A Landmark Detection and Iris Prediction Dataset for Gaze Tracking Research

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

ABSTRACT

Eye-tracking technology enables acquiring the user's gaze and using it as an input for a variety of tasks. Primarily, this is realized with external devices (i.e., eye trackers) that incorporate infrared sensors. However, requiring users to have dedicated eye-tracking equipment limits the potential applications of this technology. Therefore, in the last decade, several research groups have pursued the development of gaze tracking solutions that leverage standard RGB cameras such as the webcams embedded in laptops. Unfortunately, these systems have lower accuracy and reliability than eye trackers. Nonetheless, novel landmark detection algorithms and computer vision pipelines based on machine learning might represent a more viable alternative. In this paper, we introduce and share an annotated dataset that can be utilized for developing, evaluating, and optimizing gaze tracking solutions. Our dataset incorporates features predicted using MediaPipe and specifically, Facemesh and Iris, two models designed for real-time image segmentation and object detection. Furthermore, we labeled each sample using an eye-tracking device, which provides a benchmark for studies aimed at training and testing novel gaze tracking algorithms.

Keywords: Human-Computer Interaction, Gaze tracking, Machine Learning, TensorFlow, MediaPipe, Landmark Detection, Facemesh, Iris Prediction.

INTRODUCTION

Gaze tracking (GT) has become an established technology that enables detecting the position of the eyes and using it for estimating where the user is looking. For instance, this system is utilized in combination with a computer screen to provide users with an alternative input channel that supports controlling interaction with the eyes. Nonetheless, GT can be utilized for a variety of tasks such as evaluating the usability of websites (Aviz et al., 2019), analyzing users' behavior and response with respect to digital content (Caporusso et al., 2020) (Caporusso et al., 2019), or enabling individuals to play games hands-free (Uludağli, 2018). In the last decades, GT has been realized primarily with dedicated devices utilizing infrared (IR) sensors mounted either on the display of the computer or incorporated in wearables such as head-mounted systems (Cognolato et al., 2018) or frames (Caporusso et al., 2019). However, the requirement of adopting specific hardware has limited GT and its potential use in large-scale applications. Therefore, more recently, several research groups have pursued the development of computer vision solutions for GT solutions based on traditional RGB cameras such as webcams embedded in portable computers and mobile devices. Unfortunately, previous studies have shown that GT systems based on RGB cameras have significantly lower accuracy, are not suitable for tasks that require precise user control, and, thus, require further research and development. Nevertheless, recent advances in machine learning (ML) and the implementation of more sophisticated models have provided researchers with tools for improving the accuracy and reliability of their GT solutions. In addition to ML libraries and algorithms, the availability of annotated datasets is key for supporting new developments as well as comparative studies.

Therefore, in this paper, we introduce and discuss a dataset designed for stimulating screen-based gaze tracking research aimed at replacing traditional IR devices with standard RGB cameras. Our objective was to label the features estimated by TensorFlow's landmark detection and iris prediction model with the actual location of the user's gaze on a screen, to foster research studies aimed at evaluating the use of Machine Learning to support GT. To this end, we collected data from a group of users who were involved in two gaze tracking tasks. Mediapipe (Grishchenko, 2020). Each sample in our dataset includes all the features of the landmark detection and iris prediction model (i.e., MediaPipe). In addition to the standard annotations provided by TensorFlow, we included additional properties that entirely describe all the face geometry key points acquired by MediaPipe. Furthermore, our dataset contains two different labels representing (1) the actual gaze location acquired with a dedicated IR sensor, and (2) a reference point. After detailing the data collection procedure and describing the dataset, we discuss its potential use in advancing research on GT and other relevant Human-Computer Interaction applications.

