Comparative Analysis of RGB-based Eye-Tracking for Large-Scale Human-Machine Applications

Brett Thaman[1], Trung Cao[1], Nicholas Caporusso[1]

[1] Department of Computer Science, Northern Kentucky University,

Louie B Nunn Dr, 41099 Highland Heights, United States

{thamanb1, caot1, caporusson1}@nku.edu

ABSTRACT

Gaze tracking has become an established technology that enables using an individual's gaze as an input signal to support a variety of applications in the context of Human-Computer Interaction. Gaze tracking primarily relies on sensing devices such as infrared (IR) cameras. Nevertheless, in the recent years, several attempts have been realized at detecting gaze by acquiring and processing images acquired from standard RGB cameras. Nowadays, there are only a few publicly available open-source libraries and they have not been tested extensively. In this paper, we present the result of a comparative analysis that studied a commercial eye-tracking device using IR sensors, that is, Tobii 4C and WebGazer, a software system that uses machine learning and linear regression to estimate gaze from images acquired by a standard webcam. From our findings, we can conclude that, despite the advancements in artificial intelligence and computer vision, gaze tracking using IR sensors is still significantly more accurate than RGB webcams. Specifically, the software library tested in our work is not suitable for gaze tracking tasks that require accuracy and reliability.

Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems
https://openaccess.cms-conferences.org/#/publications/book/978-1-7923-8988-7

Keywords: Human-Computer Interaction, Eye-tracking, Gaze-tracking, Machine Learning.

INTRODUCTION

Gaze tracking (GT) consists in estimating the point at which the user is looking on a computer screen. Specifically, the ability to continuously locate the user's gaze on the display in real-time has several applications. In the last decades, gaze tracking has been effectively utilized in several research studies in computer science and, particularly, human-computer interaction (Caporusso et al., 2020) as well as in medical fields such as neuroscience, and psychology (Duchowski, 2002). Nevertheless, the feasibility of applications based on gaze tracking depends on the degree of accuracy of the acquisition device and method. To this end, nowadays, most systems implementing gaze tracking features require the use of dedicated auxiliary devices that have the purpose of acquiring eye movements using infrared (IR) projectors and sensors. IR-based gaze tracking devices take the form of either head-mounted frames or bars that are attached to the display. Unfortunately, current wearable systems are cumbersome, and they are not suitable for large-scale applications due to their high price (Cognolato et al., 2018) (Caporusso et al., 2019). The alternative option, that is, infrastructure-based devices, is more affordable and unobtrusive for the user, though it does not support pervasive interaction. Regardless of the type of eye-tracking technology, the requirement of dedicated hardware limits the deployment of large-scale applications based on gaze tracking. In the recent years, thanks to the advances in computer hardware and image processing algorithms, several groups focused on the development of gaze tracking systems based on standard RGB cameras such as the webcams commonly found embedded in personal computers. Accomplishing gaze tracking without requiring any additional hardware would enable the development of unintrusive and cost-efficient systems and would support large-scale human-computer interaction applications. Although several research studies focused on democratizing access to GT by improving the speed, accuracy, and reliability of RGB sensors, nowadays there are only a few systems available, and their performance has not been tested extensively. As a result, despite its applicability in a variety of scenarios and the potential of GT in improving Human-Computer Interaction (HCI) tasks, this technology has been integrated into a few applications, only. The goal of our work is to achieve reliable and high-performance GT using standard RGB cameras, extend the potential user base of GT and foster the development of large-scale HCI applications that do not require additional and dedicated hardware. In our research, we primarily focused on infrastructure-based GT. We analyzed currently available GT systems based on IR sensors and RGB cameras and we compared their performance. In this paper, we present a detailed report of our findings, we describe the main issues and challenges in realizing GT using RGB cameras.

