 DOI:10.1117/12.2076745

Bridging the gap between eye tracking and crowdsourcing

Pierre Lebreton[a], Toni Mäki[b], Evangelos Skodras[c], Isabelle Hupont[d] and Matthias Hirth[e]

[a]Assessment of IP-Based Applications, T-Labs, Technische Universität Berlin
[b]VTT Technical Research Centre of Finland
[c]University of Patras
[d]Aragon Institute of Technology
[e]University of Würzburg

ABSTRACT

Visual attention constitutes a very important feature of the human visual system (HVS). Every day when watching videos, images or browsing the Internet, people are confronted with more information than they are able to process, and analyze only part of the information in front of them. In parallel, crowdsourcing has become a particularly hot topic, enabling to scale subjective experiments to a large crowd with diversity in terms of nationalities, social background, age, etc. This paper describes a novel framework with the aim to bridge these two fields, by providing a new way of measurements of user's experience in a subjective crowdsourcing experiment. This study goes beyond self-reported methods, and provide a new kind of information for the context of crowdsourcing: visual attention. The results show that it is possible to estimate visual attention, in a non-intrusive manner and without using self-reported methods or specialized equipment, with a precision as high as 14.1% in the horizontal axis and 17.9% in the vertical axis. This accuracy is sufficient for many kinds of measurements that can be efficiently executed only in non-controlled environments.

Keywords: Eye tracking, Crowdsourcing, Extend measurements, WebRTC, Visual attention, Human Computer Interaction

1. INTRODUCTION

Visual attention constitutes a very important feature of the human visual system (HVS). Every day when watching videos, images or browsing the Internet, people are confronted with more information than they are able to process and analyze only part of the information in front of them.[1] Therefore, the understanding of how people scan images or videos is of high interest[2] because of the variety of possible applications ranging from perceptual coding to layout optimization of websites.

In parallel, crowdsourcing has become a particularly hot topic to scale subjective experiments to a large crowd in terms of numbers of test participants,[3,4] because it enables recruiting up to a few hundred test participants with a large diversity in terms of nationalities, social background, age,[5,6] etc. within a few hours at low costs. It further allows performing the evaluation in the environment the test participants are used to (out of the lab). The ability to reach a high number of participants behaving naturally can be a great added value of crowdsourcing, and can contribute to visual attention studies. However, due to the uncontrolled and unsupervised environment, new challenges - significantly differing from those in traditional lab environments - arise while conducting subjective user studies.[7,8]

This paper describes a novel framework with the aim to bridge these two fields, by enabling crowd-sourced eye tracking experiments. The proposed approach is able to perform eye tracking experiment using standard hardware and implements best practices for subjective crowdsourcing tests. The goal of this study is to provide a new way of measurements on user's experience in a subjective experiment conducted in a crowdsourcing context, which go beyond self-reported methods, and provide new information such as visual attention, interest or emotions.

In the proposed context of crowdsourcing, previous methods were proposed to estimate visual attention using

Further author information: (Send correspondence to Pierre Lebreton.)
Pierre Lebreton: E-mail: pierre.lebreton@telekom.de, Telephone: +49 30 835354271

self-reported methods.[9] The test participants were asked to watch a video, which was followed by the presentation of a table of letters which appeared very briefly. Test participants needed then to report which letter did they saw more clearly. This approach is interesting, but can only be used sporadically. Moreover, the questions asked during the test interrupted the flow of the experiment, affecting the participant's immersion into the task.
In a non-intrusive manner, the regular input facilities (i.e. mouse pointer or cursor) have been used to estimate the user's gaze for determining user intention and salient components of the web pages.[10] However, user inputs do not necessarily reflect user's activity or non-activity: a moving or a stationary mouse does not fully capture the participant's attention to the task.
Going beyond traditional self-reported measurements (still in the crowdsourcing context), previous work considered the evolution of facial response while looking a video.[11] Test participants were recorded, and key feature points of user's faces were measured to study participants' emotions. In this work, it is proposed to go beyond these work and provide information on visual attention.
The main contribution of this paper is to show how in the crowdsourcing context, it is possible to get extended measure of user's experience by looking into visual attention without asking him to report it. In Section 2, a novel framework enabling eye tracking in a non intrusive way at the participant's home without any specific hardware requirements is described. Section 3 will describe of the different crowdsourcing experiments conducted. Section 4 will provide the results and analysis of limits of the proposed framework. Section 5, will conclude this paper and present further work.

2. FRAMEWORK

The proposed framework involves two separate parts, illustrated in high level in Figure 1: a client and a server side. Logically the operations are divided into Capturing and Synchronization and Gaze Estimation (post-processing). During the Capturing and Synchronization, on the client side, the test participant uses his/her personal computer and regular web browser to access a web page where the face recording is done while multimedia content is presented to him. The participant is recorded using the webcam of his personal computer. The captured video is streamed to a distant server using WebRTC. On the server side, the video is stored into a file. The location of every mouse click within the page is also transmitted to the server and recorded in the database. To perform these operations, the server is based on different web technologies such as Javascript, Rails and PHP components running on an Apache HTTP server.
Subsection 2.1 provides more information on the Capture and Synchronization performed on the user side of the framework, while subsection 2.2 describes the Gaze Estimation post-processing on the server side.

2.1 Capture and synchronization of data streams

On the user side, there are two task done by the framework; (1) the capturing of the two kinds of input: the click events - including the temporal and locational information of the clicks within the webpage - and the video of the user's face, and (2) provisioning both inputs with synchronization information used later on in Gaze Estimation. This user side part of the framework is targeted for a large range of devices. The main requirement on the user device is the availability of used streaming technology WebRTC, which is currently available on most computers and Android-based mobile devices (other requirements are implementation details left out of this paper).

All the information, i.e. the captured video and click inputs, is transmitted to the server side of the eye tracker framework. Since both of these inputs are recorded and transmitted independently, they need to be synchronised later on. To this end, a method based on tags was developed: whenever the user clicks, the mouse event is captured, then a dedicated marker is added to the corresponding video frame showing the face of the test participant at this moment. Figure 2 sums up the capture process: the user is interacting with the test interface, while his face is being recorded. The events triggered by user's actions on the interface (e.g. mouse clicks) are reported into the event log and into the video stream from the webcam for synchronization. Figure 2 also demonstrates the higher temporal resolution of eye tracking based gaze estimation compared to click event based methods. When monitoring just the click events, the gaze can be estimated only during user interaction. With eye tracking the gaze can be positioned also between interactions.

