Frankfurt University of Applied Sciences

Advanced Real Time System Winter Semester 2021/2022

Express Emotions using Gaze Tracking with OpenCV

Under the supervision of
Prof. Dr.-Ing. Matthias Deegener

Prepared By

Amena Islam - 1347364

Faiaz Sharaf Uddin - 1347898

Fayza Amreen - 1348185

Table of Content

Abstract

Gaze estimate is critical in the development of assistive devices for individuals who are handicapped or impaired. There are numerous systems in use today, but none of them find a balance between accuracy and cost-effectiveness. In this work, we develop a gaze detection system aimed at enabling physically challenged persons with restricted movement to convey their feelings. We provide gaze estimation techniques in conjunction with screen calibration processes to build an effective real-time camera-to-screen gaze-tracking system.

Introduction

Human interaction involves multiple channels. While each person entails a different level of human interaction, the necessity for connection is intrinsic. Studies have shown that human interaction is not only physically helpful, but also imperative for mental health. Expressing emotion is a fundamental right of every human. But unfortunately, physically impaired people find it difficult to express their emotion. For example, disease like amyotrophic lateral sclerosis (ALS) and infantile cerebral palsy cause an irreversible loss of normal motor skills, however leaving unchanged the visual abilities. Here eye gaze plays a distinct role, as it can express emotions, desires, feelings, and intentions.

Gaze tracking is the process of determining the point-of-gaze in the physical space. Gaze tracking devices are used in various fields, i.e., the research on the visual apparatus, the field of psychology, study in cognitive linguistics, in the military sector, in designing of products assisting to the disabled people. Regrettably accurate eye gaze tracking normally requires expensive specialized hardware. Additionally, these devices have many limitations, including that of being used only by expert and "trained" users. This reduces the appeal of these systems for consumer market applications. Moreover, these solutions often require a manual calibration procedure for each new user. It is notable that gaze tracking devices suffer of some disadvantages compared to other aiming systems, in particular the so-called "Midas touch": it is not possible to establish when the user is watching a point intentionally or is simply moving gazing across the screen [1]. That is, there is no way to confirm intentionality.

In general, gaze-tracking entails calculating where a subject's gaze is directed based on images acquired by the camera. This is achieved using a gaze vector, which determines the pitch and yaw of the gaze in relation to the camera. A more comprehensive kind of gaze tracking advances it further by estimating the specific point the subject is glancing at on a display in front of the subject. This task is accomplished by estimating the location of the aforesaid screen in relative to the camera (i.e., calibration), which is not known in advance. We implement a study of gaze estimation techniques in conjunction with screen calibration approaches, with the goal of developing an efficient real-time camera-to-screen gaze-tracking system [17].

In this project, the aim is to create a system which can help physically impaired people to express their basic emotions while offering good temporal dynamics and response speed making it a real time system based on the study of Gudi *et al.* [18]. There are three kinds of primary emotions: happiness which associates with reward, sadness which associates with punishment, fear and anger which associate with stress. It is presumed that all emotions are the combination of these four primary emotions. This project focuses on four emotions that are: Happiness, Sadness, Anger and Surprise. Each of the emotion will be expressed by moving their eyes only.

Not everyone is bestowed with enough money to buy the already existing iris-based systems developed for physically impaired person. The project is made with integrated webcam from laptops and budget friendly webcams. Which makes it is cost effective and easily operable. Also, this system implements gaze estimation techniques in conjunction with screen calibration approaches and addresses the disadvantage known Midas touch of gaze tracking device and gives a solution to overcome it while expressing emotions.

State of the Art

Appearance-based CNN gaze-tracking:

The initial deep learning model for appearance-based gaze prediction was proposed by Zhang et al. [2,3]. Park *et al.* [4] created a network that integrates the hourglass [5] and DenseNet [6] networks to benefit from auxiliary supervision using the gaze-map, which is a two-dimensional binary mask of the iris and eyeball. Cheng et al. [7] introduced ARE-Net, which is divided into two smaller modules: one for determining precise directions from each eye and the other for measuring each eye's dependability. Deng and Zhu et al. [8] developed two CNNs for generating head and gaze angles, which they integrated with a geometrically constrained transform layer.

