

HIGH DEFINITION H.264/AVC SUBJECTIVE VIDEO DATABASE FOR EVALUATING THE INFLUENCE OF SLICE LOSSES ON QUALITY PERCEPTION

Nicolas Staelens^a Glenn Van Wallendael^b
Rik Van de Walle^b Filip De Turck^a Piet Demeester^a

^aGhent University - iMinds, Department of Information Technology, Ghent, Belgium

^bGhent University - iMinds, Department of Electronics and Information Systems, Ghent, Belgium

ABSTRACT

Prior to the construction or validation of objective video quality metrics, ground-truth data must be collected by means of a subjective video database. This database consists of (impaired) video sequences and corresponding subjective quality ratings. However, creating this subjective database is a time-consuming and expensive task. There is an ongoing effort towards publishing such subjective video databases into the public domain. This facilitates the development of new objective quality metrics. In this paper, we present a new subjective video database consisting of impaired High Definition H.264/AVC encoded video sequences and associated quality ratings gathered from a subjective experiment. This database can be used freely to determine impairment visibility or estimate overall quality of a video in the case of lost slices due to network impairments.

Index Terms— Video quality assessment, Subjective video quality, Objective video quality metric, Quality of Experience, Quality of Service

1. INTRODUCTION

The goal of objective video quality metrics is to predict perceived video quality automatically and reliably. These metrics can be used by, for example, video service providers to continuously measure the quality of their video streams and to verify whether the performance of their services meets end-users' requirements [1, 2]. Ensuring end-users receive adequate Quality of Experience (QoE)[3] is key for maintaining customer satisfaction. Active research is still ongoing towards the construction of objective metrics for real-time prediction of video quality.

Constructing an objective video quality metric requires a structured approach consisting of different steps as shown in Figure 1.

Once the scope for the metric has been defined, a database needs to be collected consisting of a set of representative (impaired) video sequences and corresponding quality ratings. The construction of this video database comprises the selection of a number of source video sequences, encoding them

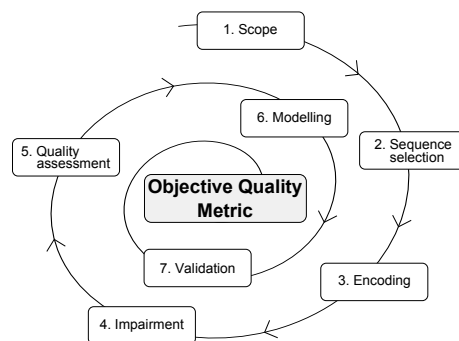


Fig. 1. In order to model perceived video quality, a structured stepwise approach is recommended.

and injecting coding and/or network impairments. Next, the quality of these sequences must be evaluated by means of a subjective video quality assessment experiment. This video database with subjective quality ratings is then used as ground-truth for modelling and validating new objective video quality metrics.

Pragmatically, subjective experiments are time-consuming. First, the video sequences to be evaluated must be created and collected, and the assessment environment must be set up. Second, it also takes time to conduct the experiment and collect results from different non-expert subjects. And third, these experiments are also expensive. This is partly due to the amount of time which is invested starting from the experiment setup up to processing the obtained results. In most cases, the test subjects are also compensated for their effort and time.

For long, subjective video databases have been kept private and secret. However, nowadays, an increasing number of databases is being made publicly available in order to facilitate the video quality research community¹. In this paper, we describe a new subjective video database for assessing the influence of packet loss impairments in the case of High Definition (HD) H.264/AVC encoded video sequences. This database contains a number of impaired encoded video

¹A comprehensive list of publicly available video databases is maintained at <http://dbq.multimediatech.cz>.



Fig. 2. Overview of the eight selected source video sequences, taken from open source movies, CDVL and TUM.

bitstream, various trace files, and associated subjective quality ratings. Our video database can be used for modelling impairment visibility and overall perceived video quality.

The remainder of this paper is structured as follows. In Section 2, we detail the creation of the impaired video sequences. An overview is presented of the selection of the source sequences, and information is provided on the encoding and impairment process. Next, the subjective quality assessment experiment is described in Section 3 and a brief data analysis of the obtained subjective quality ratings is presented in Section 4. In Section 5, we describe two objective video quality metrics which have been constructed using this subjective video database and, finally, the paper is concluded in Section 6.

2. DATABASE CREATION

Before setting up a subjective experiment, video sequences must be created with different kinds of impairments. In this section, the process of selecting, encoding and impairing the video sequences is explained in more details.

