3D Gaze Estimation
Senior Project II

Mukhit Yelemes
Department of Computer Science
Nazarbayev University
Nur-Sultan, Kazakhstan
mukhit.yelemes@nu.edu.kz

Alisher Kenzhebayev
Department of Computer Science
Nazarbayev University
Nur-Sultan, Kazakhstan
alisher.kenzhebayev@nu.edu.kz

***Abstract*—The ability to track the movements of a human body using a computer interface has applications in different areas in all kinds of day-to-day situations, from Augmented Reality (AR) to practical use in industrial production. The purpose of this project is to be used within larger Live-Feeling Communication project to create a system for intention estimation and collect a data set for future improvement of football viewers experience. In this project, we are proposing the system that will be able to estimate rough human PoG direction using a stereo camera source, by employing the 3D estimation of the user's face and depth values, as generated by the setup. And by using these estimates, providing the rough direction for the PoG of the user's point of interests on the screen.**

***Index Terms*—3D, intention point, gaze vector, eye pupil.**

I. Introduction

The ability to track the movements of a human body using a computer interface has applications in different areas in all kinds of day-to-day situations, from Augmented Reality (AR) to practical use in industrial production. As it stands, one of the successors of this approach that have showed up recently was the system tracking the movements of the eyes and head of the human subject. This advancement had enabled computer scientists to look further into matter of eye tracking and creation of a new discipline in research, called PoG (Point of Gaze) estimation. The combined approach in theory enables for ease-of-use in computers, by allowing to replace the standard mouse pointer with simple eye movements.

In this project, we are proposing the system that will be able to estimate rough human PoG direction using a stereo camera source, by employing the 3D estimation of the user's face and depth values, as generated by the setup. And by using these estimates, providing the rough direction for the PoG of the user's point of interests on the screen.

In this project we tried to find the most resource saving and real-time solution for the derivation of PoG estimation, that uses stereo image sources. There are existing papers for that describing the method at a high level, enough to get the main understanding of the subject, but not its internal structure. As well as numerous libraries designed to pinpoint the rough estimate of face edges and key points, such as MTCNN, Face Alignment and others, designed to detect or generate the depth values from stereo calibrated cameras, as OpenCV. After that our job is to try out different existing solutions in combinations and come up with our own decision that will work the best.

II. Related Work

In [5], the paper describes how the general approach for detecting human PoG works. We will be using that as well as [4] to find the working solutiong for detecting PoG using the depth values as produced by the DisparityMap in OpenCV. As a face alignment method, we used the models described in [2]. The continued work will be done with the 3rd party library as provided in [1] to produce the 3D models of face in order to detect the facial features as done in previous approch in MTCNN.

III. System Design and Architecture

System hardware is consist of two cameras and display. The cameras record Full-HD video at 30 FPS frame rate and located on top of the display. One camera is the main one and is located in the center, whereas the second is located next to it and is used to create a depth map.
Algorithm is broken down to two stages: Calibration and Working. The requirement for calibration stage was caused by the fact that the users eyeball center coordinate is initially unknown. Therefore, in the Calibration stage there will be captured an image of user looking directly to central camera with both eyes. With this image, system will create a transformation matrix that could compute the eyeball center coordinates given the other facial landmarks. As the radius of eyeball there was used an average eyeball size from medical databases, which was equal to 24 mm in diameter.
At the Working stage, system is using computing intention points in real time. In order to get the 3D gaze vector, there are computed the coordinates of eyeball center and eye pupil. Eyeball center comes out of the transformation matrix from Calibration stage applied to 3D face landmarks. Eye pupil is localized by the neural network, the architecture of which is given in figure 3. It takes cropped eye images at size of 50 by 35 and returns the 2D pupil coordinates, that further transformed to 3D coordinates.
With he combination of two camera, from stereo image there

is computed a depth map. The 3D gaze vector and depth map are passed to neural network to get the intention points. The architecture of that model is given in figure 4. It returns two dimensional point of display.

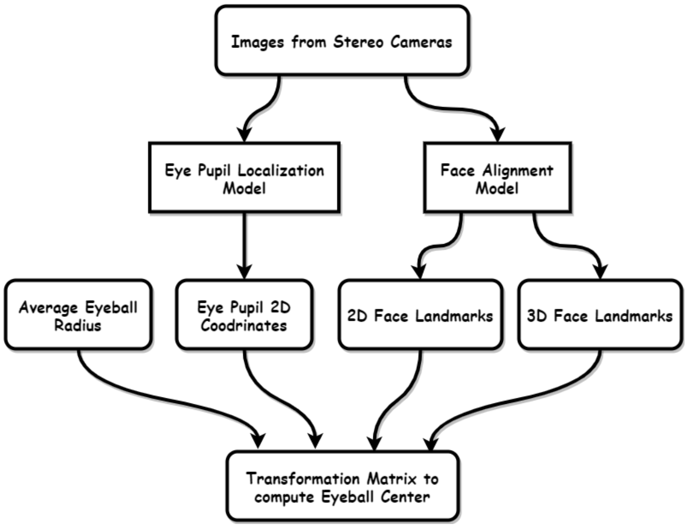
Images from Stereo Cameras
Eye Pupil Localization Model
Face Alignment Model
Average Eyeball Radius
Eye Pupil 2D Coodrinates
2D Face Landmarks
3D Face Landmarks
Transformation Matrix to compute Eyeball Center

