Influence of Viewing Experience and Stabilization Phase in Subjective Video Testing

Christian Keimel, Arne Redl and Klaus Diepold

Institute for Data Processing, Technische Universität München
Arcisstr. 21, 80333 München, Germany

ABSTRACT

In this contribution, we will examine two important aspects of subjective video quality assessment and their overall influence on the test results in detail: the participants' viewing experience and the quality range in the stabilization phase. Firstly, we examined if the previous viewing experience of participants in subjective tests influence the results. We performed a number of single- and double-stimulus tests assessing the visual quality of video material compressed with both H.264/AVC and MPEG2 not only at different quality levels and content, but also in different video formats from 576i up to 1080i. During these tests, we collected additional statistical data on the test participants. Overall, we were able to collect data from over 70 different subjects and analyse the influence of the subjects' viewing experience on the results of the tests. Secondly, we examined if the visual quality range presented in the stabilization phase of a subjective test has significant influence on the test results. Due to time constraints, it is sometimes necessary to split a test into multiple sessions representing subsets of the overall quality range. Consequently, we examine the influence of the quality range presented in the stabilization phase on the overall results, depending on the quality subsets included in the stabilization phase.

Keywords: subjective testing, viewing experience, stabilization phase, test design, influence on subjective test results, visual quality assessment

1. INTRODUCTION

Subjective testing still plays an important role in video quality assessment, as no universally accepted video quality metric reflecting the perceived visual quality exists yet. Hence it is important to be aware of the possible effects of the test setup on the results of such tests.In this contribution, we will examine two important aspects of subjective video quality assessment and their overall influence on the test results in detail: the participants' viewing experience and the quality range in the stabilization phase.

Firstly, we therefore examine if the previous viewing experience of participants in subjective tests influence the results. In order to obtain results representative of the general viewing audience, the test subjects should represent a sufficient range of different viewing experience. But especially in the academic community, the test subjects are often recruited from the student body, and furthermore mostly from within the same department. Also the viewing experience of many participants can differ widely, as some may still use a CRT for SDTV, while others may already use a large LCD for HDTV.

We performed a number of tests assessing the visual quality of video material compressed with both H.264/AVC and MPEG2 not only at different quality levels and content, but also in the different video formats 576i, 720p and 1080i. Additionally, we used single and double stimulus methods in the subjective testing. During these tests, we collected additional statistical data on the test participants. Besides information about their viewing environment at home with respect to display type and size, commonly used video format and source of the content, we also noted the field of study and semester of participating students. Overall, we were able to collect this information from over 70 different subjects. We then analysed the data using principal component regression

Further author information: (Send correspondence to Christian Keimel)

Christian Keimel	E-mail: christian.keimel@tum.de	Telephone: +49 89 289 23629
Arne Redl:	E-mail: redl@mytum.de	Telephone: +49 89 289 23601
Klaus Diepold:	E-mail: kldi@tum.de	Telephone: +49 89 289 23602

Image Quality and System Performance IX, edited by Frans Gaykema and Peter D. Burns, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 8293, 829313 · © 2012 SPIE-IS&T · doi: 10.1117/12.907967

(PCR) and partial least squares regression (PLSR), in order to determine if different viewing experiences of the test subjects have any influence on the overall results of the tests.

Secondly, we examine if the visual quality range presented in the stabilization phase of a subjective test has significant influence on the test results. A stabilization phase is usually included in subjective tests in order to provide the test subjects with hidden anchors for the overall quality range of the test. Due to time constraints, it is however sometimes necessary to split a test into multiple sessions representing subsets of the overall quality range. The reasoning is that the attention of the test subjects deteriorates with increasing duration of a session, leading to less reliable results. In such cases the question arises, if the stabilization phase of these subsets should provide anchors for the quality range of the complete test or just for the current subset?

In order to answer this question, we therefore examine in this contribution two different experimental setups. In both cases we split our test into a low quality and high quality session. In the first setup, we only include anchors representing the subset of the quality range of the current session, whereas in the second setup we include anchors representing the overall quality range of complete test. For a comprehensive overview, we evaluate in this contribution the visual quality achieved by different coding technologies at different bitrates. Besides the well known H.264/AVC[1], we also take alternative coding technologies into account by including the wavelet based *Dirac*[2] encoder into our evaluation.

Hence, we compare both cases by splitting up a test set into a low and a high quality subset, both to in combination representing a wide range of different bitrates and content. We then use either only anchors corresponding to the subset or anchors covering the complete quality range. Consequently, we examine the influence of the quality range presented in the stabilization phase on the overall results.