We expand on these foundations by evaluating the speed vs. accuracy trade-off in real-time. The size of the picture input has a substantial influence on processing performance, and we adjust the size of the input image by changing the eye/face context.

Fischer [9] described a real-time gaze tracking system based on GPUs. This was accomplished using a model ensemble technique using two eye patches and a head posture vector as input, and it performed well for person-independent gaze prediction on different datasets.

Screen calibration: Estimating point-of-gaze.

A mirror-based calibration approach can be used to establish the proper camera and screen location [10]. This technique must be reapplied for various computer and camera setups, which is cumbersome and time-consuming. During human-computer interactions, for example, mouse clicks may provide important information for screen calibration [13]. However, this is based on the assumption that when individuals click, they are always looking at the mouse pointer. Methods such as second-order polynomial regression [10] and Gaussian process regression [12] have been used to predict gaze more generically. WebGazer [14] creates regression models that directly convert pupil positions and ocular characteristics to 2D screen locations in the absence of specified 3D geometry.

These approaches benefit from the absence of thorough modeling and perform well. However, because this calibration approach was designed exclusively for a certain gaze-prediction CNN, training directly on CNN features renders it non-modular.

In this project, we utilized the data efficiency of modular screen calibration algorithms that convert gaze-vectors to gaze-points utilizing geometric modeling, machine learning, and a combination of geometry and regression. We prioritize real-world efficiency over processing speed, as measured by the number of annotations necessary to attain appropriate accuracy.

Resource Utilization

Work Process

Based on the scope, requirements, and timeline of the project, we have chosen Agile with Scrum as this is a sprint-based project management system designed to deliver the most value to stakeholders and to guide teams in the iterative and incremental development of a product.

The Scrum Framework

An organization begins the project by providing a clear vision about its goals, as well as features in order of importance. The Product Owner manages the product backlog, which includes these features. Time boxes, or iterations or sprints, refer to the set amount of time the project team has to complete the features selected.

Each sprint typically lasts between one and four weeks, and this time span is maintained throughout the project to establish a schedule. A sprint backlog is created based on the items from the product backlog that the team believes can be completed in the sprint. Feature requests and tasks are created with the sprint-planning meeting.

During the sprint backlog phase, the team develops tasks that are within the sprint backlog. This period allows the team to focus on meeting the sprint goal and shielded from interruptions. It is not possible to change the sprint backlog, but the product backlog can be changed in preparation for the following sprint.

A 15-minute scrum meeting is held daily during the sprint between the team members. Team members stand in a circle and describe what they did yesterday, what they intend to accomplish today, and what is hindering them.

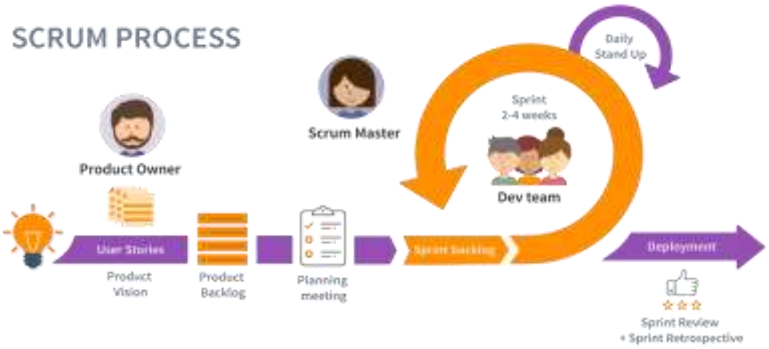


Figure 1: Agile with Scrum [16]

Team members then demonstrate their work to the stakeholders at the end of a sprint and collect feedback that will impact their work in the next sprint. A retrospective is also held to learn what they can do better. There is great importance to this meeting since it focuses on the three pillars of Scrum, namely transparency, inspection, and adaptation.

