Crowd workers proven useful: A comparative study of subjective video quality assessment

Dietmar Saupe, Franz Hahn, Vlad Hosu
Igor Zingman, Masud Rana
Department of Computer and Information Science
University of Konstanz, Germany

Shujun Li
Department of Computer Science
Faculty of Engineering and Physical Sciences
University of Surrey, UK

***Abstract*—We carried out crowdsourced video quality assessments using paired comparisons and converting the results to differential mean opinion scores (DMOS). A previous lab-based study had provided corresponding MOS-values for absolute category ratings. Using a simple linear transformation to fit the crowdsourcing-based DMOS values to the lab-based MOS values, we compared the results in terms of correlation coefficients and visually checked the relationship on scatter plots. The comparison result is surprisingly good with correlation coefficients more than 0.96, although (1) the original video sequences had to be cropped and downscaled in the crowdsourcing-based experiments, (2) the control of the experimental setup for the crowdsourcing case was much less and (3) it was widely believed that data from crowdsourcing workers are less reliable. Our result suggests crowdsourcing workers can actually be used to collect reliable VQA data in some applications.**[1]

I. Introduction

Subjective quality evaluation of speech, audio, image, video, or multimedia is typically carried out in laboratory environments that allow for controlling the testing environment as required according to the ITU recommendations [1]. It has been shown, however, that crowdsourcing can provide reliable measures of quality of experience (QoE) without such rigorous control for images [2] and also video stimuli [3]. In this contribution we report about a crowdsourcing-based experiment aimed at reproducing mean opinion scores (MOS) from a previously-conducted lab-based experiment for the visual quality of a set of high resolution (HDTV, 1920x1080) video sequences by comparing video pairs where two video sequences are shown side-by-side on the crowd workers' own (uncontrolled) displays. Although we had to reduce the video sequences' resolution to 480x400 to fit a variety of displays, we have found a very strong correlation between the high-resolution lab-based MOS values and our crowdsourcing-based low-resolution paired comparison (PC) results.

II. Video data sets and crowdsourcing platform

For our experiments we chose the IRCCyN IVC 1080i video quality database [4] that contains 24 groups of video sequences in 1920x1080 resolution and i50 (interlaced, 50 fps) format of 9 to 12 seconds, each group including 8 video sequences with different levels of visual quality ranging from excellent to bad. The video sequences with different levels of quality were obtained by encoding 24 raw source video sequences using different bitrates. To keep this pilot study more manageable as a first experiment we used only 10 of 24 sources and only 4 of 8 levels of quality, evenly spreading over the whole range of quality. The study in [4] provided the absolute category ratings using the hidden reference (ACR-HR) testing methodology. These MOS values were obtained from 24 observers. The paper also showed MOS values according to the subjective assessment methodology for video quality (SAMVIQ) and compared these results.

In order to facilitate our crowdsourcing-based study a number of video preprocessing steps were carried out using the open source cross-platform conversion software FFmpeg:

- The interlaced HDTV video sequences in raw YUV format and 1920x1080 resolution were deinterlaced.
- To allow for simultaneous, side-by-side display of a pair of video sequences on a typical crowd worker's screen size (which we assumed to have at least 1024 pixels horizontally), we cropped the video sequences (keeping the central part) to a format of 480x400 pixels.
- To reduce the bandwidth requirements for online video streaming we re-encoded the video sequences to a lower bitrate (2 to 4 MB per video sequence). To ensure that no significant compression artifacts are generated by this step, we required a minimum PSNR of 40 dB.
- For each of the 10 video groups each with 4 quality levels (L1, smallest bitrate and lowest quality, to L4, largest bitrate and best quality) we composed a series of $\binom{4}{2} = 6$ merged pairs of video sequences for the paired comparison tasks. In each pair one of the videos was randomly selected for display on the left and the other on the right side.

For the crowdsourcing platform we used CrowdFlower (http://www.crowdflower.com/) and the implementation of the tests was based on the QualityCrowd framework [5].

III. Experimental setup and data analysis

In the standard paired comparison methodology [1] subjects are asked for a binary choice, namely which of the two presented stimuli has the better quality. Here we adopted a finer scale of categories similar to the five-level Likert scale.

Value	Interpretation
-2	Left is much better
-1	Left is slightly better
0	No difference
1	Right is slightly better
2	Right is much better

[1]We thank the German Research Foundation (DFG) for financial support within project A05 of SFB/Transregio161.

In crowdsourcing-based experiments one of the key problems is to ensure the reliability of the performance of the crowd workers. Typically, in CrowdFlower this is done as follows:

- A qualification test that must be passed in order to begin with the actual experiment.
- Test questions on each page of an experiment for which ground truth answers are provided. Workers whose success rate on these test questions drops below 70% are excluded from further work on the job and their answers will not be passed on for analysis.
- A system of worker qualification levels. A crowd worker with a good track record earns upgrades of his/her level and one with poor performance will receive a level reduction. Workers with higher qualification levels have access to better paid tasks and crowd employers may limit the access to their jobs by imposing a minimum level of qualification.

To assess the effect of the degree of control we performed two crowdsourcing studies, the second one with less restricting requirements regarding reliability control in the qualification test and in the test questions during the actual work.

