ON THE USE OF REFERENCE MONITORS IN SUBJECTIVE TESTING FOR HDTV

Christian Keimel and Klaus Diepold

Technische Universität München, Institute for Data Processing, Arcisstr. 21, 80333 München, Germany
christian.keimel@tum.de, kldi@tum.de

ABSTRACT

Most international standards recommend the use of reference monitors in subjective testing for visual quality. But do we really need to use reference monitors? In order to find an answer to this question, we conducted extensive subjective tests with reference, color calibrated high quality and uncalibrated standard monitors. We not only used different HDTV sequences, but also two fundamentally different encoders: AVC/H.264 and Dirac. Our results show that using the uncalibrated standard monitor, the test subjects underestimate the visual quality compared to the reference monitor. Between the reference and a less expensive color calibrated high quality monitor, however, we were unable to find a statistically significant difference in most cases. This might be an indication that both can be used equivalently in subjective testing, although further studies will be necessary in order to get a definitive answer.

Index Terms— subjective testing, reference monitor, Dirac, AVC/H.264, HDTV.

1. INTRODUCTION

International standards on subjective testing for visual video quality often recommend the use of professional reference monitors in tests [1, 2]. The reasoning is, that these devices have only a negligible impact on the overall visual quality due to their superior built quality and their strict adherence to video standards as ITU-R recommendation BT.709 [3] for HDTV. Also their conformance to the standards is guaranteed by the manufacturers and signal processing for so-called picture enhancement found in many consumer devices is omitted. Thus the influence of the displays on the visual quality in subjective testing can be assumed to be a fixed, well known constant. Furthermore, the reproducibility of the results between different laboratories is therefore highly likely, presuming that all other parameters are also fixed.

One problem in practice is, however, that such equipment is rather expensive, even when compared to high quality computer monitors. This may not pose a problem for public and private broadcasting companies, the industry or specialized research institutes working on visual quality, but for researchers and developers working primarily on other research areas, these costs may very well be prohibitive. Imagine for example the developer of a video encoder who wants to ascertain the visual quality achieved by his encoder during development: he will be hard pressed to justify the costs for acquiring a reference monitor.

But do we really need to use reference monitors? Or might it be sufficient to use less expensive color calibrated high quality computer monitors? In order to find a answer to these questions, we will compare in this contribution the results of subjective visual tests performed using a reference monitor with the results obtained by using normal computer monitors.

We propose two different scenarios: firstly, a color calibrated high quality computer monitor to represent a sensible and reasonably priced solution. Secondly, an uncalibrated standard computer monitor as a worst case scenario. We will perform the same subjective test with the reference monitor, the color calibrated high quality monitor and the uncalibrated standard monitor in order to determine possible differences in the perceived visual quality by the test subjects. We will use the HDTV test sequences from the SVT test set [4] and encode them with two different coding technologies AVC/H.264 [5] and Dirac [6, 7]. As differences are more likely to occur at higher visual quality, we selected only bit rates on the upper end of the scale for encoding.

We do not intend to compare the visual quality of the different monitors themselves, but rather their influence on the results of subjective tests. The results achieved with the reference monitor will be considered to be the true visual quality in this context .To the best of our knowledge this is the first contribution on this topic for HDTV.

This contribution is organized as follows: firstly, we will describe the used monitors and their calibration. Then we introduce the setup of the subjective tests before presenting and discussing the results. Finally, we conclude with a short summary.

2. EQUIPMENT

In this section we will briefly introduce the LCD monitors used and the calibration process. We selected two additional monitors in addition to our reference monitor representing high quality and standard devices. Also we color calibrated the high quality monitor to get it as close as possible to our reference monitor.