Resource Management

A resource management process involves determining resources' schedule and allocation in advance to ensure maximum efficiency. The word resource refers to anything that is needed to complete a task or project, from individuals with specific skill sets to software that is adopted.

Resource Management Process

Resources are managed during the planning stage and throughout the lifespan of a project. We identify and understand some stages of resource management in order to properly manage project resources. This is called the resource management life cycle.

Resource Planning

When a project scope Is defined, we estimate what resources each task will require. We also consider resources to implement risk management strategies and manage changes.

Resource Scheduling

It is important to ensure the availability of project resources when needed. Having a solid supply chain and aligning resource schedules with the overall project schedule will do the trick.

Resource Allocation:

As a continuous process, resource allocation is about picking the appropriate resource at the right time for a task. Creating a resource schedule, for instance, involves prioritizing tasks that are critical.

Here we are presenting Individual Contribution to the project:

Individual Contribution to the project:

Start Date	End Date	Days to Complete	Project Tasks
18.11.2021	24.11.2021	6	Requirement Analysis
25.11.2021	01.12.2021	6	Project Specifications
02.12.2021	15.12.2021	13	Project Modeling
16.12.2021	05.01.2022	20	Development
06.01.2022	19.01.2022	13	Integration and Testing
20.01.2022	26.01.2022	6	Functional Evaluation and Validation
02.12.2021	16.03.2022	104	Scientific Report

Table 1: Schedule

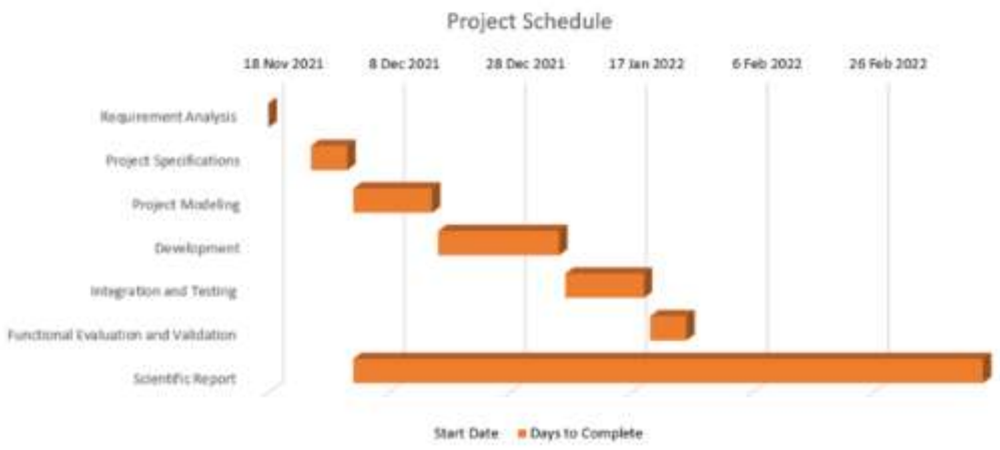


Chart 1: Project Schedule

Individual Contribution to The Project

SL No	Project Tasks	Team Members (X = Contribution)		
		Amena	Faiaz	Fayza
1	Requirement Analysis	X	X	X
2	Project Specifications	X	X	X
3	Project Modeling		X	X
4	Development		X	X
5	Integration and Testing	X	X	
6	Functional Evaluation and Validation	X		X
7	Scientific Report	X	X	X

Table 2: Resource utilization

Method and Materials

Proposed Method

This system will help physically impaired people who lacks verbal capability to express their basic four emotions: Happiness, Sadness, Anger and Surprise by moving their eyes only. The different emotions will be displayed on the screen in 2×2 window grid. The main objective of the system is to detect the zone observed by the user in a grid of cells with the least possible contact with the user, only gaze, and print out the detected emotion.

