Eye-Tracking Algorithm for Low Webcam Image Resolution Without Calibration

Allana L. dos S. Rocha
University of Pernambuco
Pernambuco, Brazil
Email: alsr@ecomp.poli.br

Leandro H. de S. Silva
University of Pernambuco
Federal Institute of Paraíba
Email: lhss@ecomp.poli.br

Bruno J. T. Fernandes
University of Pernambuco
Pernambuco, Brazil
Email: bjtf@ecomp.poli.br

***Abstract*—Applications of eye-tracking devices aim to understand human activities and behaviors, improve human interactions with robots, and develop assistive technology in helping people with some communication disabilities. This paper proposes an algorithm to detect the pupil center and user's gaze direction in real-time, using a low-resolution webcam and a conventional computer with no need for calibration. Given the constraints, the gaze space was reduced to five states: left, right, center, up, and eyes closed. A pre-existing landmarks detector was used to identify the user's eyes. We employ image processing techniques to find the center of the pupil and we use the coordinates of the points found associated with mathematical calculations to classify the gaze direction. By using this method, the algorithm achieved 81.9% overall accuracy results even under variable and non-uniform environmental conditions. We also performed quantitative experiments with noise, blur, illumination, and rotation variation. Smart Eye Communicator, the proposed algorithm, can be used as eye-tracking mechanism to help people with communication difficulties to express their desires.**

I. Introduction

Eye-tracking is the continuous process of measuring the movement of the eye in relation to the head [1]. This eye movement is also referred to as the gaze direction. In a short definition, the gaze direction is the point at which someone is looking. Eye-tracking strategies are relevant for various applications ranging from understanding human activities and behaviors to improving human interactions with the machine, as Assistive Technology (AT) for communication of disabled people [2]. In the applications of eye-tracking as AT, the goal is to improve people's quality of life [3]. These technologies are composed of assistive, adaptive, and rehabilitative devices in order to maintain the Activities of Daily Living (ADL) [4]. The alternative forms of communication arise, then, as one of the solutions, helping the interaction and social participation of these individuals.

The eye-tracking devices can be split into image-based and electrooculography (EOG) based [1]. EOG based methods assume that the eye behaves like a dipole and seeks to capture the saccadic eye movements through electrodes attached to the user's face [3]. Image-based methods make use of cameras in the visible or infrared spectrum to determine the gaze direction using strategies of digital image processing and machine learning [1]. This approach suffers from problems of camera positioning and lighting variation [5].

Some image-based eye-tracking strategies require specific hardware for operation, such as head-mounted cameras and infrared sensors or specific cameras [5]. Another issue with image-based systems is that the accurate classification of the gaze direction depends on calibration routines, since systems without calibration are more comfortable to be adopted by the user [6].

An approach is to use EYECAN [7], which includes eye calibration and tracking. The device, composed of camera, glasses, battery, heat sink and other components, is connected to a computer to use the EYECAN software. Once calibrated, the user's eye functions as a cursor and allows use of the eye writer, a digital keyboard. However, the device has some limitations such as: no head movement is allowed during the experiment, calibration must be done without blinking the eyes and even itching in eyes of a new user.

In the commercial area, Tobii® is company that produces high-precision optical mouse's using specific and proprietary hardware and software. However, this solution has a high financial cost, reducing the accessibility for many people with disabilities.

The solutions presented require extra equipment, which go beyond a computer and a webcam. Besides this, the need for calibration is present in the state-of-art of eye-tracking, which can cause discomfort to patients when they cannot make any movement, so that the accuracy and efficiency of the tools are not compromised. In addition, cost is an important weight factor and must be considered.

This work proposes a calibration-free strategy based on digital image processing to identify the gaze direction, based in five states: right, left, up, center, and closed eyes. The Smart Eye Communicator was evaluated with a database of 2,908 images obtained by the conventional webcam of a laptop (0.3-megapixel resolution 640x480) in different positioning and lighting conditions, obtaining an average accuracy of 81.8%. Because it is an algorithm that does not require calibration or specific hardware, this strategy has a low cost and can compose a system of AT to facilitate the communication of disabled people, such as people with ALS.

