Eye tracker in the wild: Studying the delta between what is said and measured in a crowdsourcing experiment

Pierre Lebreton
Technische Universität Berlin
Assessment of IP-based applications
pierre.lebreton@tu-ilmenau.de

Isabelle Hupont
Institute for Intelligent Systems and Robotics
Pierre and Marie Curie University
hupont@isir.upmc.fr

Toni Mäki
Communication Systems
VTT Technical Research Centre of Finland Ltd
toni.maki@vtt.fi

Evangelos Skodras
Wire Communications Laboratory
University of Patras
evskodras@upatras.gr

Matthias Hirth
Chair of Communication Networks
University of Würzburg
matthias.hirth@informatik.uni-wuerzburg.de

ABSTRACT

Self-reported metrics collected in crowdsourcing experiments do not always match the actual user behaviour. Therefore in the laboratory studies the visual attention, the capability of humans to selectively process the visual information with which they are confronted, is traditionally measured by means of eye trackers. Visual attention has not been typically considered in crowdsourcing environments, mainly because of the requirements of specific hardware and challenging gaze calibration. This paper proposes the use of a non-intrusive eye tracking crowdsourcing framework, where the only technical requirements from the users' side are a webcam and a HTML5 compatible web browser, to study the differences between what a participant implicitly and explicitly does during a crowdsourcing experiment. To demonstrate the feasibility of this approach, an exemplary crowdsourcing campaign was launched to collect and compare both eye tracking data and self-reported metrics from the users. Participants performed a movie selection task, where they were asked about the main reasons motivating them to choose a particular movie. Results demonstrate the added value of monitoring gaze in crowdsourcing contexts: consciously or not, users behave differently than what they report through questionnaires.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; H.5.1 [**Multimedia Information Systems**]: Evaluation/methodology; I.2.10 [**Vision and Scene Understanding**]: Video analysis

CrowdMM 2015, Brisbane, Australia

DOI: http://dx.doi.org/10.1145/2810188.2810192.

General Terms

Human factors, algorithms, experimentation, measurement

Keywords

Eye tracking, visual attention, crowdsourcing, computer vision, user studies

1. INTRODUCTION

Visual attention is an important feature of the human visual system. Every day people are faced to a lot of visual information, sometimes even more they are able to process. Visual attention allows people to selectively process the vast amount of information with which they are confronted, prioritizing some aspects of information while ignoring others by focusing on a certain location or aspect of the visual scene [2]. This capability of human beings has therefore strongly caught the interest of psychologists, psychiatrists, publicists, designers, educators, etc. in the last decades.

The emergence of accurate eye trackers in the early 2000s has opened the door to the reliable exploration of visual attention, both from a quantitative and qualitative point of view. However, traditional eye tracking hardware is high-priced, relies on dedicated equipment (such as infrared cameras), and requires local study of participants (i.e. in the laboratory). Therefore acquiring gaze data from a large number of users is rendered very problematic and time-consuming.

Applying eye tracking techniques in the crowdsourcing environments could be of great value. Firstly, such deployments would allow collecting gaze behavior of a virtually unlimited number of participants worldwide, in an ecological and low-cost manner. Secondly, gaze saccades would provide added valued insights about which visual factors influence aesthetic [8] or emotional ratings [6], thus complementing prior crowdsoucing based research. Finally, monitoring gaze behavior of crowdsourcing users would also provide additional means to identify cheaters, in addition to existing methods such as statistical evaluations of user ratings [4] or monitoring of user interactions with the task interface [3].

However, moving the eye tracking user studies to "the wild" implies some technical challenges to be solved: (1) detecting robustly the user's pupil from a standard RGB webcam, without control of scene conditions (illumination, head pose, etc.), and (2) performing the calibration process that allows matching the user pupil positions with screen positions. Many existing crowdsourcing studies have considered methods to overcome those challenges in evaluating visual attention. For example, in the work by Rudoy et al. [10] participants were asked to watch a video followed by the brief presentation of a panel of letters in different spatial positions, and report which letter they saw the most clearly in the short period of time they had to look at the panel. Huang et al. [5] make use of a different approach: they consider that the usage of the mouse by a participant is directly related to gaze in the case of web page navigation. However, mouse input does not necessarily reflect user activity or non-activity and sometimes a user can move gaze and mouse independently of each other. Alternatively, Xu et al. [13] propose the use of webcams to measure visual attention in a crowdsourcing context. In an initial phase, they ask the participant to look at specific locations on the screen to perform calibration. Once calibrated, the gaze evaluation on image is performed. This approach has however the drawback that it requires intrusive re-calibration between images due to users head movements, which implies participant's flow of attention to be frequently interrupted.