The basic algorithm of a gaze tracking device is:

i. 4-predefined emotion representation
ii. Input pre-processing by normalizing the facial images
iii. Predict the gaze vector using CNN
iv. Screen calibration to convert gaze-vectors to points on the screen

To comprehend this gaze two features of eye are examined: lateral and vertical. The lateral features measure the amount of sclera on the sides of the irises, while the vertical ones measure the white part above them. Gaze on screen then achieved using a gaze vector, which determines the pitch and yaw of the gaze in relation to the camera. Careful camera calibration to the screen is done to estimate the point of gaze on the screen.

The system also manages the striking disadvantage of gaze tracking devices Midas touch where it is not possible to define when the user is watching a point intentionally or is simply moving gazing across the screen. To confirm intentionality of the user's gaze, the system is built to prompt the emotion only if the gaze of the user is hold on to a particular grid of emotion for 3 consecutive seconds. If the gaze is moved from a grid before 3 second, the system considers it as an unintentional gaze.

Requirement Analysis

Technological Requirements

Hardware:

1. High-definition Camera: Resolution: 1920x1080 pixels
2. 64-bit CPU (Intel architecture)
3. 4 GB RAM
4. 5 GB free disk space

Software:

1. Operating System: Ubuntu 20.04.3
2. Python 3.8
3. OpenCV Library
4. NumPy Library

Functional and Nonfunctional Requirements

The requirements for this gaze tracking device to function properly and give a user-friendly real-time result are:

i. **Accurately locate the direction of the gaze**: The direction of gaze should be accurately located.
ii. **Be compatible with all subjects**: It must be compatible with all subjects.
iii. **Have no to minimal contact with the user**: System must ensure minimal invasiveness (no contact with the user).
iv. **Robust to variations in brightness**: It should be able to deal with variation in brightness.
v. **No obstruction of the user's view**: There can be no interference with the user's view.
vi. **Good precision in detecting the observed point**: It should have a high degree of precision, or to detect the observed point with a relatively low error percentage.
vii. **Timeliness and speed**: Offer good temporal dynamics and response speed making it a real time system.
viii. **Compatible**: Compatible with (relatively) large movements of the head and eyes.
ix. **Display emotions**: Express emotion with clear red mark on the image of the expression following the line of gaze.
x. **No additional hardware**: There is no requirement for additional hardware.

System Architecture and Design

The pipeline contains four parts:

A. Representation of Emotions: Human have three fundamental emotions: Happiness, Sadness and Fear. All other emotions are the combinations of these three emotions. Therefore, for this system we have selected four emotions to work with: Happiness, Sadness, Fear and Surprise.

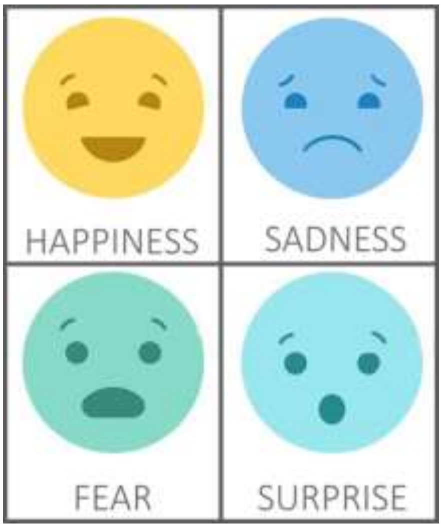
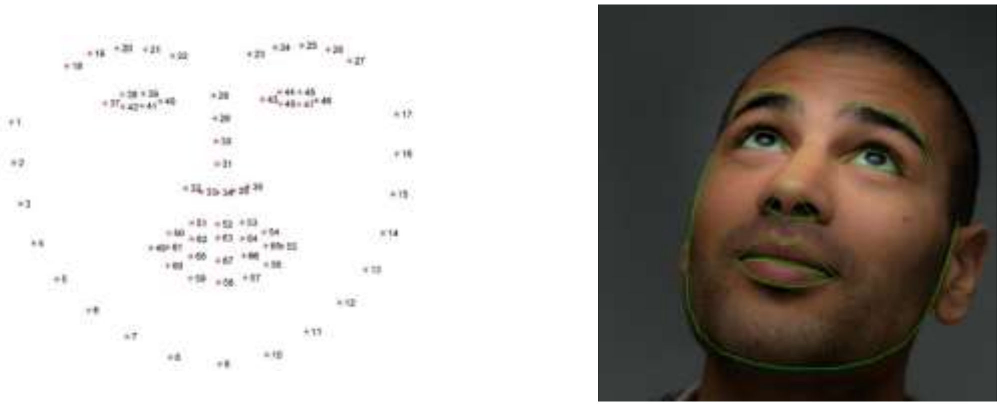


