THE TUM HIGH DEFINITION VIDEO DATASETS

Christian Keimel, Arne Redl and Klaus Diepold

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

ABSTRACT

The research on video quality metrics depends on the results from subjective testing for both the design and development of metrics, but also for their verification. As it is often to cumbersome to conduct subjective tests, freely available data sets that include both mean opinion scores and the distorted videos are becoming ever more important. While many datasets are already widely available, the majority of these data sets focus on smaller resolutions. We therefore present in this contribution the TUM high definition datasets that includes videos in both 1080p25 and 1080p50, encoded with different coding technologies and settings, H.264/AVC and Dirac, but also different presentation devices from reference monitors to home-cinema projectors including soundtrack. The datasets are made freely available for download under a creative commons license.

***Index Terms*—** HDTV, subjective testing, video quality assessment, 1080p25, 1080p50

1. INTRODUCTION

The research on video quality metrics depends on the results from subjective testing for both the design and development of metrics, but also for their verification. Unfortunately, it is often not possible to conduct own subjective tests, either because of limited time or other resources. Hence, freely available data sets that include both mean opinion scores (MOS) and the distorted videos are becoming ever more important. While many datasets are already widely available the datasets with progressive high definition content in 1920×1080, particularly for higher frame rates at 50 frames per second (fps), are still rare.

In this contribution we will therefore present the two TUM high definition datasets that includes videos in both *1080p25* and *1080p50*, respectively. In the *1080p25* dataset, different coding technologies and settings, H.264/AVC and Dirac, are included, whereas in the *1080p50* dataset H.264/AVC encoded videos presented with different devices from reference monitors to home-cinema projectors with surround sound system are included.

This contribution is organized as follows: after a description of our video quality evaluation laboratory and the parameters common to both datasets, we will discuss both in detail before describing the available downloas and concluding with a short summary.

2. SETUP

Before describing both datasets in detail, we will give a short overview of the test setup and equipment used in generating both datasets.

2.1. Room

All test were conducted in the video quality evaluation laboratory at the Institute for Data Processing at the Technische Universität München in a room compliant with ITU-R BT.500 [1]. An overview of the room's layout is given in Fig. 1. The room is equipped with a programmable background illumination system at a colour temperature of 6500 K, allowing us to illuminate the room reproducibly in a multitude of different scenarios. The walls and ceiling have mid-grey chromaticity as required by ITU-R BT.500.

The laboratory's infrastructure allows the video quality evaluation via HDMI or HD-SDI connections up to a resolution of 1920×1080 at 60 fps. Additionally, a 7.1 surround audio system enables us to assess audio-visual quality in a realistic environment e.g. for home-cinema scenarios.

2.2. Equipment

The subjective tests were performed with four different presentation devices: two reference displays, a high-quality consumer LCD TV and a LCoS projector. The viewing distance was set to two times (2H) and three times (3H) the screen height for the *1080p50* and *1080p25* data sets, respectively. An overview of the different devices is given in Table 1.

All devices are capable of presenting a spatial resolution of 1920×1080 and, except for the Cine-tal Cinemagé 2022, support a frame rate of 50 fps. The reference displays were connected via a 4:2:2 YC_BC_R HD-SDI single- or dual-link connection and both the LCD TV and the Projector were connected to the video server via a HD-SDI to HDMI converter (AJA Hi5-3G) as illustrated in Fig. 2 and Fig. 3 Before the subjective testing commenced, all displays were colour calibrated. 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 [2]. The target luminance was set to $100\,\frac{cd}{m^2}$. The background illuminations was low as required in [1].

Additionally, we used a permanently installed 7.1-hifi-system, consisting of an AV-Receiver, two front-speakers, one center-speaker, four dipole loudspeakers and one subwoofer, all of high-quality hifi grade in a subset of the *1080p50* test setup.