This contribution is organized as follows: we review related work before introducing the subjective testing methods used in this contribution. Then we first discuss the influence of the subjects' viewing experience. Following this first main part, we examine the influence of the stabilization phase on the results of subjective test, before concluding with a short summary.

2. RELATED WORK

Contributions so far discussing the influence of different parameters on the results of subjective tests mostly focused on technical issues e.g. Pinson and Wolf[3], discussed the influence of monitor resolution, whereas Keimel and Diepold[4] examined the influence of different monitor types. Others focused on test methodologies: Corriveau et al.[5] considered the effects of different rating scales on the context dependencies of the votes and Baroncini [6] suggested improvements to make double stimulus tests more interesting for test subjects.

For the second part of this contribution, this is to the best of our knowledge the first contribution explicitly examining the relationship between the visual quality of the anchor sequences in the stabilization phase and the overall achieved visual quality.

3. SUBJECTIVE TESTING

All tests were performed in the video quality evaluation laboratory of the Institute for Data Processing at the Technische Universität München in a room compliant with recommendation ITU-R BT.500[7] as shown in Fig.1.

Two professional 24 inch LCD reference displays with a native resolution of 1920 × 1080 pixels were used (Cine-tal Cinemagé 2022 and Sony BVM-L230). The decoded videos were converted to 4:2:2 YC_BC_R for output to the reference display via a HD-SDI link. Due to the screen size, only two viewers for HDTV and four viewers for SDTV took part in the test at the same time to allow stable viewing conditions for all participants. All test subjects were screened for visual acuity (Snellen chart) and color blindness (Ishihara charts). The distance between the screen and the observers was three times the picture height (3H) for HDTV and six times the picture height (6H)for SDTV.

For the double stimulus tests, we used a variation of the standard DSCQS test method as proposed by Baroncini[6]. This Double Stimulus Unknown Reference (DSUR) test method differs from the standard DSCQS test method, as it splits a single basic test cell in two parts: the first repetition of the reference and the processed video is intended to allow the test subjects to identify the reference video. Only the repetition is used by the

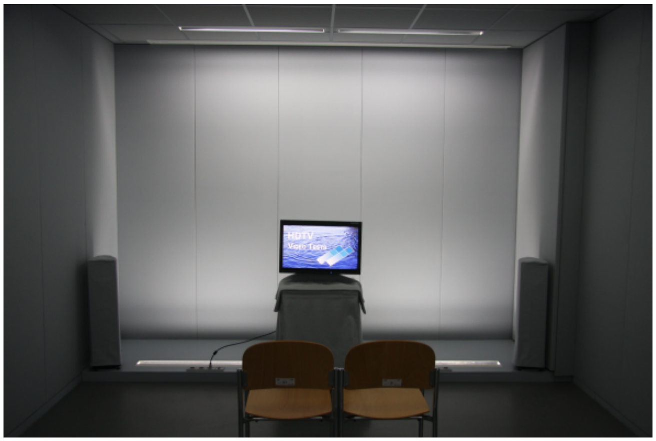


Figure 1: Test room

viewers to judge the quality of the processed video in comparison to the reference. The structure of a DSUR basic test cell is shown in Fig.2. In order to verify if the test subjects were able to produce stable results, a small number of test cases were repeated during the test. Processing of outlier votes was done according to ITU-R BT.500[7] and the votes of one test subject were removed based on this procedure. The mean opinion score (MOS) was calculated by averaging the valid votes for each test case

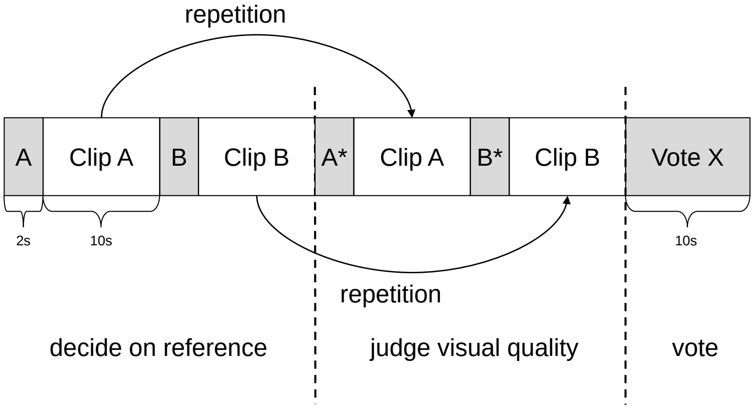


Figure 2: Basic test cell DSUR

For the single stimulus tests, we used a variation of the well known SSIS method standardized in ITU-R BT.500[7], the Single Stimulus Multimedia Method(SSMM). Instead of an impairment scale as in SSIS, SSMM uses a scale that directly evaluates the quality perceived by the test subject. This method has been used extensively in the MPEG standardization of H.264/AVC[8], its extension SVC[9] and in previous contributions[10]. The structure of a SSMM basic test cell is shown in Fig. 3.

