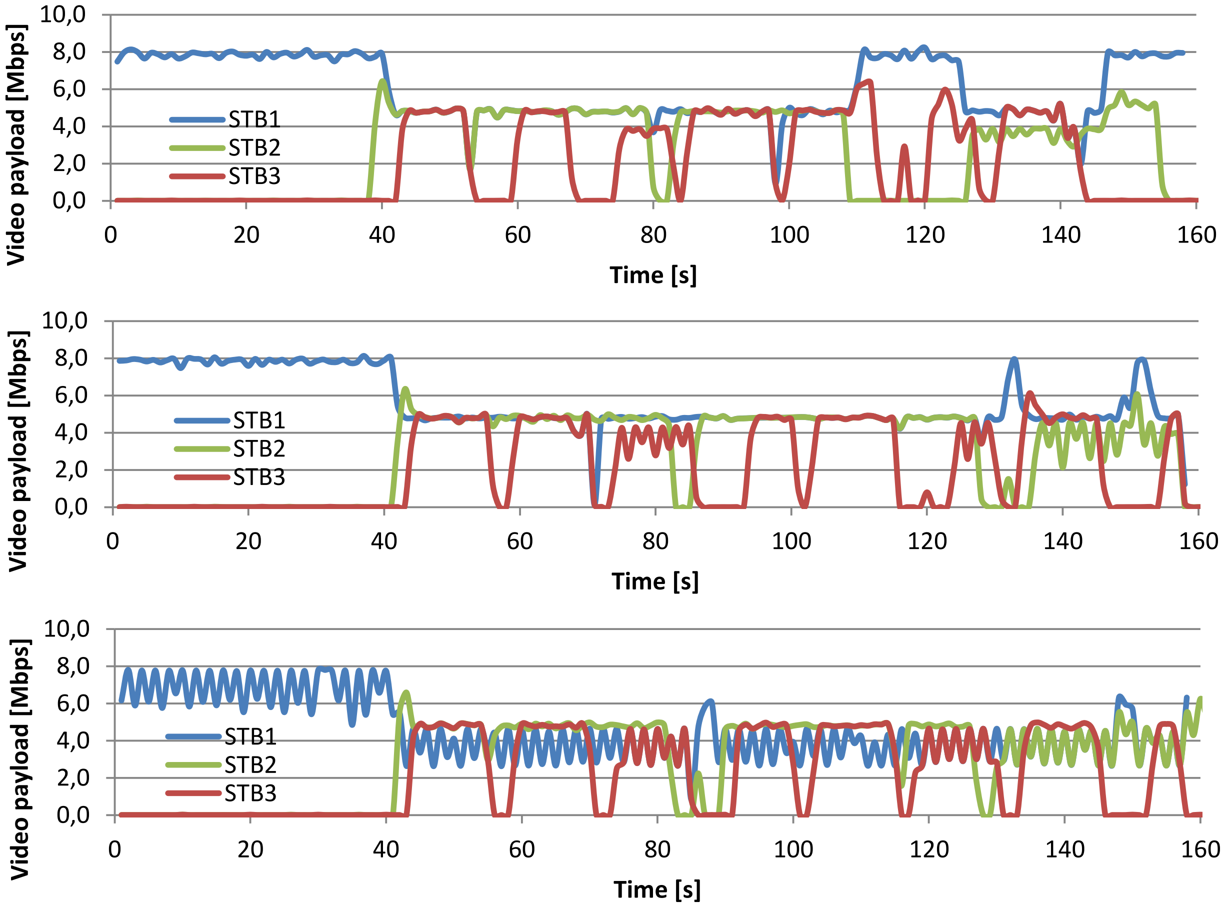

Figure 2: Temporal video stream payload profile of fast zapping patterns for use case 1 at STB1, 2 and 3. Top: action, middle: soccer, bottom: documentary

Our experimental conditions were planned to involve intervals of 15 second-periods between channel switching. By contrast, the temporal profiles show unexpectedly long channel transitions (which can be estimated by referring in the figures to the 0-values of STB 2 and 3. In accordance to this observation, low average payloads and high variances at STB2 and 3 can be observed. Even more notably, we often see sudden drops of payload volume that are synchronous to the mentioned transitions, see for example second 98 in the action sequence (top part of the figure), second 76 in the soccer sequence (middle) and second 85 in the documentary sequence (bottom). Such drops could directly cause visual impairments at STB1.

This also appears to have an impact on the zapping curve shapes of these set top boxes. In consistence to the switching plan, a 45 second transition interval of STB2 and a 15 second transition interval of STB 3 is visible. From the temporal profile, again fast bandwidth changes are visible within the documentary streams.

Content class	Action			Documentary			Soccer		
STB	1	2	3	1	2	3	1	2	3
Zapping pattern	Fast								
Average video payload [Mbps]	5,39	2,84	3,64	3,83	3,13	4,08	4,83	2,93	4,09
Variance of video payload [Mbps]	2,12	4,93	3,65	1,09	4,03	2,10	0,64	4,52	2,31
Maximum video payload [Mbps]	8,23	6,34	5,84	6,31	4,98	6,58	7,91	6,05	6,31
Zapping pattern	Slow with active 3 active STBs								
Average video payload [Mbps]	5,39	3,71	3,10	3,85	3,86	4,37	4,78	3,80	4,50
Variance video payload [Mbps]	1,69	2,69	5,09	0,49	1,31	1,79	0,35	1,29	1,28
Maximum video payload [Mbps]	8,13	6,54	5,01	7,56	5,61	4,99	7,82	4,94	5,01
Zapping pattern	Slow with connecting and disconnecting STB2 and STB3								
Average video payload [Mbps]	6,75	3,29	1,55	5,83	3,37	1,48	6,79	5,17	1,47
Variance video payload [Mbps]	2,37	5,86	5,05	3,14	6,01	4,89	2,07	1,56	4,83
Maximum video payload [Mbps]	8,14	6,24	4,99	7,85	6,30	5,01	8,05	6,50	4,98

Table 11: Video stream statistics for use case 1

Table 11 shows that the maximum payload slightly exceeded 8 Mbps SOTA in all action and one soccer sequence (encoded by OPT encoder). The encoded video streams (documentary) are always below the defined payload in the Look-up table (cf. section 1.5.2). The variance of video payload is substantially higher in all cases at STB2 and STB3. This is caused due to relative long transmission intervals between the content classes (or channels). The statistical values do not indicate considerable differences between the zapping patterns; this may be caused by significant impact of single encoder and the length of zapping interval (between two channels).

Moreover, the relatively high variance values to average video payload indicate lack of system orchestration between PDD and video encoding technology. The minimum bandwidth level (L2) was settled at 4,95 Mbps, but this value in average video payload was exceeded only in 5 test cases at STB1 (at STB1 is running continuously the same content). The average video payload at STB1 over all zapping patterns and contents (summing up to 18 min. of video streaming) is 5,27 Mbps, which should be higher. These relatively low average video payloads at STB1 indicate problems with the OptiBand system integration, and this may negatively influence the experienced quality.

Note that for the presented video streams we were not able to perform more elaborated component analyses, which could show direct relations between the influence factors, because TIs provided only three representative samples for each experimental condition.

The **slow zapping** was designed with longer zapping periods with three active STBs. The average switching period was set to 60 seconds and the whole session contained 2 transmissions (compare section 1.3.4). The resulting streams are shown in Figure 3 (the actual test session starts at second 40).

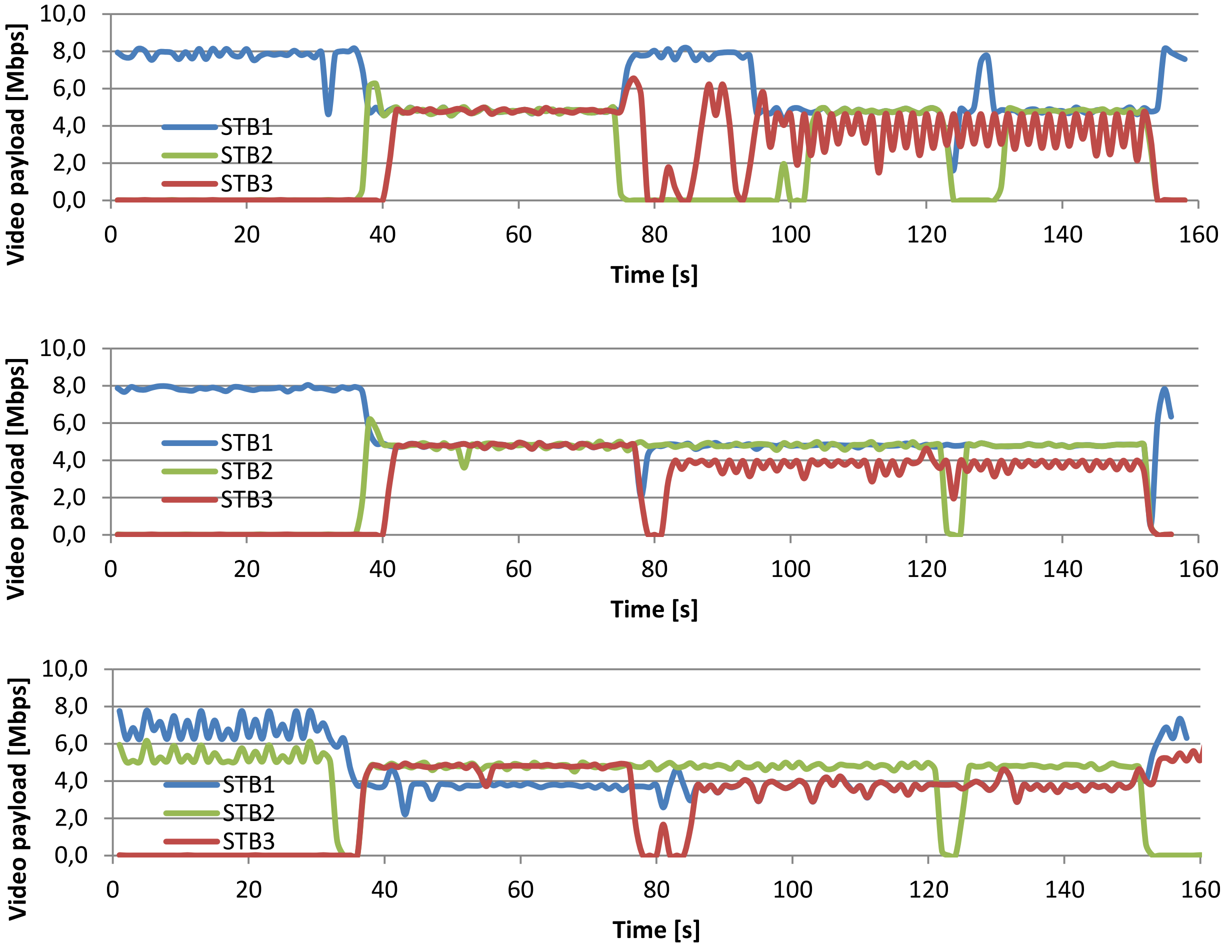

Figure 3: Temporal video stream payload profile of slow zapping patterns with active STB2 and STB3 for use case 1 at STB1, 2 and 3. Top: action, middle: soccer, bottom: documentary

The above figure and Table 11 show that there was a relatively low payload within this experimental condition (3,85 Mbps). This is lower than the recommended minimum payload (see bandwidth level 2 in Table 2).

