Crowdsourcing as Self-Fulfilling Prophecy: Influence of Discarding Workers in Subjective Assessment Tasks

Michael Riegler[1], Vamsidhar Reddy Gaddam[1], Martha Larson[2], Ragnhild Eg[3], Pål Halvorsen[1], Carsten Griwodz[1]

[1]Simula Research Laboratory, Norway, and University of Oslo, Norway
[2]Delft University of Technology, Netherlands, and Radboud University Nijmegen, Netherlands
[3]Westerdals Oslo School of Arts, Communication and Technology, Norway

***Abstract*—Crowdsourcing has established itself as a powerful tool for multimedia researchers, and is commonly used to collect human input for various purposes. It is also a fairly widespread practice to control the contributions of users based on the quality of their input. This paper points to the fact that applying this practice in subjective assessment tasks may lead to an undesired, negative outcome. We present a crowdsourcing experiment and a discussion of the ways in which control in crowdsourcing studies can lead to a phenomenon akin to a self fulfilling prophecy. This paper is intended to trigger discussion and lead to more deeply reflective crowdsourcing practices in the multimedia context.**

I. Introduction

Crowdsourcing nowadays is a widespread practice for collecting data for multimedia research. Studies are often used to label objects in images, as in, e.g., [1]. Input quality is relatively straightforward to control for such cases by using standard techniques such as majority vote or control questions.

Recently, an increasing number of studies have focused on collecting information that requires users on the crowdsourcing platform to make a 'judgement call'. Such tasks are generally referred to as *subjective*, since what is critical is the perspective of the human observer, i.e., the subject. Subjective tasks can be considered to belong to two underlying types, (i) those more closely related to human perception (in which variance is assumed to be a result of human physiology) and (ii) those more closely related to human interpretation (in which variance is related to different world views, level of expertise, backgrounds or tastes). Tasks of type (i) ask users on the crowdsourcing platform to judge the perceived quality of video and images. These tasks are used in quality of experience (QoE) research, such as [2], [3], [4]. Tasks of type (ii) ask users to interpret the content, i.e., whether or not the clothing depicted in an image is 'fashion' [5], [6], and are currently much less commonly studied. In this paper, we carry out an experiment of type (i). Our discussion starts on the basis of this study, and encompasses both type (i) and type (ii) tasks. We focus on the ways in which the practice of discarding the work of crowdsourcing platform users based on quality control methods effectively steers study results towards a foregone conclusion. We consider this phenomena similar to a self-fulfilling prophecy, and as such, in need of careful attention.

The purpose of this paper is to raise awareness of the danger of assuming that for a given crowdsourcing task it is appropriate to throw out outliers. Instead, multimedia researchers using crowdsourcing should carefully consider whether removing outliers may be self-defeating. The use scenario in which the crowd judgements are to be used is critical in this regard. QoE researchers, for example, may remove outliers in order to focus only on common variation in human perception, as related to what is considered 'normal' physiology. Commercial companies, for example, may discard outliers since their interest is to reach largest number of consumers possible. An advertising campaign would like to focus on the color scheme that is attractive to the majority of their target audience. The minority who would prefer an alternative have little influence on profit. Here, in contrast to these examples, we are interested in cases in which outlier removal has unnoticed, unintended, and undesired interactions with the use scenario. In this paper, we will have a closer look at this problem by looking at a user perception study in the use case of quality of experience for soccer/football videos. This study is chosen because of its complexity. The study involves several factors, e.g., user's response time, general perception level and threshold of annoyance or detection, that all may have had an impact on the input provided by the users on the crowdsourcing platform. This complexity makes it particularly tempting to make heavy-handed assumptions when imposing quality control.