RELATED WORK

For over two decades, using an external device was the only option for GT applications. Although the methods that use auxiliary systems such as IR sensing (Sun et al., 2015) or RGB-D cameras (Liu & Zhu, 2012) are accurate, they present an issue where the applications are available to a limited number of users who have these devices. However, in the last years, the increase in the definition of embedded cameras and computational power of laptops and personal devices resulted in the development of machine learning algorithms that are particularly suitable for image processing and computer vision tasks with standard webcams (Bevilacqua et al., 2015). This, in turn, has unlocked a new era in software-based eye tracking: nowadays, algorithms can identify the user's pose and face position in a set of images, and they can accurately draw the bounding boxes surrounding each of the eyes. Subsequently, GT requires estimating the x and y coordinates representing the point where the user is looking on the display. Nowadays, several commercial systems offer webcam-based GT. Indeed, the main challenge of webcam-based GT is achieving a level of accuracy that is comparable with IR devices. For instance, the authors of (Burton et al., 2014) realized a comparative analysis of the performance of IR-based systems and webcams in simple GT tasks such as staring at target objects (i.e., images) on the screen, using two proprietary technologies, that is, SMI infrared and Tobii's Sticky webcam eye-tracking system. The study reported a high accuracy (i.e., ranging from 81% to 100%) of webcam-based systems when using large targets (i.e., 33% of the width or height of the screen). In contrast, the performance score of the IR sensor was 100%. However, the study also reported that the average accuracy of webcam-based GT was significantly lower than that of IR sensors when using smaller targets (i.e., 10% of the width of the screen), with accuracy values ranging from 41-78% for Tobii's Sticky webcam, compared to the SMI infrared system, which scored 81-100%. As a result, the study found that webcam-based GT is not suitable for tasks that require finer control, due to their poor performance and reliability. Nonetheless, over the last years, the scientific community has realized multiple attempts at developing new systems aimed at improving the performance of GT (Xu et al., 2015). For instance, WebGazer (Papoutsaki et al., 2016) is a system that aims at rendering webcam-based GT available in web applications. It utilizes TensorFlow's landmark detection algorithms to identify the head and pose of the user. Then, it utilizes linear regression to estimate the gaze using data acquired from a calibration routine. In contrast to proprietary systems, WebGazer is distributed as an open-source library, which makes it especially suitable for large-scale HCI applications.

STUDY

Our objective was to evaluate the state of the art of GT based on RGB cameras and to study the applicability of publicly available open-source libraries to GT tasks for large-scale applications such as websites and remote collaboration software. To this

end, we compared the performance of Tobii 4C, a commercial IR-based eye tracking device, and WebGazer, one of the few web-based GT tools.

Participants

A total of 14 people, that is, 13 males (92.85%) and 1 female (7.14%), volunteered to participate in the experiment. Participants were all aged 18-26 (i.e., average age 20±2). Although some subjects reported astigmatism and color blindness, they were considered as having a normal vision, as the former condition would not impact the outcome of the experiment, and colorblindness was taken into consideration in the design of the interface of the data collection software.

Materials

In our experiment, we utilized a commercial infrastructure-based eye-tracking device, that is Tobii 4C, which has an image sampling rate of 90 Hz. Simultaneously, we used the laptop's embedded RGB webcam with a resolution of 720p (1280 x 720 pixels), which is the current standard for most commercially available webcams. In addition to Tobii's driver, the experiment software comprised a custom wrapper built around the Tobii SDK, a webpage incorporating WebGazer, and a stimulus routine consisting of a circle moving on the screen (described below), which was used as a reference. The experiment task was realized on a Full HD display (i.e., 1920×1090 resolution) running the software in full-screen mode. Moreover, we developed a software utility that logged the input acquired by the eye-tracking device simultaneously with the coordinates of the reference and the gaze position as estimated by WebGazer. The sampling frequency of our data acquisition software was 250Hz, to make sure we were able to collect all the data sampled asynchronously from Tobii and WebGazer.

Procedure

Participants were introduced to the experiment and given informed consent. Then, they were taken to a distraction-free room where they were seated in front of a computer equipped with the eye-tracking device, the webcam, and the acquisition software. Before starting the recording, subjects completed a short questionnaire that asked for their demographic information and confirmed that they had normal vision. Subsequently, participants were positioned at a distance of 60-65 cm (approximately 2 feet) from the display and the acquisition devices. Before each session, we executed the calibration routine on the eye-tracking device and the software. Tobii's calibration routine asks the subject to stare at each of seven markers on the screen for several seconds, whereas WebGazer's calibration routine utilizes nine points: as the user clicks on each marker five times, the software estimates the position of the gaze using the location of the mouse as a reference. Before starting the trial, we made sure that the calibration accuracy reported by the systems was at least 70%.

