This is a pre-print. The paper will be published in the proceedings of the 7th IEEE International Conference on Human System Interaction and will be available in IEEE Xplore.

Idiosyncratic repeatability of calibration errors during eye tracker calibration

Katarzyna Harężlak and Pawel Kasprowski and Mateusz Stasch
Institute of Informatics
Silesian University of Technology
Gliwice, Poland
Email: katarzyna.harezlak@polsl.pl, kasprowski@polsl.pl, mateusz.stasch@polsl.pl

***Abstract*—Dynamic development of high quality cameras and algorithms processing eye movement signals entails growing interests in using them in various areas of human-computer interaction. Determining subjects which user is looking at or controlling the operation of computer processes can serve as examples of these areas. However, making eye movement signal valuable requires some preparatory steps to be taken. They belong to a process called calibration aiming at creating a model for mapping output delivered by an eye tracker to user's gaze points. The quality of such model is assessed based on a calibration error defined as a difference between accurate data and this obtained from a model. The goal of the research presented in the paper was to analyse to what extent the calibration error depends on the specific participant's features - it is repeatable – or to what extent it may be avoided during the recalibration. Additionally an influence of two calibration method a polynomial and an artificial neural network (ANN) on the final results were studied as well.**

Keywords: eye movement, face recognition, data mining

I. Introduction

Eyes are one of the most complicated human organs and the analyses of eye movements may reveal a lot of information about the human being. Eye movements may be used to communicate with a computer environment, which may adapt its behavior according to the user's gaze directions [1]. The first and most important element of such communication, which highly influences all subsequent tasks, is properly realized calibration process [2]. In most gaze-directed environments it is crucial to precisely determine where a user is currently looking. Although there exist eye trackers, which don't need a calibration, their setup is very complicated. Usage of a typical eye tracker requires a prior calibration process, during which an eye tracker raw output being an eye signal of an examined person looking at a set of points with known positions (so called Points of Regard of PoRs) is collected [3]. This output is subsequently used to build a model for specifying user's gaze points for unknown areas. The quality of such model is assessed based on a calibration error defined as a difference between accurate data and this obtained from a model. If the error is sufficiently low, an assumed interaction can start. In other case - the error is too high - the user must be recalibrated or, if it does not cause expected improvement, an interaction has to be given up. The goal of the research presented in the paper was to analyse to what extent the calibration error depends on the specific participant's features – it is repeatable – or to what extent it may be avoided during the recalibration.

II. The state of the art

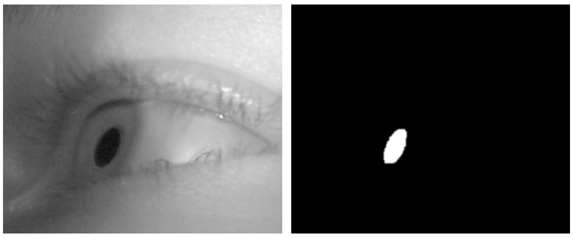
As it was stated in the previous section, properly performed calibration is an essential part of every activity involving eye movement signal. The calibration errors calculated during the verification stage, are very often used to determine whether a quality of data obtained during an experiment is sufficient. But what should be done if this is not the case? Many researchers use information about calibration results to remove samples with low quality [4]. Surprisingly, the calibration process has not attracted much attention in the eye movement research. Developing a calibration scenario in most cases is a responsibility of the eye tracker producers – researcher just analyses output from manufacturer's calibration procedure.

When preparing one's own calibration scenario it is important to consider how many points – stimuli – to take into account and in which locations they should be presented [5]. Another important choice, which can influence final results is selecting a method used to map data provided by an eye tracker to a user's gaze point [3]. There can be various interpolation functions, including polynomial ones with different degrees and number of terms [6][7][8], artificial neural networks [3][9][10] or Support Vector Regression [11] analyzed.

The results of calibration can also depend on a user [4] and, surprisingly, may even be conditional upon the operator's experience [12]. Moreover, there are users for whom is difficult to conduct an appropraite calibration process due to several factors like: glasses, contact lenses, mascara or dropping lids [12]. Most of them may be easily avoided (e.g. by getting off glasses or removing mascara). However, it is not obvious whether there are any physiological or behavioral features that make the specific person constantly reluctant to a proper calibration. This paper tries to study that problem, comparing calibration information collected during three different sessions involving the same group of the participants.

