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 majority of these datasets 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.

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 Cinemage 2022, support a frame rate of 50 fps. The reference displays were connected via a 4:2:2 YCbCr 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 cd/m². 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
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, Old-TownCross, 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

Table 1: Presentation devices used in the subjective tests

<table>
<thead>
<tr>
<th>Device</th>
<th>Category</th>
<th>Screen size</th>
<th>max. fps</th>
<th>Used in dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cine-tal Cinemagé 2022</td>
<td>LCD Class A reference monitor</td>
<td>24”</td>
<td>25</td>
<td>1080p25</td>
</tr>
<tr>
<td>Sony BVM-L230</td>
<td>LCD Class A reference monitor</td>
<td>23”</td>
<td>50</td>
<td>1080p50</td>
</tr>
<tr>
<td>Sony KDL-55X4500</td>
<td>Consumer LCD TV with RGB background illumination</td>
<td>56”</td>
<td>50</td>
<td>1080p50</td>
</tr>
<tr>
<td>JVC DLA-HD750</td>
<td>LCoS Projector</td>
<td>2.8m</td>
<td>50</td>
<td>1080p50</td>
</tr>
</tbody>
</table>
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...
Table 2: Video sequences and bitrates for different rate points (RP)

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Coding difficulty</th>
<th>Start frame</th>
<th>Used in dataset</th>
<th>Bitrates [MBit/s]</th>
<th>RP1</th>
<th>RP2</th>
<th>RP3</th>
<th>RP4</th>
</tr>
</thead>
<tbody>
<tr>
<td>OldTownCross</td>
<td>Easy</td>
<td>1217</td>
<td>1080p25</td>
<td>5.4</td>
<td>5.6</td>
<td>9.6</td>
<td>13.7</td>
<td>19.0</td>
</tr>
<tr>
<td>InToTree</td>
<td>Easy</td>
<td>5199</td>
<td>1080p25</td>
<td>5.8</td>
<td>10.4</td>
<td>13.1</td>
<td>17.1</td>
<td></td>
</tr>
<tr>
<td>FlagShoot</td>
<td>Easy (assumed)</td>
<td>6611</td>
<td>1080p50</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>CrowdRun</td>
<td>Difficult</td>
<td>7111</td>
<td>1080p25</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>TreeTilt</td>
<td>Medium</td>
<td>9077</td>
<td>1080p50</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>PrincessRun</td>
<td>Difficult</td>
<td>10429</td>
<td>1080p50</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>ParkJoy</td>
<td>Difficult</td>
<td>15523</td>
<td>1080p25</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>DanceKiss</td>
<td>Easy</td>
<td>17953</td>
<td>1080p50</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
</tbody>
</table>

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.

4.1. 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 sub-tests and at least 15 valid subjects as suggested in [11] evaluated each sub-test. The 95% confidence intervals of the subjective votes...
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 dB(A). The silence noise level i.e. the noise level with all devices running but without playing a sound was below 30 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


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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>1080p25</th>
<th>1080p50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder</td>
<td>LC</td>
<td>HC</td>
</tr>
<tr>
<td>JM 12.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JM 17.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile&amp;Level</td>
<td>Main, 4.0</td>
<td>High, 5.0</td>
</tr>
<tr>
<td>Reference Frames</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>R/D Optimization</td>
<td>Fast Mode</td>
<td>On</td>
</tr>
<tr>
<td>Search Range</td>
<td>32</td>
<td>128</td>
</tr>
<tr>
<td>B-Frames</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Hierarchical Encoding</td>
<td>On</td>
<td>On</td>
</tr>
<tr>
<td>Temporal Levels</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Intra Period</td>
<td>0.5 s</td>
<td>0.5 s</td>
</tr>
<tr>
<td>Deblocking</td>
<td>On</td>
<td>On</td>
</tr>
<tr>
<td>8x8 Transform</td>
<td>Off</td>
<td>On</td>
</tr>
</tbody>
</table>
**Fig. 7:** Basic test cells

![Diagram of basic test cells](image)

**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.

![Graphs for CrowdRun, ParkJoy, OldTownCross, and InToTree](image)

**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.

![Graphs for TreeTilt, DanceKiss, FlagShoot, CrowdRun, and PrincessRun](image)