We developed a service where the user is provided with a full panorama video and where the user can operate a virtual pan-tilt-zoom (PTZ) camera from the panorama video. However, transfer of a full panorama video is bandwidth intensive. For this reason, we used an approach that splits the video into tiles. The system only fetches high quality tiles where the user needs them, i.e., in the region of the panorama used to extract the virtual view as figure 1 shows. We then performed a study to evaluate various tiling approaches. This study revealed the danger of discarding workers, and also gave indications

978-1-4673-8695-1/16/$31.00

that not controlling (i.e., not removing outliers) yields useful results. The main contributions of this paper are: (1) to draw attention to the problem of self-fulfilling prophecy; (2) present an experiment that reveals this problem; (3) discuss our experiences, which indicate that avoiding the exclusion of any crowdworkers minimizes the influence of the experimenter's expectations on results. The exception is crowdworkers who are clearly not seriously pursuing the instructions stated in the task (commonly referred to as 'cheaters'). As such, the paper is not a conventional paper, but is rather meant to raise awareness and lead to discussion. The remainder of the paper is organized as follows. First, we cover related work in crowdsourcing that falls within the scope of the problem. Then, we present our crowdsourcing study. After that, we present the experiments and results. We go on to detailed discussion of the results, and then, to the conclusion.

II. Problem Scope

Crowdsourcing is well known as a powerful tool for multimedia researchers to collect datasets or to support their experiments. The advantages of crowdsourcing is the large number of workers that are available, but also the diversity of the workers. This makes it not only useful for objective tasks where, e.g., workers have to annotate objects in pictures or in videos like in [7], [8], but also subjective assessment tasks can be performed on crowdsourcing platforms. When presented with a given multimedia item, humans must consider both the content of the item, but also their understanding and perception of the world. In general, humans share a common understanding of the real world, but this is also strongly influenced by the interpretation and perception of individual human beings. These interpretations can be considered as individual perspectives, which do not fit one universal solution [6], [3]. Calculating the consensus of the crowd is often interpreted as the combination of several judgements of crowdworkers combined to one high quality judgement. A great deal of effort has been put into this area, which has led to a large number of different methods and techniques. An overview and also reference implementations of these methods can be found in [9]. These related papers show us that it is possible to perform subjective tasks on crowdsourcing platforms. Another problem that comes with subjective tasks is how to determine the quality of the workers' submitted work. It is common practice to use methods like majority vote, gold standard tests, and workers reliability to for example, improve the quality of the worker. These methods work very well for objective tasks that have something that can be controlled very easily like for annotation tasks. If we want someone to label a dog, a control question with a picture that contains a cat can help to identify workers that do not perform the task correctly or did not understand it [10], [11].

However, concepts such as quality of experience, affect [12] or fashion [13] are open to different interpretations. For some people, a video with a too high resolution is not perceived as pleasing as videos that they are accustomed to. Further, arteacts in videos can be disturbing for some people, whereas others do not even notice them because the overall experience is more important for them. As long the movie does not stop and the story goes on, it is perceived positively. For this reason, all these and related tasks can be considered as subject to interpretation. These interpretations can be seen as personal or individual perspectives. Nonetheless, a large amount of agreement can exist between different humans. Moreover, it has been shown that interpretation-influenced consensus falls short of being universal [6]. The affect and fashion examples open the door for another discussion about the experimenter's influence on results, and another problem that must be considered in designing a subjective crowdsourcing campaign. The experimenter creates the language through which the crowdworkers' view on the data is filtered. This may be the terms used for rating, or the phrasing of the question asked, and is an influential factor. In our experiment, this is shown by the tasks asking workers to detect quality drops and to report annoyances in the videos. In some cases, we asked one question, but could interpret the crowdworker input as responding to the other.