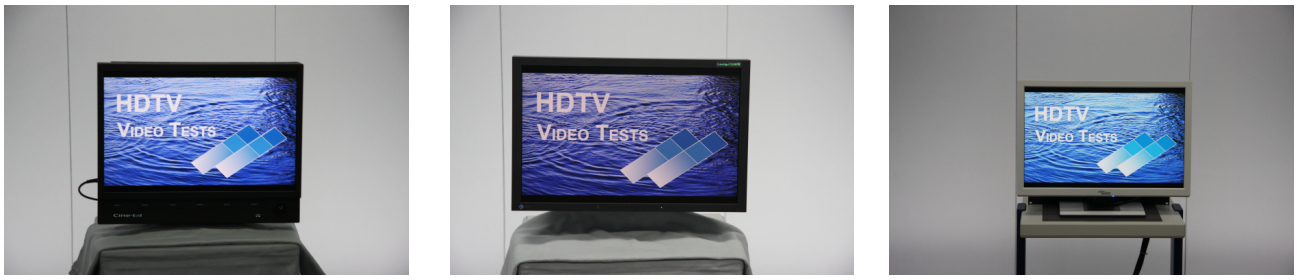
2.1. LCD monitors

In addition to our Cine-tal Cinemagé 2022 reference monitor, we selected two representatives for our proposed high quality and standard monitor scenario: a monitor aimed at professional color processing, the EIZO CG243W, representing high quality devices and a normal office display, the Fujitsu-Siemens B24W-5, representing standard devices. The first one was particularly chosen for the possibility of

Table 1: LCD monitors used in the test

	Reference	*High Quality*	*Standard*
Type	Cine-tal Cinemagé 2022	EIZO CG243W	Fujitsu-Siemens B24W-5
Diagonal	24 inch	24 inch	24 inch
Resolution	1920×1080	1920×1200	1920×1200
Input	HD-SDI	DVI	DVI

 Digital Object Identifier: 10.1109/QOMEX.2010.5518305



(a) (b) (c)

Fig. 1: The reference monitor (a), the color calibrated high quality monitor (b) and the uncalibrated standard monitor (c).

hardware color calibration i.e. not the 8 bit look-up table (LUT) in the graphic card of the computer is modified during calibration, but the internal 12 bit LUT in the display is directly modified, thus allowing a higher precision in calibrating without reducing the number of available colors. The monitors are shown in Fig.1, further details can be found in Table 1.

The reference monitor was connected directly to our video server via a HD-SDI single link. As the high quality monitor supports the desired HDTV resolution of 1920×1080, only a conversion from HD-SDI to HDMI/DVI was done using a AJA Hi5-3G converter, that also performed the expansion of the video signal from the video range (16–235) into the full range (0–255). Unfortunately the used standard monitor does not support the 16:9 input signal. Therefore a Doremi Labs GHX-10 cross converter was used to display the 1920×1080 on the native 1920×1200 screen and also to expand the video signal to the full range. On both monitors the video was shown with a 1:1 aspect ratio and letter boxing.

2.2. Calibration

For calibration we used a X-Rite i1 Pro spectrophotometer. The color gamut, white point, color temperature and gamma were chosen according to ITU-R BT.709 [3]. The target luminance was set to $100\,\frac{cd}{m^2}$, similar to most reference monitors. In Table 2 the target values for the calibration and the measured values for the high quality monitor after calibration are shown.

Table 2: Calibration target and results

	Target	High Quality	Standard[(a)]
Luminance $[\frac{cd}{m^2}]$	100	100.18	332
Gamma	2.2	2.2	2.2
Color Temperature	6504K	6541K	6900K
White point [x,y]	0.313, 0.329	0.312, 0.329	0.309, 0.329
Chromaticity			
Red [x,y]	0.640, 0.330	0.639, 0.328	0.651, 0.318
Green [x,y]	0.300, 0.600	0.298, 0.603	0.199, 0.660
Blue [x,y]	0.150, 0.060	0.152, 0.059	0.149, 0.056

[(a)] uncalibrated

The standard monitor was not color calibrated but only reset to its factory defaults with a color temperature of $6500K$ and the sRGB color gamut. We then used the spectrophotometer to measure its colorimetric properties. Table 2 shows clearly that not only the luminance is too high, but that also the primaries are not matching ITU-R BT.709 very well at its factory defaults. In particular the green primary is way off, shifting the color gamut far into the green. Our test subjects also remarked on the extremely high *brightness* compared to the other monitors.