In our previous work, we proposed a webcam-based eye tracking framework that addresses these deficiencies [7]. In our framework participants are not asked to look at specific positions on the screen for calibration, but their interaction with the system is used instead: in line with the findings by Rodden et al. [9] we consider that when a user clicks on a given user interface element, he is looking at this element. The *pupil-screen position* pairs collected offer a non-intrusive way to perform a continuous re-calibration of the system all along the test and thus to maintain accuracy of the eye tracking results in longer interaction periods.

The contribution of this paper is twofold. First, we present an enhanced version of the framework presented in [7]. This version includes an on-line monitoring of the test conditions to detect insufficient light conditions and face-to-camera positions. In case unsatisfactory conditions are detected user is notified and guided to correct the conditions. The monitoring also deploys an improved mechanism to estimate fixation point of the participants to increase the accuracy of the estimation. Second, we conduct an exemplary test during which eye tracking and self-reported metrics are both collected to study the differences between what a participant consciously and unconsciously does. The results demonstrate the added value of monitoring gaze data in a crowdsourcing context.

The structure of the paper is following. Section 2 presents our eye tracking framework for crowdsourcing, focusing on the improvements carried out for this work. Section 3 describes the experimental protocol followed for this study on visual attention. The results obtained are detailed in Section 4, while conclusions are discussed in Section 5.

2. VISUAL ATTENTION FRAMEWORK

In previous work [7], a framework for measuring visual attention in a crowdsourcing context was developed. It is a webcam-based approach which does not require any additional hardware on the end-user device, nor any additional software than a HTML5 compatible web-browser. User campaigns revealed some issues in the initial version of the framework; the bad lighting conditions and misplacement of users with respect to the camera made some cases impossible to analyse. Therefore, in this work, the framework was improved by adding an on-line monitoring of the test conditions as described in the following.

The framework records the participant's face and mouse clicks. It assumes that when a click is performed by the user, the participant will be looking at the location of the click. Therefore every interaction of the user with the platform during the entire length of the test is used as calibration data, enabling continuous re-calibration. The framework is divided into two parts; a user-side for recording user interactions and video data, and a server-side for information storage and post-processing.

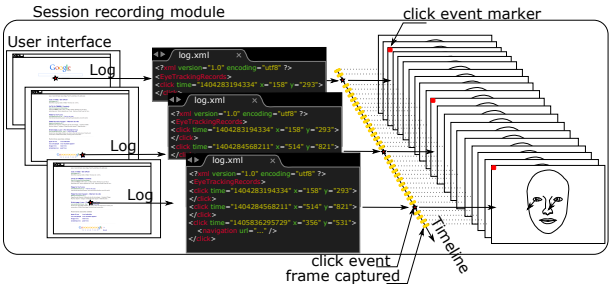
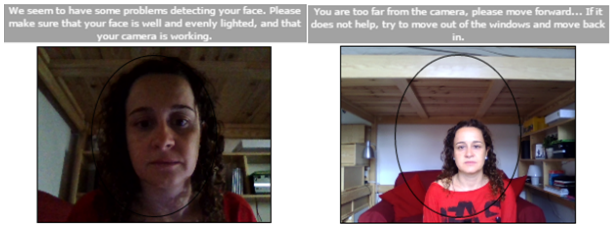


Figure 1: User-side: recording of user's interaction and face.

2.1 User-side

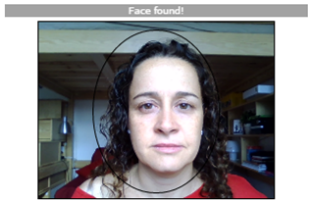
The user-side's software components are implemented as web applications. Therefore, only a modern browser supporting HTML5 and a web cam is required on the end-user device. The application concept is schematically depicted in Figure 1. During the test, the user is browsing a website and performs several clicks while navigating through the site. The position of the clicks on the screen and the time at which they occur are recorded. In the same time, the face of the participant is recorded into a video stream using Web-RTC based components. This video stream is processed by adding a click event marker into the appropriate video frame at the time of each click event, enabling synchronization between click records and the video stream from the webcam. Both of these data, video and click records, are then streamed to a distant server which performs the processing for gaze estimation.