Furthermore, there was a considerable difference of average payload of STB1 (the test participant's TV) between the content classes: action had the highest payload (5.39), but documentary and soccer only 3.84 and 4.78).

The planned zapping patterns are well visible in the temporal profiles (Figure 3), except for the action content (upper part of the figure). A critical aspect is the relatively long transition between channels, especially in case of zero Mbits periods.

The **slow zapping stream with independent connecting and disconnecting STB2 and STB3** captures are shown in Figure 4 (test session start is at second 40).

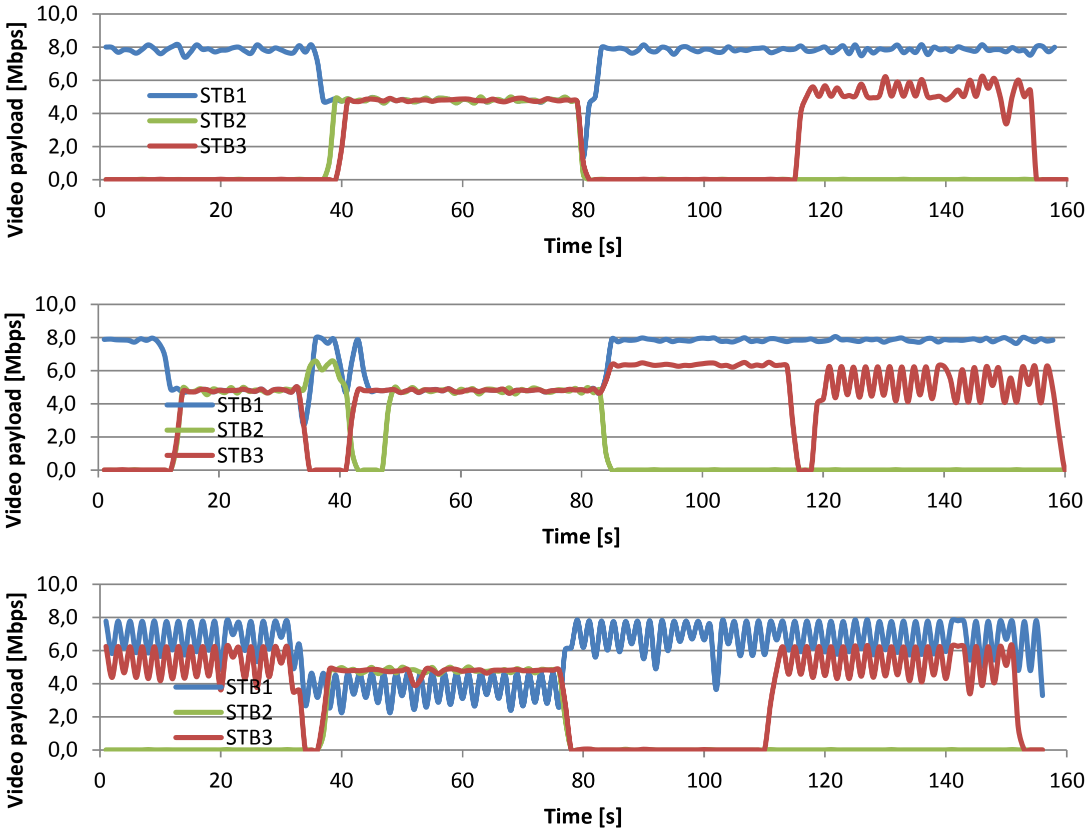

Figure 4: Temporal video stream payload profile of slow zapping patterns with connecting and disconnecting STB2 and STB3 for use case 1 at STB1, 2 and 3. Top: action, middle: soccer, bottom: documentary

The above depicted temporal profiles reflect the expected shape of zapping patterns, with one exception regarding the soccer video (middle part of the above figure. STB3 is still at 6 Mbps between second 90 and 120, although it should have been disconnected in that period. For this use case, the video payload average and variance values shown in Table 11 are not fully interpretable, because disconnecting of STB 2 and 3 has strong impact on these values. For soccer and action contents at STB1, the average values were above the optimal average payload (6.86 Mbps), as was expected. For documentary, however, the average payload was below this value.

Use case 2 (2 TVs, 12 Mbps)

The **fast zapping** stream captures for particular STBs are shown below (see Figure 5, actual test session starting at second 40).

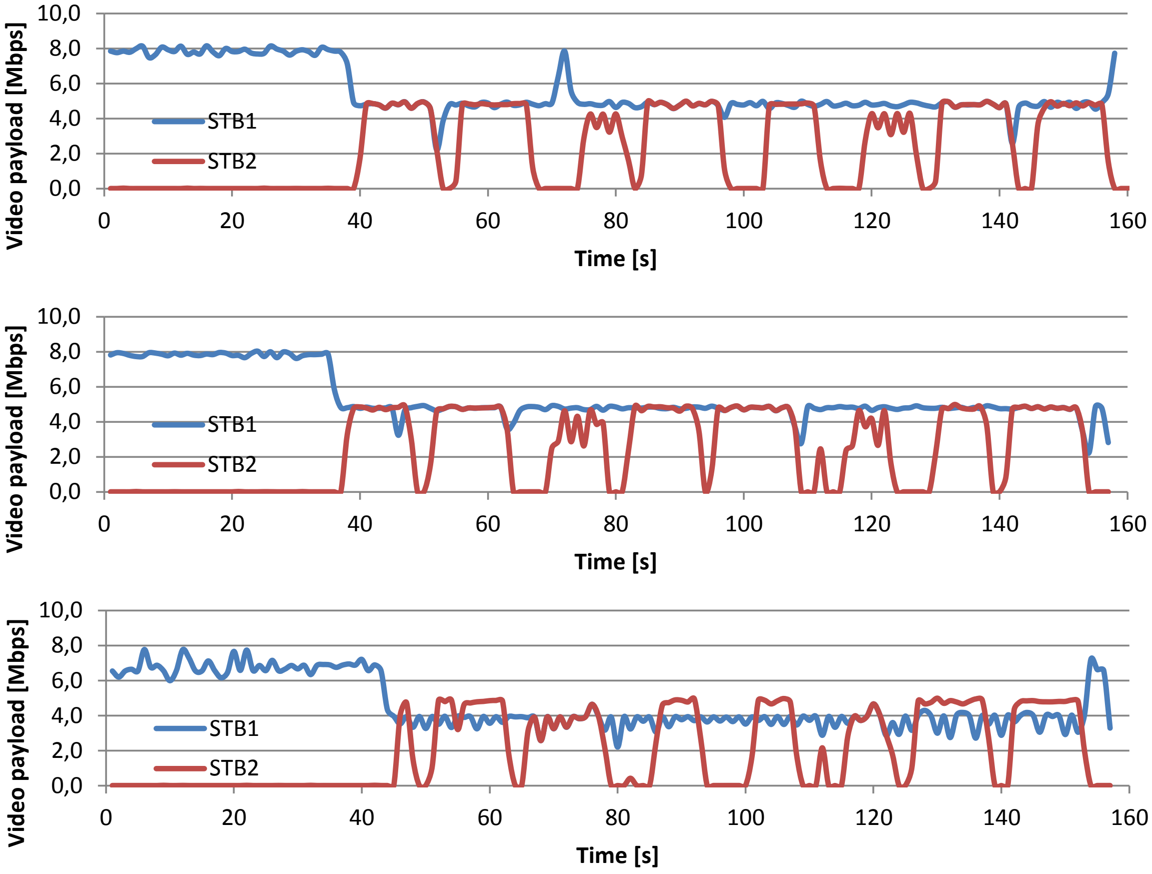

Figure 5: Temporal video stream payload profile of fast zapping patterns for use case 2 at STB1 and 2. Top: action, middle: soccer, bottom: documentary

The depicted video streams in the figure above show similarly long transition times as in use case 1. As Table 12 shows, the average bitrate in all investigated sessions at STB1 in this use case was 4,66 Mbps. These low average payload values indicate inefficient bandwidth resources allocation which indirectly may negatively influence the subjective ratings. Again, one can see the sudden drops in STB1 that appear simultaneously to the zapping transitions of STB2, especially for action and soccer.

Content class	Action		Documentary		Soccer	
STB	1	2	1	2	1	2
Zapping pattern	Fast					
Average video payload [Mbps]	4,82	3,19	3,89	2,70	4,68	3,15
Variance video payload [Mbps]	0,31	4,26	0,72	4,44	0,18	4,18
Maximum video payload [Mbps]	8,02	4,99	7,20	4,99	4,92	4,99
Zapping pattern	Slow with active 3 active STBs					
Average video payload [Mbps]	4,93	3,52	3,68	3,91	4,90	3,99
Variance video payload [Mbps]	1,60	4,00	0,97	2,59	0,35	2,50
Maximum video payload [Mbps]	8,00	5,03	6,25	4,98	8,13	4,99
Zapping pattern	Slow with connecting and disconnecting STB2 and STB3					
Average video payload [Mbps]	4,88	4,06	4,47	2,87	5,65	2,80
Variance video payload [Mbps]	0,31	1,88	2,65	4,25	1,87	4,50
Maximum video payload [Mbps]	8,01	4,99	7,78	5,01	8,14	5,00

Table 12: Video stream statistics for use case 2

The **slow zapping** stream captures for particular STBs are shown in Figure 6 (as always the test session start is at second 40).