Our contribution on subjective assessment crowdsourcing tasks comes to a different conclusion than other works for which the common understanding of the world is not individual enough. In these scenarios, subjectivity is not a factor that has to be considered, which makes them much easier to solve. For example, [14] points out that the crowdworkers are able to reproduce results close to ones generated by experts in the laboratory. Experts who consult with each other are shown to diverge from the conventional wisdom of the crowd, which leads to a loss of information and a result closer to what researchers would like to see at the end. Other related papers like [4], [15], [2], [16] try to control and limit subjective tasks conducted by crowdworkers very strict with mathematical methods such as removing outliers. This may indeed lead to results that are finally closer to the hypothesis that researchers want to prove. However, it is important to take some aspects into consideration to avoid a negative influence on the results. If we design the control mechanism, can we be sure that we do not get a self fulfilling prophecy [17], [3] at the end? Because the crowdworkers are so many, the chance is possibly high that we will find a subset of workers that agree on the researchers subjective interpretation of the world for a certain topic, e.g., which quality for a video is good. Paying them for the work potentially exacerbates this problems since workers are incentivized to try to find out what researchers expect and fulfill these expectations as well as possible so as not to risk their payment. Apart from that, we have to ask ourselves the questions: (i) Is it a good idea to remove people that do not agree with the majority? We cannot really say that people are doing incorrect work because we do not really know how they perceive the world and which experiences they have had in their lives. And, we even do not take age or cultural differences into account. (ii) Is an algorithm that reflects the quality needs of a specific cultural or life-experience-influenced subset of individuals really helpful, or is it just playing around to find some correlations to proof a certain hypothesis?

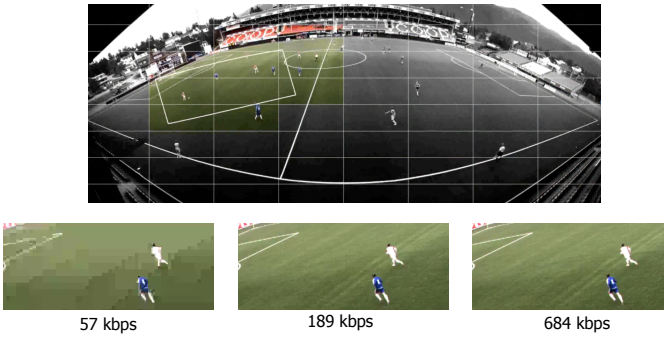


Fig. 1. Tiles in different quality.

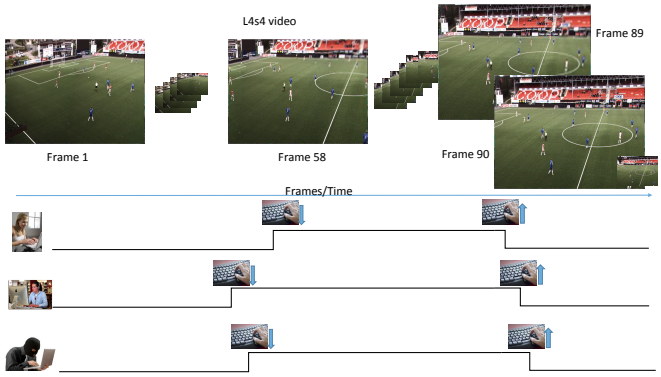


Fig. 2. Design of the crowdsourcing task. Frame 1 shows a scene with good quality. Frame 58 to 89 contain bad quality sequences. The workers press and hold a key on the keyboard to report when they perceive bad quality or when they are disturbed by video artefacts. If they think that the quality increased they release the key.

Finally, another important contributor to self-fulfilling prophecy is exactly *how* the questions are asked. This is well researched in the marketing field [18]. In this paper, we want to raise the question of whether controlling crowdsourcing for subjective assessment tasks is the right way to go, and whether it is even possible without ending up with a limited subview of a problem.