Table 3: Tested video sequences

Sequence	Frame Rate	Bit Rate [MBit/s]
CrowdRun	25 fps	19.2 / 28.5
InToTree	25 fps	13.1 / 17.1
OldTownCross	25 fps	13.7 / 19.0
ParkJoy	25 fps	20.1 / 30.9

3. SUBJECTIVE TESTING

After describing the used equipment in the last section, we will now discuss the selection of the used video sequences and encoder settings, but also the general test setup and the used methodology.

3.1. Sequences and Encoder Scenarios

We selected two different bit rates from 13 Mbit/s to 30 Mbit/s on the upper end of the reasonable bit rate scale. These two rate points represent on one hand nearly perfect quality, where the coded video is often indistinguishable from the uncoded reference, and on the other hand still a very high quality, but with noticeable artifacts. We decided to use only comparably high bit rates as one can assume that especially for very high quality either inferior signal processing e.g. smaller LUTs or dithering in the monitor introduces significant non-coding artifacts like blurring or the unnatural presentation of colors lower the perceived visual quality. Whereas for lower bit rates, the overall visual quality is already so bad, that additional degradation due to the monitor does not play such an prominent part in the overall perception of the visual quality.

The test sequences were chosen from the SVT high definition multi format test set [4] with a spatial resolution of 1920×1080 pixel and a frame rate of 25 frames per second (fps) was used. The particular sequences are *CrowdRun*, *ParkJoy*, *InToTree* and *OldTownCross*. Each of those videos was encoded at the selected bit rates. The artifacts introduced into the videos by this encoding include pumping effects i.e. periodically changing quality, a typical result of rate control problems, obviously visible blocking, blurring or ringing artifacts, flicker, banding i.e. unwanted visible changes in color and similar effects. An overview of the sequences and bit rates is given in Table 3.

Table 4: Selected encoder settings for AVC/H.264

	LC	HC
Encoder	JM 12.4	
Profile&Level	Main, 4.0	High, 5.0
Reference Frames	2	5
R/D Optimization	Fast Mode	On
Search Range	32	128
B-Frames	2	5
Hierarchical Encoding	On	On
Temporal Levels	2	4
Intra Period	1 second	
Deblocking	On	On
8x8 Transform	Off	On

The sequences were encoded using the AVC/H.264 reference software [8] version 12.4. Two significantly different encoder settings were used, each representing th complexity of various application areas. The first setting is chosen to simulate a low complexity (LC) AVC/H.264 encoder using a 'Main' profile according to Annex A of the AVC/H.264 Standard: many tools that account for the high compression efficiency are disabled. In contrast to this a high complexity (HC) setting aims at getting the maximum possible quality out of this coding technology using a 'High' profile. In addition to AVC/H.264, we used the *Dirac* encoder [6, 7] in order to investigate if different coding technologies have any influence. The development of Dirac was initiated by the British Broadcasting Cooperation (BBC) and it is a wavelet based video codec, originally targeting at HD resolution video material. For Dirac, the standard settings for the selected resolution and frame rate were used. Only the bit rate was varied to encode the videos. The used software version for Dirac is 0.7, available at [9]. Selected encoding settings for AVC/H.264 are given in Table 4. The decoded videos were converted to 4:2:2 YC_BC_R for output to the monitors via HD-SDI. This was done by bilinear upsampling of the chroma channels of the 4:2:0 decoder output.

3.2. Test Setup

The 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 [1] as shown in Fig.2. To maintain the viewing experience

Fig. 2: Test room

that can be achieved with high definition video, the distance between the screen and the observers was set to three times the picture height. Due to the screen size, only two viewers 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 and color blindness.

The tests were carried out using a variation of the standard DSCQS test method as proposed in [10]. 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 presentation of the reference and the processed video is intended to allow the test subjects to decide which is the reference video. Only the repetition is used by the viewers to judge the quality of the processed video in comparison to the reference. The structure of a basic test cell is shown in Fig.3.