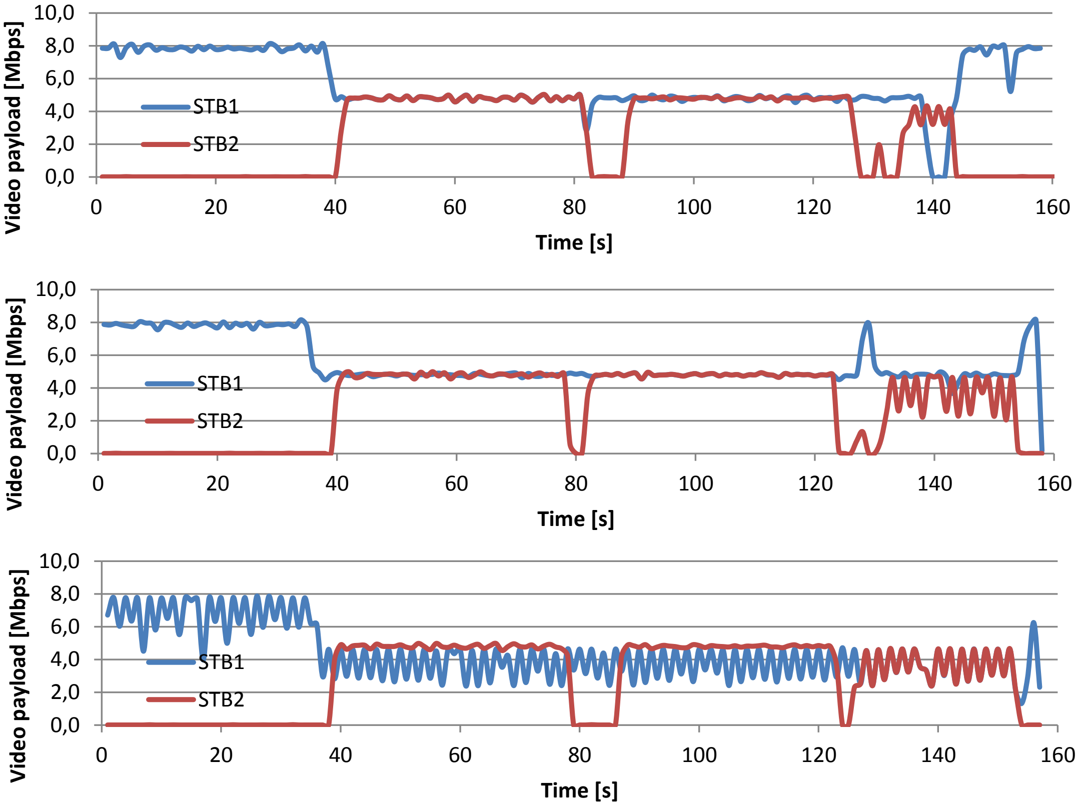

Figure 6: Temporal video stream payload profile of slow zapping patterns with active STB2 for use case 2 at STB1 and 2. Top: action, middle: soccer, bottom: documentary

The zapping pattern specified for this experimental condition is well visible. However, there are again relatively long transition times on the zero Mbps level. During the test session, both STBs apparently were only allocated to the minimum bandwidth level L2 from the look-up-table. The average payloads for action and soccer at STB1 were almost at L2 level (see Table 12). The variances for action and soccer at STB 1 were high due to the problematic last channel transitions at STB2. The documentary stream at STB1 again shows a significantly lower average value at STB1 to L2 level and its variance is relatively high.

The stream captures for **slow zapping with independent connecting and disconnecting STB2** are shown in Figure 7 (the test session starts at second 40).

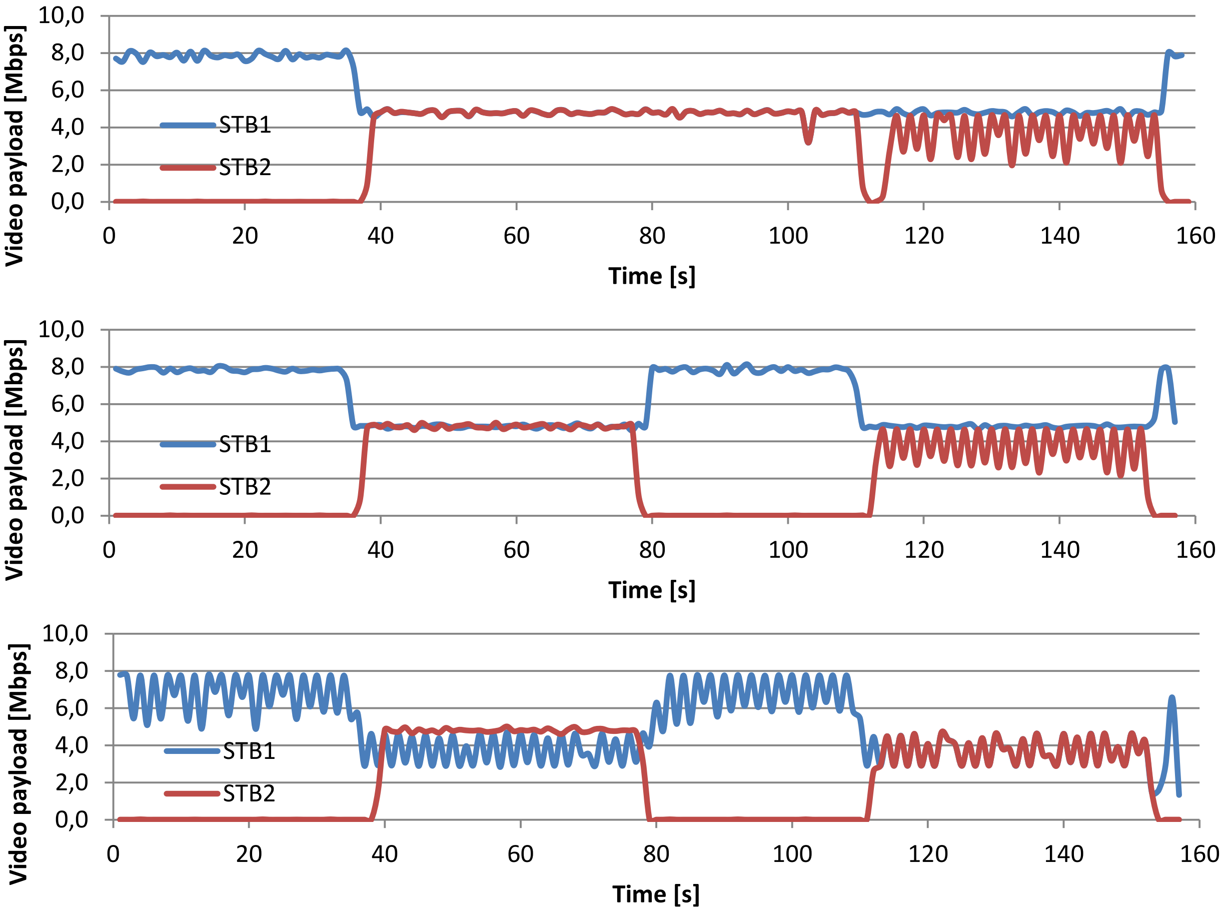

Figure 7: Temporal video stream payload profile of slow zapping patterns with connecting and disconnecting STB2 for use case 2 at STB and 2. Top: action, middle: soccer, bottom: documentary

The depicted video streams (see Figure 7) have standard shapes for soccer and documentary at STB1. In the action sequence (top part of the figure), the STB2 did not disconnect. This as well decreased the average payload at STB1 below the L2 level. On STB1, the documentary content had a significantly lower average value at STB1 to L2 level and variance is relatively high to average payload.

Similarly to the observations of use case 1, the STB2 streams have significantly lower average payloads due to the zapping pattern at these particular STBs (compare Table 12). This is caused due to the long channel switching interval. This naturally increased variance of video payload at STB2. Moreover, the average bitrate over all investigated sessions at STB1 (together 18 min of video streams) in this use case was 4,66 Mbps. Such unexpectedly low average payload values indicate a lack of system orchestration between the PDD and video encoding technology, which indirectly may negatively influence the subjective ratings. The average video payload was surpassed only for soccer content in slow zapping pattern with connecting and disconnecting STB2.

Use case 3 (1 TV, 7Mbps)

Figure 8 provides the temporal profiles for use case 3, which reflects the reduced actual state of technology conditions for IPTV. Only one TV set is considered per household, and the video payload was specified as 6.5 Mbps. As always, the evaluation was performed for three different contents action (top part of the figure), documentary (middle), and soccer (bottom).