2.3. Video Sequences

For both tests, we selected in total eight different video sequences from the well-known SVT multi format test set [3] with a resolution of 1920×1080 and a frame rate of 50 fps; the 25 fps version for the *1080p25* dataset was generated by dropping every even frame. Each video sequence has a length of 10 s. We choose this test set, as on

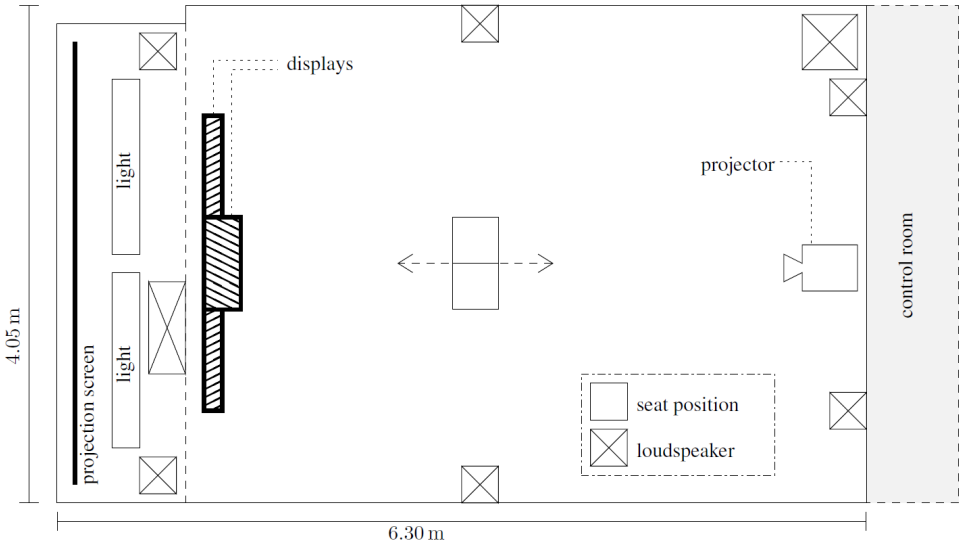


Fig. 1: Layout of the video quality evaluation laboratory at the Institute for Data Processing (not to scale)

Table 1: Presentation devices used in the subjective tests

Device	Category	Screen size	max. fps	Used in dataset
Cine-tal Cinemagé 2022	LCD Class A reference monitor	24”	25	1080p25
Sony BVM-L230	LCD Class A reference monitor	23”	50	1080p50
Sony KDL-55X4500	Consumer LCD TV with RGB background illumination	56”	50	1080p50
JVC DLA-HD750	LCoS Projector	2.8m	50	1080p50

the one hand it is one of the few available in 1080p50 and secondly because of the availability of an additional soundtrack.

We selected the following scenes from the test set: *ParkJoy*, *OldTownCross*, *CrowdRun*, *InToTree*, *TreeTilt*, *PrincessRun*, *DanceKiss* and *FlagShoot*. All scenes except of *FlagShoot* are clips proposed in [3] that cover the the whole range of coding difficulties. The selection of the specific video sequences for the *1080p50* dataset was mainly motivated by the attractiveness of the corresponding soundtrack for the given sequences. The additional sequence *FlagShoot* was selected due to its interesting sound effects for the audio-visual sub-test in the *1080p50* dataset. The start frames of the scenes are shown in Fig.5 and more details are given in Table 2.

2.4. Testing Methodology

We used two different testing methodologies for the two datasets. For the *1080p25* dataset, we used the double-stimulus DSUR method and for the *1080p50* dataset the single-stimulus SSMM method.

The Double Stimulus Unknown Reference (DSUR) method is a variation of the standard DSCQS test method as proposed in [4]. It 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 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.7a. 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 as shown in Fig. 4a.

The Single Stimulus MultiMedia (SSMM) method is a variation of the standard SSIS test method as proposed in [5]. It differs from the standard SSIS test method that instead of a impairment scale, a discrete quality scale is utilized. In order to avoid context effects, each sequence and coding condition was repeated once in a different context i.e. different predecessor sequence and different coding condition. The structure of a basic test cell is shown in Fig. 7b. 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 as shown in Fig. 4b.

Before each test, a short training was conducted in with sequences of different content to the test at similar quality levels and with similar coding artefacts to the tests, resulting in a training session of ten sequences. 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 tests. Additionally, a stabilization phase of five sequences was included in the beginning of each test.