Picture quality, contrast, viewing distance, and user head movement are crucial factors in ensuring optimal performance of the platform [7]. While the latter one, the head movement, is difficult to avoid, it was observed in [13] that within short periods of time the test participants remain most of the time stable. This phenomena enables continuous calibration, as used in our approach, to stabilize the performance of the algorithm over the timespan of the entire test. The picture quality-related factors have also been addressed by on-line monitoring of the test conditions; at the beginning of the test, a Haar Cascade classifier was used inside the user side software to detect faces and ensure that the participants are not too far from the screen by measuring the face size on the recorded video. To ensure good lighting conditions and overall image quality, the contrast



(a) The user face is not well illuminated

(b) The user is too far from the screen



(c) Image has enough quality, the test can start

Figure 2: On-line monitoring of test conditions: distance from the face to the camera, image quality and lighting conditions are checked and feedback is provided to the user.

measure proposed in [1] was deployed. The measure allows checking that the picture from the user reaches a minimum quality threshold. Before proceeding with the test, all the conditions had to be satisfied. The user was guided in form of textual feedback. Figure 2 illustrates two cases of bad conditions and one acceptable.

2.2 Server-side

On the server-side, the videos are post-processed to find the eye positions from the recordings of participants' faces. The eye positions are measured relative to stable feature points around the eye corners in highly textured locations. The relative tracking enables higher robustness in the presence of the head movements. The position of the face is initially found using the Viola-Jones face detector [12]. Next the eye center positions are detected using the eye localization algorithm from Skodras [11]. A linear model providing the mapping between the eye positions and the positions on the screen is then fitted, by applying the random sample consensus algorithm (RANSAC) to the click data. The fixation can be estimated for every frame by the means of interpolation between the calibration points (the frames marked at the times the user clicked mouse). Further details can be found in [7].

2.3 Quality control and adaptive coefficients

It is possible to continuously evaluate the performance of the fixation estimation algorithm, based on the assumption that participants are looking to the point they click with their pointing device. Figure 3 depicts the prediction error of fitting compared to the ground truth provided by the click input (i.e. estimated vs. real click positions) throughout a single test session.

Four different regression models that map eye positions to the screen positions are proposed: linear models for both the left (1) and the right (2) eye positions individually, a linear model using the average of the both individual regression models (3) and an "adaptive" (4) fitting. The "adaptive"

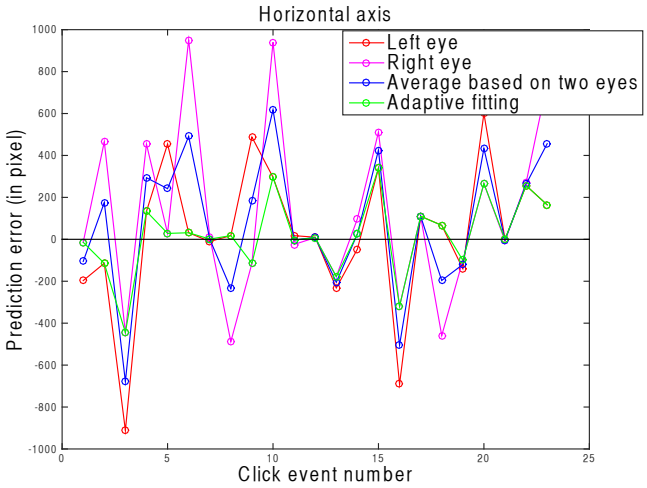


Figure 3: Fixation estimation quality along the test.

fitting is one of the contributions of this paper. It is based on the observation that estimation accuracy of each eye-center position and related fixation position varies (e.g. due to non-uniform lighting of the participant's face at a specific instant of the test) as can be seen in Figure 3. As each click provides a fixation position on the screen, it is possible to estimate which one of the linear models 1), 2) and 3) is the most reliable in temporal proximity of a calibration point. The "adaptive" fitting then always selects the best model for prediction.