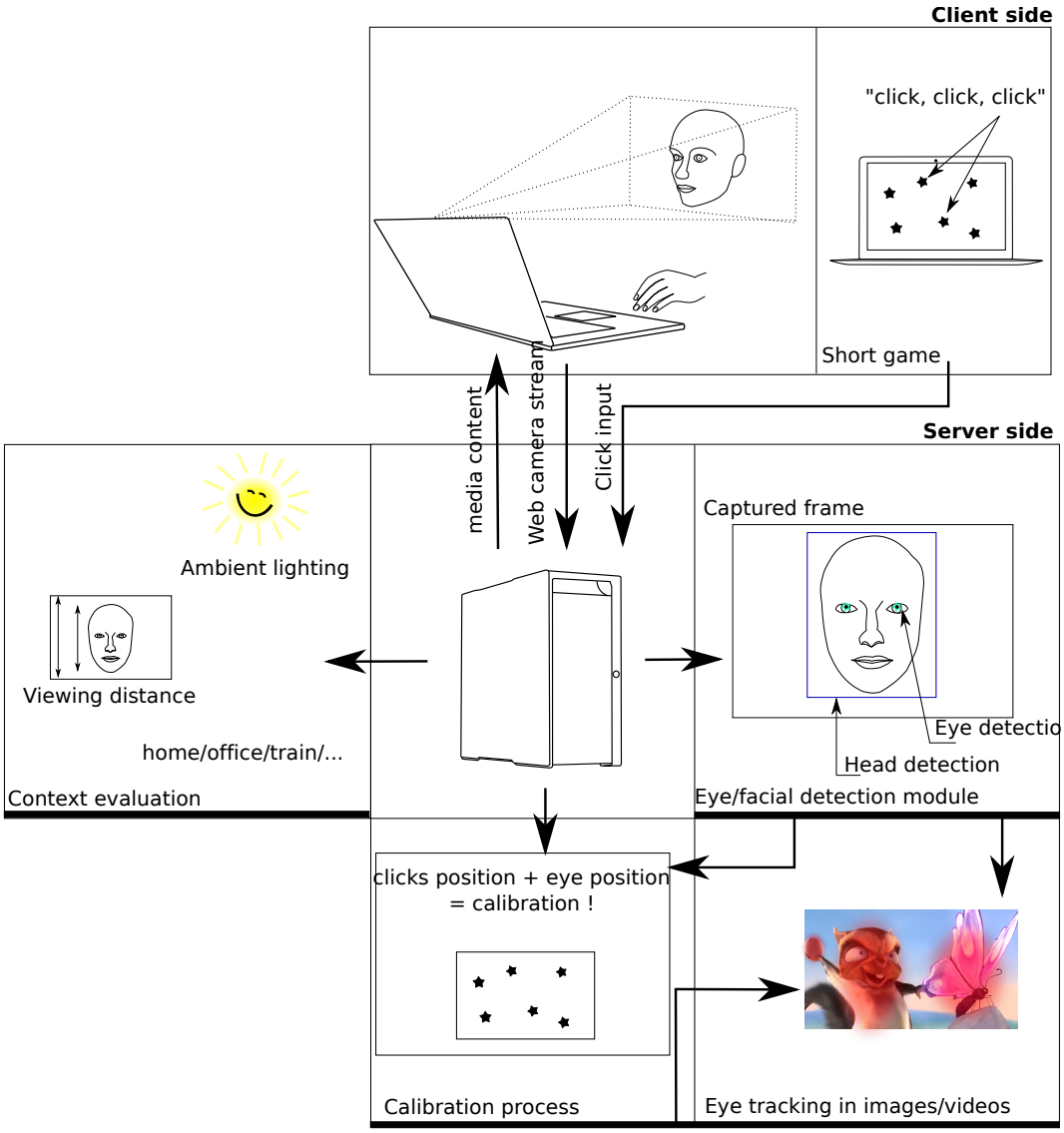

Figure 1. General overview of the framework.

2.2 Gaze estimation

Once both the video of the face of the test participant and the click inputs are received on the server side, the gaze estimation is done via post-processing in an offline manner. The direction of the gaze is estimated by modeling the movement of the eyes inside the eye socket. In order to represent the movement of the eyes, it is required to accurately detect the positions of the points that move and contribute to gaze direction, i.e. eye centers and eyelids, and the positions of "anchor" points which constitute stable face locations serving as reference points throughout the image sequence (so that we are able to measure distance vectors independently of the position of the face) (Figure 3). The feature space is formed as horizontal and vertical distances between "moving" and "anchor" points. Given the assumption of independence of gaze estimation in the two axis, two separate feature vectors are constructed for each direction.

The position of the face is initially found using the Viola-Jones face detector.[12] The position of the eye centers are detected using the eye localization algorithm of Skodras.[13] Unlike "moving" points, the "anchor" points do not need to be at specific locations of the face as their role is to serve as reference points in order

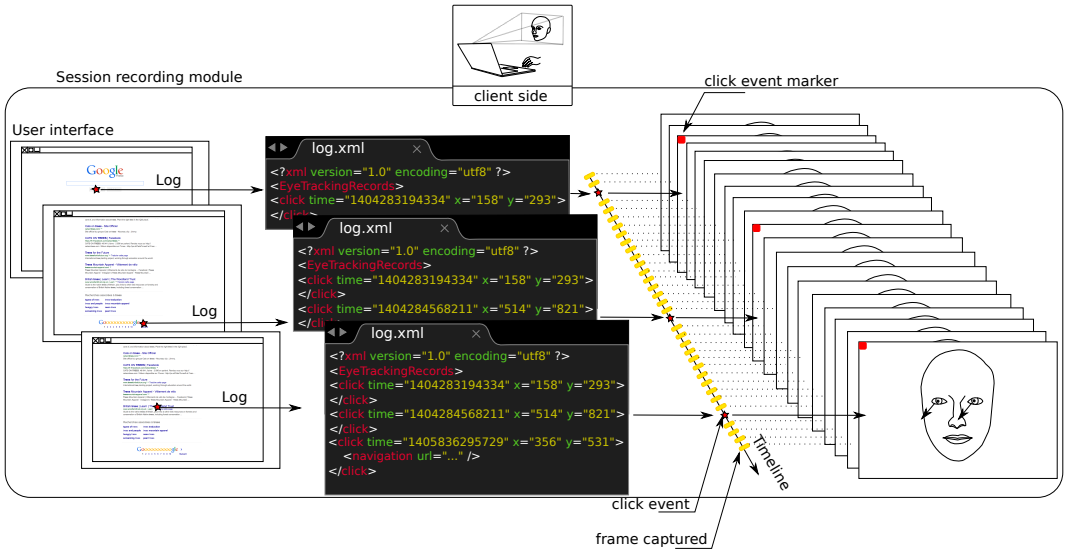

Figure 2. Recording user session.

to derive the distance vectors. In the current approach "anchor" points are arbitrarily chosen for each subject in highly textured locations around the eye corners (for an example, please see the star in Figure 3) and their positions are tracked throughout the entire video using the Lucas-Kanade inverse affine transform.[14]

Once the feature vectors are formed, a mapping between eye positions and the respective location on the screen needs to be derived. This correspondence is established during calibration (c.f. "Calibration process" of Figure 1). The calibration is based on the assumption that test participants are looking at the GUI object that they are clicking. From the moment of the click, it is then possible to identify the relationship between eye position and locations on the user's screen. During the calibration only a certain amount of eye position - gaze position correspondences are known. However, fixation points between the known points can be estimated by the means of regression. Given enough reliable samples from the different marked key frames, the fixation points within the whole screen can be interpolated.