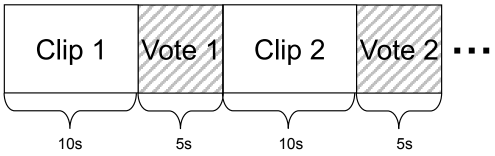


Figure 3: Basic test cell SSMM

One problem in single stimulus methods can be the influence of the context in which the test sequences are shown: the visual quality of the previous sequence influences the perceived quality of the current test case. In order to avoid context effects, every test case was therefore shown twice in a different context. Both votes of each test case are then averaged during processing of the result. During the processing of the votes, votes were rejected if the difference between the two votes for one test case was larger than three scale units. Always both votes were removed. If more than 15% of all votes of a subject had to be removed, the subject and its votes were removed completely, as this strongly indicates that the test subject is not able to reproduce its quality estimation. We then determined MOS by averaging all valid votes for each test case.

A discrete voting scale with eleven grades ranging from 0 to 10 was used for both DSUR and SSMM, to allow the test subjects to differentiate between relatively small quality differences. Before the test itself, a short training was conducted with ten sequences of different content to the test, but with similar quality range and coding artifacts. During this training the test subjects had the opportunity to ask questions regarding the testing procedure.

4. INFLUENCE OF VIEWING EXPERIENCE

Quite often a diverse group of subjects participate in subjective testing with not only a different background, but also a different viewing experience. In this section we will discuss the possible influence of these experiences on the results of the subjective testing.

4.1 Setup

During an extensive subjective testing campaign, we compiled additional statistical data about the participating subjects, not only about their education, gender and age, but also about their viewing experience i.e. the viewing environment in which the subject usually consume video.

The testing campaign included video sequences with a wide range of different content in both wide screen (16:9) and normal (4:3) 576i PAL-SDTV, 720p and 1080i HDTV. The SDTV and HDTV video sequences were encoded according to the MPEG2 and H.264/AVC video standards, respectively. Also different broadcast grade hardware encoders were used with three and five different bitrate points for HDTV and SDTV, respectively. These rate points cover a realistic range of visual quality from poor to high quality, representative of the quality levels encountered in real life broadcasting. Depending on the availability of an uncompressed reference we used the DSUR method or the SSMM method. An overview of the used sequences is given in Table 1.

Table 1: Video sequences

Format	Testing Method	Content	Encoders	Rate Points
576i (16:9)	SSMM	Movie, sport, news channel with crawl, documentation	4	5
576i (4:3)	SSMM	Cartoon, sitcom	4	5
	DSUR	Test sequence	4	5
720p	DSUR	Game show	3	3
1080i	SSMM	Movie, documentation	3	3
	DSUR	Game show	3	3

All in all, 108 and 38 different combinations of encoder, rate points and video sequences were tested with the single stimulus SSMM method and the double stimulus DSUR method, respectively. With on average 18 subjects per test case, this resulted in nearly 800 individual votes or data points. As the focus of this contribution is on the influence of the subjects viewing experience on the subjective test results and not on the comparison of different encoders, we will omit further details about the different encoders.