III. Research environment

The experimental setup of the tracking system used in the research is shown in the figures 2 and 3. The main its component was a VOG head-mounted eye tracker developed with single CMOS camera with USB 2.0 interface (Logitech QuickCam Express) with 352x288 sensor and lens with IR-Pass filter. The camera was mounted on an arm attached to a head and was pointing at the right eye. The eye was illuminated with a single IR LED, placed off the axis of the eye that causes the "dark pupil" effect, which was useful during pupil detection. The image obtained from the camera is converted to grayscale. Subsequently it is tresholded to change the darkest points of the image into white ones (figure 1). For such processed image the contour detection algorithm is applied to find the smallest convex polygon surrounding the white shape. The center of gravity of this polygon is a estimated center of a pupil.The system generates 20 - 25 measurements of the centre of the pupil per second.



(a) Grayscale image (b) Thresholded image

Fig. 1: Pupil detection algorithm

The calibration was done on a 1280x1024 (370mm x 295mm) flat screen. The eye-screen distance was equal to 500mm and vertical and horizontal gaze angles were 40° and 32° respectively. To avoid movements, the head was stabilized using a chin rest.

The experiments were repeated three times using a set of 29 points distributed over the screen as presented in the figure 4. In each session, points were displayed in the same, predefined order. Each point was displayed for 3618 msec. and it was pulsating in order to keep the user focused on it. The time interval between sessions for one user was at least three weeks to avoid the learning effect - when user learns the order of the points and is able to anticipate the next point position. All session took place in the same day in a week, the same time during a day and in the same room. Therefore, it can be said that the conditions of the experiments were comparative. Altogether 24 participants took part in the experiments. Some of them were involved only in two of three sessions, but most of users participated in all of them. Before each experiment, participants were informed about the general purpose of the experiment after which they signed a consent form. As it was said, the participants numbers differed between sessions and not all of them took part in all sessions. This is why the sets of the participants for session 1 and 2, 1 and 3 and 2 and 3 had to be made uniform. Their final numbers are presented in table I.

TABLE I: Number of the participants involved in particular sessions

Sessions	Participants number
1 and 2	20
1 and 3	20
2 and 3	18

The participants' eye positions reflecting points of regard (PoR) presented on a screen, acquired during particular sessions, were used to define two types of models mapping an eye position of a participant to PoRs for unknown samples (in the same session). The first of them was based on a commonly used the second order polynomial function [6][7], for which values of $A_x \dots E_x$ and $A_y \dots E_y$ parameters (1) were calculated using a classic Levenberg-Marquardt optimizer [13].

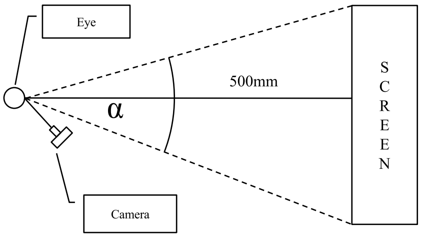


Fig. 2: The architecture of the system

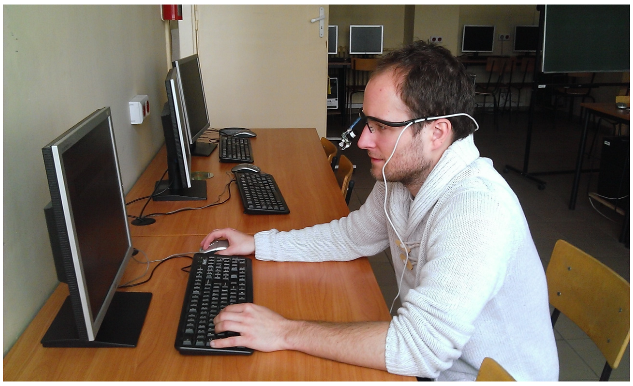


Fig. 3: Photo of the measurement setup

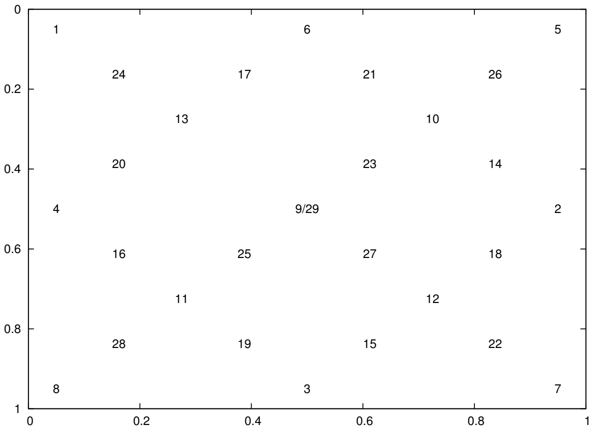


Fig. 4: Layout and order of calibration points

$$\begin{aligned} x_s &= A_x x_e^2 + B_x y_e^2 + C_x x_e + D_x y_e + E_x \\ y_s &= A_y x_e^2 + B_y y_e^2 + C_y x_e + D_y y_e + E_y \end{aligned} \quad (1)$$

where x_s and y_s are estimated gaze coordinates on a screen.