The mapping between the image data (in terms of features) and screen coordinates of user fixations is derived using regression with a linear model, as depicted in Figure 4. Drawing from Figure 4, it can be observed that the proposed linear model is appropriate considering the ratio of data over noise. Once the regression coefficients are calculated, the linear model can be applied to estimate the fixations given the unknown image data converted in terms of features. Figure 5 depicts this part of the fixation estimation process: the training is performed using

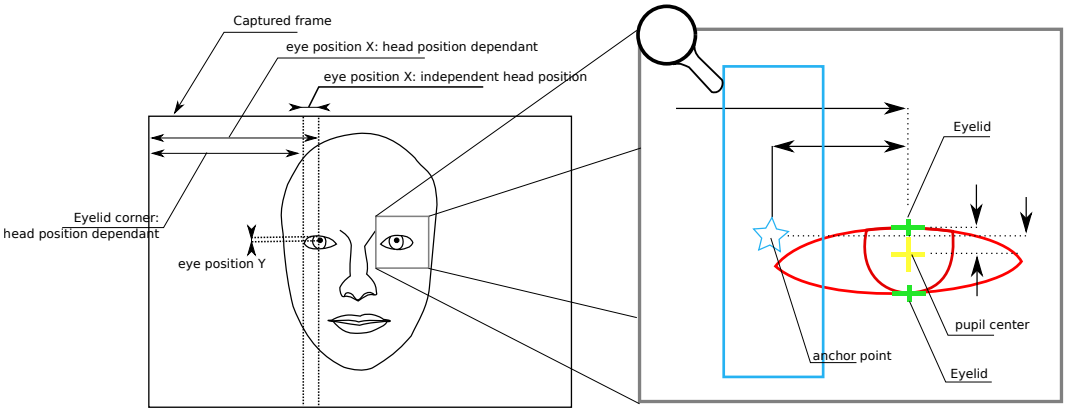

Figure 3. Normalization of eye position using stable facial points.

the red points and the fixation estimation is done for the blue points. For example, if we consider all the points circled in the Figure 5, once the regression is performed on the three red points corresponding to click events, finding the fixations corresponding to the blue points is based on the linear model trained on the red points. This simplifies the problem of gaze estimation to interpolating every fixation between known key points. Based on this, a continuous calibration between eye position and click position during the entire length of the experiment is possible, as depicted in Figure 6.

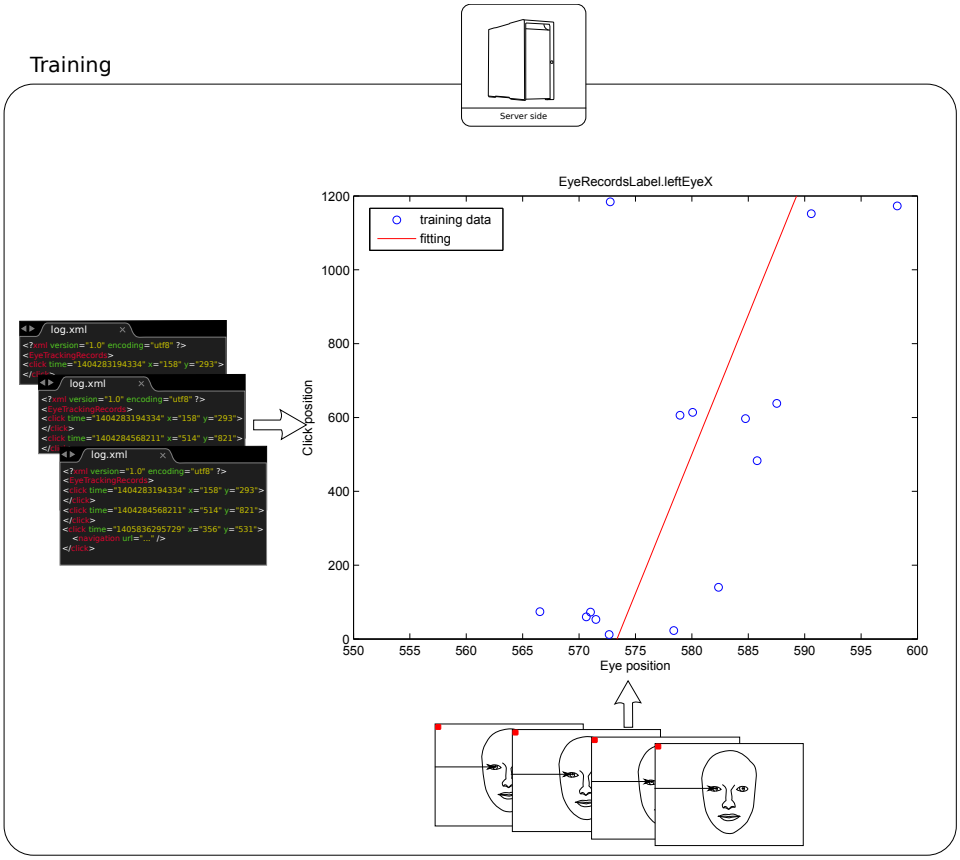

Figure 4. Training of the linear model between eye position and click position.

2.3 Scalability and robustness

The event and video capturing implementation is logically separated from the "content application" that users are interacting with during the tests. The application independent capturing allows using the framework easily with various web applications, as highlighted in Figure 7, that are accessed via an Internet browser, such as Google Chrome. Additional information collected from content application may facilitate the analysis though, especially in complex tasks.

Except from the displacement of the eyes, the movements of the head also contributes in deriving the direction of gaze. Head pose invariance in gaze estimation either requires special hardware configurations (such as helmets, glasses etc) or prior knowledge of the camera parameters and geometry, being impracticable in the context of crowdsourcing. Head pose estimation using monocular information and no special lighting still remains a challenging research topic,[15] requiring computationally demanding algorithms for modeling, while inaccurate estimations of head pose may trigger much larger errors in final gaze estimation, reducing the system's overall accuracy and robustness. In proposed approach, information of head pose is only implicitly incorporated through the mapping function. Given that minor head movements do not have a noticeable influence on the outcome, the only condition is that the head movements are constrained.