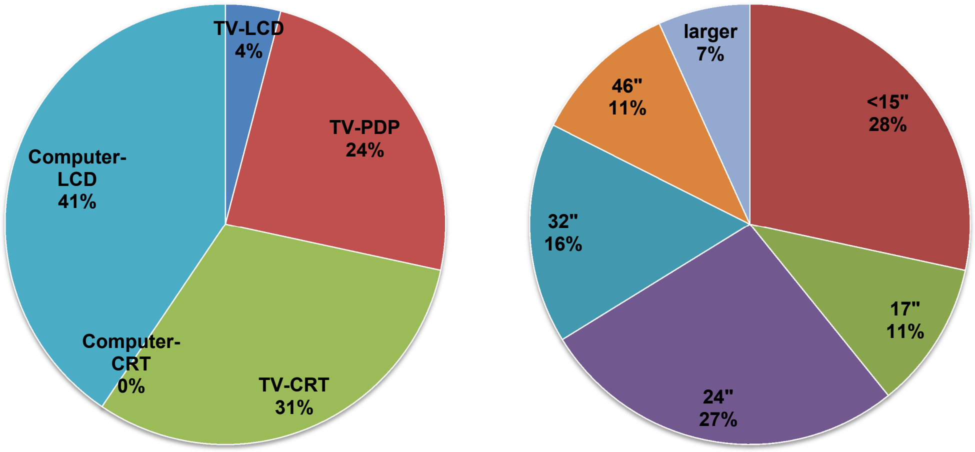
4.2 Statistical data

In total, 74 different subjects participated in the tests. All of them were student between age 20 and 30, 80% male and 20% female. Most of them (85%) were engineering or computer science students. The tests were conducted between April and July 2010.

After the subjects participated in the tests, we provided them with a questionnaire aimed at determining their usual viewing environment in which they *consume* video, with respect to the following parameters:

- **Display type**: TV-LCD, TV-PDP, TV-CRT, Computer-LCD, Computer-CRT
- **Screen diagonal** in inch
- **HDTV capability of display**: None, *HD ready*(720p), *FullHD*(1080i/p)
- **HDTV content format**: None, 720p, 1080i, 1080p, not applicable (NA)
- **HDTV content source**: Cable, Satellite, Blu-ray, Internet, not applicable (NA)

In Fig. 4, the display type and screen diagonal of the subjects are shown and in Fig. 5(a) how many of the subjects had access to a HDTV capable display, and if so what format they mainly used and where the content was obtained from in Fig. 5(b) and Fig. 5(c), respectively.



(a) Display type (b) Screen diagonal in inch

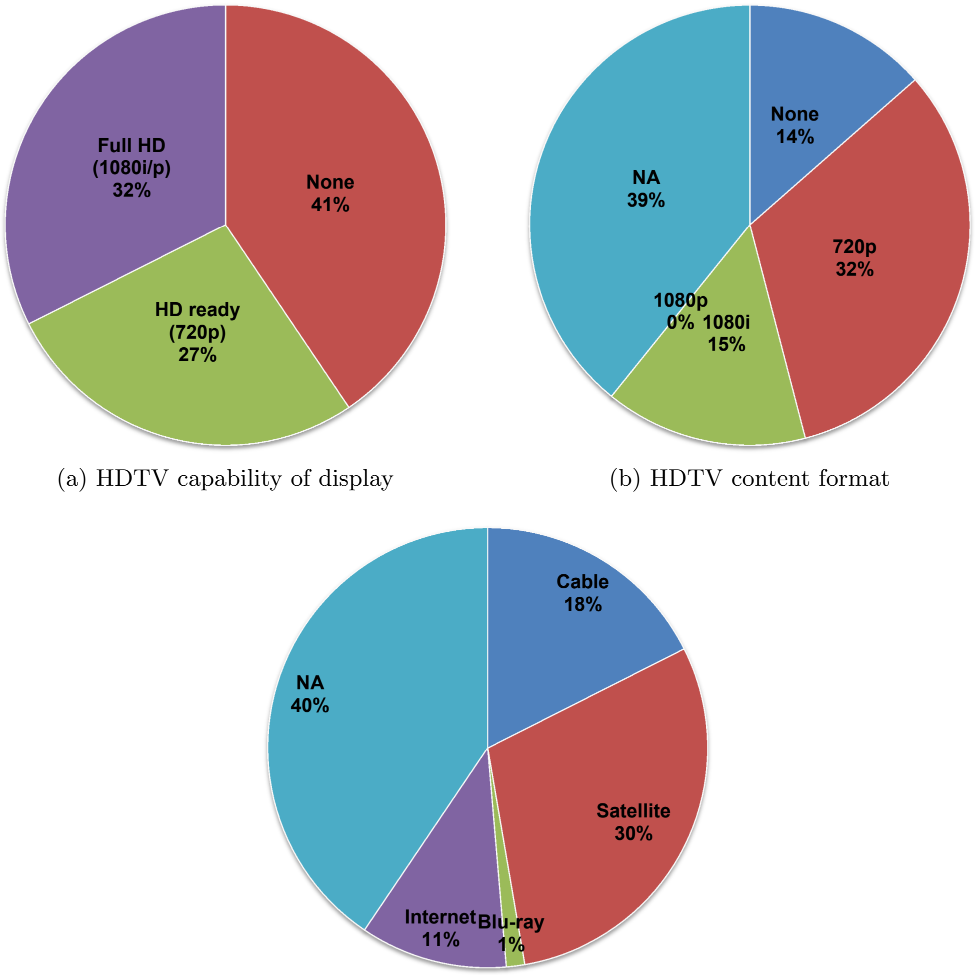
Figure 4: Display type and screen diagonal of test subjects' viewing environment

