LARGE SCALE SUBJECTIVE VIDEO QUALITY STUDY

Zeina Sinno[1], Alan C. Bovik[1]

ABSTRACT

Most of today's video quality assessment (VQA) databases contain very limited content and distortion diversities and fail to adequately represent real world video impairments. This is in part because conducting subjective studies in the lab is slow, inefficient and expensive process. Crowdsourcing quality scores is a more scalable solution. However given that viewers operate under innumerable viewing conditions (including display resolutions, viewing distances, internet connection speeds) and because they are not closely supervised, multiple technical challenges arise. We carefully designed a framework in Amazon Mechanical Turk (AMT) to address the many technical issues that are faced. We launched the largest available VQA study, collecting more than 205000 opinion scores provided by more than 4700 unique participants. We have verified that our framework provided us with results that are highly consistent with the ones obtained in a lab environment under controlled conditions.

***Index Terms*—** Video Quality Assessment, Subjective Study, Crowdsourcing.

1 Introduction

The goal of VQA research entails to develop algorithms that closely correlate with humans' perception of quality. Consequently, these algorithms need to be trained and/or tested on extensive subjective video quality data sets so that it maybe asserted that they reflect or are capable of closely replicating human judgments. Over the past decade, researchers have developed multiple VQA databases. Notable databases include the LIVE VQA Database [1], the LIVE QoE Database [2], the LIVE Mobile Video Quality Database [3], the TUM databases [4, 5], and the MCL-V [6]. These databases offer very limited content and distortion diversity and were usually conducted under highly-controlled laboratory conditions by introducing sets of graded simulated impairments (H.264/AVC, packet loss, scalng artifacts...) onto a limited number of high-quality videos that were captured using high-end cameras. Real world videos are far more diverse and complex as they have been subjected to complex, nonlinear, commingled distortions that are likely impossible to accurately synthesize. The content in the these VQA databases has been videographed by only a few users, thereby constraining the ability of learned VQA models trained on them to generalize to diverse contents, levels of videographic expertise, and shooting styles [1–7].

Among the reasons behind these limitations is that each video must be rated by a substantial of subjects [8], while recruiting participants and conducting the experiments is time consuming and expensive. The subjects need to use allocated hardware, in some reserved physical space, and they need to be instructed individually about how to take these studies. These constraints limit how many studies can be ran concurrently, making the collection of the results inefficient.

Crowdsourcing still picture quality scores [9] has proved to be an efficient and successful way of collecting the data, motivating us to attempt a similar video study. Crowdsourcing video quality scores is, however, significantly more challenging because in addition to the issues related to participant problems (distraction, reliability and a imperfect training) encountered in the case of images, serious issues need to be addressed when videos are displayed, including variations in display quality, size and resolution, display hardware speed and bandwidth conditions. For example, slow hardware or bandwidth can cause video interruptions and stalls which will highly impact (skew) the collected quality scores. These technical problems listed above have not been fully addressed in previous attempts to crowdsource video quality scores [10–16].

In this work, we present a new framework in AMT to scale up the collection of video quality scores. We have meticulously designed the framework, paying attention to the design of the user interface, the monitoring of the subjects, and the overall supporting pipeline used to execute a large-scale study. We believe that this new approach is a more scalable way of collecting subjective scores than in-lab experiments. However, we were able to verify that the crowdsourced scores were quite consistent with those obtained in the lab. Scaling up subjective studies and collecting scores more efficiently is key for creating more comprehensive and representative VQA databases, which in turn will help advance VQA research.

2 Crowdsourced Video Quality Study

Here, we present the details of our new framework.

[1]: The author is at the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin, Austin, Texas, 78712. (emails: zeina@utexas.edu - bovik@ece.utexas.edu).

Table 1. Imposed constraints in the study.