The second type of mapping was performed using an artificial neural network (ANN). An activation network with sigmoid function as an activation function was used. The network was trained using the Back Propagation algorithm with normalized samples recorded during a session. Configuration of the network consisted of two neurons in the input layer, 10 neurons in one hidden layer and two neurons as the output. The network was trained until the total train error was lower than 0.1. ANN has been already used in several eye tracing applications [10][9]. In the first step of the research the models built using these functions were based on all points presented to the participants - for each session independently. These models were subsequently tested using the same set of points. Their quality was verified by computing the distance between accurate positions of the displayed points and their locations obtained from a specific model. This factor, expressed in degrees, was calculated according to the equitation 2.

$$E_{deg} = \frac{1}{n}\sum_{i}\sqrt{(x_i - \widehat{x_i})^2 + (y_i - \widehat{y_i})^2} \qquad (2)$$

where y_i, x_i represent observed values, $\widehat{x_i};\widehat{y_i}$ represent value calculated by a model.It must be emphasized that it takes some time for an eye to react to a stimulus position change to fixate on another position. Such occurrence is called *saccadic latency* and lasts approximately 100-300 msec [14]. During earlier experiments (not published yet), it was calculated that the safest range of measurements to include for further studies is obtained between 700 msec. and 1800 msec. after the stimulus position changed. Therefore, only these measurements were taken into account in both training and validation phases.

IV. Results correlation

E_{Deg} values evaluated in the step described above were used to assess the codependency of the results acquired in particular sessions. Based on its values the correlation coefficients between session 1 and 2, session 1 and 3 and session 2 and 3 were calculated. Analyzing outcomes presented in table II, high correlation between results obtained in the first and the second as well as the first and the third sessions can be noticed, especially for models based on the polynomial function. Similarly, high correlation can be observed for the first and the third session in case of the ANN method. The correlation of results calculated by this method in case of the first and the second sessions along with results obtained by the polynomial function in case of the second and the third sessions can be considered as meaningful. Only the results related to the ANN method for the second and the third sessions turned to be uncorrelated.

TABLE II: Values of the correlation coefficients for particular sessions

Sessions	Polynomial method	ANN method
1 and 2	0,626345	0,487624
1 and 3	0,673711	0,594534
2 and 3	0,354834	0,125152

Statistical significance of the computed results was studied, taking the lack of correlation as a hypothesis H0 and p=0,05 as the significance level. The conduced tests confirmed most of the outcomes, by rejecting hypothesis H0 in 4 out of 6 instances (Table III). In case of session 2 and 3 and the polynomial function, the hypothesis H0 could not be rejected, yet it is worth emphasizing that the achieved value was close to the boundary level.

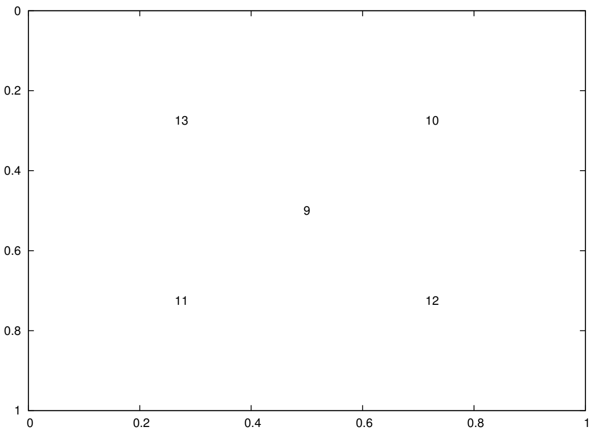


Fig. 5: Layout of the set 5-1

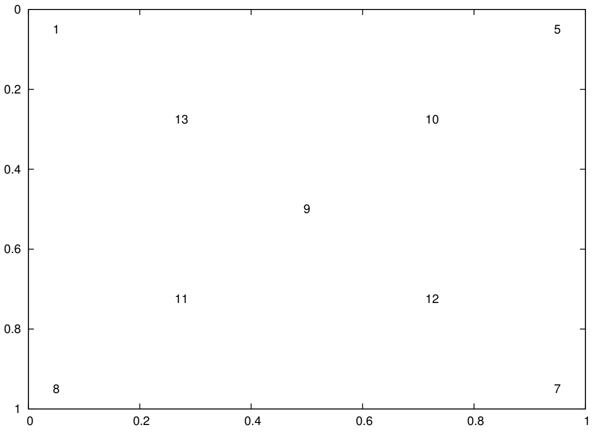


Fig. 6: Layout of the set 9-1

TABLE III: The results of the statistical tests of the calibration errors correlation