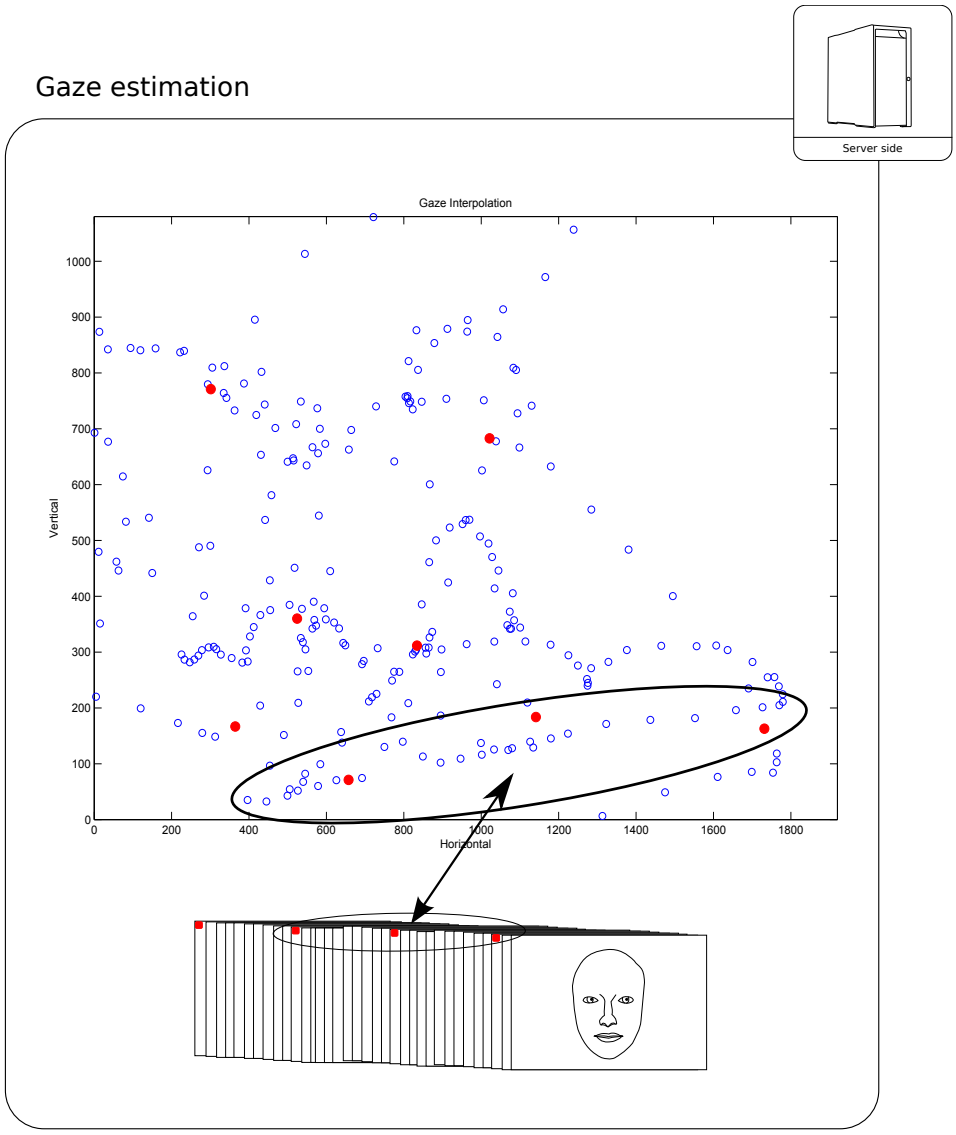

Figure 5. fixations estimation using the pre-trained linear model.

To compensate the inaccuracies caused by head moevements, an analysis method involving use of sliding window is envisaged, as depicted in Figure 7. Considering only the "fresh enough" samples to re-calibrate the system may prove to be succesful to enhance the prediction performance.

3. EXPERIMENTS

3.1 Campaigns

In order to evaluate the performance of the proposed framework, two crowdsourcing campaigns were conducted; a preliminary experiment involving volunteer online testers and a paid crowdsourcing campaign, carried out using the Microworkers platform. The structure of both campaigns was the same and it is described in next section.

In the first preliminary experiment 8 participants (collaborators and acquaintances) voluntarily completed the test, providing also feedback for improving and fine tuning the experiment. The second crowdsourcing experiment was carried out using the Microworkers crowdsourcing platform, which provided the functions to

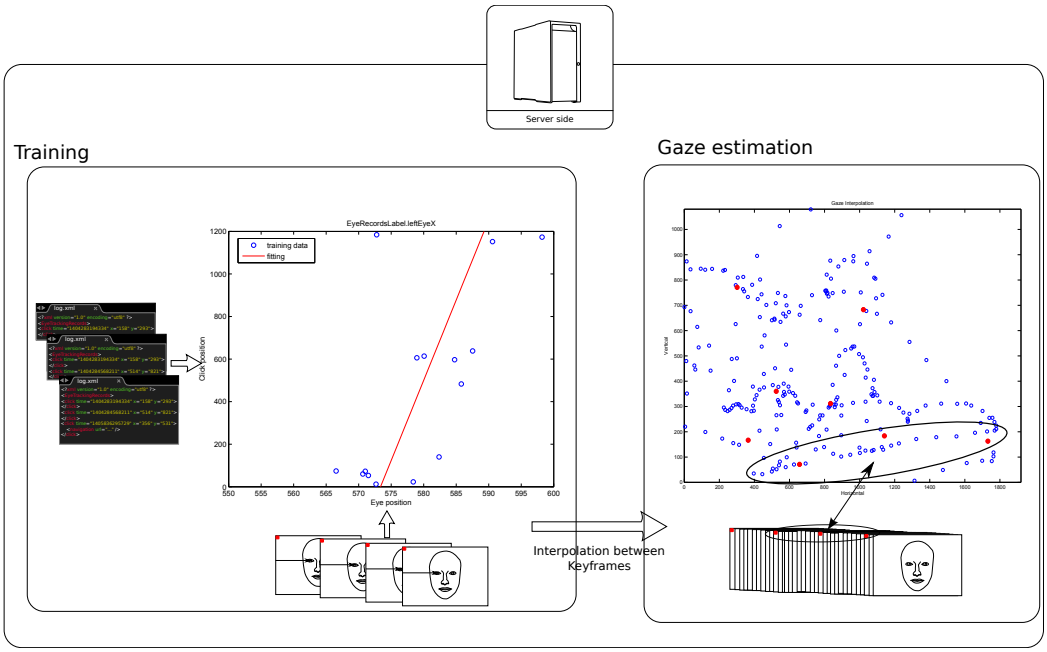

Figure 6. Gaze estimation for every frame using linear regression between frames where click event occurs.

hire the test users and approve the work executed. The campaign was available globally for Microworkers users, and each user was rewarded with 1 USD after providing the required proof. Out of the 30 workers, the input provided by 20 of them was meaningful for our experiments. The excluded videos included cases where the gaze tracking algorithm failed to provide reliable results due to very large head movements (2 cases), movement of the head out of the image (2 cases), bad illumination conditions (3 cases) and strong reflections from glasses which completely occluded the eyes (3 cases).