II. Smart Eye Communicator (SEC)

Smart Eye Communicator (SEC) is the proposed algorithm to detect the gaze direction using a facial landmarks predictor to locate the contours of the eyes and eyelids, which are illustrated by Figure 1. Those landmarks points are used to

calculate the Eye Aspect Ratio (EAR) of each video frame, according to the formula

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2 \cdot \|p_1 - p_4\|}, \quad (1)$$

where p_i is an eye landmark as shown in Figure 1b. EAR measure is a relation between the width and height of the eye [8]. The numerator of this equation computes the distance between the vertical eye landmarks while the denominator computes the distance between horizontal eye landmarks, weighting the denominator appropriately since there is only one set of horizontal points but two sets of vertical points.

(a) All 68 face landmarks (b) Eye landmarks

Fig. 1. Face landmarks and eye landmarks

To check whether the eyes are open or closed, we calculate the scalar amount of the EAR of each video frame and compare it with a threshold. In short, the proportion of the eye is approximately constant while the eye is open, but rapidly drops to zero when a blink of the eye is occurring. Thus, if the eyes are open, digital processing is performed on the input frame in order to refine and locate the center of the pupil. After its location, the center is used to calculate the distance between the lateral and horizontal extremities of the eyes, in order to determine the gaze direction.

A. Pupil center coordinates

The SEC flowchart is shown in Figure 2. Initially, the captured image from the webcam is converted from RGB to gray scale. Next, we use the face detector, implemented in the Dlib library, which is done using the classic Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid and a sliding window detection scheme [9]. After that, we cropped the image at the threshold of the face, in this way, we obtained the region of interest of the image. To make the image clearer, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied, in which 68 landmarks were found using the shape predictor [9]. To estimate the landmark locations, the algorithm examines a sparse set of input pixel intensities (i.e., the "features" to the input model), passes the features into an Ensemble of Regression Trees (ERT) and refines the predicted locations to improve accuracy through a cascade of regressors.

To decrease the calculation load during the entire processing, the right and left eyes are extracted from the image using the face landmarks. Then, the histogram equalization is applied to both eyes. To darken and eliminate some noise from the image, we subtract 70 units from each pixel and apply the histogram equalization again. Next, we get the iris based on the color range of the image (0 to 15).