2.1. Source video sequences

In order to span a wide range of different content types, eight video sequences were selected based on their amounts of spatial (SI) and temporal (TI) information [4]. These sequences were taken from open source movies, the Consumer Digital Video Library (CDVL) [5] and the Technical University of Munich (TUM). However, instead of taking the maximum SI and TI value over all the video frames, the upper quartile (Q3) value is taken as overall value for the sequence, as recommended by Ostaszewska *et al.* [6]. In Figure 2, a screenshot is presented of the eight selected video sequences. The calculated Q3.SI and Q3.TI values for each of the sequences are

Table 1. Characteristics of the eight selected sequences.

Sequence	Source	Q3.SI	Q3.TI
basketball	CDVL	62.07	29.67
BBB*	<i>Big Buck Bunny</i>	29.77	13.26
cheetah	CDVL	41.33	25.62
ED*	<i>Elephants Dream</i>	63.55	8.27
foxbird3e	CDVL	45.27	25.85
purple4e	CDVL	73.03	23.41
rush hour	TUM	23.72	9.53
SSTB*	<i>Sita Sings the Blues</i>	66.17	9.73

presented in Table 1. Marked sequences (*) are taken from open source movies.

All selected video sequences were in full HD resolution (1920x1080 pixels), with a frame rate of 25 frames per second and a duration of 10 seconds.

2.2. Sequence encoding

Different encoder configurations were obtained by analysing video content available from different online video platforms (such as YouTube and Vimeo) and by inspecting the default settings recommended by commercially available H.264/AVC encoders. Based on this analysis, the following settings were used to encode our eight source video sequences:

- Number of slices per picture: 1, 4 and 8;
- Number of B-pictures: 0, 1 and 2;
- GOP size [7]: 15 (0 or 1 B-picture) or 16 (2 B-pictures);
- Closed GOP structure;
- Bit rate: 15 Mbps.

The main focus of our subjective video database is evaluating the influence of network impairments on perceived quality.

Therefore, the encoding bitrate was set high enough in order to ensure no encoding artefacts were present in the sequences. All sequences were also visually inspected.

Based on the parameters listed above, *x264* [8] was used to encode each of our eight video sequences nine different times.

2.3. Impairment generation

As indicated in the previous section, the focus of this video database is studying the effects of network impairments (e.g. packet losses) on the quality perceived by end-users. More specifically, we are interested in assessing the influence of losing particular slices in the case of H.264/AVC encoded video sequences. Different impairment scenarios were created by considering the following parameters:

- Number of B-pictures between two reference pictures (0, 1, 2);
- Type of first lost slice (I, P, B);
- Location within the GOP of the loss (begin, middle, end);
- Number of consecutive slice losses (1, 2, 4);
- Location within the picture of the first lost slice (top, middle, bottom);
- Number of consecutive entire picture drops (0, 1).

These loss scenarios were simulated using Sirannon [9] based on the configuration depicted in Figure 3. In this setup, the

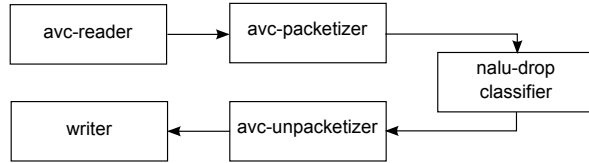


Fig. 3. RTP packets, which carry data from particular slices, are dropped using the *nalu-drop classifier* component. After unpacketizing, the resulting impaired sequence is saved to a new file.

raw H.264/AVC Annex B bitstream is packetized as specified in RFC 3984 [10]. Next, slices are dropped by discarding all RTP packets carrying data from the corresponding slice(s)². Finally, after unpacketizing the stream, the impaired H.264/AVC Annex B compliant bitstream is saved to a new file.

Creating a full factorial using all parameters listed above would result in a total number of 486 loss scenarios. It is clear that this results in a too large number of video sequences to be evaluated subjectively. Furthermore, not all combinations

of parameters are feasible; for example, B-slices can only be dropped when the encoded video sequence actually contains B-pictures. Experimental design was used to further limit the number of loss scenarios by eliminating the least unique scenarios [11]. As each loss scenario is identified by a combination of six parameters (6-tuple), the least unique scenario is the one for which each parameter value occurs the most. For example, in the following loss scenarios, the third one is the least unique:

- (0, I, end, 2, top, 0)
- (1, P, begin, 4, middle, 0)
- (1, P, begin, 1, top, 0) ← *least unique*
- (2, B, begin, 2, end, 1)

This process was repeated to select a final number of 48 loss scenarios. This resulted in a total number of 384 impaired encoded video sequences. No visual impairments occurred during the first and last two seconds of video playback.