After calibrating the hardware and software, participants were presented with the experiment task, which consisted in staring at a circle (i.e., reference) and following its movement on the screen. The circle moved smoothly along a predetermined path on a grid consisting of 20×20 markers (the grid was not shown to the subject). Each trial lasted 250 seconds, that is, the time required by the circle to move along the path. Our data collection tool continuously acquired the location of the circle and the position of the user's gaze as predicted by Tobii and WebGazer. Each subject realized one trial only.

Results and Discussion

We collected a total of 228336 data points representing: the coordinates of the reference (i.e., the position of the circle on the display), the gaze location acquired by Tobii, and the gaze location estimated by WebGazer. Then, we calculated the distance in pixels between the position of the reference and the gaze predictions (i.e., dispersion). Tobii's device had average dispersion of 60.46±29.52 pixels, whereas GazeTracker's dispersion was 507.58±297.99 pixels, on average. Figure 1 shows the position of the reference and the gaze position as estimated by Tobii and WebGazer. All the values were rescaled relatively to the position of the reference. As depicted in the Figure and summarized by our descriptive statistics, overall Tobii outperforms WebGazer even in controlled conditions that enabled to accurately calibrate the software. Subsequently, we calculated the overall performance in terms of accuracy, by normalizing the dispersion with respect to the size of the display. Our results show that Tobii is significantly more performing, resulting in an average accuracy of 97.26%±1.34% accuracy compared to WebGazer's score, that is, 76.96%±13.53%. Although WebGazer's accuracy result, that is, almost 80%, can be considered promising for an RGB camera-based GT system, its dispersion score provides a better understanding of the inaccuracy of the system. An average dispersion of approximately 500 pixels on a 1920×1080 display means that predictions are usually half-screen away from the actual gaze position, can have a significant impact on the user experience, especially in critical tasks. Figure 2 presents a more detailed comparison of the reliability of Tobii and WebGazer in terms of accuracy. Furthermore, we evaluated dispersion and accuracy in different areas of the screen. Specifically, in our analysis, we divided the display into a 3×5 matrix consisting of 15 quadrants, and we compared the performance of Tobii and WebGazer in the middle of the screen (i.e., quadrant 2,3), and in the areas at the top-left (i.e., quadrant 1,1), bottom-left (i.e., quadrant 3,1), top-right (i.e., quadrant 1,5), and bottom-right corners (i.e., quadrant 3,5), which usually are the most critical for GT systems. As shown in Table 1, Figure 3, and Figure 4, WebGazer's performance experience a statistically significant drop between the center of the screen and the areas closer to the corners. In contrast, Tobii's dispersion, which increases as gaze moves from the top-left quadrant to the bottom-right quadrant, is the result of a correction caused by the display resolution. Regardless, Tobii is one order of magnitude more accurate than WebGazer.

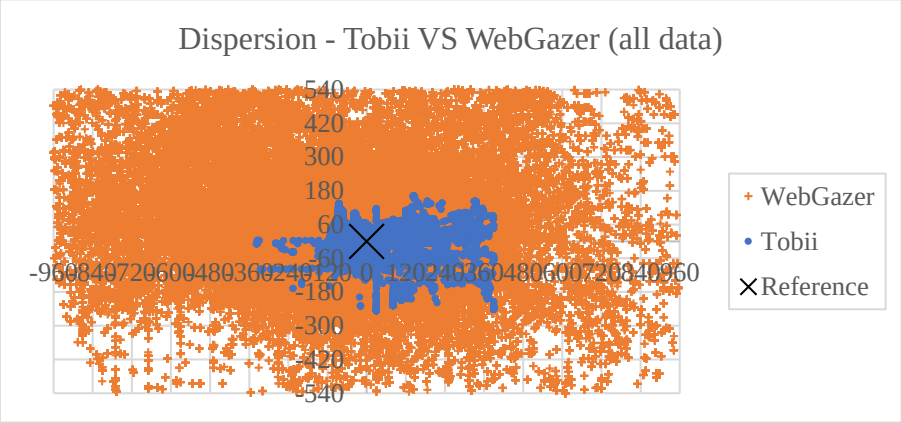

Figure 1. Dispersion of Tobii and WebGazer with respect to the reference. The Figure shows all the collected data points having a dispersion within ±960 pixels on the X-axis (half of the width of the display) and less than ±540 pixels on the Y-axis (half of the height of the display).

