Crowdsourcing experiments with a video analytics system

Eirini Takoulidou, Konstantinos Chorianopoulos
Department of Informatics
Ionian University
Corfu, Greece
rinoulit@gmail.com, choko@acm.org

***Abstract*—The need for more experimental data, but also quicker and cheaper, lead us beyond traditional lab experiments, approaching a new subject pool via a crowdsourcing platform. SocialSkip is an open system that leverages the video clickstream data for extracting useful information about the video content and the viewers. The difficulty of embedding a pre-existing system as a task demands a carefully designed interface, adjusting experiments and be aware of workers' cheating behavior. We present a replicable task design and by analyzing crowdsourced results, we highlight problems in experimental procedure and propose potential solutions for future crowdsourcing experiments. The proposed crowdsourcing methodology achieved the collection of a significant amount of video clickstream data, in a timely manner and with affordable cost. Our findings indicate that future social media analytics systems should include an integrated crowdsourcing module. Further research should focus on collecting more data by controlling the random worker behavior a priori.**

Keywords—Web Video; Crowdsourcing; User Experiment

I. INTRODUCTION

Researchers in many scientific fields of computer science profit from the internet services in order to storage and analyze their experimental results. Soon enough they also realized that the internet could be a source of experiments' subjects. Giving the users the chance to participate in experiments only by visiting a webpage can be a part of their daily routine. Furthermore, it has a major impact in saving effort to prepare locally an experiment (e.g. laboratory), sometimes interpreted in time and cost savings. Internet as a medium eliminates possible difficulties in organizing the subjects; find the facilities for large scale experiments and researcher's effort, by using Web 2.0 tools.

The Games with a Purpose (GWAP) approach leveraged the strength of the users, set their through a gamified system for various purposes, for example the task of labelling images [1]. Although their great success, obviously this approach can't be proper for every type of task. Besides, the number of users participating voluntary and their characteristics depends on task's type [2], the users don't stay loyal in a game for a long time and the experimenter doesn't have total control of the target group for his experiments.

The bargains of using internet to conduct experiments are highlighted in Horton et al. [3] focused on the crowdsourcing opportunities in crowdsourcing platforms. There, researchers can publish their experiments as tasks and ask their registered users, known as workers, give their judgments about the research issue. These workers are unknown and available to respond to a call in exchange of minor payments. Researchers benefit from platforms simply by posting their experiments for data collection and/or distribute their data for evaluation. In their study, Horton et al. used a crowdsourcing platform (MTurk)[1] in order to conduct their decision-making experiments. They realized that the workforce in online marketplaces and their collective intelligence can be the solution in this kind of research problems. Moreover, the availability of a diverse and large subject pool, directs to a major challenge to grab another point of view. They also highlighted crowdsourcing's basic concern, the malicious worker behavior. Money as a motivation usually eludes "bad" workers trying to get the reward with the least effort. Nevertheless, conducting experiments for the same research field, Paolacci et al. [4] measured the worker attention and task completion comparing to traditional subject pool approaches resulting in very promising conclusions for the quality of the crowdworkers.

The crowdsourcing opportunities, under an insecure quality environment, are a challenge, not only for the needs of the crowdsourcing systems, but for the social media research fields too. Dealing with all this cost-time issues and above all the need for more data, lead us to further investigate the use of a crowdsourcing methodology in the video research.

II. RELATED WORK

A. *Crowdsourcing Video Experiments*

Recently, on the field of multimedia, researchers dealing with the issue of cost and time reduction and also the need for more data, approached the crowdsourcing trend as a methodological tool. Considering the high cost of traditional video image annotation Vondrick et al. [5], built a system for large-scale and economical video annotation experiments via MTurk. Although they managed the cost reduction and a large amount of data, they had to collect results of three years experiments in order to discover the "experts" in video annotation from the available workforce. But this expertise is not a required qualification for all the types of experiments.