For decoding the impaired H.264/AVC video sequences, we used an adjusted version of the JM reference decoder and added frame copy as error concealment strategy in the case of entire picture drops [12].

2.4. Trace files

Different trace files are also included in the subjective video database which provide more detailed information on the loss scenario. These trace files can be used, for example, to identify the exact location of the loss in terms of missing pictures, slices, NAL units and RTP packets, or to obtain information about all RTP packets (resulting from the packetization process).

Furthermore, for each of the impaired encoded video sequences, a Hybrid Model Input XML (HMIX) file is also available. This XML-based data exchange file format has been proposed by the Video Quality Experts Group (VQEG) Joint Effort Group (JEG), to enable faster development of new objective quality metrics [13]. HMIX files contain detailed information about slices, macroblocks, motion vectors, and quantization parameters and provide an alternative way to process and extract all necessary information from the video sequence in order to model video quality.

3. SUBJECTIVE QUALITY ASSESSMENT

3.1. Environment setup and assessment methodology

As explained in the previous section, a total number of 384 impaired and 72 encoded video sequences needed to be evaluated subjectively. In order to avoid viewer fatigue, experiment duration was limited to around 25 minutes by splitting the video sequences into six distinct datasets each containing 76 sequences.

²No aggregation was used during the packetization process.

Subjective quality assessment experiments were conducted using these six datasets following the Single Stimulus (SS) Absolute Category Rating (ACR) methodology [4]. As such, all sequences were shown one-after-another. Immediately after watching each sequence, subjects had to answer the following questions:

1. Did you see any visual impairment(s)?
2. How would you rate the visual quality of the video sequence?

A 5-grade ACR scale with adjectives was used to gather the subjective quality ratings.

Before the start of the experiment, subjects received specific instructions on how to evaluate the different video sequences. Three training sequences were also used to make the observers familiar with the experiment. Furthermore, all subjects were tested for visual acuity and normal vision using Ishihara plates and a Snellen chart, respectively. Playout order of the different video sequences was randomized in each experiment so that no two subjects evaluated the sequences in exactly the same order.

All experiments were conducted inside an environment compliant with ITU-R BT.500 [14]. Test subjects were seated at a distance of 4 times the picture height (4H) of a 40 inch full HD LCD television set.

3.2. Subjects

Overall, a total number of 40 distinct test subjects participated in this experiment. As this experiment consisted of six different datasets, subjects were encouraged to evaluate more than one dataset although not necessarily on the same day. Each dataset was evaluated by exactly 24 subjects. Post-experiment screening of the test subjects, as detailed in Annex V of the VQEG HDTV report [15], was used to eliminate outliers from the response data.

Table 2. Overview of subjects per dataset.

Dataset	Male		Female	
	Count	Age range	Count	Age range
01	19	[18-34]	5	[23-30]
02	20	[18-34]	4	[23-30]
03	19	[18-34]	5	[23-30]
04	16	[20-34]	8	[22-30]
05	18	[24-34]	6	[20-30]
06	18	[24-34]	6	[20-30]

In Table 2, more information on the test subjects per dataset is provided.

4. DATA ANALYSIS

In [16], Winkler proposes a number of indicators for analysing the subjective ratings of a video database. The indicators are

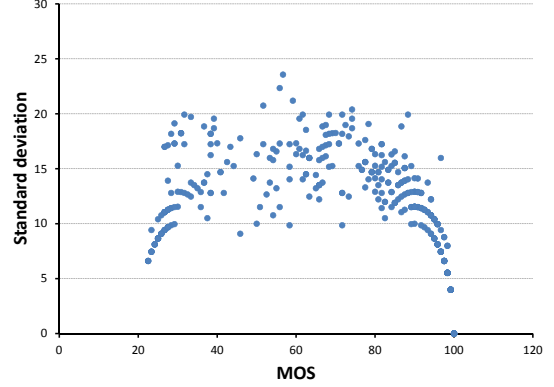


Fig. 4. MOS versus standard deviation of our subjective quality ratings.

mainly based on the Mean Opinion Scores (MOS) and standard deviations as they are indicative for the quality range and precision of the results. As recommended, we first linearly scaled our responses to a 100 scale for analysing the quality of our subjective ratings. For calculating the indicators, we only consider the impaired video sequences. Figure 4 plots the standard deviations against the MOS for each sequence. In correspondence with the results in [16], the standard deviation is typically higher around the middle of the MOS range. The inverted-U shape of the plot is largely due to the clipping of the ratings towards the ends of the scale [17].