3. 1080P25 DATASET

The test that resulted in the *1080p25* dataset aimed originally at the comparison of different coding tools and coding technologies for high definition material. Only the Cine-tal Cinemagé 2022 reference display was used in this test. In order to take into account the performance of different coding technologies for high definition content, we selected two different encoders representing current coding technologies: *H.264/AVC* [6] and *Dirac* [7]. The *1080p25* dataset was

(a) Reference monitor (b) LCD-TV (c) Projector

Fig. 2: Presentation devices: setup of displays and projector

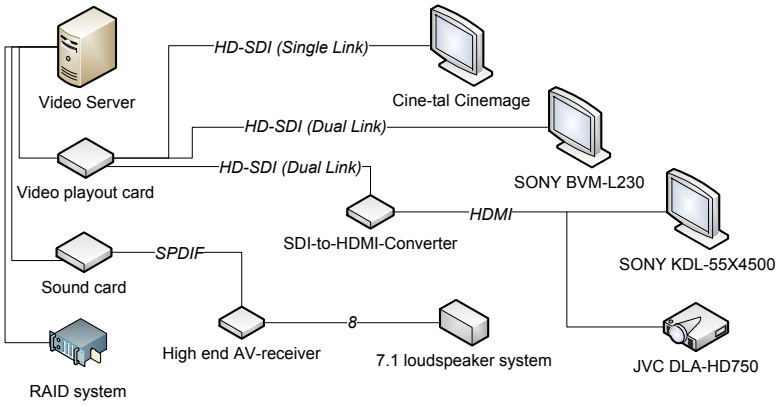
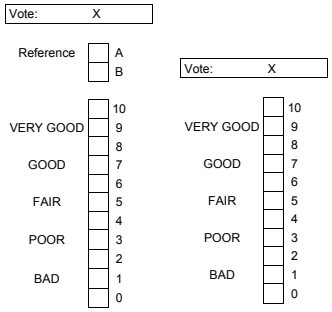


Fig. 3: Technical overview



(a) DSUR (b) SSMM

Fig. 4: Voting Sheets

first presented in [8].

3.1. Encoder Scenarios

H.264/AVC is the latest representative of the successful MPEG and ITU-T standards, while 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. Wheras it follows the common hybrid coding paradigm, it utilizes the wavelet transform instead of the usual block-based transforms e.g. DCT. Hence, it is not necessary for the transform step itself to divide the frame into separate blocks, but the complete frame can be mapped into the wavelet domain in one piece. This fundamental difference to the popular standards of the MPEG-familiy was also the main reason we chose Dirac as the representative of alternative coding technologies in this contribution. Overlapped block-based motion compensation is used in order to avoid block edge artifacts, which due to their high frequency components are problematic for the wavelet transform. Unlike the H.264/AVC reference software, the Dirac reference software version 0.7 used in this contribution offers a simplified selection of settings by just specifying the resolution and frame rate, instead of specific coding tools. Therefore, only the bitrate was varied.

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. We used the reference software [9] of H.264/AVC, version 12.4. The difference in computational complexity is also shown by the average encoding time per frame: 34 and 201 seconds per frame for the LC and the HC H.264/AVC version, respectively. Selected encoding settings for H.264/AVC are given in Table 3.

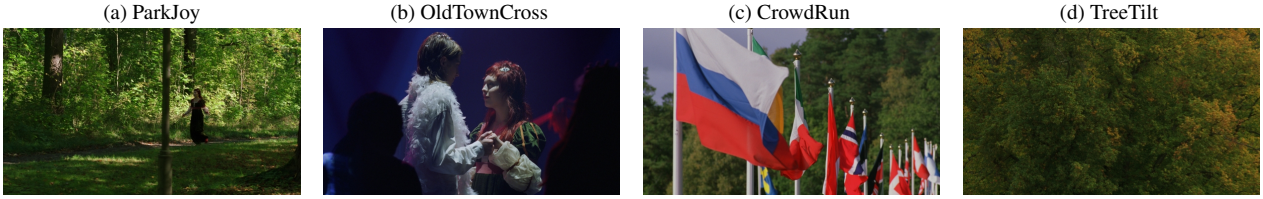
We selected four bitrates individually for each sequence depending on the coding difficulty of the sequences from the overall range of 5.4 Mbit/s to 30 Mbit/s representing real life high definition applications from the lower end to the upper end on the bitrate scale.