Presenting a crowdsourceable framework [6] to quantify the QoE (Quality of Experience) of the video content, the

[1] Mechanical Turk (MTurk) Homepage: https://www.mturk.com/

experimenters end up in very promising results. Comparing the experimental results from laboratory part-time employees and crowdsourcing workers, they found that the participants where almost equally qualified in both subject sources. Supporting the crowdsourcing methodology, they realized that they achieved wider participant diversity, a substantial parameter when trying to understand people's different perceptions, while they managed a significant cost reduction. The crowdsourcing results from QOE experiments help to improve the video quality and the viewers' experience. But the crowdworkers' perception could be beneficial to other systems research issues, handle a better understanding to viewers' behavior and/or the video content.

The questionable trust in crowdworkers and its impact in data quality concerned both above studies. Although they deal with issues from the same research field, the different purposes demanded different quality control approaches. Vondrick et al. [5] chose the "gold standard" quality control approach, rejecting the "bad" workers that failed in annotating an already known dataset. Chen et al. [6] rejected workers that their annotations didn't fit an expected quality rate. These approaches, as categorized in Chen et al. study [7], are runtime approaches that ensure quality after workers' job submission. On the other hand, focused on design-time approaches, Wu et al. [8] crowdsourced a framework, called Click2SMRY, to generate video summaries from crowdworkers by marking the video highlights. The results compared with abstracts given from experts and algorithms, showed that users can produce satisfactory summaries for different types of video. They assumed that it could be possible the users to produce summaries while watching a video, in real conditions. This study's purposes matched our research interests, but also the tempting and low-effort (for the experimenter) quality control choices. Though, in our point of view, it might be a burden for the user to work voluntarily, as expected by the authors, while already available and effortless metadata could lead to the same, and maybe even more useful information about the video.

B. *Video Clickstream Analytics*

During the last years, the intense growth of online video services stimulated internet users to watch, create and share video material between users all over the world. Besides the common types of content being available on-demand, there is a plentiful amount of user generated content. These videos, not only approach the quality but also the great cost of making a professional video. Millions of users use every day popular platforms such as YouTube in order to browse these videos and spend hundreds of hours to watch and/or share video content. In order to implicitly extract useful information about video [9] proposed a collective intelligence method that leverages user interactions while browsing the video content. Therefore, they developed a web-based system, called SocialSkip that stores all these interactions with a customized video player. To ensure their hypotheses they conducted a lab experiment with 23 subjects, using general-purpose computers. Although the cost seams minor, the effort for the researcher to synchronize and calibrate the experiment's conditions is significant. We can't quantify this cost-effort by a metric; we only can assume that this might be feasible for a few subjects, but almost impossible for large-scale experiments demanding hundreds of subjects. A further analysis of these results, to better understanding of user video activity behavior, Chorianopoulos [10] claimed that demands large-scale experimentation. Even though we try to conduct the experiments in groups, it might be time affordable and surely increases the experimenter's effort. Moreover, a subject group composed of undergratuated students in Computer Science [9]; it's definitely not a representative sample of the general population, as the Youtube viewers. Obviously, the course credit is not a general motivation to elude subjects except students.

The purpose of our study is to take advantage of the potentials of crowdsourcing, using the services of a crowdsourcing platform as a methodological medium, in order to collect video clickstream data with a video analytics system.

III. EXPERIMENTAL METHODOLOGY

In this section, we present the video analytics system and the experimental procedure based on previous lab experiments but customized as a crowdsourcing task.

A. *SocialSkip System*

SocialSkip is a cloud based and open source experimental system[2] using the Google App Engine (GAE) for data storing and the YouTube API. This video analytics system was developed to gather users' interactions while watching a video. Each time a user enters the system's webpage, a new record is created. Users of SocialSkip are getting a unique id and their interactions with the web video player are recorded and stored in Google's database alongside with their id. The time it occurs is recorded within a second's accuracy.

The SocialSkip Service[3] provides button or seekbar browsing options. In previous experiments with SocialSkip player [9], it has been used the main functionality of a typical VCR device because of its familiarity to users. They have modified the classic forward and backward buttons to Goforward and Gobackward where the first one goes backward 30 seconds and its main purpose is to replay the last viewed seconds of the video, while the Goforward button jumps forward 30 seconds and its main purpose is to skip insignificant video segments.