Figure 2: 2×2 window grid of 4 emotions to choose from by user's gaze

These four emotions are displayed in 2×2 window grid. This will be in front of the user who is unable to express emotions physically or verbally. The user will gaze upon the emotion they are feeling.

B. Input pre-processing by normalizing the facial images
The input to the system is obtained from facial images of subjects using the `OpenCV` library. Through a face finding and facial landmark detection using `NumPy` library, the face and its key parts are localized in the image.



(a) (b)

Figure 3: (a) 68 facial landmark markups used by numpy library. (b) landmarked face using the 68 facial landmark markups used by numpy library

The detected 2D landmarks are fitted onto a 3D model of the face. Therefore, facial landmarks are roughly localized in the 3D camera co-ordinate space. By comparing the 3D face model and 2D landmarks, the head rotation matrix, translation vector, and the 3D eye locations are obtained in 3D camera coordinate space. A standardized view of the face is now obtained by defining a fixed distance between the eye centers and the camera center.

C. Predict the gaze vector using CNN

To predict the point of gaze on the screen gaze vector is required. Gaze vector determines the pitch and yaw angles of the gaze vector with respect to the camera from the normalized pre-processed images.

To train the system to predict the gaze vector MPIIGaze [***] dataset is used as input of facial images. It contains 37,667 facial images from 15 different participants. The images have variations in illumination, personal appearance, head pose and camera-screen settings. The ground truth gaze target on the screen is given as a 3D point in the camera coordinate system. System predicts the pitch and yaw angles of the gaze vector with respect to the camera from the normalized pre-processed images.

The predicted virtual gaze vectors can be transformed back to the actual gaze vector with respect to the real camera using the transformation parameters obtained during image pre-processing. These vectors can then be projected onto a point on the screen after screen calibration.

D. Screen calibration to convert gaze-vectors to points on the screen

To project the predicted 3D gaze vectors (in the camera coordinate space) to 2D gaze-points on a screen, the position of the screen with respect to the camera must be known which is difficult to obtain in real world settings. The aim of screen calibration is to estimate this geometric relation between the camera and the screen coordinate systems such that the predicted gaze vectors in camera coordinates are calibrated to gaze-points in screen coordinates.

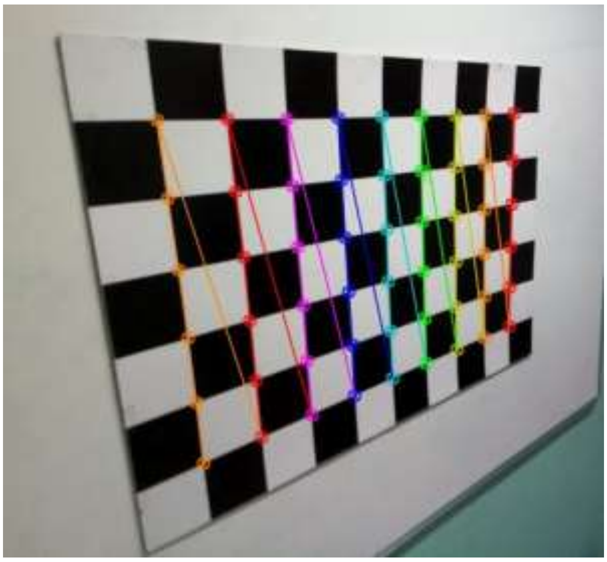


Figure 4: 9×6 chessboard pattern for screen calibration

For our system’s screen calibration, we used 9x6 chessboard pattern.