	Constraint	Reason(s)
(1)	The subject's reliability score =$\frac{\text{Accepted Tasks}}{\text{Completed Tasks}} \geq 90$.	To filter out negligent subjects.
(2)	The subject can participate only once.	To avoid any judgment biases .
(3)	Mobile phones and tablets cannot be used to participate in the study.	1) For mobile devices, preloading videos into memory is disabled. The videos are streamed instead, causing stalls. 2) Mobile browsers downscale/upscale videos, causing additional artifacts.
(4)	The minimum display resolution is 1280×720.	The majority of the videos in the database had a resolution $\geq$ 1280x720.
(5)	The supported browsers are: Google Chrome, Safari, Mozilla Firefox, and Opera. The unsupported browsers are: Internet Explorer and Edge.	Video preloading is disabled on some browsers (Internet explorer and Edge).
(6)	The browser zoom level must be set to 100%.	To avoid downscaling/upscaling artifacts.
(7)	The used device should have a good computational power.	Slow hardware introduces video stalls.
(8)	The subject's network should have a good internet capacity.	1) For fast video preloading. 2) To avoid an additional computational overhead.

2.1 Content

Our goal is to create a framework that would allow us to collect subjective scores efficiently, leading to the availability of databases that represent real world videos more closely. To demonstrate this, we built a database of videos that were captured by 80 different inexpert videographers, aged between 11- 65 years, that volunteered to provide us with the content. The volunteers used 101 different devices (43 models - 15 mobile brands). The content was very diverse featuring scenes of nature, sports games, music concerts, parades, dancers, cowboys... The content was shot in all the populated continents and in $\sim$ 30 countries. We did not provide the volunteers with any instructions, except to upload their videos just as captured, without any processing (for example by video processing 'apps' like Instagram or Snapchat).

Originally, the volunteers provided us with 1000+ videos. We removed redundant content shot by the same volunteer, disturbing content (e.g.violent bull fight), and any videos with a duration of less than 10 seconds. We cropped the remaining videos to 10 seconds, while preserving story continuity. As a result, we obtained 585 videos. The final pool of videos had 18 different resolutions and spanned a wide range of quality owing to the intrinsic nature of many distortions including poor exposures, and a variety of motion blurs, haziness, various imperfect color representations, low-light effects including blur and graininess, resolution and compression artifacts, diverse defocus blurs, complicated combinations of all of these, and much more. The interactions of multiple artifacts also give rise to very complex, difficult to describe composite impairments, that were hard to identify. The resulting database is original as it presents the largest number of unique contents, capture devices, distortion types, contributors, and combinations of distortions ever found in a VQA database.

During a study session, a subject viewed 50 different videos: 7 during training and 43 during testing. The testing videos contained 4 distorted videos drawn from the LIVE Video Quality Assessment Database [1], which we will refer to as the "golden videos." These videos were previously rated by human viewers in the tightly controlled study [1], and are used, along with the prior subjective scores from [1], as a control to validate the subjects' ratings. The remaining 39 videos were drawn our database. 4 of those videos were displayed to all users, and 31 others were randomly selected, among which, 4 were displayed at relatively displayed moments as a control. The 43 testing videos were placed in re-randomized order for each subject.

2.2 Human Subjects

4776 AMT workers took part of our study, from highly diverse age groups, about half from each gender, and from diverse backgrounds (located in 56 different countries). This participants' sample is a much more globally representation sampling than any lab experiment. The participants received a single US dollar once they completed the study. Since the subjects were unsupervised and had various viewing conditions, we had to impose eligibility constraints to guarantee that the study would be executed smoothly and to collect more consistent results. A summary of the eligibility constraints and a brief justification of each of them are found in Table 1.

2.3 Framework

The subjective study workflow is presented in Fig. 1. Workers that meet constraint (1) are able to preview the study.