RELATED WORK

Dedicated GT devices use arrays of IR sensors and rely on hardware-based processing to isolate the pupils, which results in reliable systems having high accuracy (i.e., above 90%) and speed (60-1200 Hz). Nevertheless, in the last years, several groups explored viable alternatives to dedicated eye-tracking devices using traditional RGB cameras and image processing algorithms. However, standard consumer cameras such as webcams have a lower acquisition frame rate (i.e., 24-30 frames per second), provide only a monocular view of the subject's face, and process the acquired images via software. As described in (Mounica, 2019), GT based on RGB cameras involves a sophisticated workflow consisting of multiple steps, that is, detecting and tracking the user's face, locating their eyes, identifying the position of their pupils, and subsequently estimating the coordinates of the point of the screen where the user is looking, in real-time. Indeed, each step involves a different problem, specific image processing techniques, and computational concerns and requirements. Although many groups tackled the issue using different types of ML techniques (Gudi, 2020) (Sahay and Biswas, 2017) (Zhu and Deng, 2017), currently there are no ML pipelines that process the entire workflow in an end-to-end. Moreover, significantly affects performance (i.e., speed), accuracy (i.e., the distance between the predicted gaze location and actual target), and reliability (i.e., within-subject and cross-subject accuracy). Recently, MediaPipe (Grishchenko, 2020) (Kartynnik et el., 2019) has been introduced as an efficient solution especially designed for real-time computer vision tasks. It comprises a set of models each optimized for specific image segmentation and object recognition problems such as face, hand, posture, and iris detection. Specifically, MediaPipe Iris (Ablavatski et al., 2020) focuses on identifying a subject's pupils, tracking their movement, and estimating their distance from the camera. However, the library does not contain features for tracking gaze on a computer screen. Nevertheless, the performances of the models are remarkable, as MediaPipe is designed for being incorporated into mobile applications and websites, which makes the library one of the promising candidates for exploring the feasibility of GT based on standard RGB cameras. Although the models are designed for GT, their predictions could be utilized to implement a GT system. As a result, several groups have incorporated MediaPipe and specifically, Facemesh and Iris, into their GT solutions. For instance, WebGazer (Papoutsaki et al., 2016) is a library that processes the facial geometry and iris position predicted by MediaPipe with a linear regression algorithm that, based on an initial calibration, enables estimating the position of the user's gaze on the screen. Unfortunately, previous studies that compared WebGazer and a commercial eye-tracking device based on IR sensors have demonstrated that the former solution results in significantly lower performance and it is not suitable for most GT tasks that require accuracy and reliability (Thaman et al., 2022).

THE DATASET

The objective of our work is to foster studies aimed at replacing eye-tracking devices, with specific regard to IR sensors, with standard RGB cameras such as the webcam. Specifically, we aim at incorporating MediaPipe into our processing pipeline. MediaPipe Facemesh (Kartynnik et el., 2019) is a neural network-based model that approximately predicts the surface geometry of a human face. To this end, Facemesh receives an image or video stream as an input, and it applies a face detection algorithm based on Convolutional Neural Networks (CNN) that identifies the bounding box of the facial rectangle. Then, the model predicts 468 facial landmarks each consisting of a set of coordinates representing the point in a three-dimensional space, as shown by Figure 1. As a result, the algorithm enables aligning the predicted triangular topology to the face of the subject in a variety of face postures, angles, and distances. Additionally, MediaPipe includes Iris (Ablavatski et al., 2020), a model designed to predict eye, eyebrow, and iris geometry. The model represents each eye with five coordinates that indicate the center of the pupil and the top, bottom, left, and right edges of the iris. Although the models are optimized for real-time performance, the primary intended application of both models is augmented reality for entertainment purposes such as image filters (i.e., selfie overlays). Furthermore, given their number and complexity, the features are not suitable for being directly utilized for training or calibrating a model intended for use in large-scale applications such as websites, without prior preprocessing such as dimensionality reduction or further future extraction. To this end, datasets play a crucial role because they support the development, evaluation, and optimization of new models. However, most datasets contain eye patches or the cropped bounding box with the subject's image, which are not suitable for gaze tracking research (Mounica, 2019). Thus, as a first step, we collected an annotated dataset of gaze tracking features that we plan to use in our research. In this section, we describe its content to other groups that might be interested in incorporating it in their studies.

Data Collection

The dataset was acquired with a laptop computer equipped with a Full HD display (i.e., 1920×1090 resolution) running the data collection software in full-screen mode. The video stream was acquired using the embedded RGB webcam of the computer, which had a resolution of 720p (1280 x 720 pixels), that is, the current standard for most commercially available webcams. In addition, the data collection equipment included an external eye-tracking system that acquired the gaze of the subject. To this end, we employed a commercial device, that is, Tobii 4C. The system, which predicts the gaze location at a sampling rate of approximately 90 Hz based on the subjects' pupils, was mounted at the bottom of the display of the laptop. The data collection software consisted of an application developed ad hoc. It comprised a webpage with