3.2 Description of the task

The experiment was divided into two parts (Figure 8). Testers browsed to the webpage of the test environment, using a WebRTC compatible web browser (Google Chrome). After a demographic questionnaire, the users needed

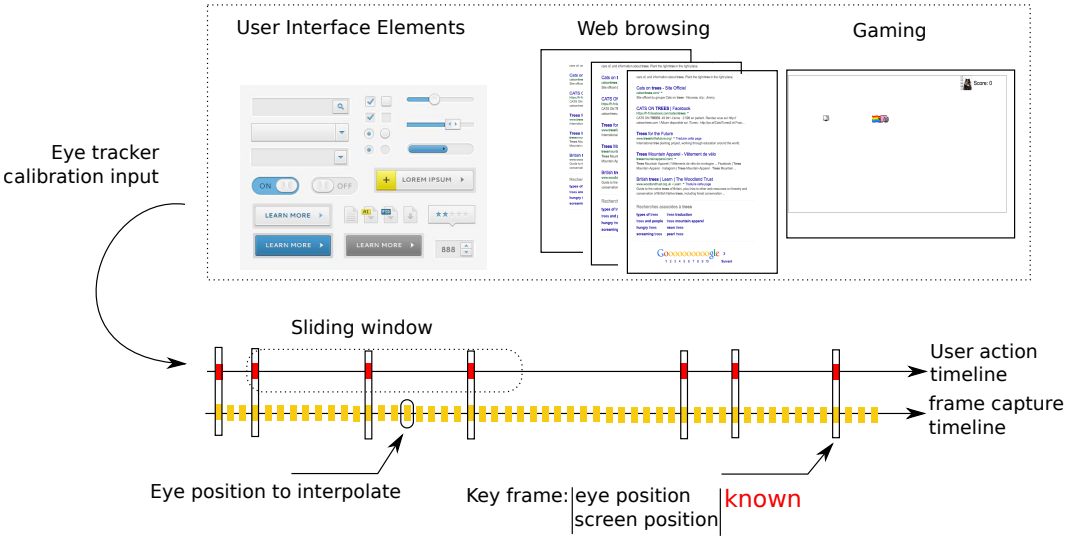

Figure 7. Gaze interpolation using a sliding window based on different kind of input for calibration.

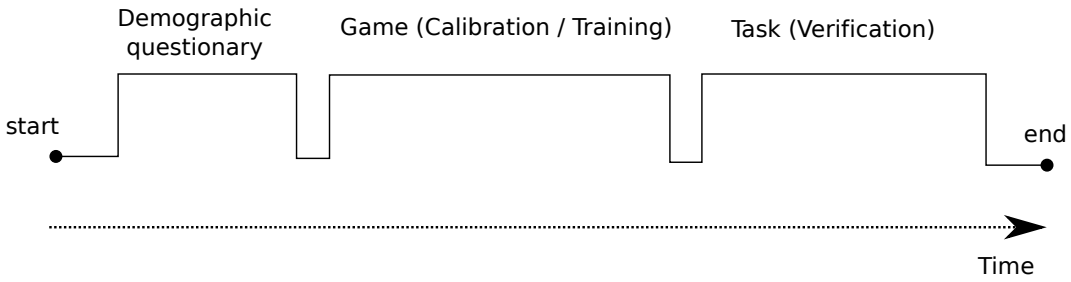

Figure 8. Experimental procedure.

to grant the application the access to the webcam and agree on recording a video of the user's face during the experiment. The first part of the actual test, illustrated in Figure 9, was a game where test participants had to click on moving objects on the screen, serving as the calibration phase, through which the mapping function was derived. The use of this game made sure that the test participant were dedicated to the task and were always looking at the mouse pointer when clicking. Moreover, the scoring system was favoring clicks with high inter-distances. The purpose of this guiding was to increase the distances between the consecutive eye positions and to cover larger area of the screen in order to achieve better calibration. Then, in the second part of the test, shown in Figure 10, the participants were presented with a web page comprising of disks of different colors. The participants were asked to click on all the red ones. This latter part was used to measure the proposed system's accuracy by following the "inverse" procedure; given the eye position data and the mapping function built, the ability of the algorithm to estimate the click positions was verified.

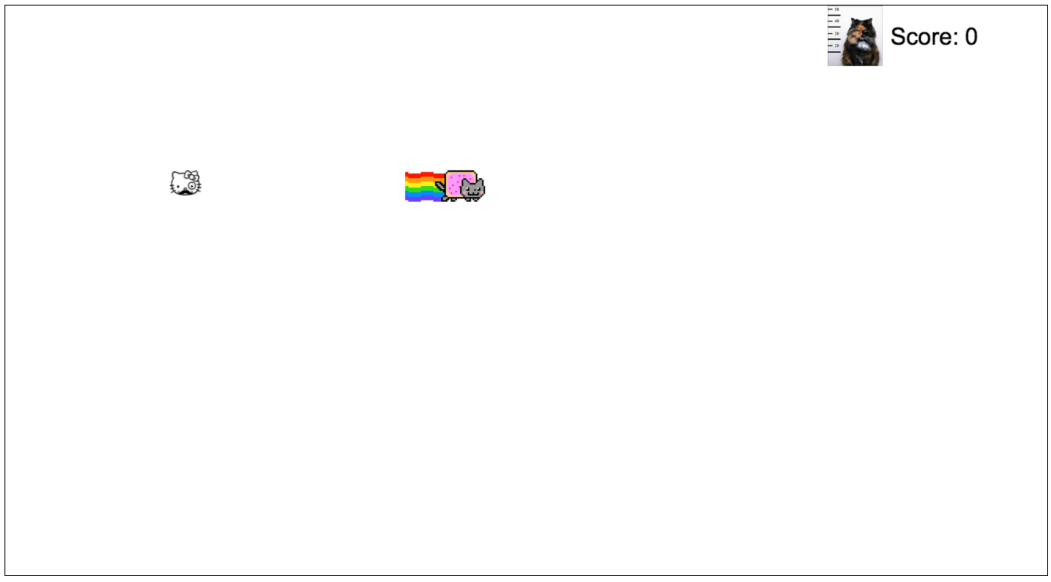

Figure 9. Calibration: User was asked to click on 15 cats as they appeared on the screen.