The promising results of this work made us extend the basic methodology. Instead of buttons, we chose the seekbar for the users' browsing activity because we wanted to examine the random guesses as it reflects the more realistic browsing behavior in Youtube videos. Furthermore, according to Li et al. [11] observations when seek thumb is used heavily, users have to make many attempts to find the desirable section of the video and thus causing significant delays, that was a challenge for more research in user video activity. We didn't take in consideration the other interactions (Play/Pause) based on the findings of previous analysis [10].

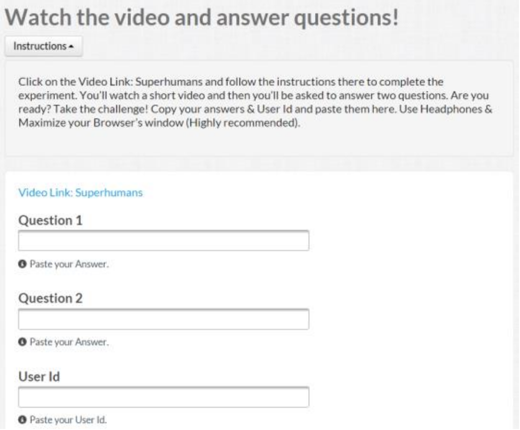
[2] SocialSkip source: https://github.com/vidoodlics/socialskip

[3] SocialSkip service: www.socialskip.org

B. *Crowdsourcing SocialSkip*

To achieve our goals we decided to publish SocialSkip experiments via a crowdsourcing platform. We chose recruiting the participants using Crowdflower[4], a meta-platform that provides quality control tools and give us access to a workforce from multiple crowdsourcing subject pools called channel-partners.

Task Design: Crowdflower offers various templates for popular tasks, but our purposes differentiate from the usual. We preferred to avoid increasing the programming effort thus we propose a simple, easily replicable task design. We informed the workers that they had to follow a link to conduct an experiment and they only had to answer some questions about a video. Considering the risk of the limited control we had on participants' engagement, task's requirement was that workers had to give as task's answers, some elements given during the experiment about the video content and also their unique id that automatically generated by SocialSkip. By this requirement, we force the workers to follow the experimental procedure, minimizing the case they would give random answers and build our metric for engagement assessment (user id). We created two tasks for two different SocialSkip experiments (Figure 1). The basic structure was exactly the same and the only difference was the link directed to a SocialSkip video experiment.



Watch the video and answer questions!

Click on the Video Link: Superhumans and follow the instructions there to complete the experiment. You'll watch a short video and then you'll be asked to answer two questions. Are you ready? Take the challenge! Copy your answers & User Id and paste them here. Use Headphones & Maximize your Browser's window (Highly recommended).

Video Link: Superhumans

Question 1

❶ Paste your Answer.

Question 2

❶ Paste your Answer.

User Id

❶ Paste your User Id.

Fig. 1. Task example as shown to the crowdworkers

Task Settings: Focused on a design-time approach; we used CrowdFlower tools for quality control. To ensure that each worker only participated once in each experiment, the same task was not available for acceptance by workers with the same IP address. Moreover, after the completion of a task, we flagged the participants to avoid their involvement in future experiments. Previous knowledge in experimental procedure would surely affect their behavior and might add some risk for random or even no answer. Because all the videos were in English, the tasks were available to workers located in certain countries where the majority is native English speaking population.

C. *SocialSkip Experiments*

The 200 workers, from various CrowdFlower partner-channels, after accepting our task directed to a different SociaSkip experiment (100 workers/ experiment). Informed by the instructions, the participants had to link with SocialSkip experiment and follow the detailed instructions of how to achieve to answer correctly the questionnaire. So we encouraged them to browse the video content using the seekbar. We carefully chose each experiment's elements that compound a different video-questionnaire set and described extensively below.