The criteria used for quantifying the characteristics of subjective ratings and corresponding values are:

- *Range of MOS (R^{MOS})*: Inter-percentile range (central 90% range)
→ **value:** 74.86
- *Uniformity of coverage (U^{MOS})*: indication of the spread of the quality levels over the whole range.
→ **value:** 0.80
- *Variability (V)*: small variability means that the MOS is more ‘reliable’ and indicates smaller confidence intervals.
→ **value:** 16.06
- *Discriminability (D)*: Indicates how well subjects are able to distinguish individual videos across the database.
→ **value:** 0.77

These values correspond with the data of other video databases analysed by Winkler. The interested reader is referred to [16] for more information on the calculation and interpretation of the indicators listed above.

5. OBJECTIVE METRIC CONSTRUCTION

The subjective video database described in this paper has been used as ground-truth for constructing two No-Reference (NR)

bitstream-based objective video quality metrics. In both cases, machine learning techniques were used to model impairment visibility and quantify the quality of impaired video sequences.

5.1. Modelling impairment visibility

As described in Section 3.1, subjects had to indicate whether they perceived a visual impairment during video playback. This way, an objective video quality metric for automatic detection of visual impairments can be constructed. In turn, this enables service providers to verify that their end-users receive adequate QoE at all times [1, 2].

In [18], we extracted different parameters from the impaired encoded video bitstreams to create an NR bitstream-based objective video quality metric. The goal of this metric is to determine whether losses in the video stream will result in impairments deemed visible to the end-users. This can be defined as a classification problem. Therefore, we used decision trees as machine learning technique for determining impairment visibility. An example tree is depicted in Figure 5.

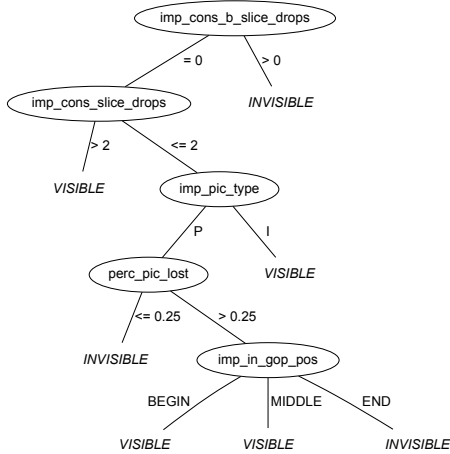


Fig. 5. Decision tree for determining impairment visibility, using only parameters extracted from the received encoded video bitstream.

Ten-fold cross validation was used to evaluate the performance of the decision tree. The tree depicted in Figure 5 has a classification accuracy of 83%.

The reader is referred to [18] for more details on using decision trees for modelling impairment visibility.

5.2. Quantify perceived video quality

A different approach was followed in [19], where we used symbolic regression to quantify the quality of the impaired video sequences. The goal was, again, to construct an NR bitstream-based metric by estimating video quality using parameters

extracted solely from the received video bitstream. Genetic programming was used to identify the most influencing parameters and to find the best fit for estimating quality.

As detailed in [19], perceived video quality (MOS_p) can be predicted according to the following equation:

$$MOS_p = 4.615 - 0.548 \cdot (20 \cdot i_loss \cdot (1.079 - perc_pic_lost) \cdot perc_pic_lost + imp_cons_slice_drops \cdot perc_pic_lost \cdot p_loss), \quad (1)$$

where i_loss and p_loss equal 1 in case the loss originates in an I- or P-picture, respectively; $perc_pic_lost$ represents the percentage of the picture lost; and $imp_cons_slice_drops$ represents the number of consecutive lost slices.

In this case, the EPFL-PoliMI video quality assessment database [20, 21] was used for validating our objective metric on unseen/untrained data. This resulted in a Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-Order Correlation Coefficient (SROCC) of respectively 0.88 and 0.88. These results outperform PSNR and VQM measurements [19].

6. CONCLUSIONS AND FUTURE WORK

In this paper, a subjective video database has been presented for assessing the influence of slice losses in the case of HD H.264/AVC encoded video sequences. Nine realistic encoder configurations and 48 loss scenarios have been used to create a total number of 456 video sequences. Real human quality ratings have been obtained by conducting several subjective video quality assessment experiments in a BT.500 compliant environment. This subjective video database can be used to model both impairment visibility and overall quality perception.