Sessions	P-value	
	Polynomial method	ANN method
1 and 2	0,000779	0,017528
1 and 3	0,000209	0,001645
2 and 3	0,069114	0,398373

V. Results repeatability over sets of points

The promising results, obtained in the previously described studies, were encouraging to more detailed analyses of captured samples. Therefore, in a collection of all 29 points used in experiments, sets differing in a number and layouts of points were distinguished. There were 61 sets defined and sets presented in figure 5 and 6 can serve as examples. Due to limited space, the detailed description of all sets is not presented here.

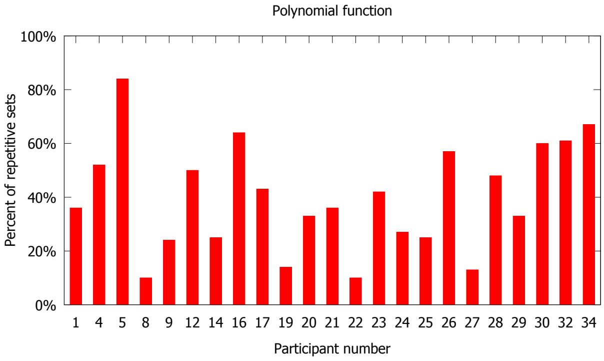


Fig. 7: Percentage of repetitive sets for various participants - The polynomial function

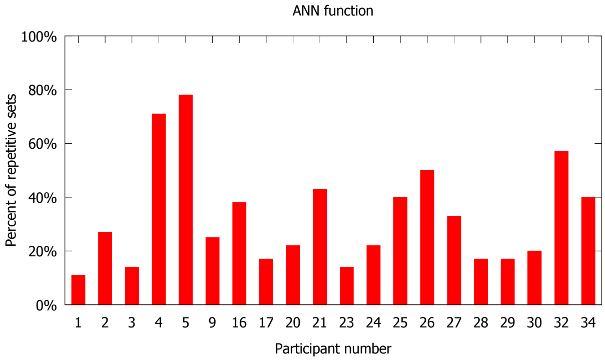


Fig. 8: Percentage of repetitive sets for various participants - The ANN function

Samples related to a particular set and a particular participant were used to define calibration models taking the polynomial (equitation 1) and ANN methods into account. All models were checked using 16 testing points from the same session.

Models assessment, once again, was performed using E_{Deg} error values (equitation 1). These values were studied to check if the best results, defined as ones with the lowest error degree, in all sessions are related to the same sets of points. Because of the fact that results of some sets were very close to each other, it was assumed that they would be treated equivalently. It regarded the results, for which the difference between ordered ascending E_{Deg} values were lower than 0.5 degree. Such rule was applied for all participants, all sessions and both methods described earlier. Selected sets of points will be further referred as the *best results set*. The first finding of the aforementioned studies was an evaluation of participants number, in case of which the same set of points was found in more than one of the *best results sets*. It amounted to 88% of the participants in case of the polynomial function usage and 77% for the second of functions. Subsequently, a percentage ratio of a number of repeated elements to a number of distinct elements in the *best results sets* obtained for all three sessions was analyzed. This analysis was conducted for each participant and for each regression function independently. The obtained results are presented in the figures 7 and 8 for the polynomial and the ANN functions respectively. The values seen on the OX axis represent a number associated with a participant, while the OY axis represents the percentage of repetitive sets for a given participant. It can be observed that in case of polynomial function results differ significantly: the minimal ratio representing a recurrence of sets is 10% for participant marked with number 22 and its maximal value 84% is related to person with number 5, averaging on 40%.

Similar results, although a little bit worse, were obtained for the second of functions. Minimal value amounts to 11% for a participant number 1, maximal value - 78% is related to person number 5 and the average value is 33%.