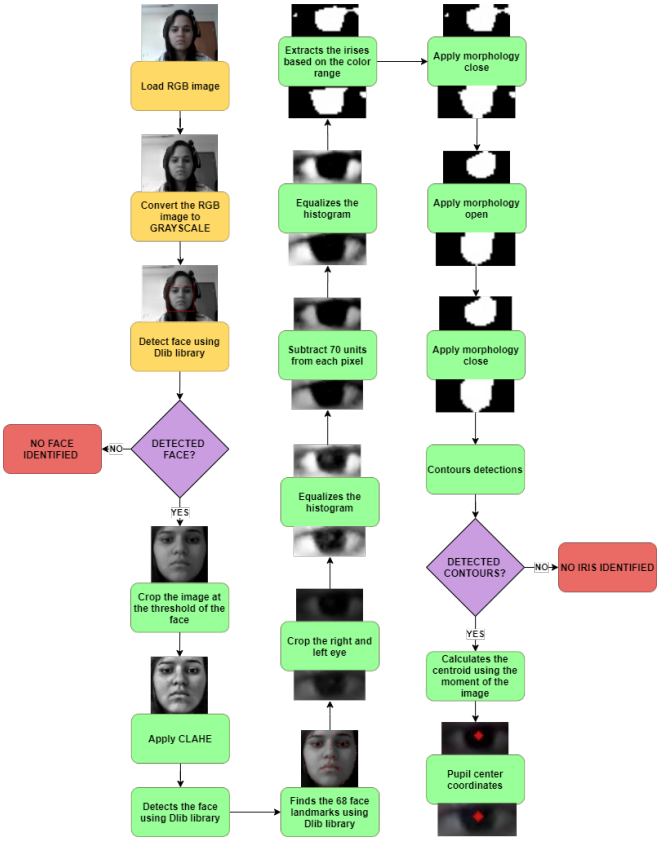

Fig. 2. Flowchart of the Smart Eye Communicator

After this, three morphological operations are performed: two closing and one opening. Dilation is necessary to gather the iris's disparate elements resulting from unwanted reflections and fill any holes in the iris contour. On the other hand, erosion is responsible for removing small portions or regions outside the boundaries.

As a result of these steps, a more polished image is obtained and used to detect contours. After the detection, we select the largest contour and calculate the centroid from the moment of the image, that is a specific weighted average of the pixel intensities of the image, with the help of which we can find some specific properties, such as the centroid. Thus, we obtain the coordinates of the pupil center.

B. Gaze direction classification

We used the vertical and horizontal displacement of the center of the pupil from the landmarks to predict the users' gaze direction, as shown in Figure 3. Given this perspective, we calculate the distance between the L (left) and C (center). If it is 50% greater than the distance between the points C and R (right), the algorithm will answer that the user's point of view is "Right". Similarly, if the distance between the points R and C is 50% greater than the distance between L and C, the answer to the point of view will be "Left". This way too, if the distance between the points D (down) and C is 50% greater than the distance between U (up) and C, the point of

view will be classified as ”Up”. Finally, if no alternative is satisfied, the direction will be classified as ”Center”.

Fig. 3. The position of the reference point in the eye (the red spot in the center), in relation to the pupil (the black circle).

III. Experiments and Results

A. Dataset

To evaluate the SEC performance, we recorded a set of videos with 12 participants (7 females and 5 males). For each frame of those videos, the correct gaze direction was human-labeled. A total of 16 videos were recorded, with different lighting conditions in different locations, using glasses or not, and in different positions in relation to a laptop webcam.

We used a Lenovo IdeaPad 320 80YH0001BR, with Intel® Core™ i7–7500u @3.5GHz processor, 8 GB DDR4 2133 MHz RAM, a 0.3-megapixel 640x480 resolution webcam and Windows 10 operating system.

These people were instructed to start capturing the video by looking at the center of the screen and then direct their gaze to the left, right, up and finally close their eyes. The captured videos have a total of 2,908 frames classified concerning the direction of the eye, with an average processing rate of 320 milliseconds per frame. The distribution of each class (gaze direction) is described in Table I.

TABLE I
Number of frames for each gaze direction.

Gaze direction	Frames
Center	899
Left	749
Right	469
Up	383
Close	408
Total frames	2908

Aiming for a quantitative evaluation of qualitative aspects of the images, we tested the performance of the algorithm in different image conditions (Figure 4).

B. Results

Regarding the evaluation with the original dataset, Figure 5 shows the F1 Score metric, which combines Precision and Recall to bring a number that indicates the overall quality of the model even with datasets that have disproportionate classes, for each gaze direction. The left direction has the highest average F1 Score equal to 0.97, followed by the right direction, with 0.90. The center, up and close directions resulted in: 0.74, 0.70, 0.66 average F1 Score, respectively. The outliers on the boxplot graph suggest that there are some videos in the dataset with very challenging lighting or positioning characteristics.

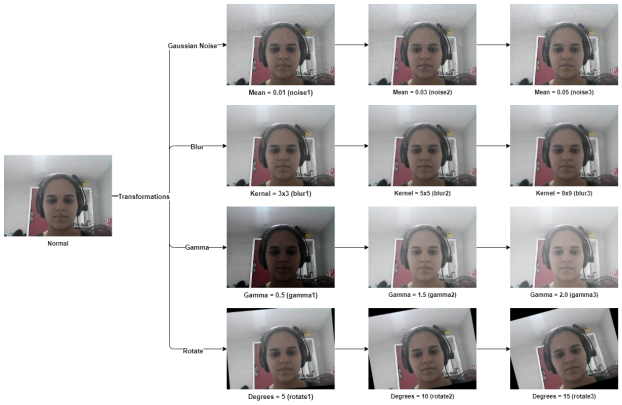

Fig. 4. Image with transformations applied.