4.3 Processing of the data

After collecting both the subjective test results in the form of MOS scores for each data point and the statistical data for each subject as described above, the data was analysed in order to determine possible relationships between a subject's votes and its viewing experience.

Firstly, the deviation of each subject's vote from the MOS score for each video sequence was calculated: a negative value representing a subject's vote below the MOS, a positive value representing a subject's vote above the MOS for the corresponding video sequence. We then averaged these values over all sequences of the same format and testing methodology for each subject. Hence, each subject's voting preferences were pooled into six values, describing the six different categories *576i(16:9) SSMM* , *576i(4:9) SSMM*, *576i(4:3) DSUR*, *720p DSUR*, *1080i SSMM* and *1080i DSUR* with respect to the voting characteristics of each subject.

Following this pre-processing, we applied different data analysis methods, namely Multi-linear Regression (MLR), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). The aim is to build a (linear) model that explains the voting characteristics of each subject with their viewing experience, where the *display type*, *screen diagonal*, *HDTV capability*, *HDTV content format* and *HDTV content source* represent the independent variables and the voting characteristic for each category represents the variable to be regressed on. For details on PCR and PLSR we refer to Jolliffe[11] and Martens and Martens[12], respectively.



(a) HDTV capability of display

(b) HDTV content format

(c) HDTV content source

Figure 5: HDTV capability of displays, viewed HDTV content and source of the HDTV content.

4.4 Results

Unfortunately, the results of the analysis are rather inconclusive. Regardless of the applied method, the best model correlation that could be achieved was lower than 0.1, indicating that the collected variables are not able to explain the observed variance in the different subjects' MOS scores adequately. Considering all variables, the *screen diagonal* seems to be the parameter with the highest possible influence in our data set, showing a negative weight, and thus indicating that the larger the screen diagonal of the subject's usual viewing environment is, the lower the visual quality of the videos sequences was judged on average.

These results suggest, that the selected statistical data is not sufficient to capture the subjects' viewing experience properly and that therefore the questionnaire needs to be modified and hence different parameters describing the subjects' viewing experience are needed.

Still, the existing, albeit minor influence of the screen diagonal indicates that the test subject's experience seems to have some influence on the overall results of the subjective testing, even if we were not able to quantify it and build a proper model at this point.

5. INFLUENCE OF STABILIZATION PHASE

In this section, we will discuss the influence of the stabilization phase on the results of the subjective tests by first providing a short introduction into stabilization in subjective testing, before describing the test setup and the results.

5.1 Stabilization Phase

The stabilization phase in visual subjective testing includes a number of n basic test cells (BTC) at the beginning of the test session. Typically, n is chosen to be between 3 to 5. Each BTC represents a set of the video sequence under test and an appropriate time allotted for voting. Depending on the methodology, one BTC for a 10s video sequence takes in total between 15s for single stimulus and 60s for double stimulus non-continuous test methodologies. Following the stabilization phase, the video sequences to be tested, from $n+1$ to m are presented in the test phase, as shown in Fig. 6. The results of the votes given in the stabilization phase are not processed and therefore not included in the overall results of the test itself.

In the stabilization phase, video sequences representing at least the worst and best quality to be encountered during the test are shown to the participating test subjects. But also sequences representing mid-level quality may be shown. The aim is to give the subjects an impression of the quality range under test. This is done in order to avoid that test subjects are either too generous or too scarce with their votes in the beginning. Otherwise, in the first case, the test subjects may give votes on the upper end of the scale in the beginning, even tough sequences with higher visual quality will occur later on in the session. But this quality difference won't be visible in the votes of the subject, as the subject already was close to the upper end of the scale and thus has much less space on scale left for differentiation; the second case is similar, just on the other end of the quality scale.

Hence these examples *anchor* the inherent quality scale of each subject on the worst and best quality in the test session. Therefore these sequences are usually called low quality or high quality anchor. Although the exact visual quality of theses anchors is of course not known a-priori, a selection of appropriate sequences is made in the test setup usually based on the experience of the test designers. Note that the subjects are not aware of the stabilization phase: from their perspective they are already doing the proper test.