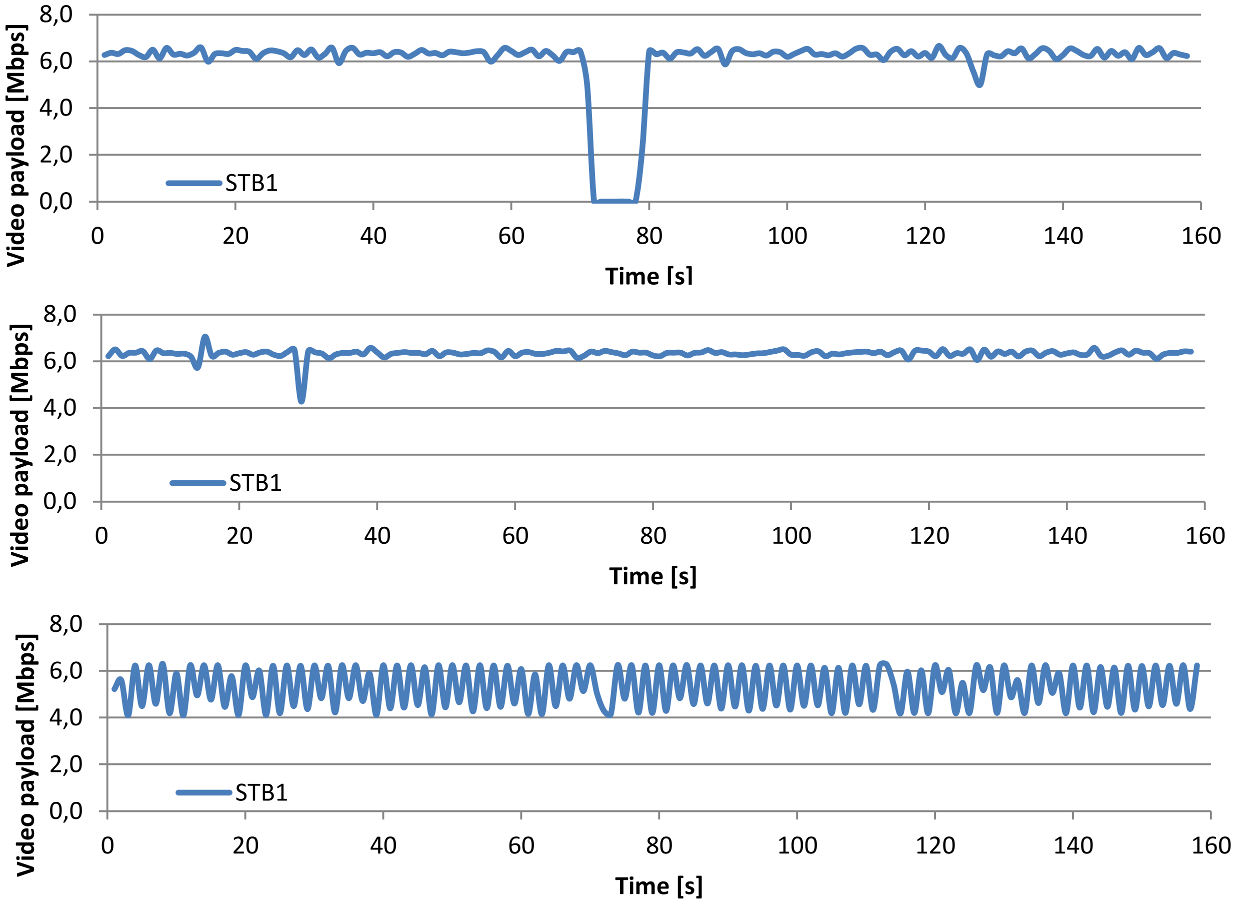

Figure 8: Temporal profile for use case 3 video stream at STB1. Top: action, middle: soccer, bottom: documentary

The depicted video streams (see Table 13) were almost identical to use case 0 except the bandwidth level which is settled at 6,5 Mbps. The planned values were achieved for soccer content at STB1. The statistical results for action were influenced by the outlier in the middle of the session.

Content class	Action	Documentary	Soccer
Average video payload [Mbps]	5,90	5,30	6,34
Variance video payload [Mbps]	2,40	0,79	0,01
Maximum video payload [Mbps]	6,66	6,30	7,05

Table 13: Video stream statistics for use case 3

2. QoE Live Test: Results and Conclusions

This section presents the results and conclusions of the OptiBand Live Test. First, an overview of the rating results for each experimental condition is provided, structured along the use cases. Then, averaged rating values with regard to the involved content classes, zapping patterns, as well as the use cases, are compared. We then conclude with a summary and some critical reflections with regard to live test preparation and conduction.

The user test data that was obtained from the TIS study was reformatted in order to make it suitable for statistical analysis. This included a transposition of participant data in the data table from variables to cases, as well as the summarization of variables for certain analysis purposes. For statistical comparisons of mean differences, we ran analyses of variance (ANOVA) for repeated samples and post-hoc pairwise comparisons (p-values Bonferroni-corrected).

2.1 Overview

Use case 0 (1 TV, 15Mbps)

Figure 9 shows the mean opinion scores (MOS) for use case 0, where participants were watching action, soccer, or documentary clips, with an available bandwidth of 15 Mbps (no other TV set was used in the household). As this use case could be regarded as the baseline condition, where no packet dropping is applied and no constraints are imposed on the network, one should expect excellent quality ratings. This expectation is only supported for the documentary clips, which achieve a MOS of 4.33 (SD=0.67). While action achieved a somewhat satisfactory mean rating of 3.94 (SD=0.94), soccer only achieves a mean of 3.71, SD=0.67. Even with the relatively low sample of 18 viewers, the rating difference between action and documentary was significant, p=0.005 (the other pairwise differences were not significant). Unfortunately, in contradiction to the test plan, participants were not asked for acceptance ratings by TIS experimenters in this use case, thus no respective results can be presented.

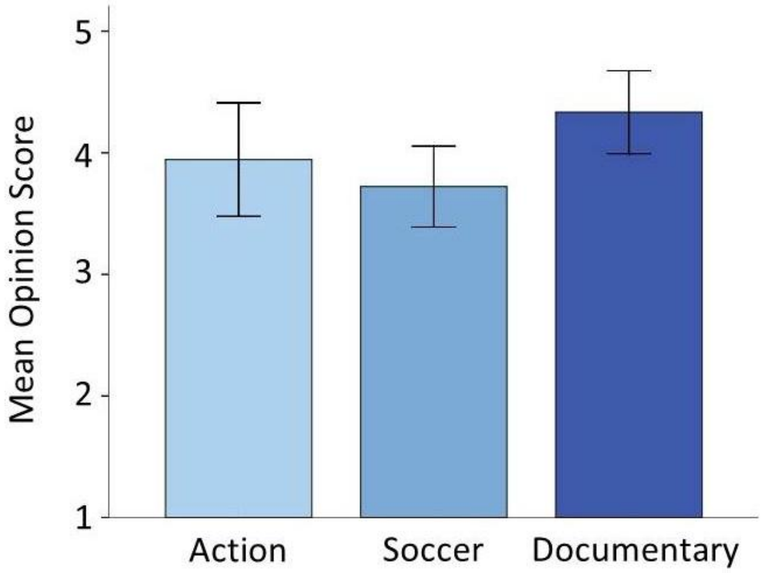

Figure 9: Quality ratings for the experimental conditions included in Use Case 0 (three content types). Error bars indicate 95% confidence intervals.

Use case 1 (3 TV, 15Mbps)

Figure 10 provides an overview of the quality rating scores for the experimental conditions within use case 1, for fast zapping, slow zapping, and slow-zapping with disconnecting STB 2 and 3 (called “Disconnecting” in the figure), realized with each of the three contents.

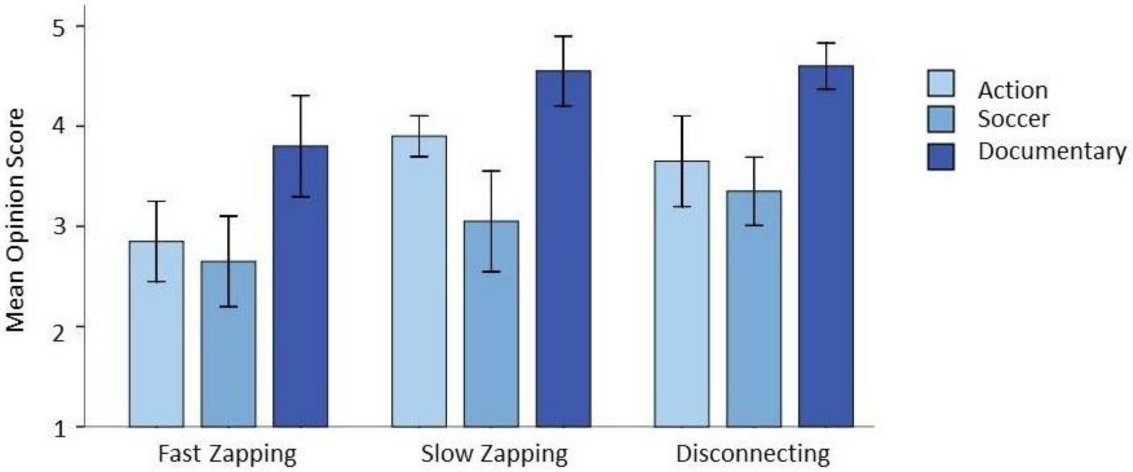

Figure 10: Quality ratings for the experimental conditions included in Use Case 1 (combinations of three zapping patterns and three content types. Error bars indicate 95% confidence intervals.

Generally it can be seen that many of the mean ratings are rather moderate. Actually, only the documentary contents received high ratings for all zapping patterns. Soccer sequences do not surpass mean opinion scores of 3.2.

The fast zapping sequences have considerably lower mean rating values than both the slow and the disconnecting pattern. This effect could have been expected, see section 1.3.2 for a systematic comparison of zapping patterns.

While for Fast Zapping and Disconnecting the action and soccer sequences had similarly low mean ratings, the Slow Zapping lead to a considerably better perceived quality of the action sequences.

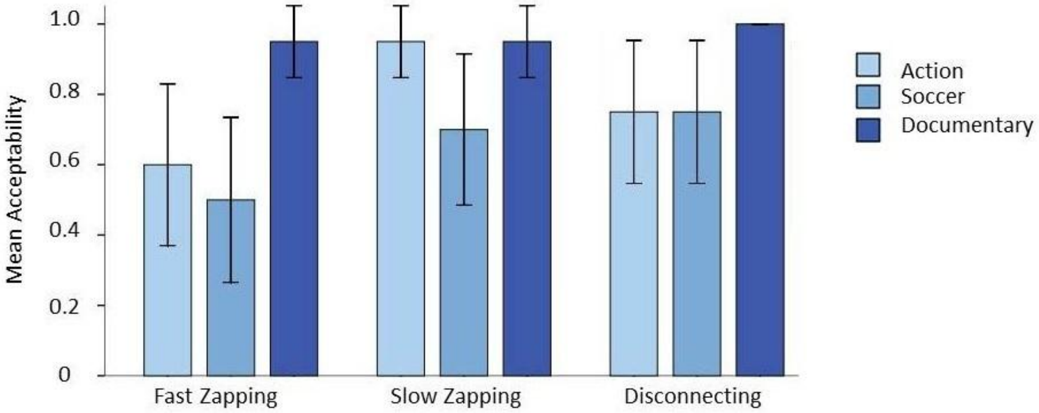

Figure 11: Share of accepted videos for the experimental conditions included in Use Case 1 (combinations of three zapping patterns and three content types).