Materials: The selection of the suitable video content is an important issue for our research. We decided to meet the challenge to examine videos with as much visually unstructed content as possible, because content-based algorithms have already been successfully implemented in videos that have visually structured scene changes. The quick spread of user-generated content, including video, made us select videos that belong to this popular category. Both videos had educational information about general issues that didn't need any expertise knowledge. Another key factor was the length of a video. We had to take under consideration that crowdsourcing workers are used to take tasks that are quick and easy. Hence each video lasts approximately 4 minutes.

The educational videos had different video production techniques and they are all available on popular Youtube Channels. The first one (Edu.A), "A future with Superhumans"[5] is an Animated video that presents the opinion of an expert in robotics about the neural implants. The second (Edu.S), called "10 Misconceptions Rundown"[6] confutes well-known myths as a narrated Slide presentation.

Measurement: This measuring process is based on the assumption of Yu et al. [12] that there are segments of a video clip that are commonly interesting to most users, and users might browse the respective parts of the video in searching for answers to some interesting questions. When enough user data are available, user behavior will exhibit similar patterns even if they are not explicitly asked to answer questions.

In order to experimentally replicate user activity we developed a multiple choice questionnaire that corresponds to two segments of the video (Ground Truth). Thus, each question corresponds to an event that could be used as hint to find the answer. We used Google Docs to create online forms for users' questionnaires and we integrated these forms in our experiment interface. The questions for each video are shown in Table 1.

TABLE I QUESTIONS THAT SET THE GROUND TRUTH

Video	Indicative questions
Superhumans (Edu. A)	As an example of motor activity, two guys are playing rock-paper-scissors. Which are the 3 colors of their clothes?
	In which case would the speaker do the surgery?

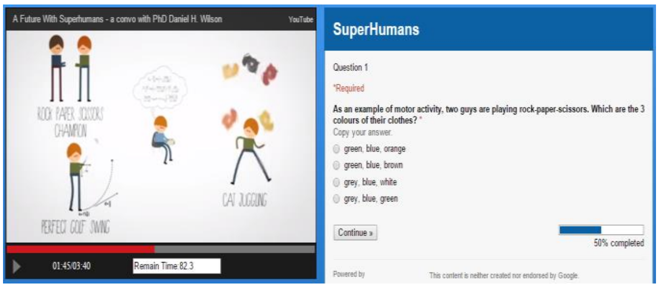
[4] Crowdflower Platform: http://www.crowdflower.com/

[5] Edu.A: https://www.youtube.com/watch?v=cgStGEDLFB4

[6] Edu. S: https://www.youtube.com/watch?v=SCzXZfNIu3A

Misconceptions (Edu S)	How many Eskimo kids have appeared?
	What is true in the belief of “fan death”?

Procedure: Starting the experimental procedure, the participants spent approximately 4 minutes to watch a video, where no browsing options were available. Next, the video player’s browsing options (seekbar) appeared, alongside the questionnaire (Figure 2). The participants had to answer the questionnaire under a 2 minutes time restriction. When this time expired, the video player was deactivated and they had only the chance to finish the job left. Although we informed the users that the purpose of the study was a challenge in finding the answers to the questions within time constraints, this restriction aimed to motivate the users to actively browse through the video. Given the proper engagement, an honest participant could complete properly the task and receive the monetary reward.


A Future With Superhumans - a convo with PhD Daniel H. Wilson
YouTube
ROCK PAPER SCISSORS CHAMPION
PERFECT GOLF SWING
CAT JUGGLING
01:45/03:40
Remain Time:82.3

SuperHumans

Question 1

*Required

As an example of motor activity, two guys are playing rock-paper-scissors. Which are the 3 colours of their clothes? *

Copy your answer.

○ green, blue, orange
○ green, blue, brown
○ grey, blue, white
○ grey, blue, green

Continue »

50% completed

Powered by

This content is neither created nor endorsed by Google.

Fig. 1. Screenshot from the experimental procedure (Questionnaire-Segment match)

IV. RESULTS

Our crowdsourced experiments completed in less than 9 hours and cost $61 for recruiting 200 participants. Maybe they were registered users, but using Crowdflower as a mediator to the available workforce of various crowdourcing marketplaces, we can’t have access in their demographics, except their location. We recruited workers from USA (49.5%), Canada (32%), Great Britain (18%) and Australia (0.5%).