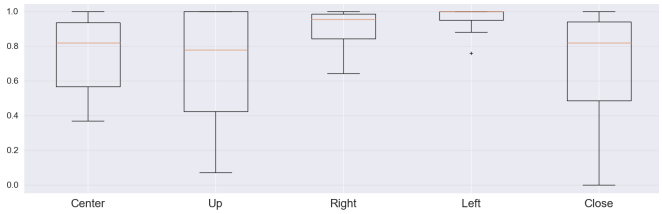

Fig. 5. F1 Score for the 5 directions obtained with original dataset.

Figure 6 reports the confusion matrix for the five directions. The SEC achieved an average of 81.8% correct answers. The largest one being the right one, with 442 correct frames out of 469 frames. The major confusion is between the center and the left direction.

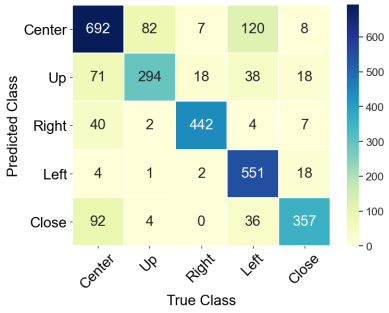

Fig. 6. Confusion matrix for the five directions.

Regarding the evaluation of the qualitative aspects of the images, Figure 7 show the boxplot graphs of the weighted average F1 Score for each applied transformation and the original dataset. Analyzing such result data, it is noted that the greatest impact factor were the applications of noise. The results of the videos rotated in 15 degrees in relation to their central axis also differ considerably from the average of the other cases. However, the other transformations maintained a variation rate close to the results without changes in the images.

The confusion matrix containing the results of the transformations made for the qualitative analysis is shown in Figure 8. Thus, the mean number of assertive frames is 75.9%, a

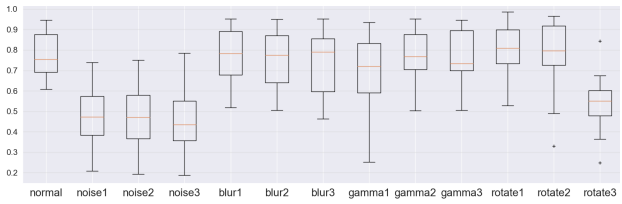

Fig. 7. Weighted average F1 Score of the transformations.

drawdown of 5.9% in relation to the original dataset. Although the confusion between center and left remains, there are also confusion between all classes and center direction.

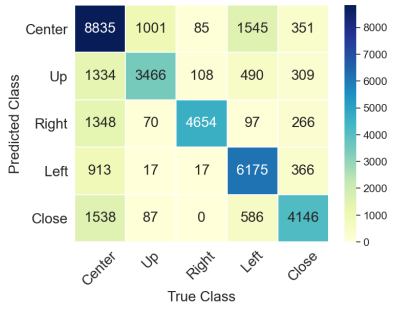

Fig. 8. Confusion matrix with transformation results.

IV. Application

With the Smart Eye Communicator software, Figure 9, the user can manipulate a graphical interface by moving the checkbox in the direction the eye is positioned. The frames offer the option of expressing the following needs: thirst, hunger, clearing saliva, neck pain, watching television, itching, pain, change of position, shortness of breath and turning BiPAP (turning on Bi-level Positive Airway Pressure).

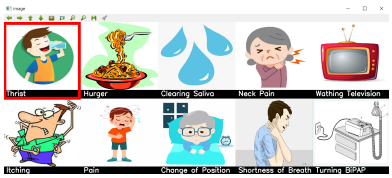

Fig. 9. Graphic interface.