Summarizing the obtained results, they were divided into four groups taking calculated ratio as a criterion. The participants with ratio falling between 0 and 20% constituted the first group. Similarly, the next four groups were defined for participants related to ratios belonging to (20% , 40%], (40%, 60%] and (60%, 80%], (80%, 100%] partitions respectively. Data presented in table IV indicates that in case of polynomial function, the highest percentage of the participants was related to ratio belonging to scopes of 20% - 40% and 40% - 60%. of the sets repetitiveness. The second of functions provided outcomes, which placed the majority of the participants in the ratio partitions defined by scopes of 0% - 20% and of 20% - 40%, which is worse results in a comparison with the first function.

TABLE IV: Percentage of users with repetitive results in a given scope

Partition scope	Percentage of users with a given scope of repetitiveness	
	Polynomial method	ANN method
$\leq$ 20%	17%	35%
>20% and $\leq$40%	35%	40%
>40% and $\leq$60%	30%	15%
>60% and $\leq$80%	13%	10%
>80%	4%	0%

VI. Analysis of the Results

To start the analysis of the results the main questions asked at the beginning of the paper: (1) to what extent the calibration error depends on the specific participant's features or (2) to what extent it may be avoided during the recalibration - have to be answered.

Experiments leading to gain such knowledge included a triple calibration procedure. The collected data was used to build calibration models using two functions – the polynomial with a second order and the ANN one. The calibration accuracy was verified using error values equal to a degree distance between an accurate and a calculated position of a point.

The values were firstly used to check the correlation of data originating from various sessions. As it was presented earlier, the estimated coefficients, confirmed by statistical tests, indicated the moderate or significant correlation between sets of values obtained for first and second sessions and first and third ones. These results regarded both of the analyzed methods. Conclusions, which can be drawn based on these studies, are twofold. On one hand they revealed relation between calibration errors. On the other hand, the strength of that correlation, not reaching highest possible values, in conjunction with weak correlation between outcomes obtained for second and third sessions, shows that there are still opportunities to improve an eye movement signal acquisition. This can be achieved, for example, by triggering recalibration process.

Such ambiguity of the research findings encouraged us to conduct further analysis. It was interesting to find out the extent to which the general results are reflected in data of particular users. To make analysis more detailed the set of 29 points used in the experiment was divided in smaller groups differing in numbers and layouts of their elements. The aim of this detailed studies was to check how repetitive is registered eye movement signal when various calibration scenarios were taken into account. The attention was focused only on this scenarios, for which values of calibration error were equal or lower than 0.5 degree.

The analysis was conducted from two points of view. At first, it was checked how many participants obtained considered calibration error for the same scenarios in different sessions of the experiment, yet these scenarios did not necessarily have to be the same for various participants. Evaluated percentage of such users for both studied methods (polynomial and ANN ones) was quite high and amounted 88% and 77% respectively. It can indicate that human's eye movement signal, if registered with appropriately adjusted scenario, can make calibration procedure highly repetitive. An influence of a chosen calibration method has to be emphasized as well. After that, recalibration can make collected data more reliable.

These conclusions seem to be confirmed by the analysis done from the second point of view. It is represented by studies of the percentage ratio describing, independently for each participant, number of repeated sets in relation to the number of distinct elements appearing in three sessions. These studies revealed high diversity of the ratio, which varied from 10% to 80% for both methods, with an average value of 40% for the polynomial function and 33% for the ANN one. Such unstable results indicate that any of the calibration processes was influenced by some problems and some studies have to be done to eliminated these obstacles. Providing the knowledge in how many cases recalibration should be done was the aim of the last type of analysis conducted during the research. As it was presented in table IV majority of results representing calibration error repetitiveness were, for polynomial function, classified in the range from 0% to 60%. For these group of users, it can be expected that recalibration process for primarily obtained high calibration errors, can improve quality of data gathered during subsequent measurements of an eye movement signal. Repetitiveness higher than 60% can suggest that the value of a calibration error can recur, which is a good information when it is low and bad news in other case.

Considering the result for the ANN methods it can be reasoned that this function, for the eye tracker used in the experiment, provided less repetitive results. This makes it a less stable method entailing, with higher probability, triggering a recalibration process more frequently.

VII. Summary

The research presented in the paper aimed at determining whether a repeated calibration process is characterized by the same accuracy in each of its occurrences. This accuracy is expressed by a calibration error being the distance between actual position and measured user's gaze point. Values of this error were used to compare the quality of data acquired for the participants taking part in three sessions of the experiment. To test various scenarios the different sets of points and two calibration methods were analyzed.

