Assessing Internet Video Quality Using Crowdsourcing

Óscar Figuerola Salas, Velibor Adzic, Hari Kalva
Florida Atlantic University
777 Glades Road
Boca Raton, FL 33431, United States
{ofiguer3, vadzic, hkalva}@fau.edu

Akash Shah
Nirma University Institute of Technology
Sarkhej-Gandhinagar Highway, Post Chandlodia
Ahmedabad, Gujarat, India
10bce086@nirmauni.ac.in

ABSTRACT
In this paper, we present a subjective video quality evaluation system that has been integrated with different crowdsourcing platforms. We try to evaluate the feasibility of replacing the time consuming and expensive traditional tests with a faster and less expensive crowdsourcing alternative. CrowdFlower and Amazon's Mechanical Turk were used as the crowdsourcing platforms to collect data. The data was compared with the formal subjective tests conducted by MPEG as part of the video standardization process, as well as with previous results from a study we ran at the university level. High quality compressed videos with known Mean Opinion Score (MOS) are used as references instead of the original lossless videos in order to overcome intrinsic bandwidth limitations. The bitrates chosen for the experiment were selected targeting Internet use, since this is the environment in which users were going to be evaluating the videos. Evaluations showed that the results are consistent with formal subjective evaluation scores, and can be reproduced across different crowds with low variability, which makes this type of test setting very promising.

Categories and Subject Descriptors
H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *evaluation/methodology, video*

General Terms
Measurement, Performance, Design, Reliability, Experimentation, Human Factors, Standardization.

Keywords
Crowdsourcing, subjective quality, quality assessment, Internet video quality, mean opinion score, MOS.

1. INTRODUCTION
Quality of video is commonly measured using objective metrics that give a measure of distortion in video frames. Commonly used objective metrics such as PSNR, SSIM, and VQM give a measure of how faithfully an encoder can represent the video pixels being encoded. The main downside to using such objective metrics is that, by focusing on pixel reproduction, and not on the perceived quality, we are likely encoding and transmitting video at a higher bitrate than is necessary.

In spite of all the advances in video quality evaluation and development of metrics that model perceived video quality, the most reliable quality metric that reflects user experience is still subjective evaluation. In subjective quality evaluation, subjects watch a video and rate its quality on a numeric scale. Subjective quality evaluation methods such as the ITU Bt.500 standard [15] are widely accepted, but their use is limited. The cost of setting up a video evaluation lab is just one factor that contributes to limited use of Bt.500. Recruiting subjects for quality evaluation is a difficult and time consuming task. During the course of developing and optimizing video communication systems, quality evaluations have to be performed a number of times to study the impact of algorithmic optimizations and content dependencies. Convening a group of subjects to evaluate every algorithmic optimization is very difficult. Even if we have willing subjects, they might develop biases and expectations as they repeat these evaluation sessions many times. Because of these complexities in conducting subjective evaluations, the video community today relies on objective metrics for algorithmic and system optimizations. In order to make subjective quality evaluation a viable alternative, we have to find reliable and scalable solutions. Crowdsourcing has the potential to transform video quality evaluation by enabling fast and low-cost evaluations. The concept of crowdsourcing consists on requesting services from a large group of people in exchange of small amounts of money, usually referred to as micropayments. There are several web platforms that provide easy access to large communities of online users, normally through some kind of task builder interface that allows posting jobs for these users.

In this paper we present a system for subjective quality evaluations on a large scale with preliminary results of a validation study currently underway. This work focuses on video coded at Internet bitrates (less than 1 Mbps). The coding format (AVC/H.264), bitrates (up to 784 Kbps), and resolutions (720p) selected are the most commonly employed for video services over the Internet today. The main goal is to determine if crowdsourced video quality evaluations produce consistent and reliable results. Another goal of this work is to understand whether video quality evaluated under ideal test conditions is reflective of the quality experienced by users under normal everyday conditions, in which they usually consume multimedia content.

2. RELATED WORK
Crowdsourcing has been employed for a variety of tasks and experiments: transcription of spoken language [11], content

Conference'10, Month 1–2, 2010, City, State, Country.

labeling for cyber bullying detection on the Internet [16], gathering parking availability information [5], evaluation of knowledge acquisition [3], and many others.

In the multimedia domain, crowdsourcing became popular for the tasks of image annotation [13][14][22], and video summarization [21][24][23]. Other applications of crowdsourcing for video tasks include manual geo-location tagging of the video sequences [4], evaluation of the privacy filters applied in video surveillance sequences [9], boredom prediction of Internet video [19], gesture annotation [20], and nutritional analysis of photographed food [12].