System Model

Pipeline of the System

The pipeline of the system architecture is represented in the following diagram

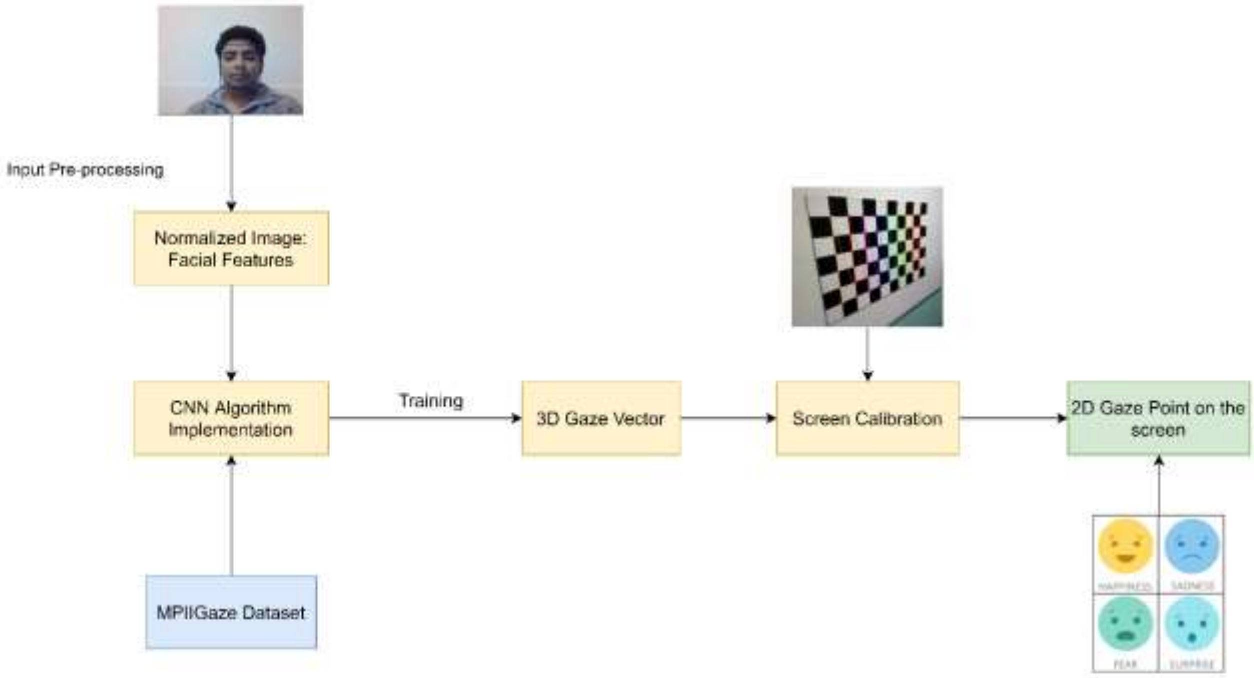
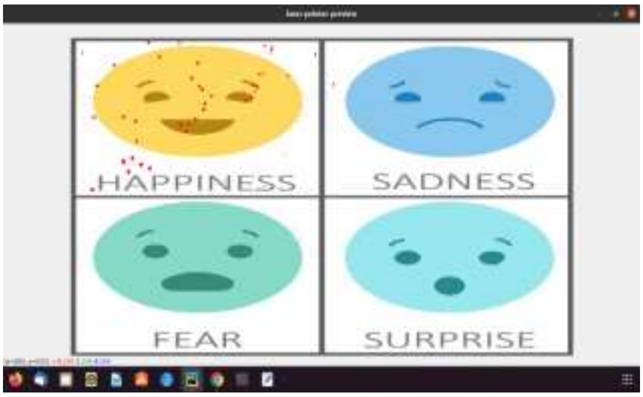