In order to estimate the fixation position on the non-marked frames, the distance between the considered frame and the closest calibration point is determined. Using this calibration point, it is possible to find the "adaptive" (optimized) coefficients and use them to estimate the fixation position for any frame (interpolation for frames with no click information). Figure 4 depicts a horizontal axis fixation predictions of a single participant during the task described in Section 3. Figure 4 also illustrates how the adaptive fitting selects the best available model at calibration points and shows the fixation prediction confidence at the calibration points in the form of the error bars.

The results considering the horizontal axis were discussed in this section. The vertical axis can be addressed similarly, but with consideration of an additional challenge: the typical position of the webcam, top of the display, is non-optimal. The angle of the view to the eyes makes measuring the vertical movement of the eyes more difficult, sometimes resulting in less precise measurements. Nevertheless, as shown in [7] also vertical accuracy was found good enough for practical deployments.

3. EXPERIMENT SETUP

In this section we discuss the applicability of our approach in an exemplary use-case in a crowdsourcing environment: collecting subjective information about the main reasons that motivate us to to watch a particular movie. The results of subjective crowdsourcing studies are difficult to evaluate because simple quality assurance techniques like gold standard data are not applicable [4]. Therefore, other techniques like consistency questions or repeated judgments are

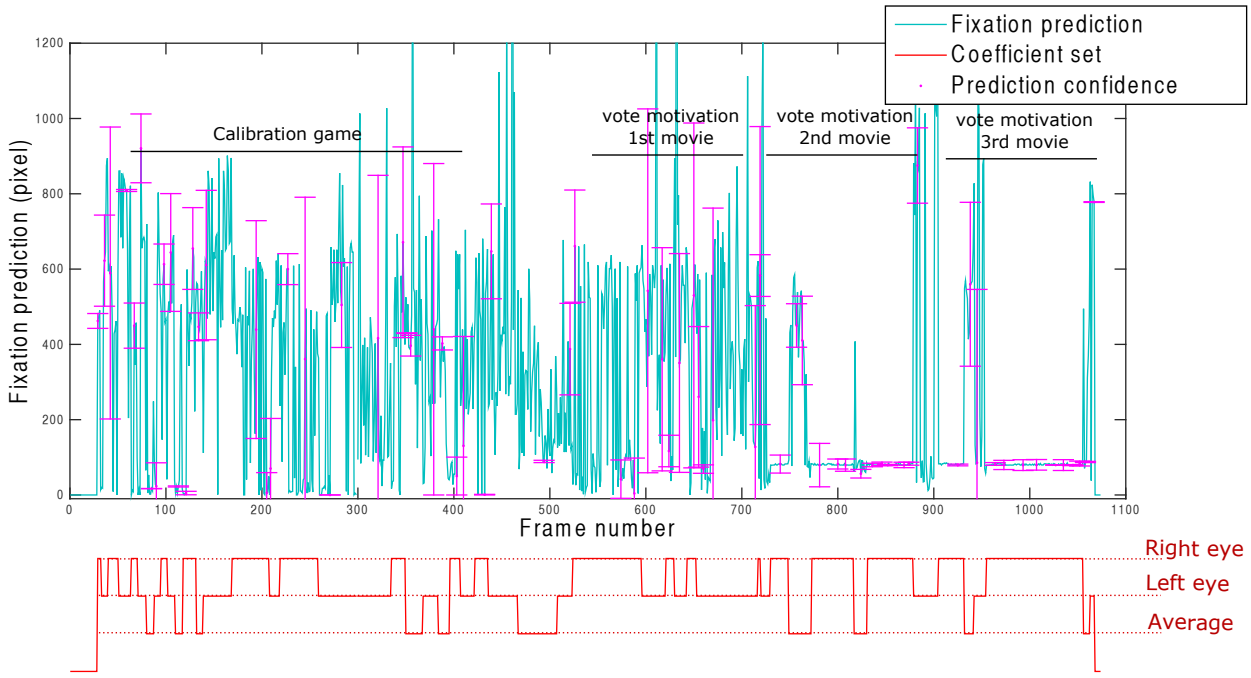


Figure 4: Measurement of the fixation on the horizontal axis over an entire test.

required, which impose an additional workload on the workers without increasing the actual number of subjective judgments. Using gaze tracking, it is possible to assess the worker's interactions with the tasks to evaluate, whether his behavior is reasonable or has properties of a random clicker.