Recently, there has been increased interest in conducting image quality evaluations using crowdsourcing platforms. Ribeiro et al. developed a metric called "CrowdMOS" [17], a crowdsourcing counterpart of the Mean Opinion Score (MOS) that is used in classic subjective evaluations. CrowdMOS is used for subjective evaluation of image quality. Experiments were conducted with Mechanical Turk, and a total of 34 workers participated in the task. Authors obtained good correlation with lab methods evaluated on the LIVE dataset [18]. Evaluation on the same dataset with similar results was conducted by Xu et al. using the "HodgeRank" method for image comparison [25]. However, this method uses pairwise comparison that differs from most common standardization recommendations.

For video quality evaluation, Keimel et al. summarized some of the challenges [7], and presented their system called QualityCrowd [8]. There are conceptual, technical, motivational, and reliability challenges linked to crowdsourcing platforms. Conceptual challenges arise from the differences between the basic concepts of crowdsourcing and the structure of subjective tests, for example, in crowdsourcing the tasks are supposed to be small so that they can be done easily and fast by the workers, whereas subjective tests are usually longer. Among the technical challenges, the most important are the setup of the testing environment, which can no longer be controlled, and the delivery of the video content via the Internet. There are also motivational challenges such as the minimum wage a worker is willing to accept. Lastly, the reliability of the results has to be controlled by both rejecting invalid input in the crowdsourcing platform, and removing outliers. QualityCrowd uses a hybrid approach in which either Adobe Flash Player or the HTML-5 video tag is used depending on the participant's browser capabilities. The videos used for the tests were encoded using H.264/AVC with the High 4:4:4 Profile, which supports lossless compression. The reported tests used 19 on-campus users connected to a high speed network. The authors reported that the results correlated well with a previous subjective evaluation study of video with packet losses conducted with 40 participants.

Evaluation of video quality under packet loss is a different problem, and doesn't necessarily validate quality evaluation using crowdsourcing. The main goal of our work is to answer the validity question definitively, and develop a scalable system that can be deployed easily. To address this, we have developed a system using HTML5 and JavaScript that can run on modern browsers. To conduct a validation study, we use the AVC/H.264 anchor bit streams used in the development of HEVC. As a part of standards development, extensive subjective evaluations were conducted by the joint ITU and ISO committee, and the subjective results are reported in [1]. Comparing the crowdsourced subjective evaluations with the results of formal MPEG evaluations should begin to provide a definitive answer to the validity question.

The platform Quadrant of Euphoria [2] also uses crowdsourcing for evaluation of multimedia content, but it is based on paired comparison instead of MOS. This approach tries to overcome the inherent problems of MOS scales. One problem is that each user can interpret the scale differently according to its own judgment. Also, the cognitive distance between MOS options may also not be the same for all the values: good (4) and excellent (5) are closer than bad (1) and poor (2).

Another video quality evaluation tool is Tally [6], a web based tool that can be used for continuous video quality evaluation. In this tool, videos are uploaded to the main server. At the time of the survey, subject scoring is done over a network, decoupling the voting control from the media player. The media is displayed on a TV or monitor, while voting is done through a web-enabled device such as a smartphone, tablet, laptop, or desktop computer. Results are stored in the main server, and can be downloaded for further use. Since each person has its own personal account on the website, many people can use the same system and have their individual history and data.

3. SYSTEM DESCRIPTION

The subjective video quality evaluation system presented in this paper is based on an HTML5 web-based tool that collects ratings of videos encoded at different bitrates compared to a reference video. Even though the user's environment cannot be controlled in the same way as in traditional subjective quality evaluation tests, some constraints must still be met. Currently, browser support is limited to Google Chrome's desktop version, which was selected because of its widespread adoption, good support of standards, and ability to play all videos with anchor profiles. The tests are run in full screen mode with a neutral black background in order to minimize distractions. To make sure the browser does not resize the videos, the subject's screen size is queried, and only the videos with resolutions smaller than the screen size are presented to the user for evaluation. Lastly, in order to ensure the videos are played continuously during the test, they are preloaded before playback.