3.3 Screening of the test conditions

Considering the crowdsourcing context, test participants can be seated at various position in front of their screen and camera: they could be too far, or not visible to the camera (for example). Therefore, to increase the performance of the proposed framework, the viewing conditions are measured online and user is guided to move to acceptable position in front of the screen, before the actual test commences. To this end, a haar cascade classifier was integrated into the test interface and was used to determine a bounding box around the participant face. The size of this bounding box was used as an indication on the viewing distance, enabling requesting

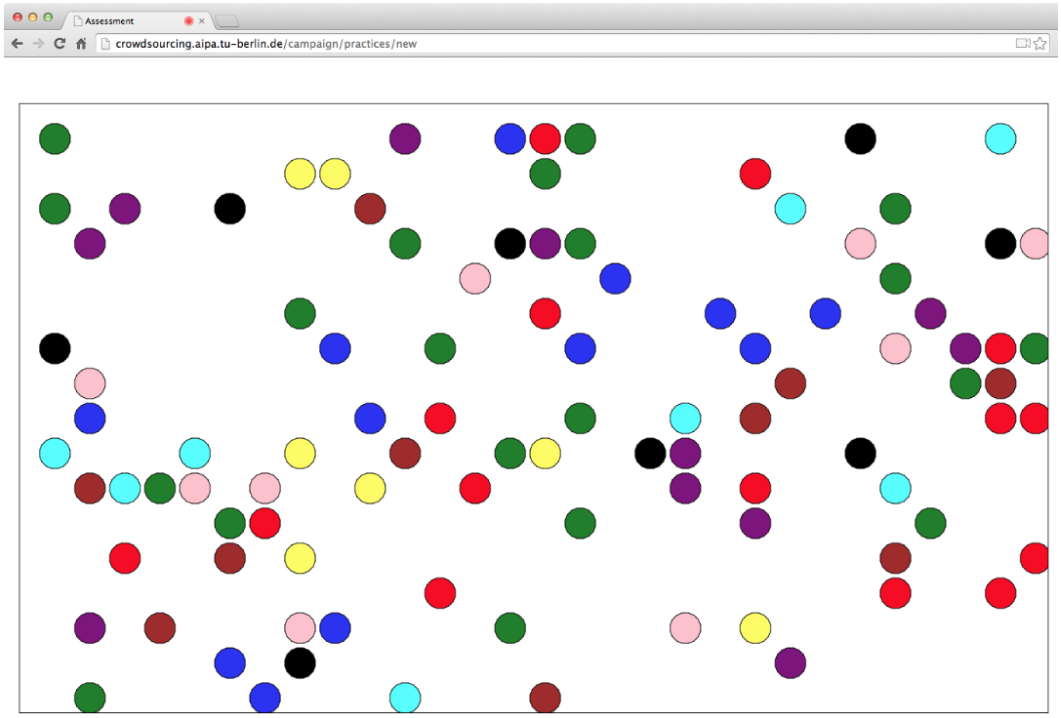

Figure 10. Second part: User was asked to click on all the red circles on the screen.

corrective user actions as required. The criteria for acceptable viewing distance is specified by following ratio:

$$Ratio = \frac{Face_{height}}{Screen_{height}} \tag{1}$$

In the current implementation, the ratio should be at least 0.375 (Figure 11 and 12).

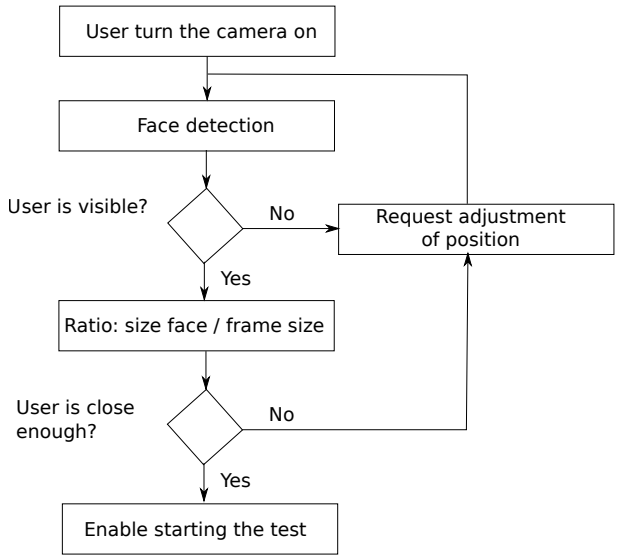

Figure 11. Test condition screening workflow.

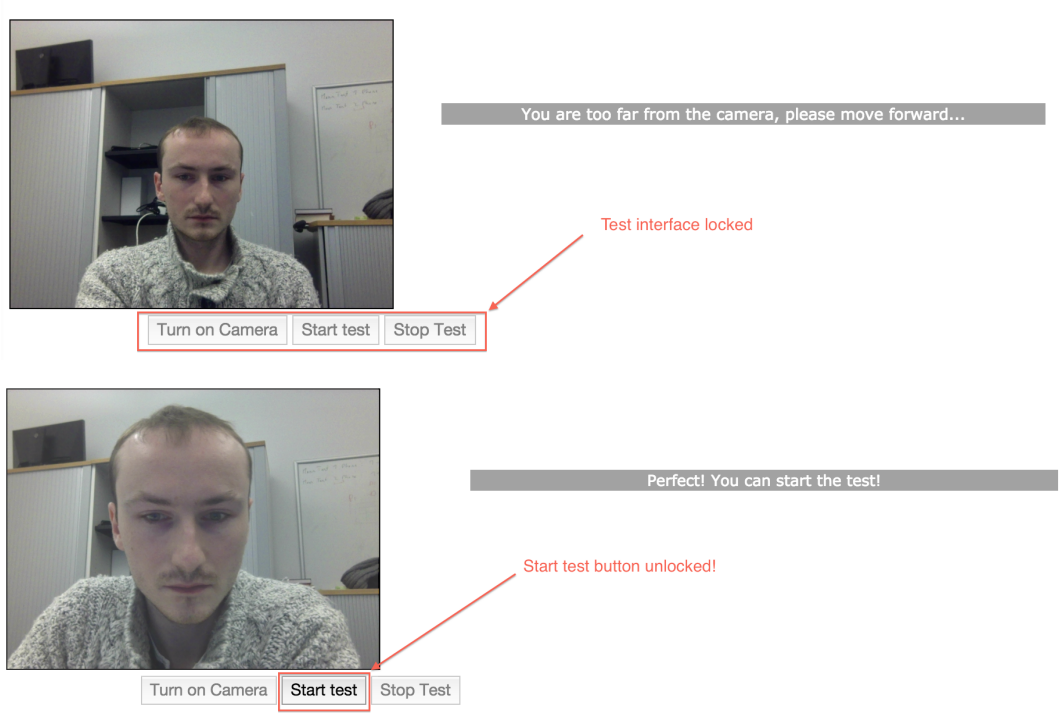

Figure 12. Test interface: Face detection, feedback to the user on his position, and activation of the test based on test conditions.

4. RESULTS

The performance of the proposed framework is promising. Even with rather low resolution of the video recorded by the webcam (640x480 pixels) and VP8 compression applied (at 1 Mbps because of networking constraints of a typical crowdsourcing user), meaningful results were acquired.