3.1 Description of the task

The task in this study is about movies selection. The user was shown information about 3 different movies, which included: the poster of the movie, a brief synopsis, the cast description and some statistics (such as its budget and the mean rating of the public) as illustrated in Figure 6. This information was queried from the IMDb movies database[1]. The user was asked to select the movie he would like to watch. He could also indicate if he had previously seen any of the movies presented. Once the movie was selected, the user needed to indicate (at least) three criteria that brought him to the final decision from the options: "title", "description", "cast", "poster", "director", "box office", "release date", "ratings" or "I knew the movie". The buttons the user needed to click in order to proceed were located as far apart from each other as possible. Such a layout allowed collecting the most valuable click-pupil pairs for eye tracking re-calibration purposes as the participants had to look at different corners of the screen.

The movie selection task had to be performed three consecutive times by the user. Before and after that, several minor phases had to be completed. The experiment pipeline is summarized in Figure 5. First, general instructions about the test were provided. Second, the users needed to grant the application access to their PC webcam and agree on recording the video of their face during the experiment. In this phase, the image quality check was also performed to ensure good facial resolution and illumination during the test. Then, a game where test participants had to click on moving objects on the screen, serving as a preliminary calibration protocol, was displayed. After that, brief specific instructions about the movie selection task were provided. Finally, a basic demographic questionnaire including questions about users film genre preferences had to be filled in and the movie selection task started. At the end of the test, the video to be uploaded to the server was shown to the users and final questions about how they felt during the experiment were asked. Again, it should be noted that every click performed by the user during the whole experiment pipeline was recorded to update eye-tracking calibration data.

3.2 Participants and campaign description

Two crowdsourcing campaigns were conducted: an experiment involving volunteer online testers recruited during a science show and a paid crowdsourcing campaign carried out using the Microworkers[2] platform. In the former campaign, 10 participants completed the test. During the latter campaign, 29 Microworkers users from English speaking countries executed the work between 16th and 20th of June 2015 and were rewarded with 1 USD after providing the required proof. The uploaded data of 13 Microworkers users could not be considered in the evaluation because the respective videos were not properly received on the server side due to a software defect. In the end, data from 16 crowdsourcing participants (13 males, 3 females) was available for evaluating the feasibility of the proposed approach.

4. RESULTS

The proposed framework allowed to estimate fixations over the entire length of the experiment. Figure 4 depicts the fixation estimations (with respective confidence intervals)

[1] `http://www.imdb.com` Accessed: Jul. 2015

[2] `http://www.microworkers.com` Accessed: Jul. 2015

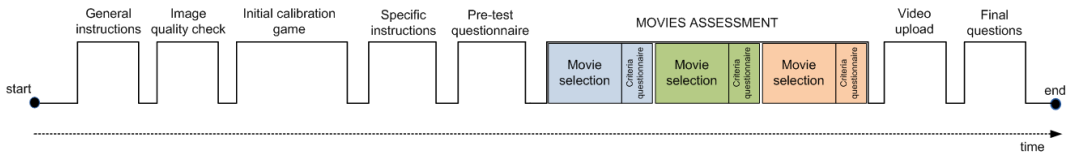


Figure 5: Experimental procedure.

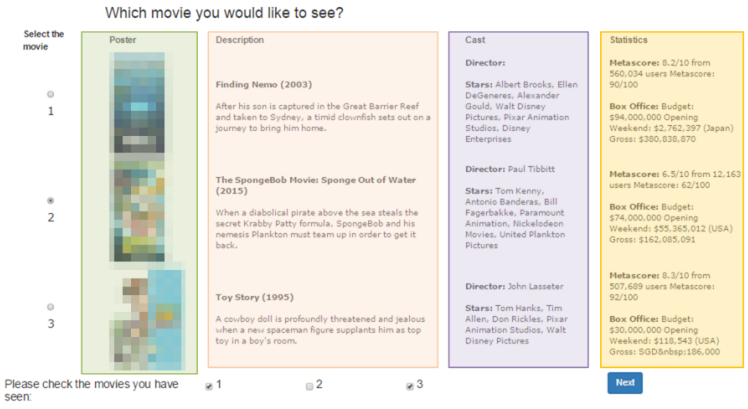


Figure 6: Test snapshot: the user is asked to select the movie he would like to see and also to indicate if he has already seen any of the movies presented. For exemplary purposes, we colored the areas of interest in this figure which were internally used to compute gaze statistics.