The test sequences were chosen from the SVT high definition multi format test set as listed in Table 2 and each of those videos was encoded at the selected four different bitrates. This results in a quality range from ‘not acceptable’ to ‘perfect’, corresponding to mean opinion scores (MOS) between 1.9 and 9.6 on a scale ranging from 0 to 10. The artifacts introduced into the videos by this encoding scheme 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. An overview of the sequences and bitrates is given in Table 2.

3.2. Processing and Results

A total of 19 subjects participated in the test, all students with no or little experience in video coding aged 20-30. All participants were tested for normal visual acuity and normal color vision with a Snellen chart and Ishihara plates, respectively. Processing of outlier votes was done according to [1] and the votes of one test subject were removed based on this procedure. The 95% confidence intervals of the subjective votes are below 1.4 on a scale between 0 and 10 for all

(a) ParkJoy (b) OldTownCross (c) CrowdRun (d) TreeTilt



(e) PrincessRun (f) DanceKiss (g) FlagShoot (h) InToTree

Fig. 5: Test sequences from the SVT test set

Table 2: Video sequences and bitrates for different rate points (RP)

Sequence	Coding difficulty	Start frame	Used in dataset	Bitrates [MBit/s] RP1	RP2	RP3	RP4
OldTownCross	Easy	1217	1080p25	5.4	9.6	13.7	19.0
InToTree	Easy	5199	1080p25	5.7	10.4	13.1	17.1
FlagShoot	Easy (assumed)	6611	1080p50	2	3	6	10
CrowdRun	Difficult	7111	1080p25	8.4	12.7	19.2	28.5
			1080p50	8	20	30	40
TreeTilt	Medium	9077	1080p50	2	3	6	10
PrincessRun	Difficult	10429	1080p50	8	20	30	40
ParkJoy	Difficult	15523	1080p25	9.0	12.6	20.1	30.9
DanceKiss	Easy	17953	1080p50	2	3	6	10

single test cases, the mean 95% confidence interval is 0.78. We determined the mean opinion score (MOS) by averaging all valid votes for each test case. The resulting MOS values are shown in Fig. 8.

4. 1080P50 DATASET

The aim of the test that resulted in the *1080p50* dataset was on the one hand to compare the perceived visual quality on different devices from a reference display to a home cinema setup, including audio, on the other hand to gain some data for 1080p50 material. In this test, the Sony BVM-L230 reference display, the Sony KDL-55X4500 LCD TV and the JVC DLA-HD750 projector were used. The test was therefore run four times resulting in four sub-tests: once for each presentation device and the forth run included the audio soundtrack in combination with the projector. The *1080p50* dataset was first presented in [10].

4.1. Encoder Scenarios

In this test, we used H.264/AVC with a encoder setting that aims at getting the maximum possible quality out of this coding technology. We used the reference software [9] of H.264/AVC, version 17.1. Selected encoding settings are given in Table 3.

We selected four bitrates individually for each sequence depending on the coding difficulty of the sequences from the overall range of 2 Mbit/s to 40 Mbit/s representing real life high definition applications from the lower end to the upper end on the bitrate scale.

The test sequences were chosen from the SVT high definition multi format test set as listed in Table 2 and each of those videos was encoded at the selected four different bitrates. This results in a quality range from 'not acceptable' to 'very good', corresponding to mean opinion scores (MOS) between 1.9 and 8.9 on a scale ranging from 0 to 10. The artifacts introduced into the videos by this encoding scheme include visible blocking, blurring or ringing artifacts, flicker and similar effects. An overview of the sequences and bitrates is given in Table 2.

4.2. Processing and Results

A total of 21 subjects participated in the test, all students with no or little experience in video coding aged 16–27, two of them female. All participants were tested for normal visual acuity and normal color vision with a Snellen chart and Ishihara plates, respectively. The two votes for each test case were compared and if the difference between these two votes exceeded three units, both individual votes for this sequence were rejected, otherwise they were averaged and considered to be valid. Additionally, it was checked, if the vote of a subject deviated more than three units from the average of the other participants and if so the individual vote for this test case was rejected for this subject. A subject was completely removed from the results, if more than 15% of his individual votes were rejected. In total, no more than four subjects were rejected for individual subtests and at least 15 valid subjects as suggested in [11] evaluated each sub-test. The 95% confidence intervals of the subjective votes

Fig. 6: Manikin for audio recording

are below 1.1 on a scale between 0 and 10 for all single test cases, the mean 95% confidence interval is 0.72. We determined the mean opinion score (MOS) by averaging all valid votes for each test case. The resulting MOS values for the reference display are shown in Fig. 9. For the other presentation devices, we refer to [10] or to the MOS scores available for download.