Indeed, participants' behavior (e.g., movement, distance from the display) during eye-tracking sessions may influence how gaze is estimated by the GT system. To this end, we compared the performance of Tobii and WebGazer across the subjects. In our analysis, we utilized all the data collected for each participant, that is, 16310 samples, on average. On average, Tobii's dispersion was 59.82±10.14 pixels, compared to WebGazer's 488.76±148.88 pixels. Moreover, Tobii's predictions have an accuracy of 97.28%±0.46%, whereas WebGazer's estimated gaze with an average accuracy of 77.81%±6.75%. As a result, we can conclude that Tobii is significantly more reliable than WebGazer in handling subject variability, also.

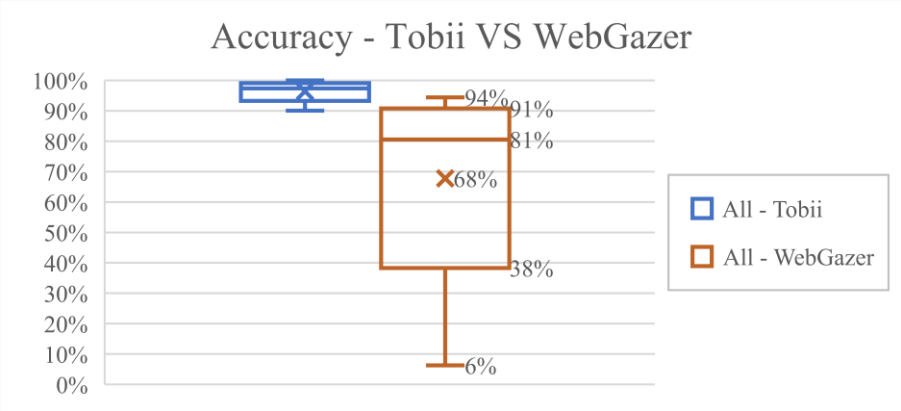

Figure 2. Accuracy of Tobii (left) and WebGazer (right).

Figure 3. Dispersion in five different areas of the screen, that is, the top-left, bottom-left, center, top-right, and bottom-right quadrants.

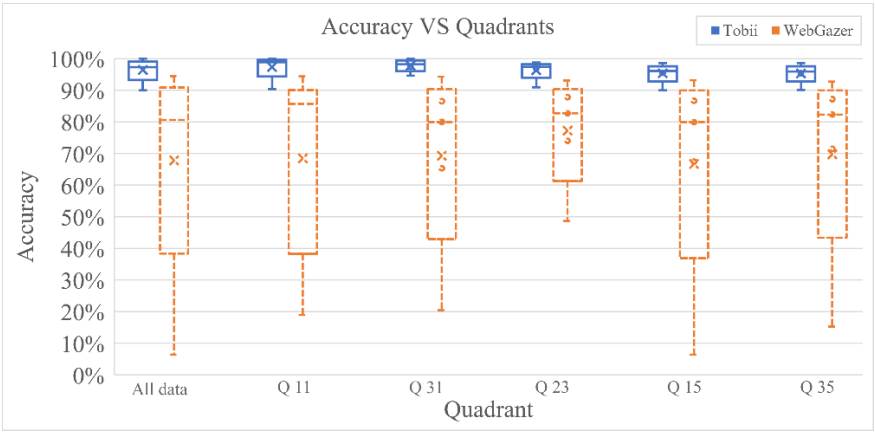

Figure 4. Comparison in the accuracy of Tobii and WebGazer in all the data (left) and each of the five quadrants described in Figure 3.

CONCLUSIONS AND FUTURE WORK

In this paper, we presented a comparative study of an IR-based GT device and a publicly available GT algorithm. The objective of our study was to evaluate the possibility of tracking gaze using standard RGB cameras (e.g., webcams) in combination with software algorithms rather than requiring dedicated devices such as IR sensors. Specifically, we focused on WebGazer, one of the few currently available open-source GT libraries available for research and commercial purposes that enable

extending their functionality and incorporating them into Human-Machine Interaction applications, with specific regard to web-based applications designed either for web browsers or for desktop use. To this end, we compared several performance aspects of Tobii and WebGazer, including the accuracy of their gaze predictions compared to a reference point representing the actual position of the user's gaze.