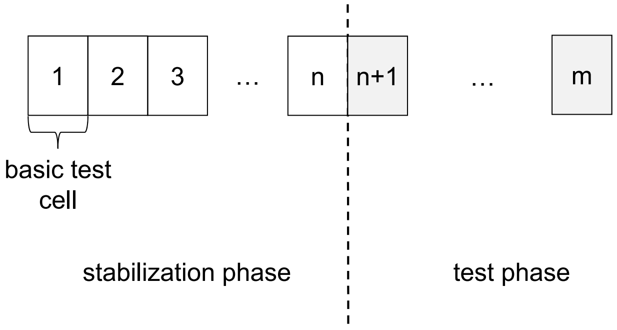


Figure 6: Stabilization phase

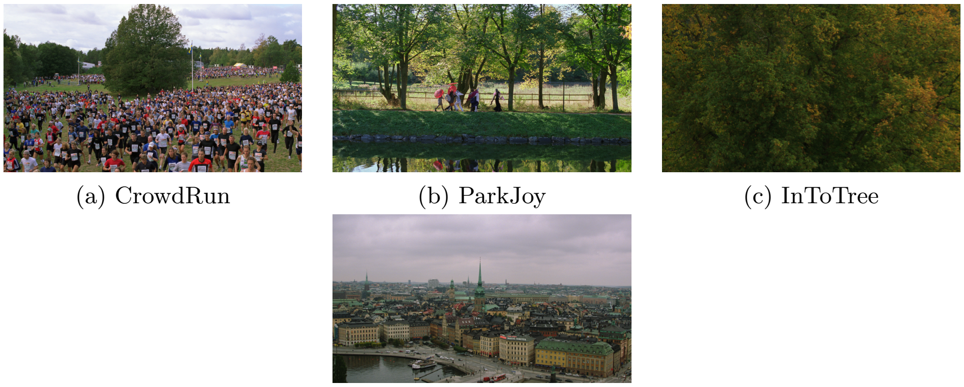
5.2 Setup and scenarios

In order to take into account the performance of different coding technologies for HDTV with respect to the visual quality, we selected two different encoders representing current coding technologies: *H.264/AVC*[1] and *Dirac*[2]. The artifacts introduced into the videos include pumping effects i.e. periodically changing quality, a typical result of rate control problems, obviously visible blocking, blurring or ringing artifacts, flicker and similar effects.

While H.264/AVC is the latest representative of the successful MPEG and ITU-T standards, Dirac is an alternative, wavelet based video codec. Its development was initiated by the British Broadcasting Cooperation (BBC) and was originally targeted at high definition resolution video material. The standard settings for the selected resolution and frame rate were used. Only the bitrate was varied to encode the videos.

For H.264/AVC, we used two significantly different encoder settings, each representing the complexity of various devices and services. The first setting is chosen to simulate a low complexity (LC) H.264/AVC encoder representative of standard devices: many tools that account for the high compression efficiency are disabled. In contrast to this, we also used a high complexity (HC) setting that aims at getting the maximum possible quality out of this coding technology, representing sophisticated broadcasting encoders. For more details we refer to our previous contribution[13].

We selected four bitrates from 5.4 Mbit/s to 30 Mbit/s representing a real life bitrate range from IPTV applications at the lower end of the bitrate scale, to high qualtiy Blu-ray discs at the upper end on the bitrate scale.



(a) CrowdRun (b) ParkJoy (c) InToTree

(d) OldTownCross

Figure 7: Test sequences from the SVT high definition multi format test set.

The test sequences were chosen from the SVT high definition multi format test set[14] with a spatial resolution of 1920 × 1080 pixels and a frame rate of 25 fps. The used sequences are shown in Fig.7: *CrowdRun*, *ParkJoy*, *InToTree* and *OldTownCross*.

Combined with the three different encoders, we therefore have a total of 48 different test conditions. Using DSUR this results in a duration of roughly one minute per BTC or 48 minutes for the complete test. Thus we split the test up into two session lasting 28 minutes each: the first test session included the two bitrates at the lower end of the bitrate scale representing the lower quality (LQ) subset and the second session included the two bitrates at the upper end of the bitrate scale representing the higher quality (HQ) subset.