All the crowdworkers completed the whole task, but they successfully asked the task’s third question, the unique user id, as follows: for the Superhumans experiment 87% and for the Misconceptions experiment 89%. We measured this success, based on the expected form of the unique id given automatically from SocialSkip system (a four-digit number, e.g 1032). The user id was the indicator for workers’ engagement; the number of workers that indeed participated and completed SocialSkip experiments.

SocialSkip system gathered 1,359 user interactions with video player. Table 2 shows the counts of these user interactions for the Forward (FW) and the Backward (BW). Counting the different stored unique ids revealed that not every participant interacted with the video player. Particularly, SocialSkip stored interactions from the total amount of participants, as follows: for (Edu. A) 73% and for (Edu. S) 56%. The most popular interaction was the Seekforward, because the users were under time restriction in order to answer the given questionnaire. Although the conditions of the experiments and the controlled motivation were the same, the interactions’ counts are different between the videos.

The counts outside the parentheses are the interactions that workers left behind, after the engagement filtering analysis. Inside the parentheses are the raw interactions we received without filtering. We chose to compare the experiments counts from lecture videos (similar video type) in previous lab experiments [9] with crowdsourcing. Considering the shorter video duration and the less segments of Ground Truth we can safely assume that we had to expect half number of interactions. Based on this approximation, we calculated the difference in expected (Exp. Dif.) and real raw data. The results show that with crowdsourcing we managed to gather significantly more data, than conducting in lab conditions.

TABLE 1 ENGAGEMENT FILTERED INTERACTIONS IN COMPARISON WITH RAW AND LAB EXPERIMENTS

Video	**FW Eng.(Raw)**	**FW Exp. Dif. (%)**	**BW Eng.(Raw)**	**BW Exp. Dif. (%)**
Edu. A	592 (708)	+490	136 (148)	+270
Edu. S	239 (283)	+135	67 (89)	+122

According to Chorianopoulos [10] the replay user activity seems suitable for modeling user interest. In Figure 3 we analyze the user activity signals based on Backward activity (Replay) and we present the visual comparison among the Raw Replay activity, the Engaged Replay activity and the Ground Truth (interesting video segments). The Figure 4 shows the same for the Edu.S video experiment.

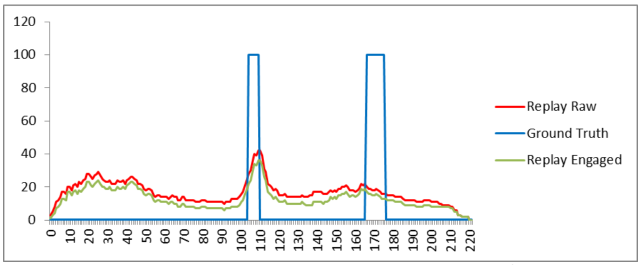


Fig. 2. Compared graph shows Replay-Ground Truth for Superhumans (Edu. A) user video activity

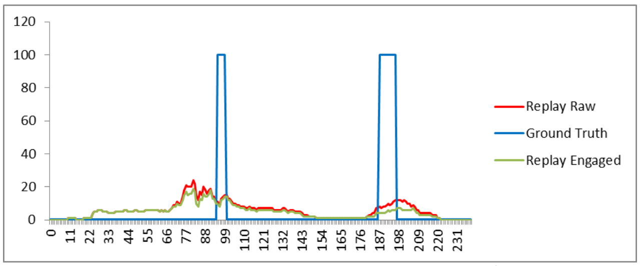


Fig. 3. Compared graph shows Replay-Ground Truth for Misconceptions (Edu. S) user video activity

In both videos the user activity matches the Ground Truth in accordance with the lab experiments’ results. Additionally, the user video activity, based on raw data, follows similar pattern with the data after engagement filtering. This means the participants that didn’t stay dedicated to the experiment, hopefully didn’t mislead our user video activity research expectations.