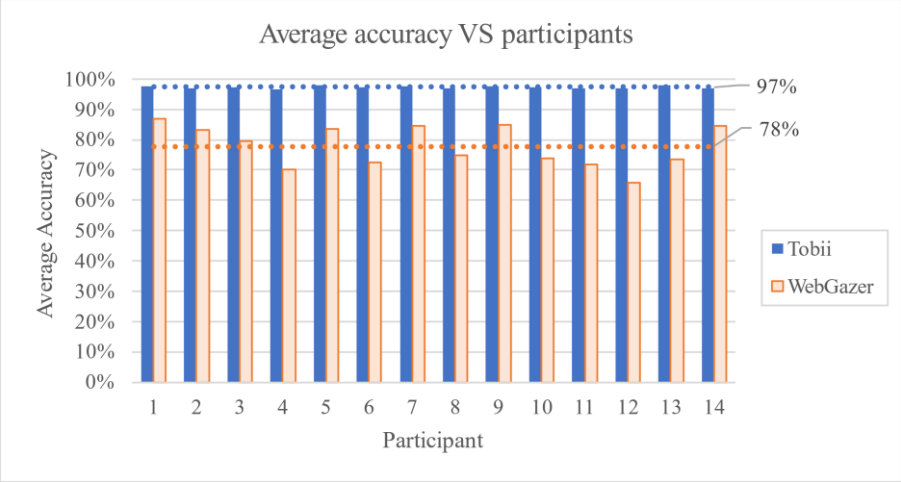

Figure 5. Comparison of the average accuracy of Tobii and WebGazer with the different participants.

Although our findings suggest that WebGazer has an accuracy of 79% compared to Tobii, it is not suitable for GT tasks that require finer control, due to the discrepancy between its predicted values and the actual position of the gaze of the user. Furthermore, we evaluated the reliability of the two systems in tracking gaze in different areas of the screen. In this regard, our data confirmed that Tobii outperforms WebGazer in terms of reliability because its predictions have a comparable level of accuracy across the entire width and height of the display. In contrast, WebGazer's accuracy significantly drops when the users move their gaze from the center of the screen toward the corners of the display. Nonetheless, WebGazer might be integrated effectively by avoiding placing UI elements along the borders of the screen, increasing their size and spacing them accurately. Furthermore, we compared the performance of the two GT systems across different subjects: from our findings, we can conclude that Tobii is significantly more accurate and reliable than WebGazer. In conclusion, RGB camera-based GT systems require further research and development before they can be incorporated in large-scale HCI applications.

REFERENCES

Caporusso, N., Zhang, K., Carlson, G., 2020. Using Eye-tracking to Study the Authenticity of Images Produced by Generative Adversarial Networks. In 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE). IEEE, pp. 1–6.

Duchowski, A.T., 2002. A breadth-first survey of eye-tracking applications. Behavior Research Methods, Instruments, & Computers, 34(4), pp.455–470.

Cognolato, M., Atzori, M., Müller, H., 2018. Head-mounted eye gaze tracking devices: An overview of modern devices and recent advances. Journal of rehabilitation and assistive technologies engineering, 5, p.2055668318773991.

Caporusso, N., Walters, A., Ding, M., Patchin, D., Vaughn, N., Jachetta, D., et al. Comparative user experience analysis of pervasive wearable technology. In: International Conference on Applied Human Factors and Ergonomics. Springer; 2019. p. 3–13.

Sun, L., Liu, Z. & Sun, M.-T., 2015. Real time gaze estimation with a consumer depth camera. Information Sciences, 320, pp.346–360. Available at: http://dx.doi.org/10.1016/j.ins.2015.02.004.

Liu, M., Zhu, Z., 2012. A case study of using eye tracking techniques to evaluate the usability of e-learning courses. International Journal of Learning Technology, 7(2), p.154. Available at: http://dx.doi.org/10.1504/IJLT.2012.047980.

Bevilacqua, V., Biasi, L., Pepe, A., Mastronardi., G., Caporusso, N., 2015. A computer vision method for the italian finger spelling recognition. In: International Conference on Intelligent Computing. Springer; p. 264–74.

Burton, L., Albert, W. & Flynn, M., 2014. A Comparison of the Performance of Webcam vs. Infrared Eye Tracking Technology. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58(1), pp.1437–1441. Available at: http://dx.doi.org/10.1177/1541931214581300.

Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J., 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:150406755.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 3839–3845). AAAI.