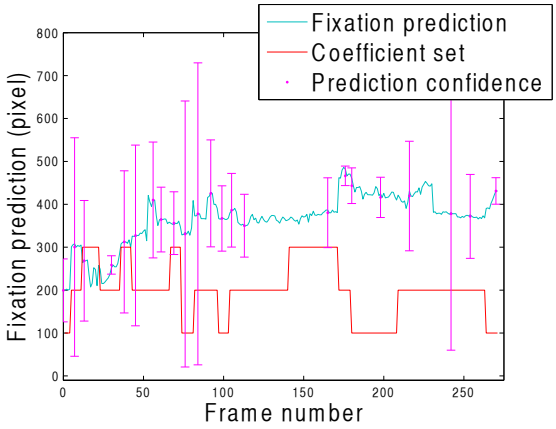


Figure 7: Fixation prediction for any frame including confidence information on the vertical axis.

along the test for one participant. As a general observation, it can be seen that the participant first scans the entire screen when there is movement (the calibration game) or new kind of pages (instructions and the selection of the first movie). When the user goes through already familiar page designs, the 2nd and the 3rd movie selection, starting from the frame 700, the dominant focus at the left part of the screen can be observed. That part corresponds to the position of the list of options he has to choose from. There is a clear difference in the user's gaze behaviour when comparing the first exposure to the movie selection page and the subsequent ones. The behavioral change, possibly caused by learning, is something that could not be captured by traditional metrics such as task completion time.

Figure 8 demonstrates a type of data illustration depicting the fixation map obtained for one participant while choosing a movie. Putting this plot in relation to Figure 6, it can be seen that this participant has mainly focused on the Poster column and to a lower extent on the meta-data and the cast information.

To study the relationship between what participants replied and how did their attention fixate, the movie selection page was divided into four categories (see Figure 6). Then the time users spent watching each category was calculated from the fixation data. The results are depicted in Figure 9. It can be observed for example that the first participant spent 40% of the time on the 1st category and 40% of the time on the second category.

After each selection, the participants were asked to report which criteria motivated their decision (at least three criteria in the order of importance). A chart based on their replies, shown in Figure 10 was then built to compare their answers to the fixation data. The selected main criteria was assigned a weight of 3, the second in importance a weight of 2, and the third a weight of 1. The criteria weights of all the selections were summed per user and per criterion to find out their main criteria used in choosing the movies. From this chart, it can be seen e.g. that the first participant based his decision mainly on the second category (the movie description) but also on the cast, indicating that the survey and the eye tracking data reveal different influence factors.

This also holds for the other participants. The participants mainly reported to have based their decision on the description of the movies, however fixation data shows that they spent most of the time looking at the meta-scores. The second most attention was drawn by posters, while the description was third when counting the fixation time. While this does not imply that the participants were lying, since it is possible that they did base their decision on the description, it is very interesting to see that they mainly spent their time at looking the meta-scores. It is possible that the meta-score may have influenced users' decisions, even though they do not admit the effect or they are not even aware of it. This study shows, in line with previous studies, that self-reported measurement does not necessarily provide the same results as physiologic measurement as self-reported measure provides mostly information on the conscious aspects of user decisions.

5. CONCLUSION

An exemplary subjective user study was conducted to demonstrate the benefits eye tracking mechanism in the crowdsourcing environment. The eye tracker developed in previous work was extended and applied to measure the user behaviour. It provided relevant information, not available via traditional usability metrics, enabling to evaluate the

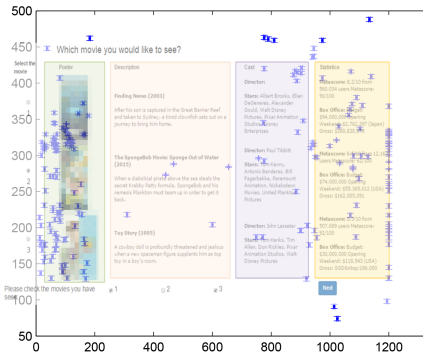


Figure 8: Example of fixation map for one participant when choosing a movie.