V. DISCUSSION & FURTHER WORK

In this study, we conducted experiments with a pre-existing video analytics system, in a crowdsourcing platform. We proposed a crowdsourcing methodology managed to achieve large-scale data collection with SocialSkip. Once our research purposes differentiate from the usual demands, such as the video annotation, we had to design and calibrate the crowdsourcing task considering the malicious workers. The task design achieved an affordable job cost, quick collection and a great amount of video clickstream data while accessing a more diverse subject pool. Thus, we consider a reliable alternative methodology compared to lab experiments methodology and focus our future experiments in this direction. The in-depth analysis of the dataset is our main concern in future publications. Nevertheless, the collected dataset and the replicable task design can be used in comparison to future experiments not only for the purposes of crowdsourcing (different reward, quality control approaches etc.) but also for the video research (different video duration, video presentation type etc.).

The id confirmation, revealed those participants that didn’t pay the appropriate engagement to the experiment. Considering the non-sophisticated design approach, the proportion of unengaged workers is relatively small. The replay user activity also confirmed that these malicious workers didn’t affect the clickstream activity. Thus, this kind of participants in SocialSkip experiments could not be defined as “bad” even if he didn’t interact with the system at all. This worker behaviour can be characterized better as random, but the workers’ categorization is a total different research issue. Nevertheless, the fact that they didn’t interact with the player at all was a risk we took from the beginning because the task’s purpose was to answer the questionnaire and so they did.

Moreover, we suppose that the IP blocking and the flagging procedure excluded the repeated recruitment. Although it is a feasible action for one experimenter; but we have to consider the major loss in interactions when other experimenters replicate this methodology. The available workforce is not infinite and SocialSkip must support a list of previous participants. Generalizing this observation, we propose that embedding the crowdsourcing options in social media analytics could not only benefit large-scale data collection, but data quality too.

ACKNOWLEDGMENTS

We are grateful to Konstantinos Pardalis for the technical support in SocialSkip experiments.

REFERENCES

[1] L. Von Ahn and L. Dabbish, “Labeling images with a computer game,” in Proceedings of the SIGCHI conference on Human factors in computing systems, 2004, pp. 319–326.

[2] J. J. Horton and L. B. Chilton, “The labor economics of paid crowdsourcing,” in Proceedings of the 11th ACM conference on Electronic commerce, 2010, pp. 209–218.

[3] J. J. Horton, D. G. Rand, and R. J. Zeckhauser, “The online laboratory: Conducting experiments in a real labor market,” Exp. Econ., vol. 14, no. 3, pp. 399–425, 2011.

[4] G. Paolacci, J. Chandler, and P. G. Ipeirotis, “Running experiments on amazon mechanical turk,” Judgm. Decis. Mak., vol. 5, no. 5, pp. 411–419, 2010.

[5] C. Vondrick, D. Patterson, and D. Ramanan, “Efficiently scaling up crowdsourced video annotation,” Int. J. Comput. Vis., vol. 101, no. 1, pp. 184–204, 2013.

[6] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, “A crowdsourceable QoE evaluation framework for multimedia content,” in Proceedings of the 17th ACM international conference on Multimedia, 2009, pp. 491–500.

[7] M. Allahbakhsh and B. Benatallah, “Quality Control in Crowdsourcing Systems,” 2013.

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

[9] C. Gkonela and K. Chorianopoulos, “VideoSkip: event detection in social web videos with an implicit user heuristic,” Multimed. Tools Appl., vol. 69, no. 2, pp. 383–396, 2014.

[10] K. Chorianopoulos, “Collective intelligence within web video,” Human-centric Comput. Inf. Sci., vol. 3, no. 1, pp. 1–16, 2013.

[11] F. C. Li, A. Gupta, E. Sanocki, L. He, and Y. Rui, “Browsing digital video,” in Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 2000, pp. 169–176.

[12] B. Yu, W.-Y. Ma, K. Nahrstedt, and H.-J. Zhang, “Video summarization based on user log enhanced link analysis,” in Proceedings of the eleventh ACM international conference on Multimedia, 2003, pp. 382–391.