III. Crowdsourcing Task

The crowdsourcing campaign described here was originally performed to answer another question, namely the best tile delivery strategy for soccer videos. In the course of this study, we decided to investigate the influence of controlling workers' contributions. Recalling the original motivation, the use case is to test a service where the users are provided with a full panorama video, and can operate a virtual PTZ camera from the panorama video. Since the transfer of a full panorama video is bandwidth intensive, we developed an approach to split the video into tiles and fetch only high quality tiles where the user really needs them (1). The workers were instructed to report how changes in the quality of the video and disturbance by artefacts influenced their quality of experience while they where watching a soccer video. In order to find out how workers perform under different filter methods, we compared the results of objective metrics with subjective evaluations across a range of tiling approaches. The initial user study was designed to investigate two aspects of the subjective perception of quality. We consider the noticeability of quality distortions and the experienced annoyance related, yet distinct. We ran two consecutive experiments, one to address the detection of tiling distortions, and one to address the annoyance resulting from the distortions. In addition, we included five-point absolute category rating scales for subjective assessments of overall video quality, adhering to ITU-T P.911 [19]. Figure 1 shows an example frame of a panorama, a virtual view and the tiles that are required in high quality at that instant. The problem, however, is that, the tiles are encoded in time segments (typically 1s to 3s in length) which implies that one cannot change quality of a tile during the segment. This some times creates output frames, in which the user may see some parts fetched from poor quality tiles. Once the user moves the virtual camera to the new region, the corresponding tiles are fetched in high quality in the immediately following segment. One can observe that there are two kinds of visual experiences associated to such a process, there is a possibility of seeing a steadily degrading quality as the user pans/tilts/zooms into poor quality tiles and then the abrupt jump from poor quality to higher quality due to the change of segment during the operation. In both experiments, participants watched sequences with a duration of 12 seconds extracted from soccer game videos. These were originally chosen as representations of different football scenarios and hence provided variety of stimuli to participants, which served to increase generality. Since all sequences included pre-recorded camera panning and zooming movements, our final stimulus collection contained sequences with frequent tile shifts and varying changes in compression rate and video quality. Figure 2 gives an overview and description of the task and its design, including some example frames from videos shown to the workers.

In the detection experiment, we instructed participants to pay attention to the the quality of the presented sequences and to push down the spacebar the moment they noticed a change for the worse, holding it down for the entire duration of the quality drop. The annoyance experiment followed the same procedure, only changing the instructions to ask participants to push down the spacebar while they experienced annoyance due to low video quality. At the end of each sequence, participants rated the overall video quality on a 5-point scale ranging from "bad" to "excellent". Before the participants started the experiment, they had the chance to get familiar with the task and the mechanics by training with two short test videos. The reason we opted for crowdsourcing was to get a sufficient number of participants. It is important to point out, that we collected all work submitted, and we did not apply any filtering beforehand. This was necessary to be able to show how different filtering would lead to different strong correlations between the automatic metrics and the crowdsourcing metric, and therefore how filtering of the submitted work would support our hypothesis that the automatic measurements can model the quality of experience of humans. This approach requires some extra methodological

Exeperiment	Raw	F_1	F_2	F_3
AE1	112	87	108	65
AE2	134	96	127	37
DE1	110	83	108	51
DE2	132	98	126	20

TABLE I
NUMBER OF INLIERS BASED ON DIFFERENT FILTERS. WE APPLIED THREE DIFFERENT FILTERS IN OUR EXPERIMENT.

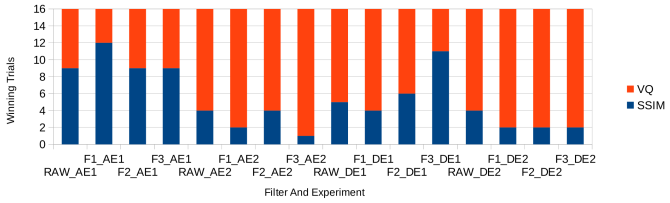


Fig. 3. Effect of filtering in various experiments.