We then chose two test setups for the stabilization phase:

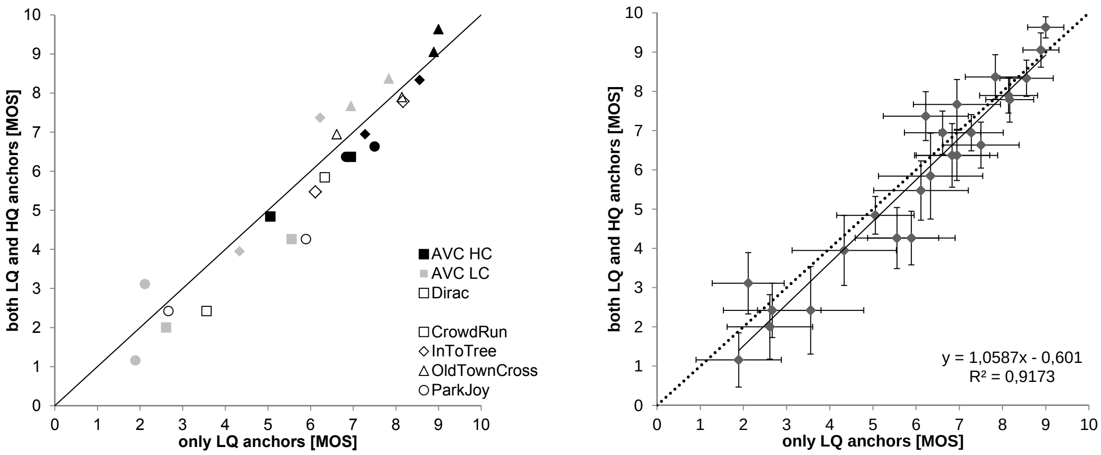
1. Only LQ or HQ anchors for the corresponding subset.
2. Both LQ and HQ anchors for all subsets.

Hence we conducted two tests with two test sessions each: firstly, a test were the stabilization phase included only anchors from the same subset i.e. only video sequences from the LQ subset as anchors for the LQ subset and only video sequences from the HQ subset as anchors for the HQ subset, in each case only representing the quality range of the current session. Secondly, a test were the stabilization phase for both the LQ and HQ subset included video sequences from both the LQ and HQ subset as anchors, representing the quality range of the complete test.

5.3 Results

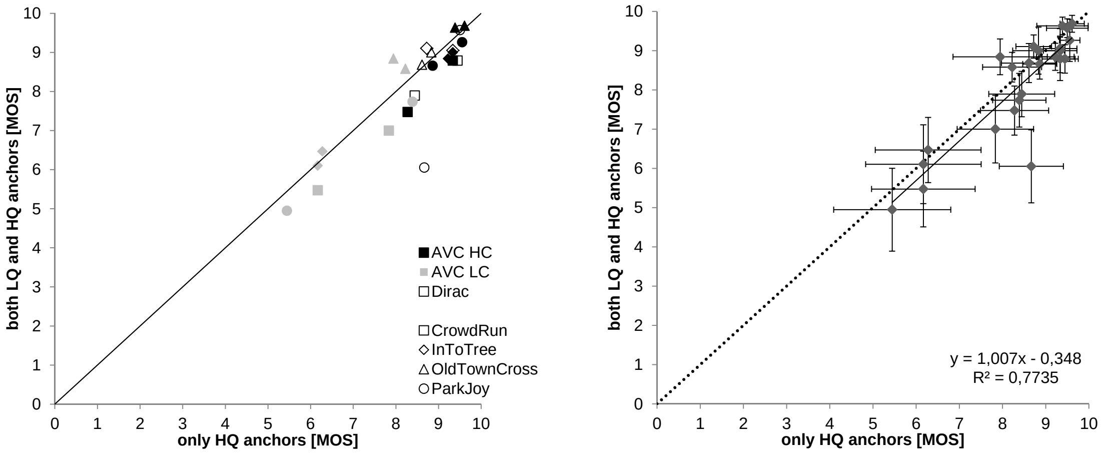
In total 21 test subjects took part in the subjective tests. The test subjects were mostly students between 20–30, with no or very little experience in video coding. After processing of the votes, two test subject were rejected, as they were unable to reproduce their results.

In Fig. 8 and Fig. 9 the results of the two test setups for different video sequences and encoders are shown. Fig. 8 shows the LQ session and Fig. 9 the HQ session.



(a) (b)

Figure 8: LQ test session with only LQ anchors vs. LQ test session with both LQ and HQ anchors. Details on sequence and codec are shown in (a), 95% confidence intervals, linear regression line and coefficient of determination R^2 in (b).