All original, encoded and impaired video sequences, trace files, and corresponding subjective quality ratings can be obtained from <http://avchd.intec.ugent.be>.

As part of future work, similar subjective video databases are being created for High Efficiency Video Coding (HEVC) and Multiview Video Coding (MVC) encoded video sequences.

Acknowledgment

The research activities that have been described in this paper were funded by Ghent University, iMinds and the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT). This paper is the result of research carried out as part of the OMUS project funded by iMinds. OMUS is being carried out by a consortium of the industrial partners: Technicolor, Televis, Streamovations and Excentis in cooperation with iMinds research groups: IBCN & MultimediaLab & WiCa (UGent), SMIT (VUB), PATS (UA) and COSIC (KUL).

Glenn Van Wallendael would also like to thank the Institute for the Promotion of Innovation through Science and

Technology in Flanders for financial support through this PhD grant.

7. REFERENCES

- [1] DSL Forum Technical Report TR-126, “Triple-play Services Quality of Experience (QoE) requirements,” DSL Forum, 2006.
- [2] ITU-T Recommendation G.1080, “Quality of Experience requirements for IPTV services,” International Telecommunication Union (ITU), 2008.
- [3] ITU-T Recommendation P.10/G.100 Amd 2, “Vocabulary for performance and quality of service,” 2008.
- [4] ITU-T Recommendation P.910, “Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union (ITU), 2008.
- [5] M.H. Pinson, S. Wolf, N. Tripathi, and C. Koh, “The consumer digital video library,” *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM-10)*, January 2010.
- [6] A. Ostaszewska and R. Kloda, “Quantifying the amount of spatial and temporal information in video test sequences,” in *Recent Advances in Mechatronics*, pp. 11–15. Springer Berlin Heidelberg, 2007.
- [7] Gerard O’Driscoll, *Next Generation IPTV Services and Technologies*, Wiley-Interscience, New York, NY, USA, 2008.
- [8] “x264,” <http://www.videolan.org/developers/x264.html>, available online.
- [9] A. Rombaut, N. Staelens, N. Vercammen, B. Vermeulen, and P. Demeester, “xStreamer: Modular Multimedia Streaming,” in *Proceedings of the seventeenth ACM international conference on Multimedia*, 2009, pp. 929–930.
- [10] S. Wenger, M.M. Hannuksela, T. Stockhammer, M. Westerlund, and D. Singer, “RTP Payload Format for H.264 Video,” February 2005.
- [11] K. Crombecq, E. Laermans, and T. Dhaene, “Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling,” *European Journal of Operational Research*, vol. 214, no. 3, pp. 683 – 696, 2011.
- [12] N. Staelens and G. Van Wallendael, “Adjusted JM Reference Software 16.1 with XML Tracefile Generation Capabilities,” VQEQ_JEG_Hybrid_2011_029_jm with xml tracefile_v1.0, Hillsboro, Oregon, US, December 2011.
- [13] N. Staelens, I. Sedano, M. Barkowsky, L. Janowski, K. Brunnström, and P. Le Callet, “Standardized toolchain and model development for video quality assessment - the mission of the joint effort group in VQEG,” in *Third International Workshop on Quality of Multimedia Experience (QoMEX)*, September 2011.
- [14] ITU-R Recommendation BT.500, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union (ITU), 2012.
- [15] Video Quality Experts Group (VQEG), “Report on the Validation of Video Quality Models for High Definition Video Content,” June 2010.
- [16] S. Winkler, “Analysis of public image and video databases for quality assessment,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 616 –625, October 2012.
- [17] 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.
- [18] N. Staelens, G. Van Wallendael, K. Crombecq, N. Vercammen, J. De Cock, B. Vermeulen, R. Van de Walle, T. Dhaene, and P. Demeester, “No-Reference Bitstream-based Visual Quality Impairment Detection for High Definition H.264/AVC Encoded Video Sequences,” *IEEE Transactions on Broadcasting*, vol. 58, no. 2, pp. 187–199, June 2012.
- [19] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, “Constructing a No-Reference H.264/AVC Bitstream-based Video Quality Metric using Genetic Programming-based Symbolic Regression,” *Circuits and Systems for Video Technology, IEEE Transactions on*, 2013, to appear.
- [20] F. De Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, “Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel,” in *First International Workshop on Quality of Multimedia Experience (QoMEX)*, July 2009, pp. 204 –209.
- [21] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, “A H.264/AVC video database for the evaluation of quality metrics,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, March 2010, pp. 2430 –2433.