For the evaluation of the system's performance, the mean error percentage in relation to the screen width and height (in terms of pixels) was used. The use of the most common metric of angular deviation was precluded as it requires the distance of the user from the screen, as well as the screen size, which are very difficult to obtain in a crowdsourcing scenario.

The results for the first experiment involving the 8 test participants are depicted in Table 1. The average gaze estimation error is 13.9% of the user's screen width (horizontal axis) and 15.3% of the user's screen height (vertical axis) with a respective standard deviation of 11.3% and 13.2%. For the second crowdsourcing experiment the mean gaze estimation error is 14.1% in the horizontal axis and 17.9% in the vertical axis with respective standard deviations of 16.3% and 18.1% (Table 2). The improved performance of the first experiment was mainly attributed to the more optimal illumination conditions and the constrained head movements. As the experiments of the proposed gaze estimation system revealed, one of the most influential factors which greatly affects performance is the movement of the head. Although small head movements do not have very big impact on the overall accuracy, larger head movements can make the performance drop significantly (e.g subjects 7, 9, 12, 20 in Table 2). The presence of glasses (subjects 4, 6, 8 in Table 1 and subjects 5, 8, 11, 15 in Table 2) did not hinder the precise localization of the eye centers and thus did not influence negatively the overall system performance (not referring to cases of very strong reflections which very excluded in advance). In Figure 13, the 4th test participant illustrates the issue where bad lighting condition have resulted in poor performance. Likewise, the influence of the degree of head movement, one of the most influential factors which determines the

Subject #	1	2	3	4	5	6	7	8	**mean**
axis X	13.3 ±	15.2 ±	11.0 ±	18.0 ±	4.9 ±	16.8 ±	10.6 ±	20.7 ±	**13.9 ±**
	6.8 %	15.9 %	9.4 %	23.1 %	4.4 %	10.8 %	5.9 %	10.0 %	**11.3 %**
axis Y	10.8 ±	20.3 ±	17.3 ±	24.2 ±	14.4 ±	12.5 ±	8.9 ±	14.3 ±	**15.3 ±**
	7.2 %	23.0 %	10.4 %	27.9 %	10.4 %	5.5 %	7.3 %	9.7 %	**13.2 %**

Table 1. Result first dataset

Subject #	X	Y	Subject #	X	Y	Subject #	X	Y
1	11.1 ±	14.0 ±	7	17.6 ±	37.2 ±	13	3.8 ±	8.7 ±
	8.0 %	12.6 %		16.0 %	36.7 %		3.4 %	5.4 %
2	10.1 ±	19.6 ±	8	20.9 ±	17.0 ±	14	5.7 ±	10.3 ±
	10.5 %	13.2 %		11.3 %	19.3 %		3.9 %	10.0 %
3	21.8 ±	10.3 ±	9	39.4 ±	21.5 ±	15	6.1 ±	12.6 ±
	19.0 %	7.2 %		42.3 %	23.0 %		5.1 %	14.1 %
4	10.1 ±	23.3 ±	10	14.9 ±	19.0 ±	16	5.6 ±	18.0 ±
	9.1 %	35.8 %		10.8 %	21.1 %		6.2 %	21.6 %
5	12.4 ±	13.1 ±	11	9.6 ±	24.6 ±	17	12.1 ±	11.2 ±
	11.7 %	7.7 %		6.0 %	13.8 %		9.8 %	6.4 %
6	17.1 ±	27.3 ±	12	18.0 ±	22.5 ±	18	9.6 ±	23.3 ±
	15.5 %	19.9 %		14.9 %	18.0 %		9.5 %	22.2 %
19	12.2 ±	8.2 ±	20	23.5 ±	18.0 ±	**Mean**	**14.1 ±**	**16.3 ±**
	16.0 %	9.7 %		27.0 %	17.0 %		**16.3 %**	**18.1 %**

Table 2. Result second dataset

system's accuracy, is illustrated in Figure 14, for the 2nd test participant who had continuous and large head movements.

The system's performance is very satisfying considering that the evaluation was done on extrapolated data: during the validation period, no click information was captured from the user to update calibration data, whereas in the optimal application scenario it is planned to continuously update calibration key points (gaze/clicks pairs). Such condition was motivated by the fact that in some scenarios, user clicks may be rather sporadic. A clear limitation of the framework is the lack of control regarding the lighting conditions on the user side, as proper lighting is important in ensuring good eye-tracking performance. Exploiting computer vision techniques for monitoring lighting is envisaged for the future studies.

The average error reported demonstrates that the proposed approach is suitable for practically estimating the visual attention of a user in an unconstrained environment. Although the proposed webcam based gaze tracker presents inferior performance compared to active light approaches, it can be applicable to situations where accuracy can be traded off for low cost, large sample size, simplicity and flexibility.

5. CONCLUSIONS AND CONTINUATION

In this paper it was shown that extended measure of user's experience could be achieve in a crowdsourcing context. A novel framework enabling studying attention was presented. Result have shown that, in a non-intrusive manner and without specialized equipment, it is possible to estimate visual attention with a precision as high as 14.1% in the horizontal axis and 17.9% in the vertical axis. This shows how new measurements can be done efficiently in non-controlled environments.

Different topics will be addressed in future work:

- The methodology will be validated by applying it to realistic web content.
- The methodology will be applied to facial expression recognition (keeping the analysis on server side). To that end, we will introduce emotional contents in our webpage, such as emotional video sequences, to elicit strong facial emotions. This will allow us to obtain emotional feedback globally from the variety of demographically differing groups.

- Some enhancements to the platform are also envisaged. Namely, the monitoring of lighting conditions is expected to improve the accuracy of the system and also to facilitate evaluating the user's environment.

6. ACKNOWLEDGMENTS

The authors would like to thank Microworkers for funding the different experiments which were conducted and Qualinet COST action for facilitating the collaboration.

REFERENCES

[1] Roth, S. P., Tuch, A. N., Mekler, E. E., Bargas-Avila, J. A., and Opwis, K., "Location matters, especially for non salient features an eye tracking study on the effects of web object placement on different types of websites," *International Journal of Human-Computer Studies* **71**, 228–235 (2013).

[2] Jacob, R. J. K., "The use of eye movements in human-computer interaction techniques: what you look at is what you get," *ACM Transaction Inf. Syst.* **9**(2), 152–169 (1991).

[3] Chen, K. T., Chang, C. J., Wu, C. C., Chang, Y. C., and Lei, C. L., "Quadrant of euphoria: a crowdsourcing platform for qoe assessment," *Network, IEEE* **24**(2), 28–35 (2010).

[4] Hossfeld, T., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P., and Schatz, R., "Quantification of youtube qoe via crowdsourcing," in [*Multimedia (ISM)2011 IEEE International Symposium on*], (2011).