(a) (b)

Figure 9: HQ test session with only HQ anchors vs. HQ test session with both LQ and HQ anchors. Details on sequence and codec are shown in (a), 95% confidence intervals, linear regression line and coefficient of determination R^2 in (b).

We can notice that there is a nearly perfect linear relationship between both setups as the slope of linear regression line is close to the desired one. In fact, the overall Pearson correlation coefficient between the results of both test setups is 0.95, for the LQ session and the HQ session alone the correlation coefficient is 0.95 and 0.88, respectively. Still, there is a noticeable offset. For most test conditions, the visual quality seems to be overestimated if only anchors from the same session are used compared to when anchors from both sessions are used in the stabilization phase. Only for one video sequence, OldTownCross, the visual quality seems to be underestimated.

This indicates that if the complete range of visual quality is provided in the stabilization phase, the visual quality in the test phase is perceived consistently differently. On the other hand, however, the confidence intervals suggest that the observed difference is not significant in the statistical sense.

6. CONCLUSION

We examined two different aspects of subjective video quality assessment and corresponding testing methodologies: the influence of subjects' viewing experience and the quality range of anchors presented in the stabilization phase on the results of subjective testing.

In the first part, where we discussed how test subjects' viewing experience influences the results of subjective testing, the results were inconclusive, as we were unable to determine a proper relationship between the subjects' viewing experience and their corresponding MOS scores. Still, there is some indication that some kind of relationship exists and further studies should be conducted in order to determine if the previous experiences of participants' in subjective tests influence the results, and if so, what parameters have what influence.

Additionally, our results in the second part suggest that the choice of the anchors in the stabilization phase do have an influence on the visual quality perceived by the test subjects. If anchors for the quality range of the complete test are provided, the perceived visual quality is offset consistently for different content. We must note, however, that the observed results are not significantly different enough from a statistical point of view. Nevertheless, it is in our view sensible to not only include anchors representing the visual quality of the current test session, but also anchors representing the overall visual quality of the complete test into the stabilization phase of each test session.

REFERENCES

[1] ITU, ISO, "ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG4-AVC), advanced video coding for generic audiovisual services," (July 2005).

[2] Borer, T., Davies, T., and Suraparaju, A., "Dirac video compression," Tech. Rep. WHP 124, BBC Research & Development (Sept. 2005).

[3] Pinson, M. H., Wolf, S., and Gallagher, M. D., "The impact of monitor resolution and type on subjective video quality testing," tech. rep. (2004).

[4] Keimel, C. and Diepold, K., "On the use of reference monitors in subjective testing for HDTV," in [*Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*], 35 –40 (2010).

[5] Corriveau, P., Gojmerac, C., Hughes, B., and Stelmach, L., "All subjective scales are not created equal: The effects of context on different scales," *Signal Processing* **77**(1), 1 – 9 (1999).

[6] Baroncini, V., "New tendencies in subjective video quality evaluation," *IEICE Transaction Fundamentals* **E89-A**, 2933–2937 (Nov. 2006).

[7] ITU, "ITU-R BT.500 methodology for the subjective assessment of the quality for television pictures," (June 2002).

[8] MPEG Test Subgroup, "Report of the formal verification tests on AVC (ISO/IEC 14496-10 ITU-T Rec. H.264)," Tech. Rep. N6231, ISO/IEC JTC1/SC29/WG11 (Dec. 2003).

[9] Oelbaum, T., Schwarz, H., Wien, M., and Wiegand, T., "Subjective performance evaluation of the SVC extension of H.264/AVC," in [*Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*], 2772 –2775 (2008).

[10] Keimel, C., Redl, A., and Diepold, K., "Comparison of HDTV formats in a consumer environment," in [*Image Quality and System Performance VIII*], Farnand, S. P. and Gaykema, F., eds., *Proceedings of SPIE* **7867**, 786716–1 – 786716–7, SPIE (Jan 2011).

[11] Jolliffe, I., [*Principal Component Analysis*], Springer (2002).

[12] Martens, H. and Martens, M., [*Multivariate Analysis of Quality*], Wiley & Sons (2001).

[13] Keimel, C., Habigt, J., Habigt, T., Rothbucher, M., and Diepold, K., "Visual quality of current coding technologies at high definition IPTV bitrates," in [*2010 IEEE International Workshop on Multimedia Signal Processing*], 390–393 (Oct 2010).

[14] SVT, "The SVT high definition multi format test set," (Feb. 2006).