Before the test starts subjects are presented with a set of instructions explaining how to proceed, and how to rate. The user also has to complete a survey, intended mainly for data collection about her viewing and gaming habits, including amount of video watched on TV and online, and the amount of time spent playing games. Upon completion of the survey, the user is allowed to continue to the actual test. The basic structure of the test is shown in Figure 1. Each test session consists of a playlist of 10 randomly selected pairs of videos, and takes about 5 minutes to complete. The video sequences are 10 seconds long, and the pairs are composed of the reference video for that sequence (labeled as A), and the same sequence encoded at a lower bitrate (labeled as B). After every pair of videos, a scale with values that go from 0 to 10 is shown so users can vote on the quality of the encoded video compared to the reference video.

Figure 1. Structure of the test

3.1 Methodology

The methodology used for the test is a variation of the Double Stimulus Impairment Scale (DSIS) using a scale of 0 to 10, instead of an "impairment" quality scale. The decision for this kind of change was made in order to be able to compare our results with the ones in the report from JCT-VC [1]. The video sequences used for the tests are the same 11 sequences from Classes C, D, and E used in [1]. Each of the sequences is encoded using AVC/H.264 High Profile at 3 different bitrate points R1, R2, and R3 kbps as shown in Table 1. In fact, we used the same AVC/H.264 bitstreams encoded and distributed by the MPEG committee for HEVC evaluations, referred to as beta-anchors [1]. Even though the beta-anchors were encoded at five bitrates, the evaluations were limited to three bitrates for two reasons: not all the subjective evaluation scores were publicly available, and these bitrates match with the quality that is used to deliver video content over the Internet. The highest bitrate (R5) anchor is used as reference, since lossless video cannot be delivered fast enough to conduct the tests online. The quality of the reference video is very high, thus any coding artifacts are largely imperceptible.

Table 1: Sequences Used in the Study

Class	ID	Name	FPS	Res.	R1	R2	R3	R5
C	S08	BasketballDrill	50	832x 480p	384	512	784	2000
	S09	BQMall	60					
	S10	PartyScene	50					
	S11	RaceHorses	30					
D	S12	BasketballPass	50	416x 240p	256	384	512	1500
	S13	BQSquare	60					
	S14	BlowingBubbles	50					
	S15	RaceHorses	30					
E	S16	Vidyo1	60	1280x 720p				
	S17	Vidyo2	60					
	S18	Vidyo3	60					

3.2 Crowdsourcing Platform Integration

The platforms selected for the experiment were CrowdFlower, and Amazon's Mechanical Turk. The former was selected in order to use the credit awarded to the idea that was presented to the crowdsourcing competition organized by the ACM's Workshop on Crowdsourcing for Multimedia (CrowdMM) in 2012. The later is one of the most popular crowdsourcing platforms nowadays. Instead of providing its own workers, CrowdFlower has been integrated with several external sources of workers, which are called channels. Mechanical Turk on the other hand provides its own crowd, and is also available as a channel through CrowdFlower.

Using the previously described platform, job batches were set up in CrowdFlower in order to collect ratings from the crowd in exchange of small amounts of money. Following the guidelines presented in [7], workers were always paid a wage over $1.38/h. A series of questions about the content of the videos were prepared in order to filter out possible missuses. Participants were presented with 4 questions selected randomly by CrowdFlower from the given set. One of the questions determines whether the worker gets paid or not (known as gold question on CrowdFlower), the worker has to answer this question correctly or otherwise it is rejected and does not receive any reward.

The flow of the task can be seen in Figure 2. Participants log in with their account on one of the external crowd channels, from where they can select our task. They are redirected to our evaluation system to do the job from a link to the video quality survey posted in the job's description. Subjects complete the survey, and then go back to CrowdFlower to answer questions based on the content of the videos they have seen.

Although Mechanical Turk was included as one of the channels of workers in the CrowdFlower tests, we also published a different set of jobs using this second platform, trying to compare the suitability of both for this type of tests. In this case, workers were presented with a small questionnaire of 4 questions at the end of the test, and they were given a completion code. The main difference with the previously described platform is that Mechanical Turk allows to the approve payments manually, thus being able to check the completion code and accuracy of the answers given before proceeding with the payment. Workers who did not answer correctly at least 3 questions were rejected. This gave us more control over what results were accepted, making a more efficient use of our resources.