the stimulus routine. Also, the page had access to the webcam so it could feed the video as an input to MediaPipe Facemesh and Iris. The data collection software recorded the entire output predicted by the MediaPipe models, that is, both the facial geometry and the key points of the irises, which in our dataset represent the features. The position of the stimulus, the key points estimated by Facemesh and Iris, and the gaze location predicted by Tobii 4C, were collected using a custom acquisition software that synchronized and logged the data at a sampling rate matching that of Tobii's device. As a result, the data collection setup enabled us to utilize two systems for labeling the features. First, we recorded the position of the stimulus itself, that is, the coordinates of a symbol on the screen, which subjects were asked to stare at. In addition, we utilized the eye-tracking device to acquire the gaze location of the subject. By doing this, our dataset enables a more versatile approach to the design and study of GT models using ML algorithms, as described below. Furthermore, the prediction realized with the eye-tracking device can be utilized as a benchmark in evaluating the accuracy of GT models built using the dataset discussed in this paper.

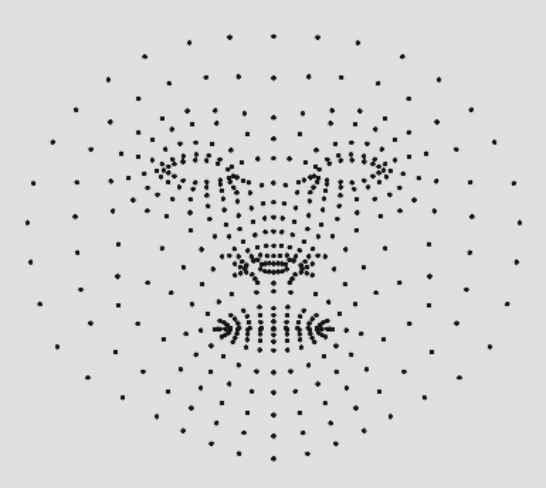

Figure 1. The map of the 468 key points in MediaPipe Facemesh.

The datasets were collected with two different groups of individuals and on different days. However, they were acquired in the same conditions: subjects were taken to a distraction-free room and seated in front of the computer equipped with the data collection hardware and software, and they were initially positioned at a distance of 60-65 cm (approximately 2 feet) from the display and the acquisition devices (i.e., webcam and Tobii), at the beginning of the data collection process. Before starting the acquisition from each participant, we executed the calibration routine on the eye-tracking device, and we made sure that the accuracy was at least 90%. Then, participants were presented with the gaze tracking tasks, which mainly consisted in staring at a circle (i.e., stimulus) and following its movement on the screen, though each dataset involved a specific task (described below). Although MediaPipe supports

detecting multiple faces in the same video stream, for our work we were interested in involving one individual only in each data collection session.

Our dataset is divided into two sub-datasets based on the GT task realized by the participants involved in the study. In dataset one, subjects followed a circle moving at random on the screen while seating at a distance of 60 cm from the acquisition devices. Also, they did not receive any instructions about moving their head. We realized one acquisition for each of the subjects. In dataset two, subjects were presented with a sequence of stimuli placed in 15 different locations of the screen (see Figure 2), and they were asked to stare at each of them while maintaining each of five predefined head postures (i.e., chin up, chin down, head turned left, head turned right, and straight, as shown in Figure 3) for three seconds. We asked subjects to repeat the task at three different distances, that is, 30, 60, and 90 centimeters from the acquisition devices (i.e., camera and IR sensor) and in five different alignments with respect to the camera (i.e., center, top-left, top-right, bottom-left, and bottom-right), as shown in Figure 2. As a result, we acquired a total of 1125 different configurations in each session. Although we realized one session per individual, we recorded multiple samples of the same configuration. This is to provide researchers with different options for using the dataset in their work. For instance, samples could be filtered, windowed, or averaged depending on the purpose of the study and intended use of the dataset. As per the file format, acquisitions are stored into separate files, each corresponding to a different subject. In each file, every line contains one sample formatted as a JSON object. The dataset is publicly available in the project repository (Caporusso, 2021). In addition to the datasets, the repository contains useful data collection, processing, and visualization tools.