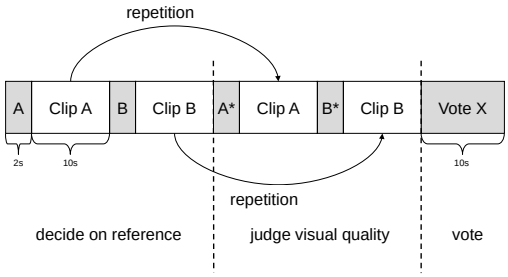


Fig. 3: Basic test cell DSUR

To allow the test subjects to differentiate between relatively small quality differences, a discrete voting scale with eleven grades ranging from 0 to 10 was used. 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. 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 Annex 2 of [1]. The mean opinion score (MOS) was determined by averaging all valid votes for each test case.

4. PROCESSING OF THE VOTES

In total 19 test subjects took part in the subjective test with the reference monitor and 21 test subjects each in the tests with the other two monitors. The test subjects were mostly students between 20–30, with no or very little experience in video coding. After processing of the votes, one test subject for the reference monitor and two test subjects for the other two monitors were rejected, as they were not

Table 5: Processing of the votes

	Reference	High Quality	Standard
Test subjects			
total	19	21	
rejected	1	2	
considered valid	18	19	
95% confidence interval			
mean	0.337	0.328	0.406
maximum	0.678	0.684	0.693
standard deviation			
mean	1.46	1.46	1.80
maximum	2.94	3.00	3.04

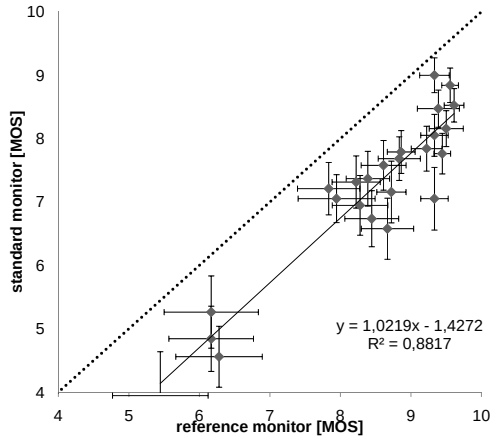


Fig. 4: Reference monitor compared to standard monitor including 95% confidence intervals and linear regression line.

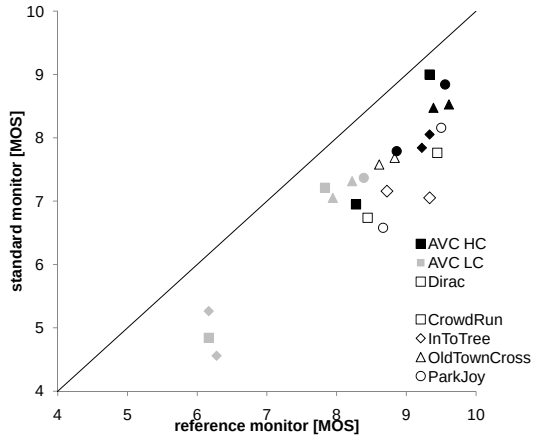


Fig. 5: Reference monitor compared to standard monitor with details on sequence and codec.

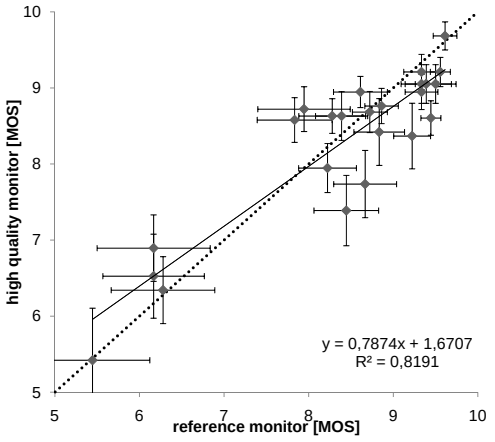


Fig. 6: Reference monitor compared to high quality monitor including 95% confidence intervals and linear regression line.