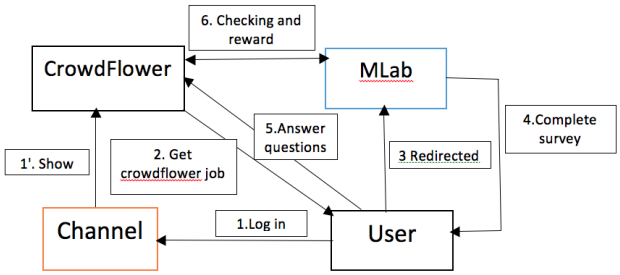

Figure 2. CrowdFlower integration

4. RESULTS

A total of 137 users participated in our experiment over a period of one week through CrowdFlower and Amazon's Mechanical Turk, and 1096 valid ratings were collected. The age of the participants ranged between 14 and 70 years old. The ratio of female and male subjects was evenly distributed, with 53% of male, and 47% of female users. Out of the 137 participants, only 20% claimed to be experts in video coding, processing, or production. The minimum display resolution used by the participants to run the test was 800x600, and the maximum 2560x1440 (see Figure 3). The number of votes collected for each individual sequence is showed in Figure 4, where can be observed that the videos with higher resolution have a lower number of ratings, this is due to the fact that around 25% of the people have displays with smaller resolution, thus these videos were not included in those tests.

In the initial questionnaire we also collected data about the user's video viewing and gaming habits. Participants reported to spend watching video content per week an average of 12.5 hours on TV, 11 hours on a computer, and 3.5 hours on a tablet or mobile device, being movies and TV shows the preferred type of content.

Regarding the gaming habits, subjects spend per week an average of 8 hours playing video games on a computer, and 3.5 hours on a tablet or mobile device, being strategy and action the two preferred game genres. There is evidence to suggest that game play, especially action games, changes perception [10]. The influence of game play and other media consumption on quality evaluations will be studied in the follow up work.

Figure 5 shows the average MOS values for all the 11 videos at three different bitrates obtained from the crowd. The MOS values increase monotonically with bitrate and confirms that the crowd was able to consistently see the increase in quality with bitrate. The relative increase in MOS value varies for the same increase in bitrate because of content dependencies.

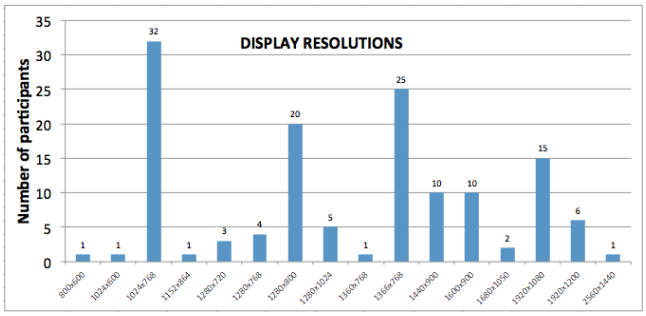

Figure 3. Variation in display resolution of the participants

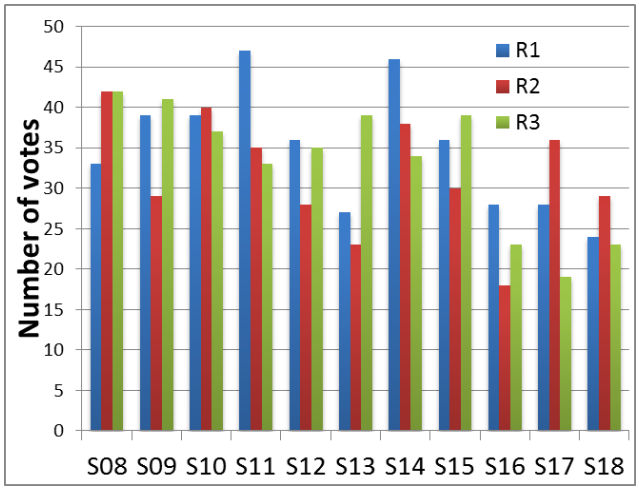

Figure 4. Number of evaluations for each video sequence tested

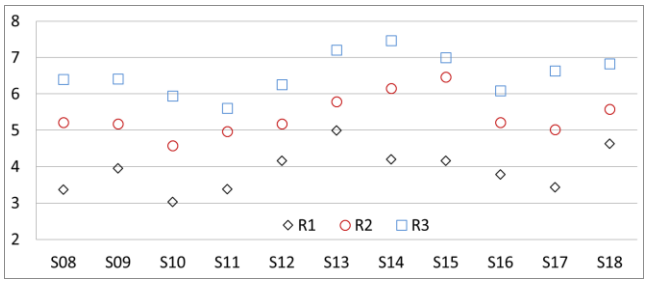

Figure 5. Average MOS values (y-axis) vs sequence number (x-axis) reported by the study