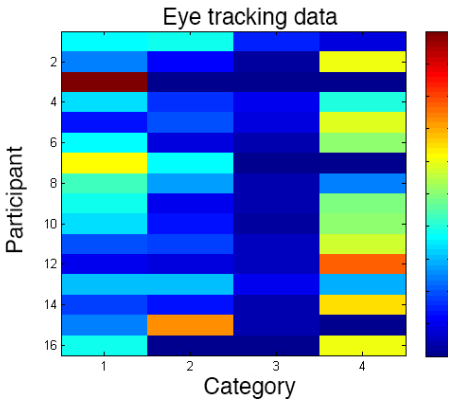


Figure 9: Eye tracker data: time spent by each participant at looking each category

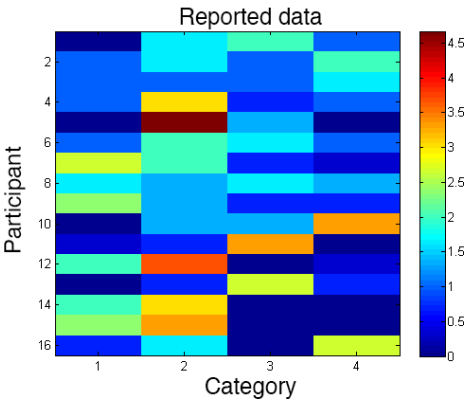


Figure 10: Survey data: category used as the main factor for the decision

evolution of the understanding of the participant during the task by measuring the evolution of how he watches the test. A second main result is to compare the replies of questionnaires and the gaze data. It was observed that they do not fully agree. This shows that consciously or not, users behave differently than what they actively report. Performing eye tracking in the crowdsourcing context enable then to get additional relevant data on understanding participants' decision.

Additionally, to improve the eye tracker performance, a new robust approach for pooling fixation prediction from each of the eyes position was presented, and it was demonstrated how it is possible to continuously measure the eye tracker accuracy along the test.

6. REFERENCES

[1] S. S. Agaian, K. P. Lentz, and A. M. Grigoryan. A new measure of image enhancement. In *Proceedings of the International Conference on Signal Processing & Communication*, 2000.

[2] M. Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

[3] M. Hirth, S. Scheuring, T. Hoßfeld, C. Schwartz, and P. Tran-Gia. Predicting result quality in crowdsourcing using application layer monitoring. In *Proceedings of the International Conference on Communications and Electronics*, 2014.

[4] 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. *Transactions on Multimedia*, 16, 2014.

[5] J. Huang, R. W. White, and G. Buscher. User see, user point: Gaze and cursor alignment in web search. In *Proceedings of the Conference on Human Factors in Computing Systems*, 2012.

[6] I. Hupont, P. Lebreton, T. Mäki, E. Skodras, and M. Hirth. Is affective crowdsourcing reliable? In *Proceedings of the International Conference on Communications and Electronics*, 2014.

[7] P. Lebreton, T. Mäki, E. Skodras, I. Hupont, and M. Hirth. Bridging the gap between eye tracking and crowdsourcing. In *Proceedings of the SPIE 9394, Human Vision and Electronic Imaging XX*, 2015.

[8] J. A. Redi, T. Hoßfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel. Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In *Proceedings of the Workshop on Crowdsourcing for Multimedia*, 2013.

[9] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *Proceedings of the Conference on Human Factors in Computing Systems (Extended Abstracts)*, 2008.

[10] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor. Crowdsourcing gaze data collection. In *Proceedings of the Collective Intelligence Conference*, 2012.

[11] E. Skodras and N. Fakotakis. Precise localization of eye centers in low resolution color images. *Image and Vision Computing*, 36, 2015.

[12] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57, 2004.

[13] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

APPENDIX

A. ACKNOWLEDGMENTS

The authors thank Microworkers.com for supporting the crowdsourcing experiments. Toni Mäki's work was partially funded by Tekes the Finnish agency for research innovation, in the context of the CELTIC+ project NOTTS. The research leading to these results received funding from the Deutsche Forschungsgemeinschaft (DFG) under Grants HO4770/2-1, TR257/38-1.