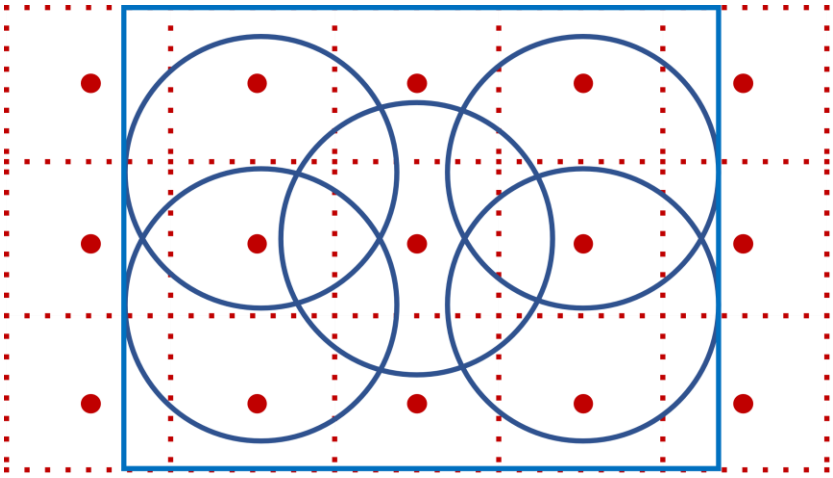

Figure 2. The data acquisition protocol utilized in Dataset 2 represented as the 15 stimuli (i.e., small solid dots) distributed on the screen in a 3×5 grid arrangement (i.e., large rectangle) and the five face alignment configurations (i.e., large blue circles) with respect to the camera (i.e., inner blue square).

Feature Specifications

The samples in Dataset 1 and Dataset 2 contain three sets of features predicted by MediaPipe. First, they include the two-dimensional bounding box of the subject's face represented as the bi-dimensional coordinates of the top-left and bottom-right corners of the rectangle enclosing the detected face, relative to the size of the image acquired from the camera. This enables estimating useful information such as the distance of the individual from the camera and the position of the face with respect to the entire frame (i.e., alignment). As a second set of features, we stored the mesh, that is, the predicted facial topology represented as an array of the 478 key points, that is, the 468 vertexes described before and shown in Figure 1, plus five key points for each eye. Each vertex is associated with a set of three-dimensional coordinates based on the size of the cropped bounding box (i.e., 256×256 or 128×128 pixels). The third set of features is the rescaled mesh object, which includes a list of 58 properties each identifying one component of the face (e.g., silhouette, nose tip, cheeks, and iris). Among them, 32 components are the basic properties included in Facemesh, which describe 238 vertexes, only. In addition, we defined 26 new properties that describe all the remaining key points. Each property consists of a series of three-dimensional coordinates representing its key points rescaled with respect to the size of the original input image. By doing this, our dataset can be utilized in studies that focus on the content of the bounding box, that is, facial and pupil geometry. Simultaneously, the rescaled mesh provides useful information about the alignment of the subject with respect to the camera, which is relevant for the last step of the processing pipeline, that is, predicting where the user is looking based on their pupils and face posture, rotation, and alignment.

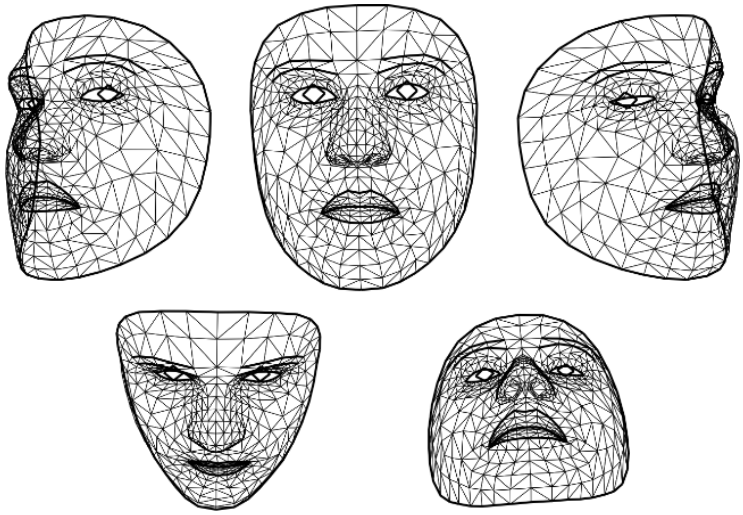

Figure 3. The five standard face postures acquired in Dataset 2.

In addition, Dataset 1 and Dataset 2 contain two labels, that is, (1) the position of the stimulus on the display and (2) the gaze location predicted by Tobii 4C, both are represented as bi-dimensional coordinates. The first set can be utilized to evaluate where the subject was expected to look, whereas the second one provides a more accurate estimate of the point on the screen where the individual was actually looking during the acquisition. This information enables discarding samples in which the subject was not staring at the stimulus. Alternatively, depending on the purpose of the study, it can be utilized as the reference target, regardless of the location of the stimulus presented to the subject. Moreover, each sample in dataset 2 also contains information regarding the distance of the subject (i.e., 30, 60, or 90 cm) and their face posture and alignment.