considerations due to challenges that concern lack of task adherence and comprehension, and in turn, reduced data consistency [20], [21]. Thus, we initially conducted 3 pilot studies to ensure that the experiments were presented in a succinct but understandable format. The first was completed by colleagues and students, the following two on crowdsourcing platforms Microworkers and Crowdflower. Following each pilot, we adapted the experiments according to the feedback we received. For the final study, we used Microworkers and collected data from a total of 200 participants. We paid $\$1.50$ per task which is more than the standard amount for the time that workers spend on our task (a single worker needed less than 15 minutes to complete the experiment). Although we implemented quality measures such as gold samples and majority votes, we decided that the highly subjective nature of the task was better solved without any filtering at all. We excluded only participants who failed to complete the experiment, and on manual inspection we removed participants who had obviously tried to circumvent the experimental tasks, altogether 15%. All other possible exclusion criteria were found to potentially exclude valid human perceptions as well.

IV. EXPERIMENTS

During the study, each user was presented with 16 videos, and their time response was recorded along with the overall perceived quality for each video. This provides us with plenty of data. Table I presents the number of users in total along with the inliers using different filtering methods. We ran two experiments per question. The annoyance experiments are termed AE1 and AE2, and the detection experiments are termed DE1 and DE2. We used three filters F_1, F_2 and F_3. The set of subjects that pass through various filters are *inliers* in our case.

Table I: No response filter (F_1). In this filter, we try to eliminate the users that have not recorded any time-response on at least 14 trials out of the 16 trials. This is a rather relaxed constraint, as this demands the user to provide some input on at least 2 trials. This is reflected in the number of inliers from table I, who sum to atleast 80% of the total users.

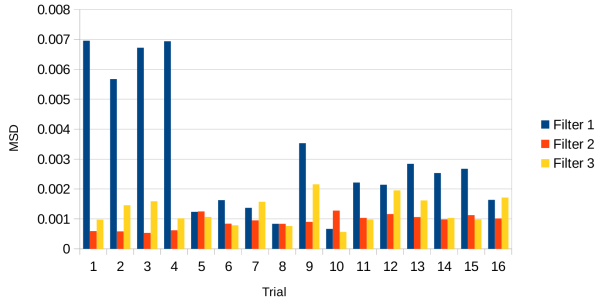
Table I: Random response filter (F_2). This filter tries to detect extremely short presses (at least 10 presses of 30 ms in a single trial) and extremely long presses (10 s). We have empirically chosen the values to relax the filter significantly. Moreover, we remove user's input only if the user has done one of these for at least 7 trials.

Table I: Inter user agreement (F_3). This filter keeps only the set of users that agree with each other. This is a rather strict requirement and we use quality ratings, asked to the users after each video sequence, for computing the agreement. The set of inliers implies that each user has to have at least 40% agreement in ratings with all the users in the set. The number of inliers demonstrate the rigidity of this filter.

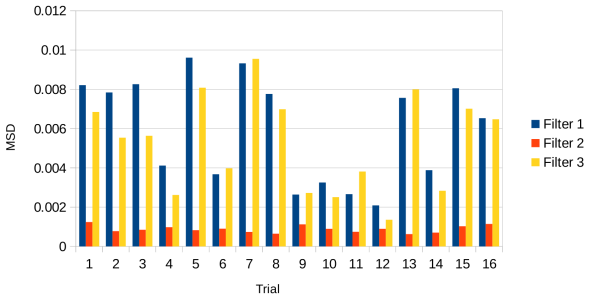
Performing user studies like these for evaluating tiling strategies is expensive and cumbersome due to the complex design. Hence, we decided to study the divergence between various commonly used objective evaluation metrics like PSNR [22], SSIM [23] and OpenVQ. OpenVQ is our independent open-source implementation of the ITU-T J.247 Annex B [24]. Results from the experiments show, that SSIM and OpenVQ diverge least from the user input. Figure 3 presents the number of times the metric is closest to the user provided data out of 16 trials and also the effect of filters on the outcome. It can be observed that the effect is quite random. For example, considering the effect of F_3. In experiment DE1, it leans more towards SSIM, and in experiments AE2 and DE2, it leans more towards OpenVQ. However, in experiment AE1, it has no effect. This behaviour can be observed from other filters as well. In summary, simply by modifying the filtering mechanism, we are able to suggest that one metric is performing slightly better than the other approach in subjective evaluation.

