Crowdsourced Subjective Video Quality Assessment

Krešimir Šakić [1], Emil Dumić [2], Sonja Grgić [2]
[1] Croatian Post and Electronic Communications Agency (HAKOM)
Radio Communication Department, Roberta Frangeša Mihanovića 9, HR-10110 Zagreb, Croatia
[2] University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Wireless Communications
Unska 3, HR-10000 Zagreb, Croatia
kresimir.sakic@hakom.hr, emil.dumic@fer.hr, sonja.grgic@fer.hr

Abstract - **This paper describes a crowdsourced subjective video quality method which evaluates various degradation types including wireless packet losses, transmission errors and compression artefacts. In recent times crowdsourcing gained a lot of momentum in various fields. The crowdsourcing method could be used in situations where a larger set of results is needed i.e. in Video Quality Assessment. The aim of this paper is to evaluate the usage of crowdsourcing for subjective video quality assessment and compare it with conventional subjective video quality assessment. To achieve this comparison we used an existing video database. Additionally, description of the crowdsourcing application design and the application usage is given, and its further development is envisaged.**

Keywords – **Crowdsourcing; video quality; subjective assesment; video database; LIVE video database**

I. Introduction

Subjective video quality assessment is often used as it is the most accurate reflection of user experience which is a complex combination of texture, colour, motion, audio and context. In subjective assessment, test subjects watch several video sequences and rate its quality on a numeric scale.

Traditional subjective quality evaluation methods have high costs, as they imply the usage of a video evaluation laboratory. One of these methods is the ITU-R BT.500-11 [1]. Additionally, the other setback is the recruiting of observers. It is difficult to recruit and motivate test subjects to participate in the subjective evaluation test. Reimbursement of their time helps solve this, but also raises costs. Often the observers are engineers and students, whose perception of video quality does not accurately represent the perception of the general public whom the video services target.

According to the ITU-R BT.500-11 [1] subjective quality test methods have been divided in:

- general subjective test methods:
  - double stimulus impairment scale (DSIS);
  - double stimulus continuous quality scale (DSCQS);
- alternative subjective test methods:
  - single stimulus methods;
  - stimulus-comparison methods;
  - single stimulus continuous quality evaluation (SSCQE);
  - simultaneous double stimulus for continuous evaluation method (SDSCE).

During the course of designing a video communication system there is a constant need for assessment of various algorithmic optimizations and content variations. Additionally, even if an organization manages to recruit a good number of observers, the problem of repetition and maintaining user interest remains. Over the course of time and test repetitions observer could develop biases and expectations which could lead to inaccurate results. All of these setbacks in conducting subjective video quality assessment are the reason why objective measures are used in system and algorithmic optimization, although there are no universally accepted objective measures.

Objective measures such as PSNR give a measure of how accurately an encoder can represent encoded video pixels. It uses all video pixels in order to asses a quality without taking into account whether the user perceived all of the pixels or not. Because of that, different image (e.g. Structural Similarity index, SSIM, [2]) and video (e.g. Video Quality Measure, VQM, [3]) quality measures have been developed. Their goal is to approximate the human quality perception (or Human Visual System, HVS) as much as possible, and consequently to correlate well with subjective measures (Mean Opinion Score, MOS).

Recently, with the introduction of virtual lossless compression [4], encoders started to exploit human visual perception. This is achieved by discarding portions of the video signal that are not perceived by the user. In spite of all advances the most reliable user experience quality measure is subjective evaluation.

Recent development in crowdsourced subjective testing [5]-[6] proves that there is a possibility to reduce costs of subjective video assessment. Conducting the subjective video quality assessment over the internet on a crowdsourced platform enables fast and low-cost evaluations. In the recent history, crowdsourced platforms have been used in various fields including the multimedia domain for the tasks of image annotation [7]-[9] and video summarization [10]-[12]. Additionally, further crowdsourcing applications were used for manual geo-location tagging of the video sequences [13], evaluation of the privacy filters applied in video surveillance sequences [14], gesture annotation [15], and nutritional analysis of photographed food [16].