4.3. Audio

Additionally, we recorded the playback of the soundtrack on our 7.1 hifi system with a manikin consisting of a head and torso. The position of this manikin as shown in Fig. 6 was the same as the participant's in the audio-visual sub-test. It was equipped with a pair of microphones in its ears, allowing us to reproduce the surround sound experience including the room acoustic via headphones. The sound system was adjusted to a maximum loudness of $80\ \mathrm{dB(A)}$. The silence noise level i.e. the noise level with all devices running but without playing a sound was below $30\ \mathrm{dB(A)}$. For further information, we refer to [10]

5. DOWNLOAD AND LICENSE

Both datasets, *1080p25* and *1080p50*, are available for download at `www.ldv.ei.tum.de/videolab`. The provided files include:

- H.264/AVC and Dirac bitstreams for both datasets
- Encoder log files from the encoding of the bitstreams
- Excel files with the (anonymised) votes of the subjects, overall MOS scores and PSNR values for all presentation devices
- Audio files with the recording from the audio-visual sub-test

The datasets are made available under a *Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Germany License*. This licence allows the use of the datasets for non commercial activities, to freely modify and share the datasets, as long as this publication is cited. Also modifications must be shared under the same license. For more details about the license, we refer to [12].

6. CONCLUSION

We described in detail how both datasets, *1080p25* and *1080p50*, were generated. We hope that these two high definition datasets will be helpful to others in the development of video quality metrics or other applications.

7. REFERENCES

[1] *ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for Television Pictures*, ITU-R Std., Rev. 12, Sep. 2009.

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

[3] Lars Haglund. (2006) SVT Multi Format Test Set Version 1.0.

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

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

[6] *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.

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

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

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

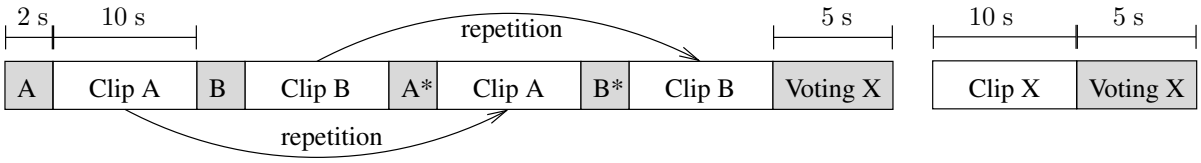
[10] A. Redl, C. Keimel, and K. Diepold, "Influence of viewing device and soundtrack in HDTV on subjective video quality," F. Gaykema and P. D. Burns, Eds., vol. 8293, no. 1. SPIE, Jan. 2012, p. 829312.

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

[12] Creative Commons. (2012, Mar.) Creative commons attribution-noncommercial-sharealike 3.0 germany license. [Online]. Available: http://creativecommons.org/licenses/by-nc-sa/3.0/de/deed.en

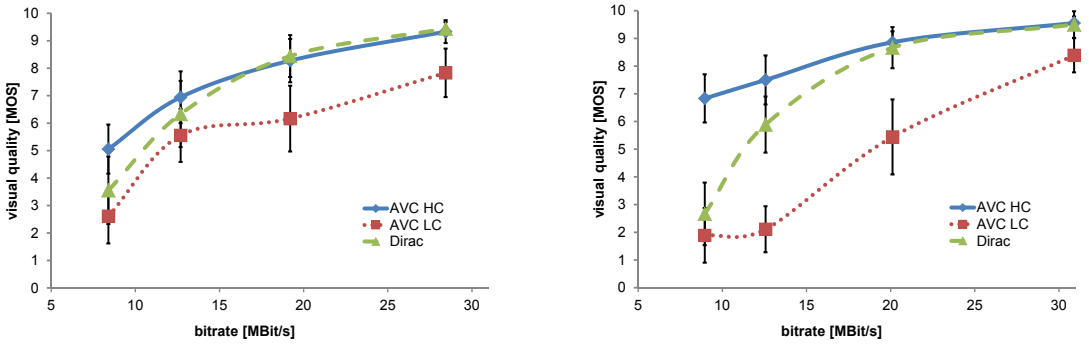
Table 3: Selected encoder settings for H.264/AVC