The main goal of these activities was to check whether a participant reproduces the stable data over various sessions or collected samples vary between trials. Repetitive results of the calibration error indicate that one calibration attempt is sufficient to decide whether particular user should take part in the subsequent tasks of the experiment or not. In the second case retriggering calibration process can lead to improvement of calibration results. However this recalibration should not be repeated infinitely. It is not a big problem when eye movements are collected under specialized operator supervision. In case of too high error values an operator may suggest ways to improve results such as, for example, removing mascara. But the results presented in this paper are very important when developing human computer interfaces that are intended to be used without any supervision. Improperly developed interface may try to constantly recalibrate user for which it is just impossible to obtain correct results. In such case, some overlay algorithms should be proposed.

To conclude the research, it is important to wonder if the studies provided answers for questions asked at the beginning of the paper. The response, which can be given is not as clear as one would like to obtain. On one hand, they proved that there are some people, who are able to keep their eye movement signal on almost the same level when an appropriate scenario and method is used. It simplifies making decisions whether a user should or not be excluded for experiment. On the other hand, the majority of the participants were related to less, but still meaningful, repetitiveness of the results. For them it can be useful to restart calibration process with the same setup or with a changed calibration scenario.

Finally, it has to be emphasised that during the research an eye movement signal was collected using a simple eye-tracking environment, accessible for ordinary users. It can be expected that usage of higher quality eye-trackers can provide better results than these presented in the paper. Yet, because of the cost necessary to bear, they are inaccessible for many users and for those people outcomes presented in the paper may seem to be useful.

Findings of the described research indicate that consecutive studies should be realized. The idea of a recalibration algorithm has to be elaborated. Additionally, existence of people for whom collaboration with eye trackers is difficult, entitles a necessity of searching for reasons and methods, which solve this problem.

Another possible extension of the current work may be based on the fact that there were participants, representing the fourth part of the population, for whom repeatability between sessions reached 60% - 100%. This finding could be used as the indicator that eye movements are personally distinctive and - what is even more important - repeatable. However, all of the intended studies require involving more participants.

References

[1] R. J. K. Jacob, "The use of eye movements in human-computer interaction techniques: What you look at is what you get," *ACM Transactions on Information Systems*, vol. 9, pp. 152–169, 1991.

[2] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 4, pp. 455–470, 2002.

[3] P. Kasprowski, K. Harężlak, and M. Stasch, "Guidelines for the eye tracker calibration using points of regard," *Information Technologies in Biomedicine*, 2014.

[4] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer, *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[5] X. Brolly and J. Mulligan, "Implicit calibration of a remote gaze tracker," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, 2004, pp. 134–134.

[6] P. Blignaut and D. Wium, "The effect of mapping function on the accuracy of a video-based eye tracker," in *Proceedings of the 2013 Conference on Eye Tracking South Africa*, ser. ETSA '13. New York, NY, USA: ACM, 2013, pp. 39–46.

[7] J. J. Cerrolaza, A. Villanueva, and R. Cabeza, "Taxonomic study of polynomial regressions applied to the calibration of video-oculographic systems," in *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ser. ETRA '08. New York, NY, USA: ACM, 2008, pp. 259–266.

[8] N. Ramanauskas, "Calibration of video-oculographical eye-tracking system," *Electronics and Electrical Engineering*, vol. 8, no. 72, pp. 65–68, 2006.

[9] Z. Zhu and Q. Ji, "Eye and gaze tracking for interactive graphic display," *Machine Vision and Applications*, vol. 15, no. 3, pp. 139–148, 2004.

[10] K. Essig, M. Pomplun, and H. Ritter, "A neural network for 3d gaze recording with binocular eye trackers." *IJPEDS*, vol. 21, no. 2, pp. 79–95, 2006.

[11] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[12] M. Nyström, R. Andersson, K. Holmqvist, and J. van de Weijer, "The influence of calibration method and eye physiology on eyetracking data quality," *Behavior research methods*, vol. 45, no. 1, pp. 272–288, 2013.

[13] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.

[14] J. H. Darrien, K. Herd, L.-J. Starling, J. R. Rosenberg, and J. D. Morrison, "An analysis of the dependence of saccadic latency on target position and target characteristics in human subjects," *BMC neuroscience*, vol. 2, no. 1, p. 13, 2001.

[15] C. H. Morimoto and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications," *Comput. Vis. Image Underst.*, vol. 98, no. 1, pp. 4–24, Apr. 2005.

[16] A. Villanueva and R. Cabeza, "Models for gaze tracking systems," *Journal on Image and Video Processing*, vol. 2007, no. 3, p. 4, 2007.