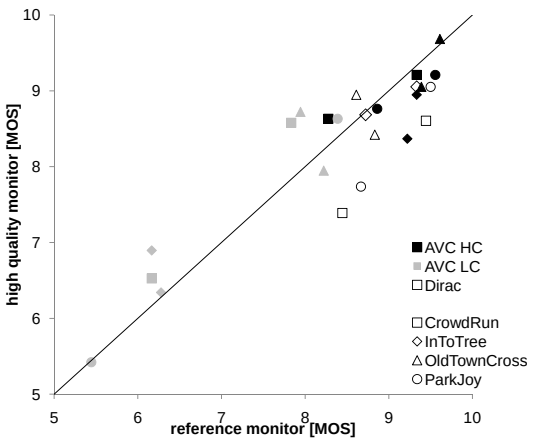


Fig. 7: Reference monitor compared to high quality monitor with details on sequence and codec.

able to reproduce their own results. All votes of these subjects were removed from the data base. Hence we considered 18 test subjects for the reference monitor and 19 test subjects for the other two monitors in the further processing of the votes.

Some of the results for the reference display have already been used in [11, 12]. The mean and maximum of the 95% confidence intervals and the standard deviation of the subjective votes over all single test cases, separated according to the different tests is shown in Table 5. We can already see now from Table 5 that the standard monitor exhibits a larger variance of the votes.

5. RESULTS

The results of the subjective test are shown in detail in Fig. 8 to Fig. 11. Unfortunately the results do not show a obvious general tendency regarding the influence of the used monitors on the visual quality. One thing we notice is, that the standard monitor apparently leads to a statistical significant, consistent underestimation of the perceived visual quality by the test subjects. Also the uncertainty is reduced at the higher rate point as shown by the reduced confidence intervals. But between reference monitor and high quality monitor, there is often no statistical significant difference noticeable between the votes.

In Fig. 4 we can see more clearly that the standard monitor leads to an underestimation of the visual quality. If we perform a linear regression, we notice that the slope is close to the desired 1, while we have a constant offset of $-1,43$. Thus the visual quality is always perceived lower. Additionally we can see in Fig. 5 that this underestimation occurs regardless of sequence or codec. This seems to confirm our earlier assumption that a standard monitor reduces the perceived quality in particular at high bit rates. If this also holds true in general for lower quality video is an open question.

The results for the high quality monitor, however, do not exhibit such a obvious behavior as we can see in Fig. 6. If we once again perform a linear regression, we get a slope of 0.78 and an offset of $+1.67$. Note that the coefficient of determination R^2 is lower than for the standard monitor, suggesting that the linear model in this case is not able to describe the variance of the data as well as before. In general there does not seem to be a statical significant difference between the high quality and reference monitor in most cases. This might be caused by the low statistical sample size of only 19 different samples. Even tough in [13] the lower bound of 15 test subjects was

shown to be sufficient, it may be that due to the apparently small quality difference between the results from the two different tests, more test subjects are needed in order to further reduce the variance.

Nevertheless, we can notice that there are small differences not only depending on sequence, but especially on the used video codec. If we look on the comparison between reference and high quality monitor in detail in Fig. 7, we notice that the visual quality on the high quality monitor seems to be underestimated for AVC HC and Dirac, but overestimated for AVC LC. This shows that it is not only important to use different sequences, but also to use different encoders as proposed in [14].

6. CONCLUSION

We compared a reference monitor to a color calibrated high quality monitor and a standard monitor with regards to their use in subjective testing for HDTV. In order to achieve this goal, we performed extensive subjective tests using different sequences and codecs. We selected two different rate points at the upper end of the bit rate scale.

Our results show, that if we use a uncalibrated standard monitor in subjective testing, the visual quality is usually underestimated by the test subjects compared to the reference monitor. Between a reference monitor and color calibrated, less expensive high quality monitor, however, we were not able to determine a statistical significant difference between the results from subjective tests conducted with either one in most cases. But we should keep in mind that we only have a rather small sample size, so this might only be an indication that a reference monitor and a high quality monitor are equivalent in their use in subjective testing.

Moreover, we have seen that not only the different sequences i.e. different content influenced the perceived visual quality on the different monitors, but also that the different coding technologies made a difference. Therefore it is sensible to not only include different sequences, but also different codecs in subjective testing. Especially if general questions regarding subjective testing are to be considered.

In future work we will aim at further determining what difference –if any at all– between reference and hight quality monitors exists with regard to subjective testing.