CONCLUSIONS

In this paper, we have introduced and detailed an annotated datasets that can be utilized to evaluate different strategies for the design of GT models. To this end, we designed two GT tasks and data collection software that acquired us to record input using a standard RGB camera and feed the video stream to the MediaPipe Facemesh and Iris models, which enabled us to obtain predictions about the facial geometry and pupil location of the subject. Simultaneusly, we acquired user's gaze with Tobii 4C, a commercial eye tracker, and we utilized the predictions to label the samples. Specifically, research groups can leverage it for studies on dimensionality reduction (e.g., use a minimal subset of relevant vertexes) and feature extraction (e.g., calculate rotation angles from the facial geometry). Furthermore, the dataset can be utilized to compare of different strategies, including hybrid strategies. Alternatively, it can serve as a source of simulated input to evaluate the performance of GT models under different conditions and analyze their performace with respect to the benchmark provided by the eye-tracking device. Furthermore, the dataset can be utilized to pre-train a general model that can be utilized by different users without the need of calibration (i.e., zero-shot classification). The dataset is shared in the project repository (Caporusso, 2021), which will contain updates about our future work.

REFERENCES

Aviz, I. L., Souza, K. E., Ribeiro, E., de Mello Junior, H., & Seruffo, M. C. da R. (2019). Comparative study of user experience evaluation techniques based on mouse and gaze tracking. In Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (pp. 53–56).

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) (pp. 1–6). IEEE.

Caporusso, N., Zhang, K., Carlson, G., Jachetta, D., Patchin, D., Romeiser, S., … Walters, A. (2019). User discrimination of content produced by generative

adversarial networks. In International Conference on Human Interaction and Emerging Technologies (pp. 725–730). Springer.

ULUDAĞLI, M. Ç., & Acartürk, C. (2018). User interaction in hands-free gaming: a comparative study of gaze-voice and touchscreen interface control. Turkish Journal of Electrical Engineering & Computer Sciences, 26(4), 1967–1976.

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, 2055668318773991.

Caporusso, N., Walters, A., Ding, M., Patchin, D., Vaughn, N., Jachetta, D., & Romeiser, S. (2019). Comparative user experience analysis of pervasive wearable technology. In International Conference on Applied Human Factors and Ergonomics (pp. 3–13). Springer.

Grishchenko, I., & Valentin Bazarevsky, R. (2020). MediaPipe Holistic—Simultaneous Face, Hand and Pose Prediction, on Device. Retrieved June, 15, 2021.

Mounica, M. S., Manvita, M., Jyotsna, C., & Amudha, J. (2019). Low Cost Eye Gaze Tracker Using Web Camera. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 79–85). IEEE.

Gudi, A., Li, X., & van Gemert, J. (2020). Efficiency in Real-Time Webcam Tracking. In European Conference on Computer Vision (pp. 529–543). Springer.

Sahay, A., & Biswas, P. (2017). Webcam Based Eye Gaze Tracking Using a Landmark Detector. In Proceedings of the 10th Annual ACM India Compute Conference (pp. 31–37).

Zhu, W., & Deng, H. (2017). Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3143–3152).

Kartynnik, Y., Ablavatski, A., Grishchenko, I., & Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile GPUs. ArXiv Preprint ArXiv:1907.06724.

Ablavatski, A., Vakunov, A., Grishchenko, I., Raveendran, K., & Zhdanovich, M. (2020). Real-time Pupil Tracking from Monocular Video for Digital Puppetry. ArXiv Preprint ArXiv:2006.11341.

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.

Thaman, B., Cao, T., Caporusso, N.: Comparative Analysis of RGB-based Eye-Tracking for Large-Scale Human-Machine Applications. In: Proceedings of the 5th International Confer-ence on Intelligent Human Systems Integration: Integrating People and Intelligent Systems (IHSI) (2022)

Caporusso, N., 2021. A Landmark Detection and Iris Prediction Dataset for Gaze Tracking Research, GitHub. Available at: https://github.com/NicholasCaporusso/A-Landmark-Detection-and-Iris-Prediction-Dataset-for-Gaze-Tracking-Research