2.3.1 Overview

An overview containing a description of the task is presented along with some instructions on how to rate a video, and a few example videos to give them a clearer sense of the task. The worker was instructed to rate the videos based on how well s/he believes the presented video quality compares to an ideal, or best possible video of the same content. Several example videos were then played to demonstrate exemplars of some of the video distortions such as under exposure, stalls, shakes, blur and poor color representation. The worker was informed that other types of distortions exist and would be seen, so the

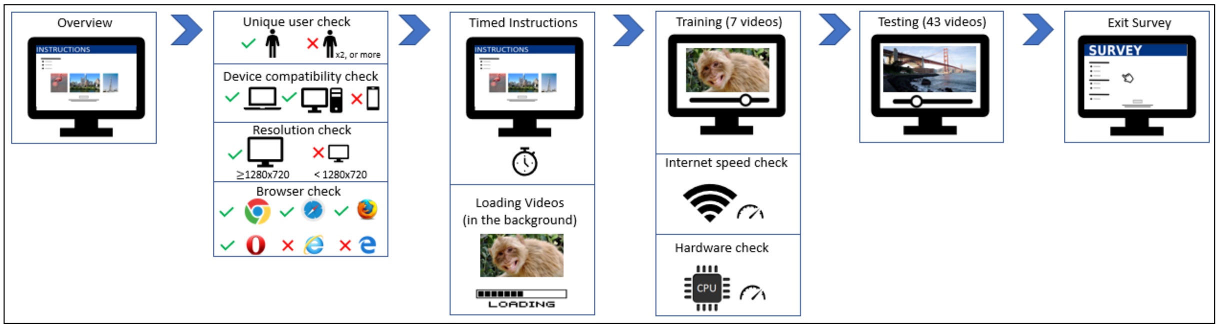

Fig. 1. Subjective study workflow.

worker would not supply ratings based only on the exemplar types of distortions, but would instead rate all distortions.

2.3.2 Constraints' Check

Once a worker accepted our hit, constraints (2)-(5) were checked. If the worker did not meet any, a message was displayed informing which constraint was not met. The worker was also encouraged to user a different display device and supported browser and to try again in case constraints (3)-(5) were not met. We automatically adjusted the browser's zoom level to 100% if constraint (6) was not met.

2.3.3 Timed Instructions

Next, the instructions were repeated again, with a countdown timer of one minute. While the instructions were being repeated, the first three videos began loading in the background, and the videos that were to be displayed during the testing phase were determined. Once the countdown timer reached zero, a *Proceed* button would appear at the bottom of the page, thereby allowing the worker to move forward.

2.3.4 Training

Afterwards, the training phase began, which consists of 7 videos. This phase was designed to feature various resolutions and videos suffering from multiple distortions and distortion levels, that span all the range of quality. The videos were displayed one at a time. The video controls were disabled and hidden to prevent less dedicated workers from pausing, replaying or skipping the videos. Before a video was fully loaded, a message was displayed showing the loading progress. Once the video was fully loaded, a message informed the user that "Video loaded and ready to be played." At this moment, an external *Play* button appeared, once clicked, the video was played in entirety (while being muted).

Once each video finished playing, it disappeared, revealing the rating interface consisting of a continuous bar that allowed the workers to rate the quality of the videos, where a Likert-like scale with 5 marks; Bad, Poor, Fair, Good, and Excellent is displayed. The initial position of the cursor was randomized. Once a change in the cursor location was detected, a *Next Video* button became clickeable. Once clicked, the worker moved to a new page, with a new video to be rated and the process continued until the last video had been rated.

During the training process, the play duration of each video was measured to assess the workers' play capability, to check constraints (7-8). There are many ways that stalls could occur while a video is playing. If a worker's hardware CPU was slow, or if other programs were running in the background (CPU is busy) then stalls or frame freezes could (and did) occur. Required background tasks (such as loading the videos to be played next) added processing overhead, while slower Internet bandwidths required increased processing overhead, further impacting foreground performance. During the training process, 7 videos of 10 seconds duration each were played. Importantly, the workers were not able to proceed further if it took more than 15 seconds to play any of the 7 videos or if any 3 of the 7 videos each required more than 12 seconds to play. Adopting this strategy guaranteed that most of the training videos were played smoothly, and also allowed us to eliminate workers who were unlikely be able to successfully complete the 'hit.'