The share of acceptable videos shown in Figure 11 is strongly correlating with the quality ratings presented before. Acceptance values above a threshold of 80% were only achieved by the documentary clips and the action clip in the slow zapping condition. Again, no differences were visible for the zapping patterns and the use cases (compare sections 0 and 2.2 for a related systematic analysis).

Use Case 2 (2 TVs, 12 Mbps)

Figure 12 provides an overview of the quality ratings for the experimental conditions in use case 2, i.e. all possible combinations of the three zapping patterns and three content types.

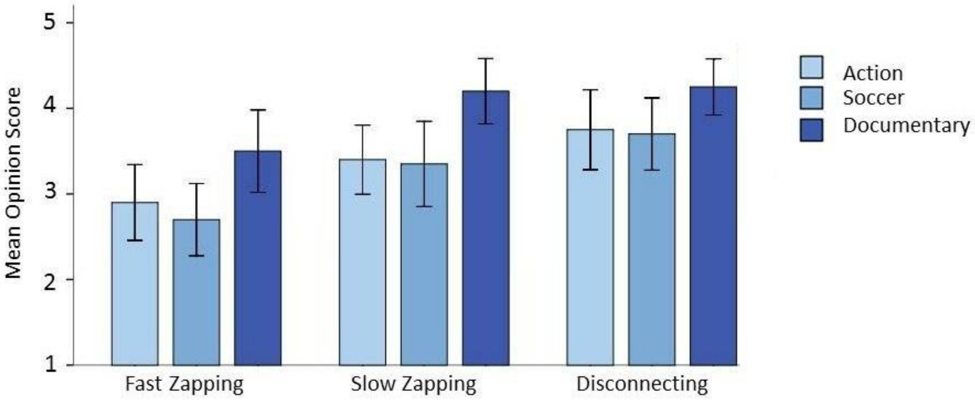

Figure 12: Quality ratings for the experimental conditions included in Use Case 2 (combinations of three zapping patterns and three content types. Error bars indicate 95% confidence intervals.

In principle, the results were quite consistent with those from use case 1. Again, the fast zapping sequences were receiving lower ratings than the slow zapping and disconnecting sequences.

As in use case 1, only the documentary contents were reaching values above a threshold of 3.7, but not any more for the fast zapping. Unlike in use case 1, the action sequence within the slow zapping was not any more a positive outlier.

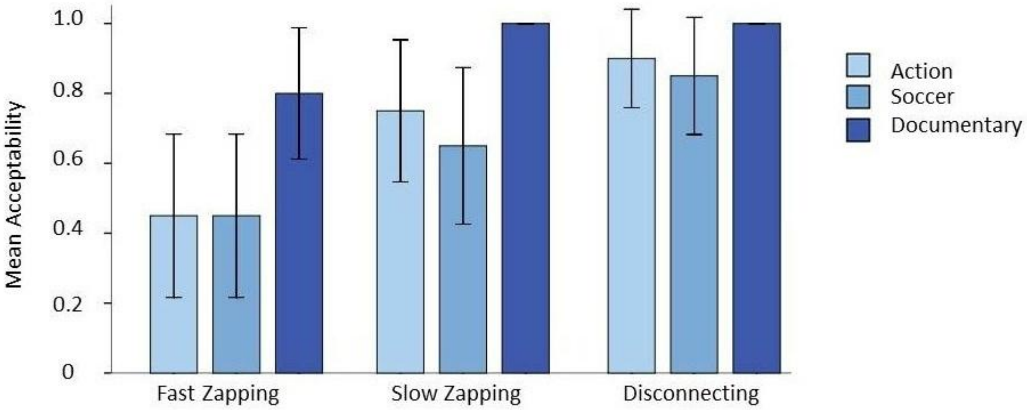

Figure 13: Acceptance ratings for the experimental conditions included in Use Case 2 (combinations of three zapping patterns and three content types). Error bars indicate 95% confidence intervals.

Figure 13 shows the acceptability results for the experimental conditions of use case 2. These basically look consistent with the quality rating scores shown before. However, hardly any of the shown differences are consistent. Note that the documentary contents in the slow and the disconnecting scenarios received 100% acceptance!

Use Case 3 (1 TV, 7Mbps)

The obtained results from use case 3 shown in Figure 14 show a similar pattern to those of use case 0. While use case 3 is also based on only one TV set per household, the difference is that a lower but still substantial bandwidth of 7 Mbps is applied and that the OptiBand packet dropping solution is applied. The

MOS scores for documentary and action are above 4.0 (M=4.11, SD=0.67; M=4.11, SD=0.9). This time, soccer ratings were not much above at the defined MOS threshold of 3.7 (3.44, SD=0.78), these were significantly lower than for action and documentary (p<0.01). Unfortunately, in contradiction to the test plan, no acceptance ratings have been performed for use case 3, thus no respective results can be presented.

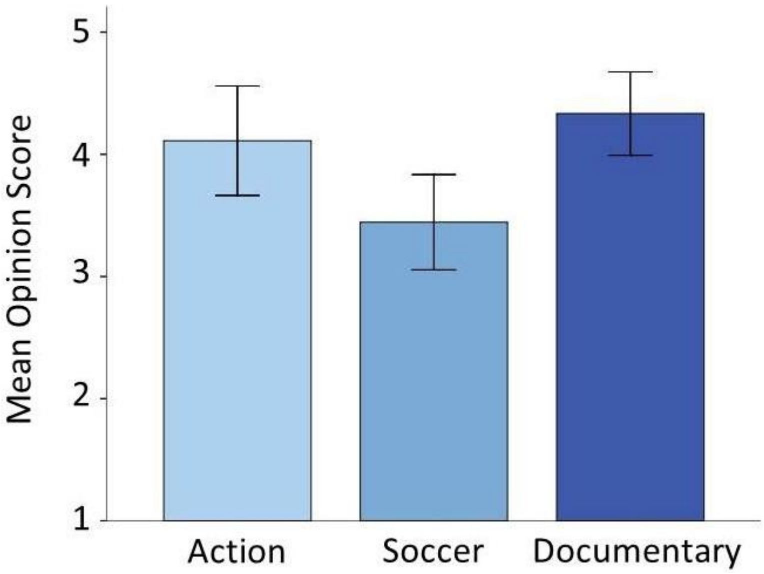

Figure 14: Quality ratings for the experimental conditions included in Use Case 3 (three content types). Error bars indicate 95% confidence intervals.

2.2 Zapping patterns

Figure 15 shows the quality rating and acceptability comparisons between the three zapping patterns, averaged over use cases 1 and 2 and the three content types.

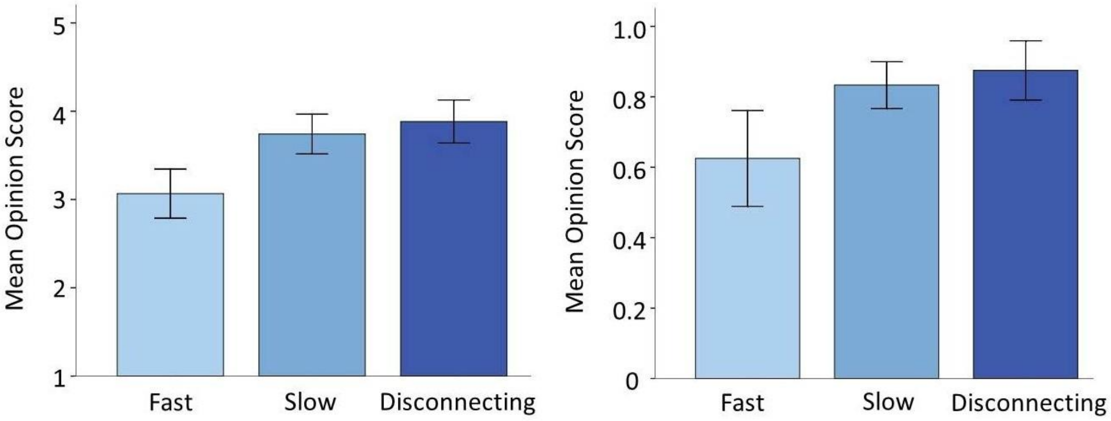

Figure 15: Comparison of Zapping Patterns (averaged over use cases and content classes)

The figure suggests a clearly lower MOS and a slightly lower mean acceptability score of the fast zapping pattern, as compared with the other two zapping patterns. Indeed, a highly significant difference between the zapping patterns could be found for quality and acceptance ratings, $F_{2,58}$ = 35.23 and $F_{2,58}$ = 12.88. The pairwise differences between fast zapping and the other zapping patterns were highly significant, both with the quality ratings and the mean acceptance scores, p < 0.000.

2.3 Content classes

Figure 16 compares the quality ratings and acceptance ratings of the three content classes, as averaged over use cases and zapping patterns.