Another interesting aspect that we observed is the effect of filtering on the data itself. Figure 4 presents the Mean Squared Differences (MSD) of the filtered data with raw data for each trial. In both experiments, F_2 filters data and still remain close to the raw data. However, in experiment 1 (figure 4(a)), we can see that filter F_3 is closer to the raw data and farther away from F_1. The exact opposite behaviour can be observed from experiment 2 (figure 4(b)). It must be noted, that even though it is not necessarily required to have a similar profile to get similar MSD values, we have observed that in this data they actually have similar profiles. Even though the filters seem sensible to apply to the data, one can definitely raise human-centric arguments against them. For example, there is a definite chance that the individual did not get annoyed during any of the trial. Filter F_1 removes all such individuals. Similar arguments can be brought up for filters F_2 and F_3 as well.

These arguments play an important role in the filtering of data and thus on the analysis of the subjective study. To get an idea about how strong users agree in the not filtered data, we also calculated the agreement on quality ratings for each sequence using the Fleiss Kappa statistic [25]. Fleiss Kappa depends on the number of participants for each ques-



(a) Experiment 1



(b) Experiment 2

Fig. 4. Mean Square Difference (MSD) of signals using various filters with the raw signal.

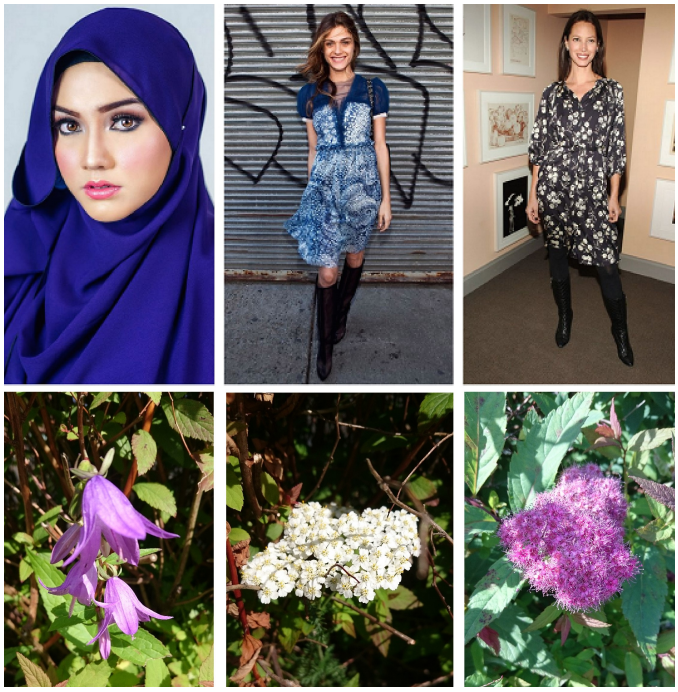


Fig. 5. Examples of subjective responses that can be collected from the crowd. Top row: Does the image depict fashion? Bottom row: Which flowers are most beautiful?

tion asked [26]. Therefore, we had to calculate the possible minimum and maximum values for our study. These values are -0.80 for the minimum and 1 for the maximum. Fleiss Kappa for the annoyance of workers lies in average between 0.24 and 0.39. Applied on the detection part of the study the inter-worker agreement varies between 0.22 and 0.37 on average across all sequences and quality conditions. This is a strong indication that not filtering the data still leads to a result where workers agree on a good level amongst each other. Which further implies, that the non filtered data is not merely noise and therefore usable to determine user preferences without getting influenced by cheating or rushing workers or only relaying on strict filtering methods that might have a self fulfilling prophecy as an outcome.