4.1 Volunteer Crowds

A subjective evaluation study was also conducted at the university level over a period of two months. This study was conducted by seeking volunteers to take subjective evaluation tests online. The study was conducted by first distributing an email announcement to students in the college of Engineering. The student body is made up of 15% female students. The student body in the College has 2200 students and is diverse with 46% White, 14% Blacks, 24% Hispanics, 6% Asian, and 6% international. The participants were given an option to enter in a gift card raffle as an incentive to participate. A total of 493 evaluations from 54 participants (2.5% response rate) were collected, with age ranging from 18 to 67. The lower response rate is due to two reasons: 1) the evaluation period started during the final weeks of the semester, and 2) incomplete evaluations were not included (e.g., when participants stop before completing the session). Using IP addresses we were able to determine that 43% of the evaluations reported were from within the university campus, and 57% participated in the tests from an off campus location. The test participants had a minimum display resolution of 800x600, and 90% had a display size of 1280x768 or higher.

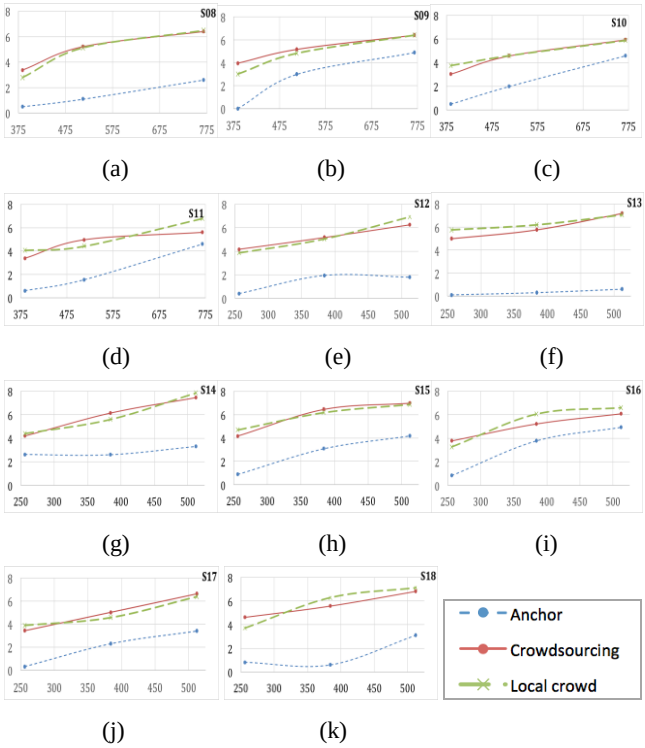

Figure 6. MOS values (y-axis) vs bitrate (x-axis) of crowdsourced and beta-anchors for bitrates R1, R2, and R3

4.2 Comparison between Paid and Volunteer Crowds

The results of both experiments are showed in Figure 6 along with the values of the beta-anchors. The values of the crowdsourcing data are very similar, indicating that results are consistent and reproducible across different crowds – paid and volunteer.

The collected ratings do not match the anchor values, being in all cases greater. This tendency of the users to rate the videos higher can be explained by the difference between the evaluation environments in crowdsourced and lab-based tests. While formal quality testing uses ideal lighting conditions, high quality monitors, and even recommended viewing distance, the crowdsourcing environment is uncontrolled and would look like anybody's home or office. Less than ideal conditions that were common in crowdsourcing could lead to subjects overlooking small compression artifacts, and hence higher MOS ratings. Other factors that have influence on MOS ratings are the lack of prior training on video quality assessment, and also the fact that the reference videos (bitrate R5) are of lower quality than the original reference videos used in the MPEG's evaluations. Although the results are not always shifted by the same value when compared to the anchor values, on average they are shifted by 3.2 points. It is worth to mention that this value exactly matches the difference between MOS values of anchors used for crowd tests and lab tests. Further studies are needed to understand the significance of this relationship and whether this relationship can be generalized. The relationship is expressed below:

Average MOS(crowd) – Average MOS(lab) =

= Average MOS(lab ref) - Average MOS(crowd ref)

This could mean that on average the values collected from the crowd correlate consistently, even though individual cases may present different trends due to content dependency, i.e. motion, background, etc. It is important to note also that the perceived video quality of Internet users may be higher than the one set by traditional tests, thus video providers may be offering videos at higher bitrates than necessary.