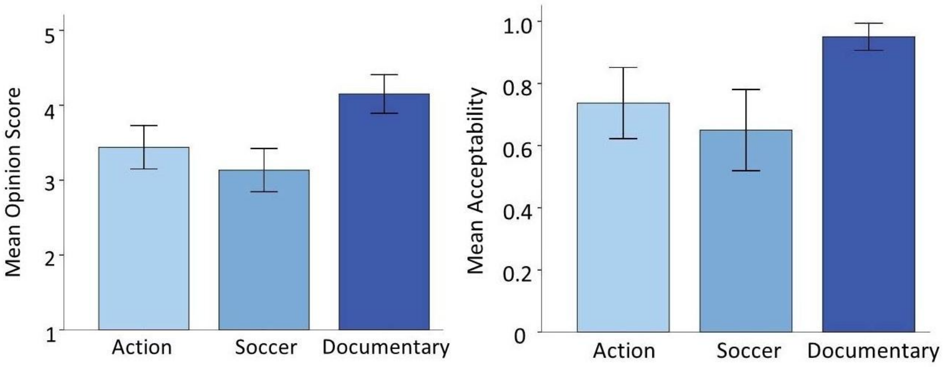

Figure 16: Comparison of content classes (averaged over use cases and zapping patterns). Left: quality ratings; right: acceptance ratings

An ANOVA resulted in a significant difference between the three content types, for both rating values, $F_{2,18}$ = 8.36, p=0.003, F2,18 = 5.04, p=0.18. As can be seen in the figure, the mean values for the documentary contents were highest. However, when applying Bonferroni corrections, documentary was only significantly higher than soccer in the quality ratings and then action in the acceptability ratings.

2.4 Use cases

In this section, we provide a comparative analysis of the OptiBand system performance with regard to the chosen use cases 1 (1 TV, 15Mbps) and 2 (2 TVs, 12 Mbps). Figure 17 provides the quality and acceptability rating values for the four use cases, as averaged over zapping patterns and content classes.

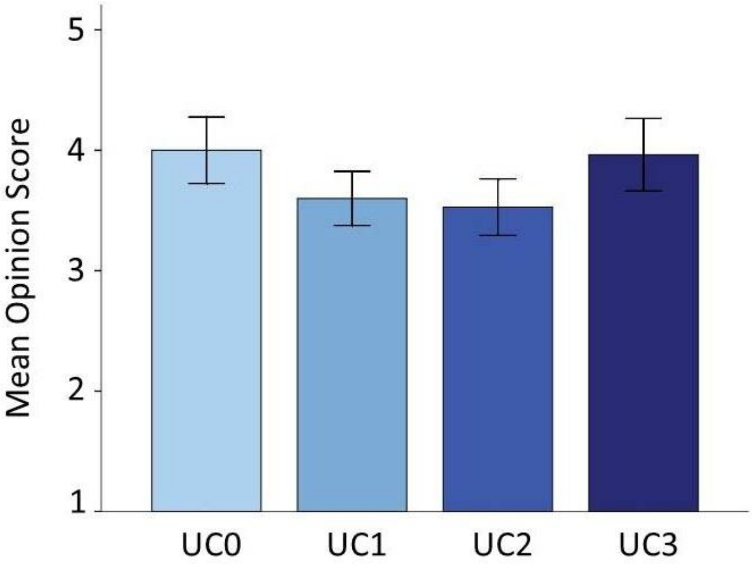

Figure 17: Comparison of quality ratings for the four investigated use cases, averaged over content types and zapping patterns

The profiles in the figures suggest that the MOS scores between the single-TV use cases 0 and 3 and the multiple TV use cases 1 and 2 differed (~ 4.0 vs. 3.7). An analysis of variance supports this assumption, $F_{\mathit{3,23.81}}$=4.18, p=0.4. However, Bonferroni-corrected pairwise comparisons did not result in any significant difference, but the difference between use cases 1 and 0 was near significant. Note that for use cases 0 and 3 only 18 participants were included in the sample and thus only very robust effects in would become significant in related comparisons.

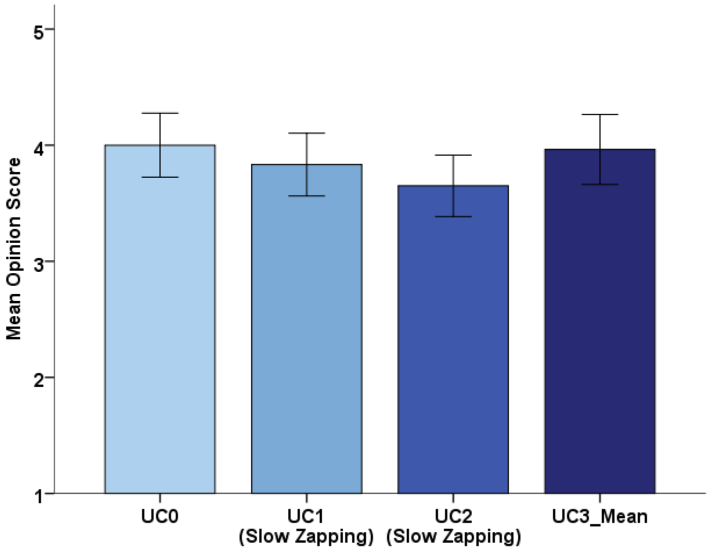

Figure 18: Comparison of quality ratings for the four investigated use cases (for use cases 1 and 2, the slow zapping conditions have been included)

As has been shown in section 2.2, rating values for the zapping patterns strongly differed. As a consequence of this differentiated picture, it is not sufficient to simply plot the average values for use cases 1 and 2. Figure 18 provides again an overview of the use cases, but for use cases 1 and 2 only the slow zapping condition is shown. We can see here that relative degradations are not very large, and also that absolute values are satisfactory.

2.5 Conclusions

In the following, the above presented results are summarized and interpreted, in order to provide answers to the formulated research questions.

Absolute quality in the baseline condition

For the interpretation of the performance and experience of the OptiBand solution, it is of course important to first review the results for use case 0, which represented 'optimal' system conditions with only 1 TV set and available 8.5 Mbps. Here, the average experienced quality of 4.0 MOS was good. However, in comparison with the scores for the state-of-the-art sequences within the previous testing iteration, MOS scores were about 0.2 lower (4.0 vs. 4.2 for action, 3.7 vs. 3.9 for soccer). This means that only with slight relative impairments of 0.3 MOS, the absolute MOS threshold would be reached.

In principle, as an explanation for these issues, the reference videos employed for the test could have been of bad quality and thus be responsible for these lower subjective ratings. However, as the same video sources had received good scores in previous QoE studies [7], this cannot serve as a valid explanation in this case. However, we do not know whether the transcoding of the videos conducted at TIS premises have resulted in potential damages. Anyway, the same equipment and transcoding method were in use for use-case 0 and for the rest of the use-cases, hence resulting with same video symptoms if any existed.

Relative quality impairments in challenging conditions

One of the main OptiBand objectives was to reduce the necessary bandwidth of an HD stream by more than 33% while preserving acceptable perceived quality. To this end, the comparison of use case 3 with use case 0 show that a reduction by 1.5 Mbps did not materialize in noteworthy degradations of the mean opinion score (compare section 0). In the more challenging use cases 1 and 2, where the available bandwidth was decreased by more than 33% per TV set, average quality values tended to be lower by about 0.3 MOS. However, we also found that with slow zapping patterns, the decline of perceived quality is almost invisible (compare the very small effect size in Figure 18).

It is the place to mention that due to PDD implementation resources considerations it was decided to use strict GOP size requirement of 2 seconds, without keeping the option for configuration. This has the potential to affect the QoE during zapping scenarios.

Thus, we can conclude that with optimized tuning and zapping strategies, small relative impairments can be achieved with the OptiBand solutions.

Absolute quality in challenging conditions

While the focus of the OptiBand QoE investigations is on assessing relative quality degradations, it is of course also important to look at the absolute achieved quality. In order to interpret these results, we had defined a threshold of 3.7 MOS and applied throughout the project. The result of the previous testing iteration, as reported in D2.4, was that all 1080i sequences exceeded this threshold, with values of 3.8–4.0 MOS, even with reductions of 55%. Based on these experiences, our expectation was that with the selected bandwidth levels the subjective service perception should not have been below 3.9 MOS, thus comfortably surpassing the 3.7 MOS threshold.

The results of the live test for use case 3 show that a reduction by 1.5 Mbps using the OptiBand packet dropping solution resulted in a satisfactory absolute level of subjective quality (~4.0 MOS). The results for the more demanding use cases 1 and 2 are more differentiated, and they seem to be especially dependent on the underlying zapping scenario. We can conclude that absolute quality is satisfactory in slow zapping situations, as the quality threshold of 3.7 MOS is slightly surpassed in the respective test conditions (with an average MOS score of 3.7, compare Figure 18 in section 0). Also, disconnections and reconnections of other set top boxes do not seem to have a strongly disturbing impact. This picture is supported by the acceptance testing results, where the lowest acceptable threshold of 0.80 was slightly surpassed in slow zapping conditions (0.83). We expect that a higher absolute perceptual quality of baseline conditions of use case 0 would also have resulted in better quality within use cases 1 and 2.

An open point for the OptiBand solution is still how to deal with fast zapping conditions. Here, relatively low rating values of 3.1 MOS and an acceptance ratio of 0.63 have been obtained. The selected bandwidth levels for H.264 streaming services should have resulted in excellent streaming performance, however this was not the case in the presented test. Possibly, this may be related to problems with the co-ordination of involved system components (PDD, network, and encoding technology), or more likely, due to implementation constrains in terms of resources and time for system stabilization. As can be seen in the analysis of the video stream characteristics (section 1.5.3), the drop of STB 1, the too long channel switching intervals, the high variance on many streams, and the high variance on the documentary streams may have caused interferences and impairments resulting in limited quality.

Content classes

Regarding the content classes, we found that the documentary sequences received the best MOS and acceptability ratings, although they usually had lower average payloads. Action and soccer contents were very similar to each other. This result cannot be explained in a straight-forward manner. On the one hand we could assume that documentary contents might have been more appealing to users, which could have resulted in higher ratings. However, the strength of the effect is still surprising, especially when considering the fact two previous studies had not found significant differences between content classes within the same original video materials.

Another possibly decisive factor may have been the applied encoder: TVN provided the encoder for the documentary materials, and OPT provided the encoder for the action and soccer sequences. Thus, in principle, the encoders may have had different efficiency levels with regard to achieving high QoE levels.

However, as mentioned before, there are several missing information layers that make an interpretation difficult. Most importantly, we do not know about the effects of the transcoding that had to be executed by TIS for the integrated prototype. Also, we do not dispose of the video recordings of the test session, so that we do not know about the actual artefacts that participants were experiencing.