2.3.5 Testing

A message was displayed informing the worker that testing was about to start. This phase was very similar to the training phase; the videos were displayed, controlled and rated in the same way. However, the testing phase required 43 videos to be rated, instead of 7.

2.3.6 Survey

Once the worker finished rating all of the videos, s/he was directed to the exit survey so that information regarding the following factors could be collected: the display, viewing distance, demographic information (gender, age and location of the worker), and whether the worker needed corrective lenses, and if so, if s/he wore them. The subjects were also asked whether they had any additional comments or questions.

3 Processing of the Results

On average, it took a worker 16.5 minutes to complete the study. Once the study was completed, after about a month, we had a total of 205 000 subjective scores.

While we hid the control bar of the videos, to prevent the subjects from skipping through them, we found that 2% of the workers were able to re-enable the controls by re-configuring the browser settings. So we excluded their results and we did not compensate them. We also excluded the results of the workers (2.5%) who indicated in the survey that they needed corrective lenses, but mentioned that they were not wearing them at the time of the study.

As mentioned earlier, we adopted a strategy to identify, during the training phase, the subjects that were the most susceptible to experiencing video stalls. While we were able to substantially mitigate the video stall problem, we were not able to eliminate it entirely owing to the fact that the performance of the processor changes over time, which might introduce hardware related stalls. Also internet connectivity can slow down, which might add an extra overhead (e.g. if the video in the background fails to load, it needs to be requested again). We observed that 77% of the videos were played without any stalls at all, while the rest of the played videos mostly suffered overall stall durations of 1-9 secs, all of which were rejected. We observed that in 95%, stalls were not favorable and led to a drop in the provided quality scores. We excluded the results of 11.5% of the subjects who suffered from stalls in 75% or more in the videos, because this can lead to a loss of focus. Next, on the remaining population we applied the guidelines of BT. 500-13 (Annex 2, section 2.3) [8] for subject rejection on the portion of non-stalled videos watched by the subjects and found that 0.5% of the subjects were outliers.

This number seemed low, so we also studied the intra-subject consistency. By design, each subject viewed 4 repeated videos during the test phase; we examined the differences in these pairs of scores, as follows. The average standard deviation of all non-stalled videos was about 18. We used this value as a threshold for consistency: given a non-stalled video that was viewed twice, the absolute difference in MOS of the two videos was computed. If it was below the threshold, then the rating for the video was regarded as consistent. Otherwise, it was not. We repeated this analysis across all the 4 videos across all subjects, and found that the majority ($\sim$99%) of the subjects were self-consistent at least half of the time. It is important to emphasize that we excluded the stalled videos from the consistency analysis and when applying the subject rejection [8], because the presence of any stalls rendered the corresponding subject ratings non-comparable.

Finally, we computed the mean opinion scores (MOS) of the videos. We noticed that the distribution of MOS spanned nearly all the range of possible qualities, with a greater density in the range 60-80.

4 Validation of the Results

4.1 Golden Videos

During the testing phase of each subject's session, 4 distorted videos from the LIVE VQA Database [1] - the aforementioned "Golden Videos" - were displayed at random placements to each worker to serve as a control. The mean Spearman rank ordered correlation (SROCC) values computed between the workers' MOS on the gold standard videos and the corresponding ground truth MOS values from the LIVE VQA was found to be 0.99. The mean absolute difference between the MOS values obtained from our study and the ground truth MOS values of the "Golden Videos" was 8.5. We also conducted a paired-sampled Wilcoxon t-test, and found that the differences between these to be insignificant at $p < 0.05$. The excellent agreement between the crowdsourced scores and the laboratory MOS significantly validates our experimental protocol.