This paper is organized as follows. Section II gives an overview of the application design for a crowdsourced subjective video quality assessment platform and describes the video database that was used. Section III presents and examines the application usage. Section IV presents our results and finally section V gives conclusions.

II. Application Design

In order to evaluate the usage of crowdsourcing for subjective video quality assessment a web application was developed. The first step in development was choosing the appropriate video sequences database. In this process we have chosen an existing video database which has already been used for conventional subjective video quality assessment. The LIVE video quality database [17] was used (later called LIVE video) in order to develop and test how much our crowdsourcing method differs from the DMOS (Difference Mean Opinion Score) results obtained in the controlled conditions. The LIVE video database consists of 10 original video sequences each having resolution of 768 x 432 pixels and 150 distorted video sequences (15 distorted sequences per one original) with 4 distortion types (details about their generation can be found in [17]):

- Wireless distortions (four test videos per reference);
- IP distortions (three test videos per reference);
- H.264 compression (four test videos per reference);
- MPEG-2 compression (four test videos per reference).

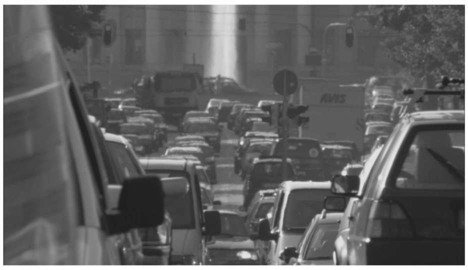

Figure 1. Original video sequence (*Rush hour*, frame 32)

Original video sequences have 8 bit planar YUV 4:2:0 format, while distorted video sequences have been converted back to the same format as the original. Six sequences have 250 frames (25 fps), one has 217 frames (25 fps) and three have 500 frames (50 fps). The frame example from the original video sequence is shown in Fig. 1. The frame examples from distorted sequences are IP distortion (Fig. 2) and H.264 distortion (Fig. 3), however they represent the worst frame (according to the PSNR) between all IP and H.264 degraded sequences, for *Rush hour* video sequence.

The original subjective study for LIVE video database was conducted using a single stimulus procedure and the observers indicated the video quality on a continuous scale. Subjects viewed each of the reference videos to facilitate computation of difference scores using hidden reference removal. Each video in the original study was viewed by 38 observers. 9 observers out of 38 were unreliable according to specifications in ITU-R BT. 500-11 and the subjective data is provided from 29 valid observers in the form of DMOS scores.

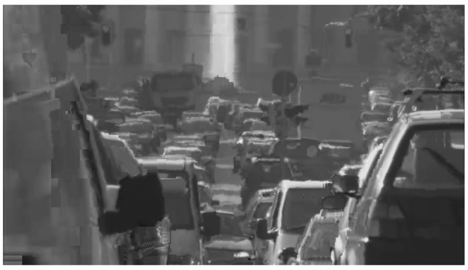

Figure 2. IP based distorsion (*Rush hour*, frame 32)

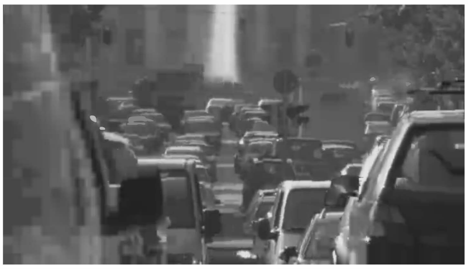

Figure 3. H.264 based distorsion (*Rush hour*, frame 32)

The original sequences from the LIVE video database were in uncompressed YUV format and their overall size was 23.8 GB. As our application is web-based, the original sequences needed to be compressed before being implemented in our application. We have used H.264 compression and ffmpeg [18], version N-50515-g28adecf [19]. Following settings were used: -vcodec libx264, -preset very slow and constant quality mode -crf 13, so that PSNR between uncompressed (YUV) and compressed sequences was between 39.7 - 49.4 dB and SSIM [2] between 0.981 - 0.997. Average PSNR was 45.9 dB and average SSIM was 0.994, which may be considered as near lossless. Taking all that into account, it can be presumed that the newly introduced compression will not influence on the tested degradation. It should be also noted that presented PSNR and SSIM values should not be considered for analysis or comparison with other codecs, because higher values could be achieved by using ffmpeg with other presets (specifically for PSNR or SSIM), but are here used just as an indicator of similarity between uncompressed and compressed sequences. At the end of the compression process the compressed LIVE video database had the size of about 1 GB.