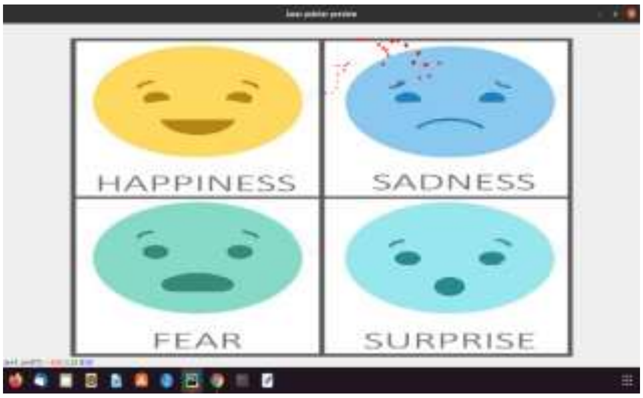
Figure 5: Pipeline of the system

Result

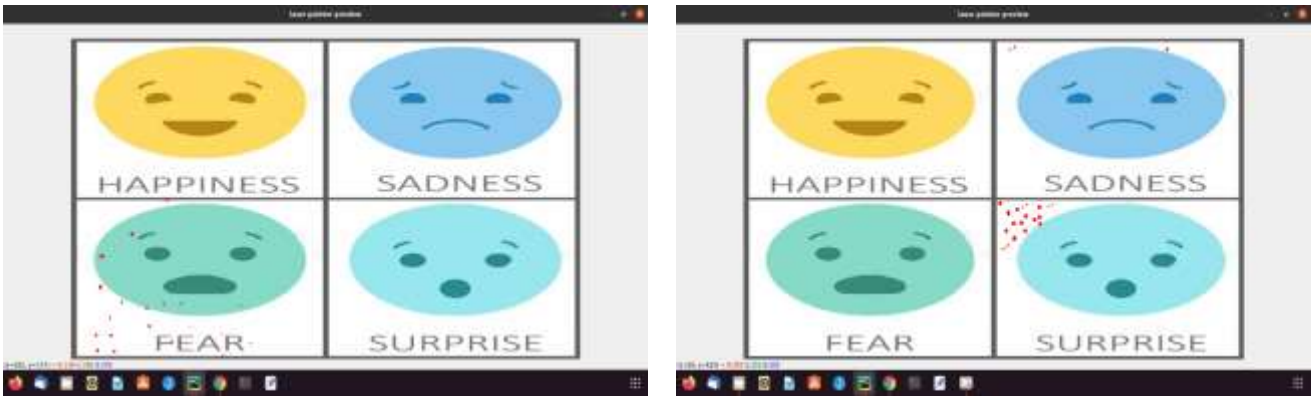
After presenting the grid window of emotions the, the user will look at the emotion that they want to express. We have required the following result after looking at each of the emotion:



(a) HAPPINESS



(b) SADNESS



(c) FEAR

(d) SURPRISE

Figure 6: (a) Result when looking into grid HAPPINESS. (b) Result when looking into grid SADNESS. (c) Result when looking into grid FEAR. (d) Result when looking into grid SURPRISE.

Discussion

Considering the impact of the system in expressing the emotions of physically imparted people, this is an effective and feasible system. This system expresses the emotion by tracking gaze over the window grid of emotions. It also tackle the issue of Midas touch by tracking the gaze and waiting 3 seconds to present the emotions.