7. REFERENCES

[1] *ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for Television Pictures*, ITU-R Std., Rev. 11, Jun. 2002.

[2] *ITU-R BT.710 Subjective assessment methods for image quality in high-definition television*, ITU-R Std., Rev. 4, Nov. 1998.

[3] *ITU-R BT.709: Parameter values for the HDTV standards for production and international programme exchange*, ITU-R Std., Rev. 5, Apr. 2002.

[4] SVT. (2006, Feb.) The SVT high definition multi format test set. [Online]. Available: http://www.ldv.ei.tum.de/lehrstuhl/team/Members/tobias/sequences

[5] *ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG4-AVC), Advanced Video Coding for Generic Audiovisual Services*, ITU, ISO Std., Rev. 4, Jul. 2005.

[6] T. Borer, T. Davies, and A. Suraparaju, "Dirac video compression," BBC Research & Development, Tech. Rep. WHP 124, Sep. 2005.

[7] T. Borer and T. Davies, "Dirac - video compression using open technology," BBC Research & Development, Tech. Rep. WHP 117, Jul. 2005.

[8] K. Sühring. (2007) H.264/AVC software coordination. [Online]. Available: http://iphome.hhi.de/suehring/tml/index.htm

[9] C. Bowley. Dirac video codec developers' website. [Online]. Available: http://dirac.sourceforge.net

[10] V. Baroncini, "New tendencies in subjective video quality evaluation," *IEICE Transaction Fundamentals*, vol. E89-A, no. 11, pp. 2933–2937, Nov. 2006.

[11] C. Keimel, T. Oelbaum, and K. Diepold, "No-reference video quality evaluation for high-definition video," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 1145 –1148, April 2009.

[12] ——, "Improving the prediction accuracy of video qualtiy metrics." *Acoustics, Speech and Signal Processing, 2010. ICASSP 2010. IEEE International Conference on*, pp. 2442–2445, Mar. 2010.

[13] S. Winkler, "On the properties of subjective ratings in video quality experiments," *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pp. 139 –144, July 2009.

[14] C. Keimel, T. Oelbaum, and K. Diepold, "Improving the verification process of video quality metrics," *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pp. 121 –126, July 2009.

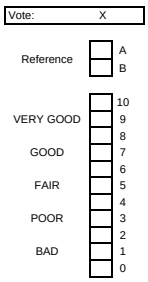
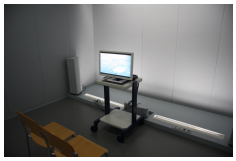


Fig. 12: Discrete eleven point voting scale as used in the tests.