The subjective test was done by using single stimulus procedure SSCQE (Single Stimulus Continuous Quality Evaluation) according to the ITU-R BT.500-11 [1] and the observers indicated the quality of the video on a continuous scale 0-10 (step size 0.1). The testing procedure is shown in Fig. 4.

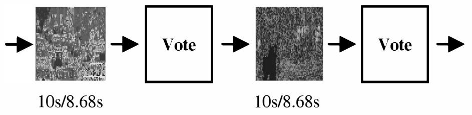

Figure 4. Single Stimulus Continuous Quality Evaluation procedure used in our subjective test

Observers were mostly students and co-workers, non-experts, between 20 and 40 years old. Additionally, due to the nature of crowdsourcing, tests were not done in a room with similar lighting conditions. Also, observers graded video sequences on different sized monitors with different calibration. Out of the around 200 invited 70 observers successfully ended the subjective test. Each observer graded 40 or 50 different video sequences (with all degradation types present) and 5 test sequences (qualification test) at the beginning, to understand how to grade other sequences. These 5 test grades were removed from further calculation. Observers also viewed each of the reference videos to be able to calculate DMOS scores, using hidden reference removal method. Overall duration of the test was around 15 to 20 minutes. Additionally the voting time was not limited, and in a few cases the overall duration exceeded 20 min. Depending on the duration of the distraction these observers experienced, they would stop at a certain sequence and then continue with the test several minutes later. In the future, we should consider limiting the overall duration of the test.

III. Application Usage

The development of the application for crowdsourced video quality assessment is planned in two stages. By the time of preparing this paper the first stage was conducted. The first stage included testing the application on college students and engineers at the University of Zagreb, Faculty of Electrical Engineering and Computing and on co-workers in the Croatian Post and Electronic Communications Agency. This step gave us feedback for development of the second stage which will be released to the general public. The incentive used in the first stage is extra credit for participating students in their final grade and free coffee for co-workers participating in the test.

The main goal of the second stage is releasing the application to the general public and testing it as a true crowdsourcing platform. There are several parameters that could be changed in the design of the second stage application. One of the most important parameters is the duration of the testing. The feedback from the first of several observers implied that the test was too long and that their concentration was dropping near the end of the test. When the application will be released to the general public, the motivation of the observers will be difficult to maintain. Therefore, we will consider shortening the overall duration of the test from 20 minutes to around 5 to 10 minutes. Equally important, we need to include adequate incentive/remuneration for the observers. Additionally we should consider different motivational methods for them. A production of a new video database could also be assessed.

IV. Results

The results from the subjective test have been written by the application in the result database. After the conclusion of the test, the results from the database have been obtained, averaged and compared with DMOS results from the LIVE video database. The first five sequences (qualification test) were removed from further calculation. Screening of the observers was performed according to the ITU-R BT.500-11 to discard observers who differ too much from the average value. Each residual (difference between reference and degraded video sequence's grade from the same observer) was converted to z-score according to the:

$$z_{nl} = \frac{d_{nl} - \mu_n}{\sigma_n} \tag{1}$$

In (1) z_{nl} are z-scores from each observer n, for video sequence l, d_{nl} are residuals from each observer n, for video sequence l, μ_n is mean score from observer n and σ_n is standard deviation from observer n (over all tested sequences l for that observer). According to [17], this is done to account for any differences in the use of the quality scale (differences in the location and range of values used by the observer).