Dataset	1080p25		1080p50
	LC	HC	-
Encoder	JM 12.4		JM 17.1
Profile&Level	Main, 4.0	High, 5.0	High, 5.0
Reference Frames	2	5	5
R/D Optimization	Fast Mode	On	On
Search Range	32	128	128
B-Frames	2	5	5
Hierarchical Encoding	On	On	On
Temporal Levels	2	4	4
Intra Period	0.5 s		0.5 s
Deblocking	On	On	On
8x8 Transform	Off	On	On

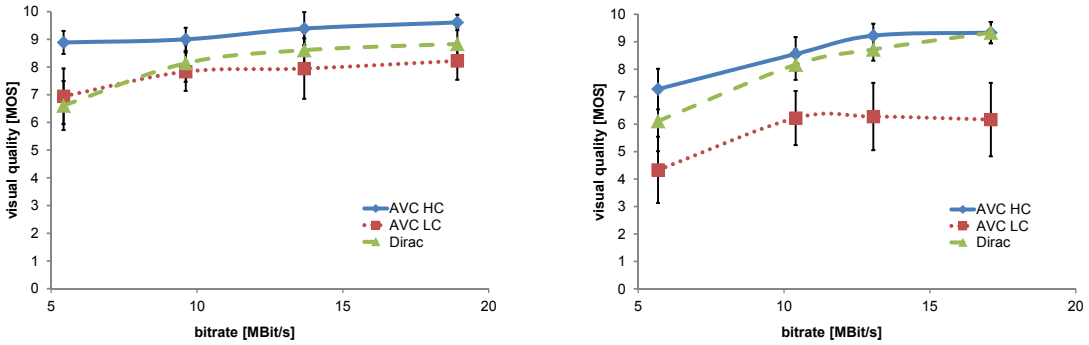


(a) DSUR (b) SSMM

Fig. 7: Basic test cells



(a) CrowdRun (b) ParkJoy



(c) OldTownCross (d) InToTree

Fig. 8: *1080p25* dataset: visual quality at bitrates from 5.4 Mbit/s to 30 Mbit/s for different video sequences and encoders. The whiskers represent the 95% confidence intervals of the subjective test results for the visual quality

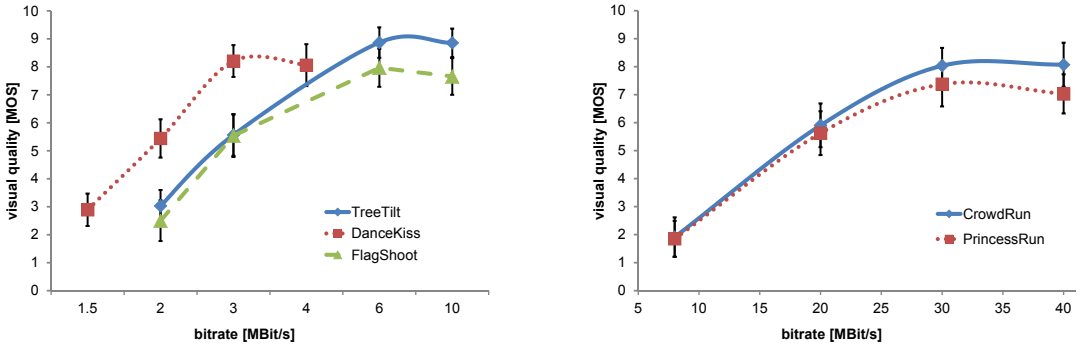


Fig. 9: *1080p50* dataset: visual quality at bitrates from 1.5 Mbit/s to 40 Mbit/s with the reference display for different video sequences. The whiskers represent the 95% confidence intervals of the subjective test results for the visual quality