V. Discussion

Figure 5 shows two examples of subjective tasks that are related to human interpretation (i.e., type (ii) tasks discussed in Section I) for which variance is related to different world views, level of expertise, backgrounds or tastes). The top row contains images that were used in a task asking workers whether the image depicts fashion [5]. The second row of images depicts a hypothetical task on multimedia aesthetics, which asks workers to judge the beauty of the flowers. This is also a subjective task since colors, form and the type of the flower, and if workers even know the flower can influence the decision. What these examples also show is that, one can not easily find a valid reason to discard workers because they do not find a certain image fashionable, or a flower not beautiful. Moreover, it is quite certain that a sub-majority of workers can be found that agrees on one aspect. However, there is no principled way to chose a sub-majority that would best represents how humans in general perceive or interpret images.

The field of human perception and its measurement faces a perennial struggle to identify noise. However, calling data that does not fit a simple fit-for-all model ‘noise’, is a controversial practice [27]. This practice should be questioned critically, especially in subjective assessment tasks. Moreover, even with a basic model, we have seen that the filtering approaches do not necessarily provide a consistent output. This property of general, but inconsistent, filtering can be tuned to obtain misleading results.

Since the design of the outlier detection methodology lies in the hands of the researchers who want to prove a hypothesis by carrying out a crowdsourcing study, a conflict of interest arises. Intended or not, such situations can lead to a self fulfilling prophecy. Defining the control mechanism based on the experience and view of a few people working in a laboratory leads to the injection of the researchers’ opinion and favored outcome into the study. Since crowdsourcing platforms give access to an almost unlimited number of users, it is easy to find a subset that agrees with the opinion of the researcher. The opinion of the researcher can easily, unwittingly, become mirrored in the design of the control mechanism. Gold standard tests and control questions written by researchers reflect researchers point of view. Without realizing it, researchers can easily steer the input provided by the crowd

Compared to crowdsourcing platforms, it is much harder to find a sizable number of people that agree with a specific opinion, if an experiment is conducted in a laboratory where study

participants must be physically present. Outlier elimination in controlled laboratory experiments comes at great cost, since usually the cost per study is so high. The high cost can serve to reduce the temptation to classify data points as outliers and remove them. Crowdsourcing is relatively less costly, and, as a result, indirectly encourages a relaxed approach to discarding outliers. The result is sometimes outlier detection approaches that eliminate a large number of data points. We must be on guard for applying methods designed to eliminate outliers in laboratory conditions to the much freer environment of a crowdsourcing platform. It is important to point out, that the self-fulfilling prophecy effect may arise in different ways for different type of tasks. The specifics of the effect depend on how human judgements are made: Fashion is cognitive, QoE is perceptual, and affect (depending on which theory) is a combination of the two.

Our experiments have suggested the potential of the solution of not applying any control methods at all to the data. Of course, this will naturally lead to results that will not appear, on the surface, as optimal. However, under other perspectives they may be more useful because they are closer to real-world patterns. In cases where experimenters do not want to eliminate control and filtering a very strong self critical view is necessary during the design process. The task design should be made careful, public available and tested by several pilots and as many participants as possible. We point to the advice of 'Trust the crowd, not ourselves' that was put forward by [6]. These measures will help ensure that the final outcome is not a self fulfilling prophecy.

VI. Conclusion

Our aim has been to raise awareness about subjective crowdsourcing tasks and the challenges related to quality control of crowdsourcing work. We have found reason to criticize the practice of eliminating some of the work that is submitted by users of the crowdsourcing platform from the final dataset that is used. This practice can lead to an undesirable result, akin to a self-fulfilling prophecy. Our crowdsourcing study was related to tiling in soccer videos. We have shown that not controlling the experiment can be a viable and useful alternative for subjective crowdsourcing tasks, enhancing the real-world value of the results.