In order to compare the lab-based MOS ratings of [4] (values range from 1 to 5) with our five-level paired comparisons (values from -2 to 2) one must either transform the paired comparisons to the MOS, i.e., one must reconstruct absolute quality levels from the relative comparisons of video qualities or, vice versa, transform the MOS values of the ACR ratings to pairwise ratings of differences in quality. The first approach of reconstructing absolute quality ratings from differences is well researched only for binary judgements ('right is better or worse than left'), additionally allowing for a tie, see e.g. [6]. For simplicity, we therefore convert the lab-based MOS of the left and right video stimuli, say S_l and S_r, to a differential MOS (DMOS) by linearly mapping the difference $S_r - S_l \in [-4, 4]$ of the MOS to $\frac{5}{8}(S_r - S_l) \in [-2.5, 2.5]$, so that by rounding to the nearest integer one would get a value in the range of our five-level Likert scale $\{-2, -1, 0, 1, 2\}$.

In total the experiment had 626 participants, 589 (94.1%) of which passed the quiz, and 576 (92%) of which passed the quality assurance testing during the experiment. On average 5.2 judgments were performed per participant. Each stimulus was rated 50 times.

IV. Results

The MOS values from the lab-based study [4] together with their standard deviations, both averaged over 10 stimuli in each quality level L1 to L4, are presented below.

Level 1	Level 2	Level 3	Level 4
1.6±0.6	2.8±0.8	3.6±0.8	4.5±0.6

For each of the 60 video pairs in both crowdsourcing studies, about 50 DMOS values were collected from crowd workers. The following table gives the 6 comparative scores averaged over the 10 video groups together with the corresponding computed DMOS values from the lab-based study.

	L1–L2	L1–L3	L1–L4	L2–L3	L2–L4	L3–L4
Crowd I	0.9±0.4	1.1±0.5	1.5±0.3	0.3±0.5	0.9±0.4	0.6±0.3
Crowd II	0.8±0.3	1.0±0.5	1.4±0.3	0.4±0.4	0.8±0.4	0.5±0.2
Lab [4]	0.9±0.4	1.3±0.3	1.7±0.3	0.5±0.5	1.0±0.4	0.6±0.2

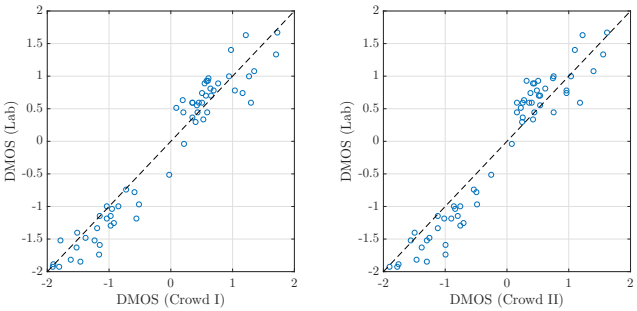

Fig. 1. Scatter plots comparing three assessments of DMOS values for 60 paired comparisons of video quality. Left: Crowd study 1 (strict quality control) versus DMOS derived from lab-based MOS values. Right: Crowd study 2 (mild quality control) versus lab. The Pearson correlation coefficients are 0.9687 (left), 0.9661 (right).

The scatter plots in Figure 1 for the 60 comparisons by different methods show very strong correlation between our crowdsourcing-based results and the lab-based ones.

V. Conclusion and Future Work

The results show that the crowdsourcing-based DMOS values are strongly related to the MOS values obtained in a strictly controlled study in a lab [4]. In fact, the lab-based study also had compared results between two closely related methodologies, namely ACR and SAMVIQ. The correlation coefficient was 0.8993. Compared with our results of 0.9687 and 0.9661, one can see that DMOS estimates using crowdsourcing can be as precise as lab-based studies even though there was severe processing of the video sequences and the control of testing conditions and worker reliability are considered much weaker in crowdsourcing studies. Another result is that MOS values of the lab-based study were linearly correlated with DMOS values from the crowd workers. These findings hold for our particular case study, but a generalization to all video quality assessment scenarios is not straighforward.

In our future work we will extend the study to include all source stimuli of the IRCCyN IVC 1080i video quality database (24 instead of 10), all quality levels (8 instead of 4) and a more elaborate analysis of the data including the reconstruction of MOS values from paired comparisons on the Likert scale.

References

[1] ITU-T, "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, 04/2008.

[2] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Proc. of 2011 18th IEEE Int. Conference on Image Processing (ICIP 2011)*. IEEE, 2011, pp. 3097–3100.

[3] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia qoe evaluation: A trusted framework," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1121–1137, 2013.

[4] S. Péchard, R. Pépion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: a resolution dependent paradigm," in *Proceedings of 2008 International Workshop on Image Media Quality and its Applications (IMQA 2008)*, 2008.

[5] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd—a framework for crowd-based quality evaluation," in *Proceedings of 2012 Picture Coding Symposium (PCS 2012)*. IEEE, 2012, pp. 245–248.

[6] H. A. David, *The method of paired comparisons*. Charles Griffin & Co. Ldt, 1988.