4.2 Overall inter-subject consistency

To study overall subject consistency, we divided the opinion scores obtained on each video into two disjoint equal sets, then we computed MOS values on each set. We conducted on all the videos, then computed the SROCC between the two sets of MOS. This experiment was repeated 100 times, and the average SROCC between the halves was found to be 0.984.

5 Conclusion

We have described the construction of a new crowdsourced framework which we used to collect more than 205000 online opinion scores of the quality of 585 videos. The scores were provided by AMT workers of various backgrounds operating under highly variable viewing conditions. The significant geographic diversity of the subject pool raised many technical challenges related to user bandwidth and computing resources. The framework we built proved to be robust against the many variables affecting the video rating process, and we demonstrated that the data that we collected are in excellent agreement with these obtained in a laboratory setup. Our approach is faster, more efficient and less expensive. An extensive analysis of the results, including the impact of the different viewing parameters on the responses of the subjects, as well as an evaluation of the performance of leading VQA algorithms on the newly created dataset will follow in [17]. We believe that scaling up video subjective studies will be a driving factor in advancing VQA algorithm design and will help motivate the creation of VQA databases of more diverse content, and distortions.

6 References

[1] 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, no. 6, pp. 1427–1441, 2010.

[2] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, and A. Bovik, "Modeling the time varying subjective quality of HTTP video streams with rate adapta-

tions," *IEEE Trans. on Image Process.*, vol. 23, no. 5, pp. 2206–2221, 2014.

[3] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Select. Topics Sign. Process.*, vol. 6, no. 6, pp. 652–671, 2012.

[4] C. Keimel, J. Habigt, T. Habigt, M. Rothbucher, and K. Diepold, "Visual quality of current coding technologies at high definition IPTV bitrates," *IEEE Int'l Wkshp. Multidim. Sign. Process.*, pp. 390–393, 2010.

[5] C. Keimel, A. Redl, and K. Diepold, "The TUM high definition video datasets," *Int'l Wkshp. Qual. Multim. Exper.*, pp. 97–102, 2012.

[6] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, "MCL-V: A streaming video quality assessment database," *J. Vis. Commun. Image Repres.*, vol. 30, pp. 1–9, 2015.

[7] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H. 264/AVC video database for the evaluation of quality metrics," *Intern. Conf. Acous. Sp. Sign. Process. (ICASSP)*, pp. 2430–2433, 2010.

[8] "Methodology for the subjective assessment of the quality of television pictures." ITU-R Rec. BT. 500-13, 2012.

[9] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, 2016.

[10] K. T. Chen, C. J. Chang, C. C. Wu, Y. C. Chang, and C. L. Lei, "Quadrant of euphoria: A crowdsourcing platform for QoE assessment," *IEEE Net.*, vol. 24, no. 2, 2010.

[11] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," *Qual. Mult. Exp. (QoMEX)*, pp. 1–6, 2017.

[12] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multim.*, vol. 16, no. 2, pp. 541–558, 2014.

[13] Ó. Figuerola Salas, V. Adzic, A. Shah, and H. Kalva, "Assessing internet video quality using crowdsourcing," *Proc. ACM Int'; Wkshp. Crowd. Multim.*, pp. 23–28, 2013.

[14] M. Shahid, J. Søgaard, J. Pokhrel, K. Brunnström, K. Wang, S. Tavakoli, and N. Gracia, "Crowdsourcing based subjective quality assessment of adaptive video streaming," *Wkshp. Qual. Multim. Exper.*, pp. 53–54, 2014.

[15] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Comm. Surv. Tutorials*, vol. 17, no. 2, pp. 1126–1165, 2015.

[16] B. Rainer and C. Timmerer, "Quality of experience of web-based adaptive HTTP streaming clients in real-world environments using crowdsourcing," *ACM Wkshp. Desig. Quality Deployment Adapt. Video Streaming*, pp. 19–24, 2014.

[17] Z. Sinno and A. C. Bovik, "Large scale study of perceptual video quality," *IEEE Trans. Image Process.*, submitted.