Confirmation is given by looking up and, once the need is met, the selected phrase is played in audio. When the user finishes expressing the intended need, he can blink his eyes three times in a row to end the program (the project will be available online).

V. Conclusion

Besides other applications, eye-tracking devices are beneficial for Assistive Technology (AT). This sort of eye-tracking application suffers from variations in lighting, positioning, and low quality of personal computer webcam. Smart Eye Communicator (SEC), the proposed method, detects the pupil center and classifies user's gaze direction, using a low resolution webcam, without calibration. To overcame the lightning and head positioning issues, SEC classifies the gaze direction into five classes: right, left, up, center, and eyes closed. These gaze directions can be used as user input to control software for communication purposes.

SEC uses a landmark detector to get eyes region of interest from the image and employs a sequence of digital image processing techniques to find the pupil center. In sequence, the gaze direction is classified based on the distance between the pupil center and the eye boundaries.

The algorithm evaluation was performed in a database with 2908 images obtained through the conventional webcam. The results showed an average accuracy of 81.8% for the five directions, reaching an accuracy of 94.24% in the right direction. Besides, in order to measure the robustness with respect to qualitative image aspects, several transformations were applied to the image, among them are: noise applications, increase and decrease of brightness and contrast, and also rotation in relation to the axis and blurring applications. The algorithm's results in these transformations showed that the most significant impact factor on the results was the noise application.

Acknowledgment

The authors would like to thank the PDTE – POLI/UPE for their support through the scholarship granted. We also thank CNPq and FACEPE for their support.

References

[1] D. Venugopal, J. Amudha, and C. Jyotsna, "Developing an application using eye tracker," in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, may 2016, pp. 1518–1522. [Online]. Available: http://ieeexplore.ieee.org/document/7808086/

[2] W. Chareonsuk, S. Kanhaun, K. Khawkam, and D. Wongsawang, "Face and Eyes mouse for ALS Patients," in *2016 Fifth ICT International Student Project Conference (ICT-ISPC)*. IEEE, may 2016, pp. 77–80. [Online]. Available: http://ieeexplore.ieee.org/document/7519240/

[3] A. Lopez, I. Rodriguez, F. J. Ferrero, M. Valledor, and J. C. Campo, "Low-cost system based on electro-oculography for communication of disabled people," *2014 IEEE 11th International Multi-Conference on Systems, Signals and Devices, SSD 2014*, pp. 1–6, 2014.

[4] M. Arbesman and K. Sheard, "Systematic Review of the Effectiveness of Occupational Therapy-Related Interventions for People With Amyotrophic Lateral Sclerosis," *American Journal of Occupational Therapy*, vol. 68, no. 1, pp. 20–26, jan 2014. [Online]. Available: http://ajot.aota.org/Article.aspx?doi=10.5014/ajot.2014.008649

[5] H. M. Elahi, D. Islam, I. Ahmed, S. Kobashi, and M. A. R. Ahad, "Webcam-based accurate eye-central localization," *Proceedings - 2013 2nd International Conference on Robot, Vision and Signal Processing, RVSP 2013*, pp. 47–50, 2013.

[6] J. R. Khonglah and A. Khosla, "A low cost webcam based eye tracker for communicating through the eyes of young children with ASD," *Proceedings on 2015 1st International Conference on Next Generation Computing Technologies, NGCT 2015*, no. September, pp. 925–928, 2016.

[7] R. Kaushik, T. Arora, Sukanya, and R. Tripathi, "Design of Eyewriter for ALS Patients through Eyecan," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, oct 2018, pp. 991–995. [Online]. Available: https://ieeexplore.ieee.org/document/8748520/

[8] J. Cech and T. Soukupova, "Real-Time Eye Blink Detection using Facial Landmarks," *Center for Machine Perception, Department of Cybernetics Faculty of Electrical Engineering, Czech Technical University in Prague*, pp. 1 – 8, 2016.

[9] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014.