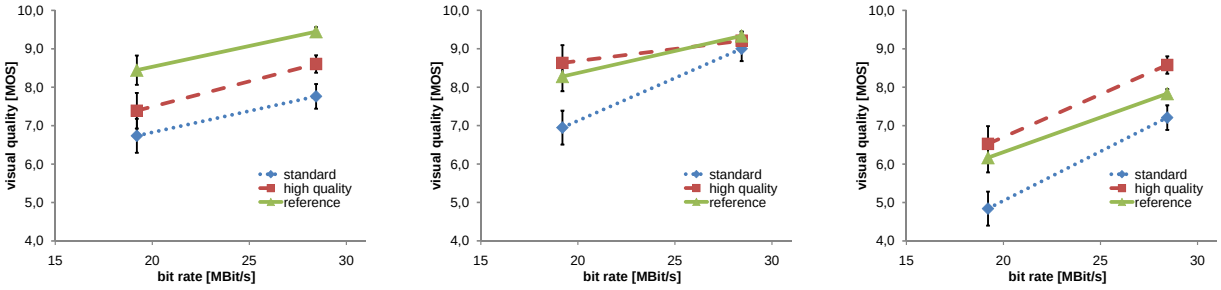
(a) Reference monitor

(b) High quality monitor



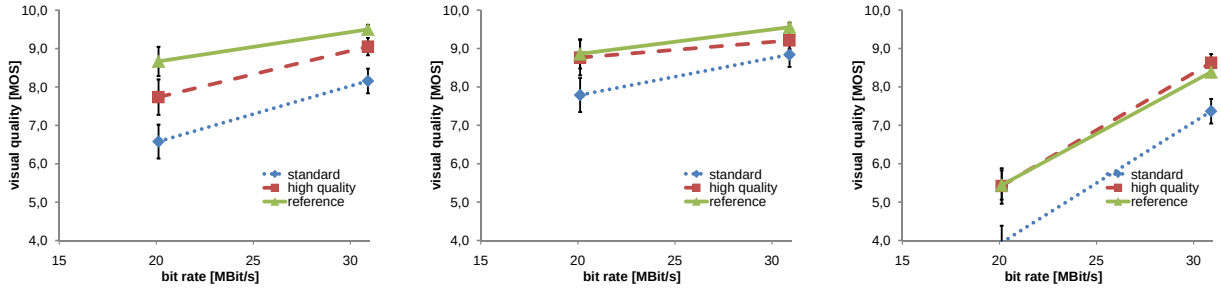
(c) Standard monitor

Fig. 13: Test setups for the different monitors.



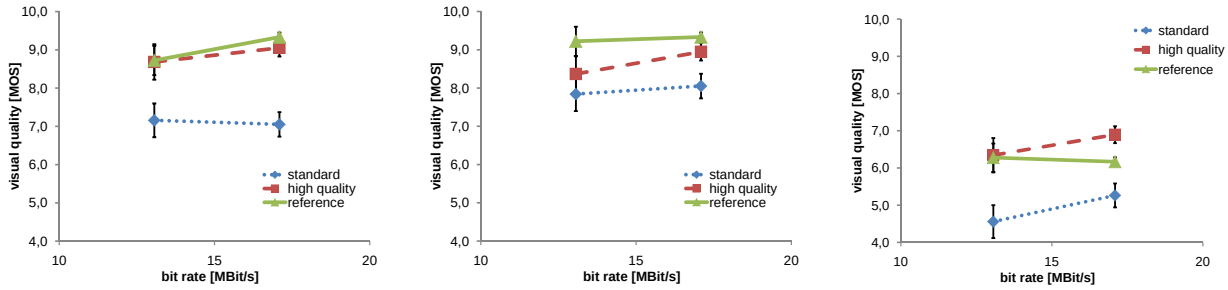
(a) Dirac (b) AVC HC (c) AVC LC

Fig. 8: Results for the subjective tests with the reference, high quality and standard monitor for *CrowdRun* .



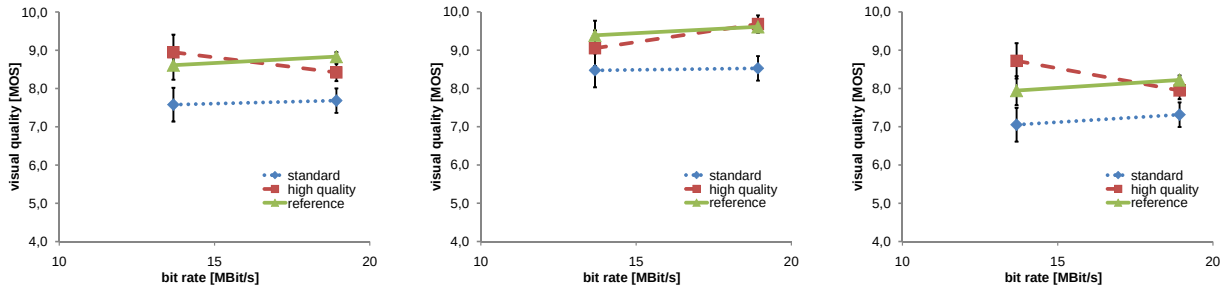
(a) Dirac (b) AVC HC (c) AVC LC

Fig. 9: Results for the subjective tests with the reference, high quality and standard monitor for *ParkJoy* .



(a) Dirac (b) AVC HC (c) AVC LC

Fig. 10: Results for the subjective tests with the reference, high quality and standard monitor for *InToTree* .



(a) Dirac (b) AVC HC (c) AVC LC

Fig. 11: Results for the subjective tests with the reference, high quality and standard monitor for *OldTownCross* .