3. The Influence of Personal Characteristics on QoE Perception

As specified in the OptiBand DoW, one goal of WP2 was to understand the influence of context on QoE perception. As outlined in D2.1, the context of viewing can be regarded as one of the main blocks of influence factors, along with expectations towards the offered service and the technical system properties (compare Figure 19). This impact of the viewing context on quality perception and rating behaviour has largely been neglected in perceptual quality research. The lack of knowledge about the needs of certain customer segments may undermine the value of QoE study results as a solid basis for marketing decisions. If the influence of contextual factors is unknown, it will also not be possible to design personalized QoE management schemes. In particular, we identified the following challenges that we found worthwhile addressing to improve the state-of-knowledge:

Personal characteristics: The impact of the user's personal characteristics, such as gender, age, prior experiences with HD services and educational background on the quality rating are so far unknown. Insufficient understanding of related influences could undermine the generalizability of QoE tests and limit the validity of design recommendations. Within the OptiBand project, we took advantage of the many user tests conducted under a consistent framework, which provided us with more trustable evidence on impacts especially of gender and age.

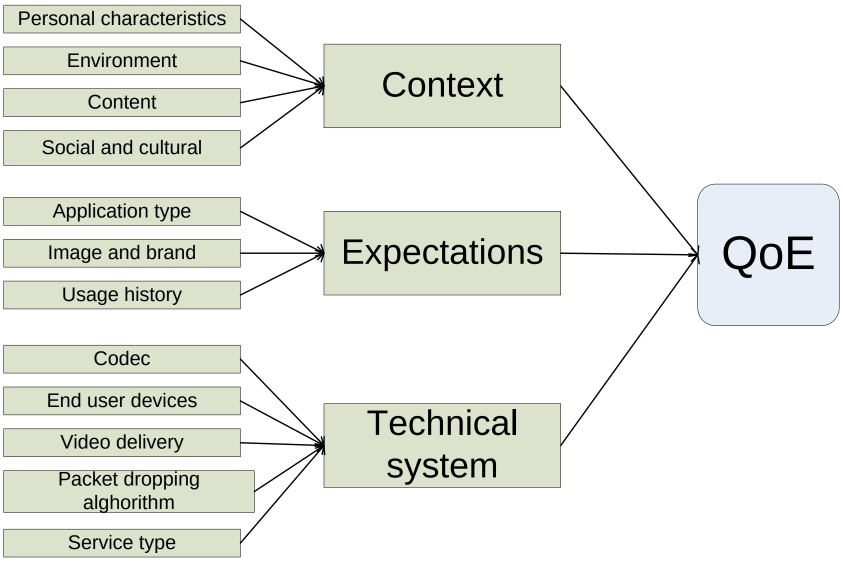

Figure 19: Influence factors on QoE (from OptiBand deliverable D2.1)

Environment: Due to the focus on standardized, laboratory-based viewing environments in QoE evaluations, it is not clear whether people in natural viewing situations at home or elsewhere would qualify a given video sequence similar as in lab situations. Only recently, QoE evaluations have been extended to more natural environments, such as through field studies. A particularly innovative methodology is 'QoE crowdsourcing', that is, the remote evaluation via electronic questionnaires, distributed via channels such as Facebook. This aspect will only be fully touched in this report, but interested readers are directed to respective research papers ([7],[1]), which have been co-authored by FTW and benefitted from the knowledge exchange and co-operation with OptiBand.

Content: There is agreement that the viewed content has a high impact on quality perception, regardless the bandwidth and other performance parameters. We have gained findings within the OptiBand project with regard to content types ([4],[5]), clip lengths [6] and clip quality distribution [8].

Social and cultural aspects may also have a strong impact on the way video quality is perceived. However, related effects are almost impossible to gather in typical laboratory-based QoE evaluations. As the OptiBand project did not involve systematic QoE studies in different countries, this aspect could not be covered in this report.

3.1 Research Questions

As mentioned above, we were focusing in this report on the impact of certain selected personal characteristics. The main variables of interest for researching the impact of personal characteristics were as follows:

Gender: While gender is one of the standard demographic variables considered important in the design of QoE studies, and while it is a common wisdom to balance the sample between male and female participants, hardly any knowledge is available on actual systematic differences in rating behaviour. One may hypothesize that males tend to a stronger affinity to technology innovations (such as HD television), and that therefore quality improvements or impairments may be monitored with more interest than by females (who might concentrate more on the content). Furthermore, due to different content preferences (such as lower interest of females for soccer), subjective quality expectations and requirements may be different (e.g. impairments in soccer are valued less critical).

Age is another key demographic variable that is balanced in proper QoE studies, as obviously varying quality ratings across age groups are assumed. As also here no systematic investigations are known, one can only speculate that younger people tend to have better visual perceptual capabilities and therefore are more critical towards poor contents. Another correlation could exist to the available budget for TV equipment, where middle-aged persons might be in the position to acquire and consume HD services, which may make them more critical towards possible quality impairments.

Use of HD services: One of the personal characteristics for QoE research that appear to be directly related to QoE perception may be the prior experience with regard to HD services and content. Possibly, people with more HD watching experience may have higher quality expectations and requirements, and thus be more susceptible to quality impairments. On the other hand, persons with HD watching experience are already used to typical impairments, such as blockiness or stalling, and thus be more indulgent in cases where they appear.

Other demographic data typically collected in QoE studies are the educational status (school and university degrees) and the professional status (i.e., current employment type). However, these variables do not seem to be directly enough related, in order to form clear hypotheses, and thus they were not in the focus of our analysis.

3.2 Method

As the basis for analysing the personal characteristics, we analysed, compared and pooled several data sources to gain a higher test participants sample.

Included data:

- **OptiBand Main QoE study, 1st project iteration [4]:** The most important data sources were the OptiBand main laboratory studies from the first two iterations. The study compared the QoE of video sequences with different bandwidth levels, packet dropping approaches, and content classes. As is shown in Table 14, participants participated in the test, with a balanced gender and age distribution.
- **OptiBand Main QoE study, 2nd project iteration [5]:** The second iteration test had exactly the same laboratory conditions, samples, and test procedures. However, due to improvement of the packet dropping algorithms, the overall quality of the presented video materials improved, so that MOS ratings were higher in the second iteration test. The number of participants of these tests was 50 (for more details).
- **OptiBand content duration study [6]:** This study comprised 65 users. The selection of presented videos and chosen bandwidth levels was different to the main studies, and as a consequence these were analysed separately.
- **QoE Crowdsourcing studies I and II ([1],[7]):** These two studies have been conducted at the University of Zilina in co-operation with FTW, using consistent video materials and methods. The data gathered from the crowdsourcing studies was reviewed with regard to gender and age (HD usage was not included in the questionnaire).

We would have been interested to analyse the data from the live test as well (see sections 1 and 2 in this report). However, we were not able to obtain the required data from TIS.

	N of participants	Gender Distribution	Age Distribution
Main study (first iteration)	28	male: 15 female: 13	AG1: <30: 15 (53.6%) AG2: 30-45: 9 (32.1%) AG3: 45+: 4 (14.3%)
Main study (second iteration)	50	male: 22 female: 28	AG1: <30: 26 (52%) AG2: 30-45: 13 (26%) AG3: 45+: 11 (22%)
OptiBand Content Durations Study	65	male: 33 (50,8%) female: 32 (49,2%)	AG1: <30: 35 (53.8%) AG2: 30-45: 19 (29.2%) AG3: 45+: 11 (16.9%)
Crowdsourcing Study I	33	male: 24 (72.7%) female: 9 (27.3%)	AG1: <30: 23 (69.7%) AG2: 30-45: 8 (24.2%) AG3: 45+: 2 (6.1%)
Crowdsourcing Study II	76	male: 64 (84.2%) female: 12 (15.8%)	AG1: <30: 42 (55.3%) AG2: 30-45: 16 (21.1%) AG3: 45+: 4 (5.3%)

Table 14: Sample characteristics of analysed user studies

Analysis approach:

All data sets were analysed separately with regard to the research questions specified above. For the analysis of each of the personal characteristic, the data pools were split into the respective personal characteristics. Potential effects of bandwidth levels were also taken into account, that is, high- and low bandwidth levels were compared, in order to identify specific rating patterns in dependence of the quality level. Pairwise statistical differences were tested by means of T-Tests for independent samples.

Furthermore, a ‘meta-analysis” was conducted with the pooled data. For this purpose, in each of the five samples, the quality rating data was normalized. These normalized values were then used for the analysis of the pooled data. In the following, the results for the characteristics of interest are presented.

3.3 Results

Table 15 provides comparisons within the five analysed studies between male and female perceptual quality ratings. We see that overall there are no significant gender-related differences. When looking only at low quality sequences we see that in the first OptiBand main study, male participants were more critical than female participants (M=2.58 vs. 3.05, p=0.017). In the content durations study we also found a significant difference for low quality sequences, but the effect size was very small (0.05 MOS), and the error probability of p=0.45 cannot be regarded as robust enough, if more strict statistical procedures are applied (such as Bonferroni corrections). There were no gender-related significant differences for the high quality files.