5. CONCLUSION

We presented a system and methodology for crowdsourcing subjective video quality evaluations. The results of the study were compared against the formal and thorough evaluations conducted as a part of MPEG standardization. The evaluations were conducted for video bitrates that are typical for Internet video services. The results show that subjects watching videos under normal conditions are more tolerant to coding artifacts than subjects evaluating videos under ideal conditions in test labs. The study presented was conducted using a volunteer crowd made up of university students, as well as using paid crowds from multiple channels on CrowdFlower. We found that collecting results from crowdsourcing platforms is faster and reliable. It took significantly longer to get the evaluation results from volunteer crowds. We saw no significant difference between evaluations by volunteer and paid crowds. The comparative evaluations show that crowdsourcing can be a reliable tool for subjective evaluations. Further data collection is necessary to understand the offset in the results obtained. The system developed uses standard HTML5 and JavaScript, and runs on mobile devices with minimal changes.

6. ACKNOWLEDGMENTS

This work was partly conducted using the credit awarded to this idea by the ACM Workshop on Crowdsourcing for Multimedia 2012 committee as part of its Crowdsourcing Competition.

7. REFERENCES

[1] Baroncini, V. et al. 2010. Report of subjective test results of responses to the joint call for proposals (cfp) on video coding technology for high efficiency video coding (HEVC). *JCT-VC document JCTVC-A204, Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG*. 16, (2010).

[2] Chen, K.-T. et al. 2010. Quadrant of euphoria: a crowdsourcing platform for QoE assessment. *IEEE Network*. 24, 2 (2010), 28–35.

[3] Gordon, J. et al. 2010. Evaluation of commonsense knowledge with Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (Stroudsburg, PA, USA, 2010), 159–162.

[4] Gottlieb, L. et al. 2012. 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, 2012), 23–28.

[5] Hoh, B. et al. 2012. TruCentive: A game-theoretic incentive platform for trustworthy mobile crowdsourcing parking services. *2012 15th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (2012), 160–166.

[6] Jain, A. and Bal, T.N. 2013. TALLY: A Web-Based Subjective Testing Tool. *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)* (2013).

[7] Keimel, C. et al. 2012. Challenges in crowd-based video quality assessment. *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)* (2012), 13–18.

[8] Keimel, C. et al. 2012. QualityCrowd - A framework for crowd-based quality evaluation. *Picture Coding Symposium (PCS), 2012* (2012), 245–248.

[9] Korshunov, P. et al. 2012. Crowdsourcing approach for evaluation of privacy filters in video surveillance. *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia* (New York, NY, USA, 2012), 35–40.

[10] Li, R. et al. 2009. Enhancing the contrast sensitivity function through action video game training. *Nature Neuroscience*. 12, 5 (May. 2009), 549–551.

[11] Marge, M. et al. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (2010), 5270–5273.

[12] Noronha, J. et al. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. *Proceedings of the 24th annual ACM symposium on User interface software and technology* (New York, NY, USA, 2011), 1–12.

[13] Nowak, S. and Rüger, S. 2010. 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, 2010), 557–566.

[14] Rashtchian, C. et al. 2010. 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, 2010), 139–147.

[15] Recommendation, I. 2002. 500-11, Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union, Geneva, Switzerland*. (2002).

[16] Reynolds, K. et al. 2011. Using Machine Learning to Detect Cyberbullying. *2011 10th International Conference on Machine Learning and Applications and Workshops (ICMLA)* (2011), 241–244.

[17] Ribeiro, F. et al. 2011. CROWDMOS: An approach for crowdsourcing mean opinion score studies. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2011), 2416–2419.

[18] Sheikh, H.R. et al. 2005. *LIVE image quality assessment database release 2*.

[19] Soleymani, M. and Larson, M. 2010. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010)* (2010), 4–8.

[20] Spiro, I. et al. 2010. Hands by hand: Crowd-sourced motion tracking for gesture annotation. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2010), 17–24.

[21] Steiner, T. et al. 2011. Crowdsourcing event detection in YouTube video. (2011).

[22] Su, H. et al. 2012. Crowdsourcing Annotations for Visual Object Detection. *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012).

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

[24] Wu, S.-Y. et al. 2011. Video summarization via crowdsourcing. *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2011), 1531–1536.

[25] Xu, Q. et al. 2012. Online crowdsourcing subjective image quality assessment. *Proceedings of the 20th ACM international conference on Multimedia* (New York, NY, USA, 2012), 359–368.