[5] Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B., "Who are the crowdworkers?: shifting demographics in mechanical turk," in [*CHI'10 Extended Abstracts on Human Factors in Computing Systems*], 2863–2872 (2010).

[6] Hirth, M., Hossfeld, T., and Tran-Gia, P., "Anatomy of a crowdsourcing platform using the example of microworkers. com. in innovative mobile and internet services," in [*Ubiquitous Computing (IMIS) Fifth International Conference on*], 322–329 (2011).

[7] Kittur, A., Chi, E. H., and Suh, B., "Crowdsourcing user studies with mechanical turk.," in [*SIGCHI conference on human factors in computing systems*], 453–456 (2008).

[8] Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., and Tran-Gia, P., "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *Multimedia, IEEE Transactions on* **16** (2013).

[9] Rudoy, D., Goldman, D. B., Shechtman, E., and Zelnik-Manor, L., "Crowdsourcing gaze data collection," in [*Collective Intelligence conference*], (2012).

[10] Huang, J., White, R. W., and Buscher, G., "User see, user point: Gaze and cursor alignment in web search," in [*CHI '12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*], 1341–1350 (2012).

[11] McDu, D., Kaliouby, R. E., and Picard, R. W., "Crowdsourcing facial responses to online videos," *IEEE Transactions on Affective Computing* **3**, 456–468 (OCTOBER-DECEMBER 2012).

[12] Viola, P. and Jones, M. J., "Robust real-time face detection," *International Journal of Computer Vision* **57**, 137154 (2004).

[13] Skodras, E. and Fakotakis, N., "An accurate eye center localization method for low resolution color imagery," *IEEE 24th International Conferenceon Tools with Artificial Intelligence (ICTAI)* **1**, 994997 (2012).

[14] Baker, S. and Matthews, I., "Lucas-kanade 20 years on: A unifying framework," *International Journal of Computer Vision* **56**(3), 221255 (2004).

[15] Murphy-Chutorian, E. and Trivedi, M. M., "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(4), 607626 (2009).

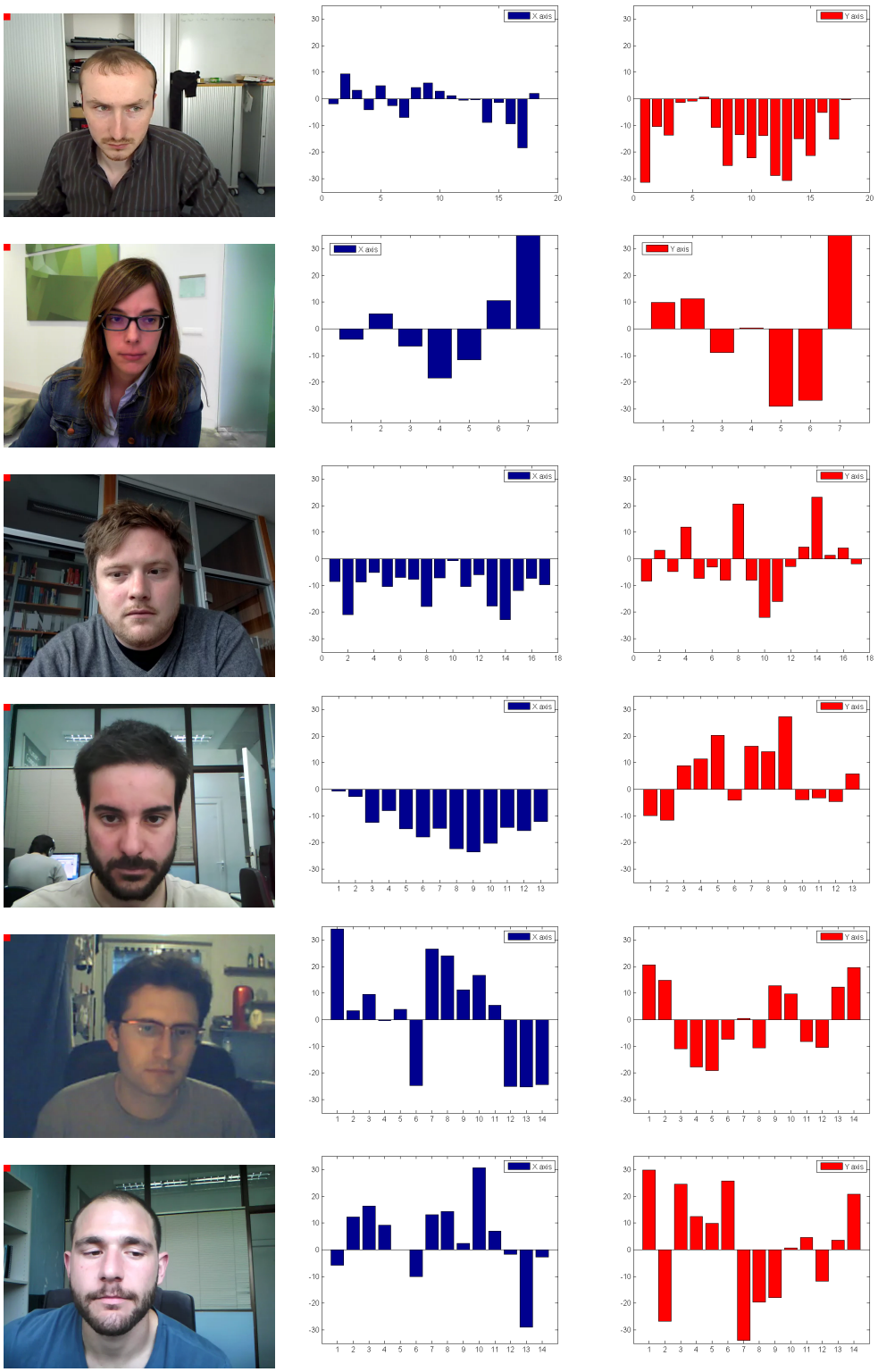

Figure 13. Different cases of performance results (with explicit permission from users in first campaign).

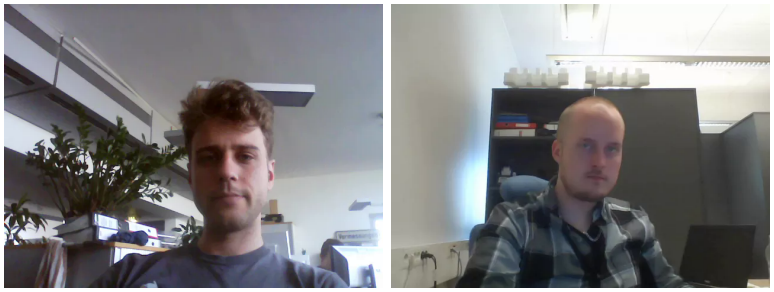

Figure 14. Case of failures. The participant on the left show an example of bad illumination conditions (uneven illumination) and the one on the right showed large head movements.