	Overall Difference	**Influence of quality level**	
		Low quality	**High quality**
Main study (first iteration)	Male: M=3.150; SD=0.623 Female: M= 3.345; SD=0.509 P= 0.103	Male: M=2.580; SD= 0.695 Female: M= 3.049; SD= 0.586 **P=0.017**	Male: M=3.640; SD=0.602 Female: M= 3.608; SD= 0529 P=0.689
Main study (second iteration)	Male: M=4.028; SD=0.472 Female: M=4.067; SD=0.580 P=0.974	Male: M=3.950; SD=0.562 Female: M=3.99; SD=0.573 P=0.978	Male: M=4.067; SD=0.509 Female: M=4.108; SD= 0.589 P=0.819
OptiBand Content Durations Study	Male: M=3.191; SD=0.426 Female: M=3.028; SD=0.433 P=0.251	Male: M=1.578; SD=0577 Female: M=1.497; SD=0.346 **P=0.045**	Male: M=4.287; SD=0.621 Female: M=4.004; SD= 0.813 P=0.775
Crowdsourcing Study I	Male: M=3.972; SD=0.167 Female: M=4.011; SD=0.222 P=0.890	Male: M=2.584; SD=0.280 Female: M=3.000; SD=0.373 P=0.380	Male: M=4.131; SD=0.216 Female: M=4.774; SD= 0.287 P=0.085
Crowdsourcing Study II	Male: M=3.752; SD=0.725 Female: M=3.275; SD=0.753 P=0.289	Male: M=3.192; SD=0.948 Female: M=3.000; SD=1.054 P=0.383	Male: M=4.083 SD=0.787 Female: M=3.550; SD= 0.797 P=0.312
Meta Analysis (normalized values)	Male: M=0.029; SD=1.007 Female: M=-0.008; SD=0.994 P=0.480	Male: M=0.039; SD=0.969 Female: M=-0.068; SD=1.047 P=0.226	Male: M=-0.029; SD=1.015 Female: M=0.063; SD=0.925 P=0.845

Table 15: Comparison between male and female participants' ratings in five QoE user related to the OptiBand project, with regard to averaged mean opinion scores (M), standard deviations (SD), and statistical significance comparisons (P).

The bottom row of the above table shows that pooling the data from the 252 participants within the five user studies did not result in significant differences between male and female participants.

Table 16 below provides the mean values, standard deviations and statistical significance values for the comparison between the three age groups (AG1, AG2, and AG3, compare section 3.2).

	Overall Difference	**Interaction with quality level**	
		Low quality	High quality
Main study (first iteration)	AG1: M=3.548; SD= 0.124 AG2: M=2.772; SD= 0.158 AG3: M= 3.120; SD= 0.235 P= AG1, AG2=1.000 AG1, AG3=0.767 AG2, AG3=0.482	AG1: M=3.043; SD= 0.157 AG2: M=2.401; SD= 0.199 AG3: M=2.750; SD= 0.297 P= AG1, AG2=0.057 AG1, AG3=1.000 AG2, AG3=1.000	AG1: M=3.987; SD= 0.108 AG2: M=3.080; SD= 0.138 AG3: M= 3.469; SD= 0.206 P= **AG1, AG2=0.000** AG1, AG3=0.109 AG2, AG3=0.390
Main study (second iteration)	AG1: M=4.006; SD= 0.113 AG2: M=3.923; SD= 0.144 AG3: M= 4.244; SD= 0.172 P= **AG1, AG2=0.003** AG1, AG3=0.366 AG2, AG3=0.696	AG1: M=3.941; SD= 0.119 AG2: M=3.829; SD= 0.151 AG3: M= 4.182; SD= 0.181 P= AG1, AG2=0.057 AG1, AG3=1.000 AG2, AG3=1.000	AG1: M=4.029; SD= 0.118 AG2: M=3.954; SD= 0.150 AG3: M= 4.279; SD= 0.180 P= **AG1, AG2=0.000** AG1, AG3=0.109 AG2, AG3=0.390
OptiBand Content Durations Study	AG1: M=3.161; SD= 0.075 AG2: M=3.038; SD= 0.116 AG3: M= 3.068; SD=0.149 P= AG1, AG2=1.000 AG1, AG3=1.000 AG2, AG3=1.000	AG1: M=1.630; SD= 0.081 AG2: M=1.410; SD= 0.126 AG3: M= 1.479; SD= 0.161 P= AG1, AG2=0.459 AG1, AG3=1.000 AG2, AG3=1.000	AG1: M=4.261; SD= 0.117 AG2: M=4.104; SD= 0.181 AG3: M= 3.755; SD= 0.233 P= AG1, AG2=1.000 AG1, AG3=0.172 AG2, AG3=0.726
Crowdsourcing Study	AG1: M=3.952; SD= 0.526 AG2: M=3.857; SD= 0.241 AG3: M= 4.142; SD= 0.101 P= AG1, AG2=1.000 AG1, AG3=1.000 AG2, AG3=1.000	AG1: M=2.891; SD= 0.818 AG2: M=2.609; SD= 0.631 AG3: M= 2.500; SD= 0.000 P= AG1, AG2=1.000 AG1, AG3=1.000 AG2, AG3=1.000	AG1: M=4.456; SD= 0.456 AG2: M=4.083; SD= 0.880 AG3: M= 4.500; SD= 0.707 P= AG1, AG2=1.000 AG1, AG3=1.000 AG2, AG3=1.000
Meta Analysis (normalized values)	AG1: M=0.144; SD=0.087 AG2: M=-0.306; SD=0.135 AG3: M=0.036; SD=0.187 P= **AG1, AG2=0.033** AG1, AG3=1.000 AG2, AG3=0.838	AG1: M=0.144; SD=0.087 AG2: M=-0.349; SD=0.134 AG3: M=-0.129; SD=0.186 P= **AG1, AG2=0.014** AG1, AG3=1.000 AG2, AG3=1.000	AG1: M=0.112; SD=0.086 AG2: M=-0.195; SD=0.132 AG3: M=0.013; SD=0.183 P= AG1, AG2=0.314 AG1, AG3=1.000 AG2, AG3=1.000

Table 16: Comparison of age groups among the user studies, with regard to averaged mean opinion scores (M), standard deviations (SD), and statistical significance comparisons (P).

When looking at the test results for the pooled data, one can see that 'middle-aged' participants (30-45 years) provided significantly lower ratings than younger participants (below 30 years). While this trend is visible in most of the studies, the effect size is rather small (at about 0.1-0.2 MOS), and this difference appears to materialize only with large sample sizes.

Table 17 shows that neither of the comparisons between participants using HDTV versus participants not using HDTV resulted in a significant difference. Also the comparison with the pooled data (143 participants) did not result in a significant difference.

	Overall Difference	**Low Quality**	**High Quality**
Main study (first iteration)	Using HD: M=2.983; SD=0.368 Not Using HD: M=3.268; SD=0.120 P=0.469	Using HD: M=2.318; SD=0.714 Not Using HD: M=2.845; SD=0.472 P=0.093	Using HD: M=4.148; SD=0.446 Not Using HD: M=3.979; SD=0.411 P=0.511
Main study (second iteration)	Using HD: M=4.115; SD=0.112 Not Using HD: M=3.944; SD=0.107 P=0.275	Using HD: M=3.603; SD=0.802 Not Using HD: M=3.426; SD=0.639 P=0.389	Using HD: M=4.148; SD=0.419 Not Using HD: M=4.035; SD=0.651 P=0.478
Content duration study	Using HD: M=3.278; SD=0.306 Not Using HD: M=3.105; SD=0.418 P=0.182	Using HD: M=1.786; SD=0.509 Not Using HD: M=1.663; SD=0.549 P=0.654	Using HD: M=4.152; SD=0.687 Not Using HD: M=4.046; SD=0.721 P=0.492
Meta Analysis	Using HD: M=0.149; SD=0.715 Not Using HD: M=-0.079; SD=0.864 P=0.153	Using HD: M=0.012; SD=0.859 Not Using HD: M=-0.086; SD=0.812 P=0.545	Using HD: M=0.096; SD=0.805 Not Using HD: M=-0.094; SD=0.938 P=0.280

Table 17: Comparison of HD usage vs. no HD usage, with regard to averaged mean opinion scores (M), standard deviations (SD), and statistical significance comparisons (P).

3.4 Conclusions

We analysed the results of 5 QoE studies with 252 participants with regard to correlations of perceived quality with personal characteristics. We did not find evidence for systematic differences related to gender. A practical implication for QoE tests with a constrained budget could be that experimenters may prioritize other aspects in the experimental design and that they do not necessarily need to perfectly balance a test sample by inviting exactly the same number of males and females.

Also, QoE ratings did not differ significantly between people using HDTV and people not using HDTV. We think that especially this aspect needs to be investigated with more rigour in future research studies, in order to get a more fundamental understanding of the influence of prior TV watching experience. Informal reports of QoE testing experts suggest that rating results are usually influenced by the extent to which viewers have been exposed to advanced TV services, such as HD IPTV. Our study was somewhat constrained in providing final answers to this questions, as we did not recruit participants for systematic comparisons with regard to this factor.

Regarding age, we found indications that persons between 30 and 45 might be slightly more critical than persons below 30 years. Further studies are needed to replicate these effects and correlate them with other potentially relevant attitudes and personality traits of viewers. For example, it would be interesting whether the observed lower ratings by people between 30 and 45 are correlated with better TV watching equipment. In order to test this, demographic questionnaires should be extended accordingly and the test sample needs to be composed such that participants are distributed to all possible combinations between age groups and prior experience.

We think that pooling the data of similar tests, as we have presented it in this report, should be further intensified. Such efforts will help to determine the relevance of certain QoE context factors, and they should thus be intensified on an international scale.

4. References

[1] Gardlo, B., Ries, M., Rupp, M., and Jarina, R., “A QoE evaluation methodology for HD video streaming using social networking,” in Multimedia (ISM), 2011 IEEE International Symposium on, dec. 2011, pp. 222 –227.

[2] 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.

[3] Fröhlich, P., Ries, M., Egger, S., D2.2 - Detailed research plan. Internal OptiBand project document, 2011.

[4] 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.

[5] Fröhlich, P. Ries, M., Fuchs, M., Stürmer, T., Masuch, K., Schatz, R. „D2.4 – Intermediate QoE research recommendations report. Available at the OptiBand project homepage: www.optiband-project.eu”, 2011.

[6] Fröhlich, P., Egger, S:, Schatz, R., Mühlegger, M., Masuch, K., and Gardlo, B. QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment In Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshopon, Australia, July 2012.

[7] Gardlo, B., Hoßfeld. T., Schatz, R. and Ries M.. Microworkers vs. Facebook: The impact of crowdsourcing platform choice on experimental results. In Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on, Australia, July 2012.

[8] Fröhlich, P., Ries, M., Masuch, K., Schatz, R.(2012). Investigating the Effects of Test Clip Quality Distribution in HD Video Quality-of-Experience Studies Proc. QoMEX 2012: International Workshop on Quality of Multimedia Experience.