What this system lack is computational delay. As the calibration and system training require much computational power, this system can be improved significantly.

Future Research

The system's accuracy is currently insufficient. In addition to the MPIIGaze dataset, the EYEDIAP [14] dataset may be used for gaze estimation. This dataset handles distant RGB and RGB-D (generic vision and depth) cameras for gaze estimation [14]. We would want to incorporate dataset into our system in order to do a comparison assertion of the datasets in relation to the system's efficiency. We also intend to integrate Hybrid eye-tracking, which combines active and passive eye-position measurement t [15]. With corneal imaging and iris contour, Hybrid Estimation aids in the recovery of precise eye location [15].

Conclusion

Physically handicapped people, particularly those with poor movement abilities, have found it challenging to communicate their feelings. The expression of human emotions may be greatly aided by the use of eye gaze. We have designed a system using gaze tracking to let impaired persons communicate their basic emotions. The technology is inexpensive and may be embedded into any mobile or portable device. Substantial levels of improvement in terms of effectiveness and precision, on the other hand, are required. Even though accuracy is not ideal, the system has established the groundwork for our future endeavors. We are certain that our continued efforts will bring a considerably more efficient result.

References

[1] Jacob, R. J. What you look at is what you get: eye movement-based interaction techniques. In Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 11-18 (1990, March).

[2] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4511–4520 (2015)

[3] Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(1), 162–175 (2017)

[4] Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation. In: European conference on computer vision (2018)

[5] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)

[6] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)

[7] Cheng, Y., Lu, F., Zhang, X.: Appearance-based gaze estimation via evaluation guided asymmetric regression. In: Proceedings of The European Conference on Computer Vision (2018)

[8] Deng, H., Zhu, W.: Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In: Proceedings of the International Conference on Computer Vision. pp. 3162–3171 (2017)

[9] Fischer, T., Jin Chang, H., Demiris, Y.: Rt-gene: Real-time eye gaze estimation in natural environments. In: Proceedings of the European Conference on Computer Vision (2018)

[10] Rodrigues, R., Barreto, J.P., Nunes, U.: Camera pose estimation using images of planar mirror reflections. In: European Conference on Computer Vision (2010)

[11] Kasprowski, P., Harezlak, K., Stasch, M.: Guidelines for the eye tracker calibration using points of regard. In: Information Technologies in Biomedicine, Volume 4. pp. 225–236. Springer International Publishing (2014)

[12] Tripathi, S., Guenter, B.: A statistical approach to continuous self-calibrating eye gaze tracking for head-mounted virtual reality systems. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 862–870. IEEE (2017)

[13] Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., Hays, J.: Webgazer: Scalable webcam eye tracking using user interactions. In: Proceedings of the International Joint Conference on Artificial Intelligence. pp. 3839–3845 (2016)

[14] Mora, K.A., Monay, F., & Odobez, J.: EYEDIAP Database: Data Description and Gaze Tracking Evaluation Benchmarks. (2014)

[15] Plopski A, Nitschke C, Kiyokawa K, Schmalstieg D, Takemura H.: Hybrid eye tracking: combining iris contour and corneal imaging (2015)

[16] Tuleap. 2022. *Understanding Agile Scrum in 10 minutes • Tuleap*. [online] Available at: <https://www.tuleap.org/agile/agile-scrum-in-10-minutes> [Accessed 15 March 2022].

[17] Velichkovsky, Boris, Andreas Sprenger, and Pieter Unema. "Towards gaze-mediated interaction: Collecting solutions of the “Midas touch problem”." *Human-Computer Interaction INTERACT’97*. Springer, Boston, MA, 1997.

[18] Gudi, Amogh, Xin Li, and Jan van Gemert. "Efficiency in real-time webcam gaze tracking." *European Conference on Computer Vision*. Springer, Cham, 2020.

Project Link: https://drive.google.com/drive/folders/1BvjjbHCfxvjHlKJPFpTHvysTFieqP-81?usp=sharing