For each time window (10s/8.68s per video sequence, reference or distorted) it was determined if z-scores were normal by using kurtosis β, over the span of all z-scores from the particular video sequence. Depending on the kurtosis, each observer was screened on deviation σ_l from the mean value $\bar{z}_l$ of each video sequence l. According to the ITU-R BT.500-11, process of discarding observers can be described according to the (2):

$$\begin{aligned}
&\forall l \in L \text{ where } L \text{ stands for number of video sequences}\\
&\forall n \in N \text{ where } N \text{ stands for number of observers}\\
&\left.\begin{aligned}
&\text{if } z_{nl} \geq \bar{z}_l + 2\cdot\sigma_l \text{ then } P_n = P_n + 1\\
&\text{if } z_{nl} \leq \bar{z}_l - 2\cdot\sigma_l \text{ then } Q_n = Q_n + 1
\end{aligned}\right\}\text{for } 2 \leq \beta \leq 4 \text{ (normal)}\\
&\left.\begin{aligned}
&\text{if } z_{nl} \geq \bar{z}_l + \sqrt{20}\cdot\sigma_l \text{ then } P_n = P_n + 1\\
&\text{if } z_{nl} \leq \bar{z}_l - \sqrt{20}\cdot\sigma_l \text{ then } Q_n = Q_n + 1
\end{aligned}\right\}\text{for } \beta \notin [2,4] \text{ (not normal)}
\end{aligned} \tag{2}$$

In addition, P and Q values were determined for every observer and if any of the values were greater than 5% of the number of tested degraded video sequences (30 or 40), that observer was discarded. Using this method, 15 observers were removed from further analysis.

Afterwards, results for every observer were rescaled to the full (and same) range of 0-100, according to the:

$$dmos_{n,l} = \frac{100}{\max(z) - \min(z)} \cdot (z_{n,l} - \min(z)) \tag{3}$$

In (3) max(z) and min(z) represent maximum and minimum z-scores over all observers and all video sequences and $dmos_{n,l}$ represents rescaled grades of the same viewer. At the end, average DMOS grade was calculated for each of the distorted video sequence as an arithmetic mean of all grades for each sequence (there were 12-17 grades per each video sequence).

At the end, DMOS values obtained by our method were compared with those in the LIVE video database. Comparison was made using linear Pearson's correlation coefficient.

We obtained Pearson's correlation of 0.8330. However, it is possible to obtain higher correlation by removing calculation of z-scores in (1) (and then changing z-scores into residuals d in (2) and (3)). Also, additional screening of the observers is possible by removing those with average reference grades below threshold (we have chosen threshold=3) prior standard screening described in (2). In this case, we removed overall 19 observers (4 in first and 15 in second, standard screening) and obtained correlation of 0.8923, Fig. 5. With only z-scores calculation removed, without additional screening, correlation was 0.8579 (with 16 observers removed in screening from (2)). It is possible that with higher number of observers correlation could be higher than 0.9.

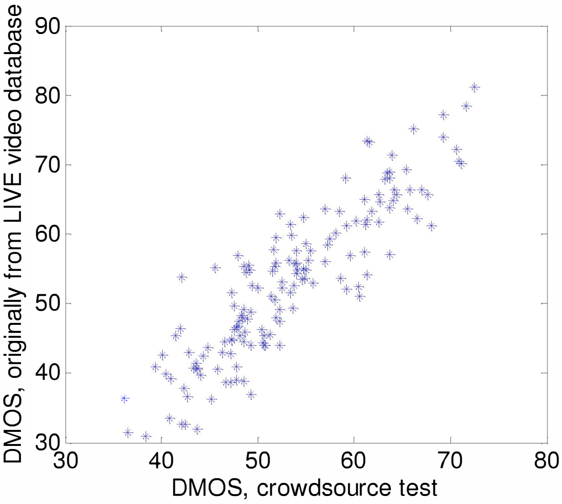

Figure 5. Comparison of DMOS results, using crowdsource testing versus original DMOS results (Pearson's correlation is 0.8923)

V. Conclusion

Traditional subjective image and video quality assessment is quite expensive as it implies the setup of a testing laboratory. In recent times the crowdsourcing platform has been used in various fields including multimedia domain.