We hope our observations will trigger broad discussion, and help other researchers in designing their crowdsourcing studies. We believe that reflective design will yield data that are more useful, and go beyond representing a subset of possible results. In the future, we would like to scale up our investigation of subjective crowdsourcing tasks, with the aim of finding more general proof that including all crowdworkers in the final data minimizes the influence of the experimenter's expectations on the results. Crowdsourcing is a powerful tool but must be used carefully, and with a healthy dose of self criticism, in order to avoid its degradation to a means to prove anything, given the the choice of quality control.

Acknowledgment

This work is founded by the FRINATEK project "EONS" (#231687) and by the EC FP7 project CrowdRec (#610594).

References

[1] G. Can, J.-M. Odobez, and D. Gatica-Perez, "Is that a jaguar?: Segmenting ancient Maya glyphs via crowdsourcing," in *Proc. of CrowdMM*. ACM, 2014.

[2] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of youtube QoE via crowdsourcing," in *Proc. of ISM*. IEEE, 2011.

[3] 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," *Proc. of MM*, 2014.

[4] T. Zinner, O. Abboud, O. Hohlfeld, T. Hossfeld, and P. Tran-Gia, "Towards QoE management for scalable video streaming," in *Proc. of MMATPQoE*, 2010.

[5] B. Loni, L. Y. Cheung, M. Riegler, A. Bozzon, M. Larson, and L. Gottlieb, "Fashion 10000: An enriched social image dataset for fashion and clothing," in *Proc. of MMSys*, 2014.

[6] M. Larson, M. Melenhorst, M. Menéndez, and P. Xu, "Using crowdsourcing to capture complexity in human interpretations of multimedia content," in *Fusion in Computer Vision*, ser. Advances in Computer Vision and Pattern Recognition, B. Ionescu, J. Benois-Pineau, T. Piatrik, and G. Quénot, Eds. Springer, 2014.

[7] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Computer Vision*, 2013.

[8] N. J. Gadgil, K. Tahboub, D. Kirsh, and E. J. Delp, "A web-based video annotation system for crowdsourcing surveillance videos," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014.

[9] A. Sheshadri and M. Lease, "SQUARE: A Benchmark for Research on Computing Crowd Consensus," in *Proc. of AAAI on HCOMP*, 2013.

[10] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds." in *Proc. of NIPS*. C.A., Inc., 2010.

[11] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua, "Assistive tagging: A survey of multimedia tagging with human-computer joint exploration," *Proc. of CS*, 2012.

[12] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. on AC*, 2012.

[13] B. Loni, L. Y. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, and M. Larson, "Fashion 10000: An enriched social image dataset for fashion and clothing," in *Proc. of MMSys*, 2014.

[14] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation," in *Proc. of MIR*, 2010.

[15] M. Riegler, P. N. Olsen, and P. Halvorsen, "Work like a bee-taking advantage of diligent crowdsourcing workers," *Proc. of MediaEval Benchmark*, 2014.

[16] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Robust evaluation for quality of experience in crowdsourcing," in *Proc. of MM*. ACM, 2013.

[17] R. K. Merton, "The self-fulfilling prophecy," *TAR*, 1948.

[18] S. Sudman and N. M. Bradburn, "Asking questions: a practical guide to questionnaire design." 1982.

[19] ITU-T, "P.911: Subjective audiovisual quality assessment methods for multimedia applications," 1998.

[20] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *Proc. of ICIP*, 2011.

[21] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proc. of CHI*, 2008.

[22] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, 2004.

[24] ITU-T, "Objective perceptual multimedia video quality measurement in the presence of a full reference," *ITU-T Recommendation J.247*, 2008.

[25] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, 1971.

[26] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: use, interpretation, and sample size requirements," *Physical therapy*, 2005.

[27] H. M. D, *Identification of Outliers*. Springer, 1980.