In this paper we have described a crowdsourced subjective video quality method which evaluates various degradation types. In order to test this method a web crowdsourcing application was developed. The results from testing this method were compared to the conventional subjective video quality assessment. To achieve this comparison we used an existing video database (the LIVE video quality database) and obtained maximal Pearson's correlation of 0.8923. It is possible that with higher number of observers, correlation will be even higher.

In the future, different parameters can be adjusted to check if it is possible to calculate higher Pearson's correlation. Also, comparison with conventional objective measures will be made, to compare correlation between objective and subjective measures (from the LIVE video database and from our subjective experiment). Results will show if crowdsourced subjective tests can replace usually much more expensive laboratory tests in strictly controlled conditions.

References

[1] ITU-R BT.500-11 "Methodology for the subjective assessment of the quality of television pictures", International Telecommunication Union/ITU Radiocommunication Sector, 2002.

[2] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE Trans. on Image Proc., Vol. 13, No. 4, pp. 600-612., 2004.

[3] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality", IEEE Trans. Broadcast, vol. 50, no. 3, pp. 312–322, Sep. 2004

[4] V. Adzic, H. Kalva, and L.-T. Cheok, "Adapting video delivery based on motion triggered visual attention", Applications of Digital Image Processing XXXV, pp. 84991L–84991L, 2012.

[5] Ó. Figuerola Salas, V. Adzic, H. Kalva, "Subjective Quality Evaluations Using Crowdsourcing", 30th Picture Coding Symposium, pp.418-421, 2013.

[6] C. Keimel, J. Habigt and K. Diepold, "Challenges in Crowd-Based Video Quality Assessment", 4th International Workshop on Quality of Multimedia Experience (QoMEX), pp. 13-18, 2012.

[7] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation", Proceedings of the international conference on Multimedia information retrieval, New York, NY, USA, pp. 557–566., 2010.

[8] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon’s Mechanical Turk", Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, Stroudsburg, PA, USA, pp. 139–147., 2010.

[9] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing Annotations for Visual Object Detection", Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, pp. 40-46, 2012.

[10] T. Steiner, R. Verborgh, R. Van de Walle, M. Hausenblas, and J. G. Vallés, "Crowdsourcing event detection in YouTube video", 2011.

[11] S.-Y. Wu, R. Thawonmas, and K.-T. Chen, "Video summarization via crowdsourcing", CHI ’11 Extended Abstracts on Human Factors in Computing Systems, New York, NY, USA, pp. 1531–1536., 2011.

[12] A. Tang and S. Boring, "#EpicPlay: crowd-sourcing sports video highlights2", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 1569–1572., 2012.

[13] L. Gottlieb, J. Choi, P. Kelm, T. Sikora, and G. Friedland, "Pushing the limits of mechanical turk: qualifying the crowd for video geo-location", Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia, New York, NY, USA, pp. 23–28., 2012.

[14] P. Korshunov, S. Cai, and T. Ebrahimi, "Crowdsourcing approach for evaluation of privacy filters in video surveillance", in Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia, New York, NY, USA, pp. 35–40., 2012.

[15] I. Spiro, G. Taylor, G. Williams, and C. Bregler, "Hands by hand: Crowd-sourced motion tracking for gesture annotation", in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 17–24., 2010.

[16] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos, "latemate: crowdsourcing nutritional analysis from food photographs", in Proceedings of the 24th annual ACM symposium on User interface software and technology , New York, NY, USA, pp. 1–12., 2011.

[17] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video", IEEE Trans. Image Process., vol. 19, pp. 1427–1441, Jun. 2010.

[18] http://www.ffmpeg.org/index.html

[19] http://ffmpeg.zeranoe.com/builds/win64/static/