Fig. 1. Algorithm of System Calibration Stage

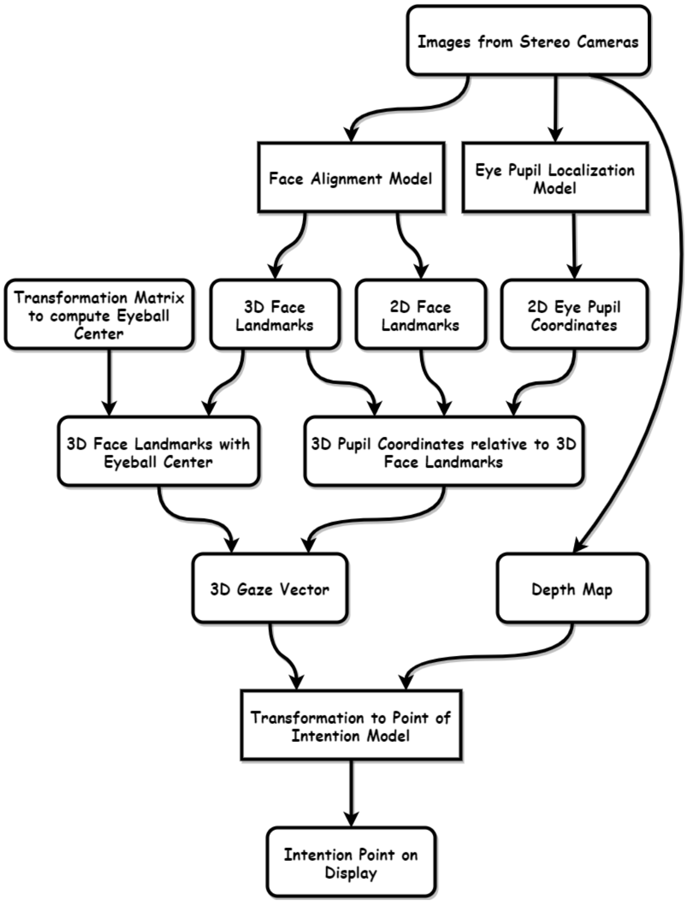
Images from Stereo Cameras
Face Alignment Model
Eye Pupil Localization Model
Transformation Matrix to compute Eyeball Center
3D Face Landmarks
2D Face Landmarks
2D Eye Pupil Coordinates
3D Face Landmarks with Eyeball Center
3D Pupil Coordinates relative to 3D Face Landmarks
3D Gaze Vector
Depth Map
Transformation to Point of Intention Model
Intention Point on Display

Fig. 2. Algorithm of System Work Stage

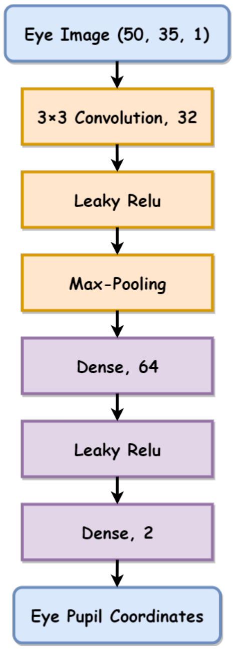

Fig. 3. Architecture of Model for Pupil Localization

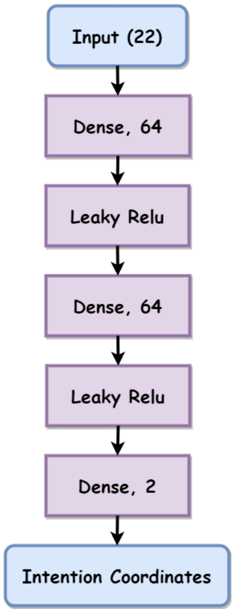

Fig. 4. Architecture of Model for Transformation to Intention Point

IV. System Features and Functionalities

The main functionality of the system is to track the eye movements and transform the to intention point at the display. As part of Live-Feeling Communication Project, that feature is going to used to identify the relation between different situations during a football match and viewers most interested area on the screen. So, the system could be run with active football match on display and collect the data of users intention points in all periods of time throughout the match and create a data set from that.

The feature of the system's algorithm is that it creates the 3D eye landmarks relative to face. Because of this, the system is able to recognize the direction of the gaze even with large rotations of the head, so there will be no limitations for the user in terms of very specific angle, very close distance or strict head pose to make the system work.

The interaction with the user requires the initial calibration stage. The user will be asked to look directly to central camera to capture the image. After that, the system will quickly complete the remaining steps of calibration and will be ready to work. At the Main stage, the system will be computing the point where user is looking at the display in a real time without any intervention on his part.

In addition to the main purpose of use in LFC Project, there are many other applications of the system. For example, it could replace the mouse for disabled people and make their interaction with their PC or other device easier and more efficient. Also, it can be used by streamers to display their intention points during the game, so that viewers could understand his way of thinking.

V. Evaluation

The initial prototype was made with using Image Processing techniques to localize eye pupil and k-nearest neighbours model to estimate the intention point. The testing of that system showed too big error values on both x and y axis, accounting to 1/3 of the display in height. Therefore, there was done analysis of system in a search for improvement methods.

Table 1: Intention Point Estimation methods

Methods vs MAE in pixels	MAE, x axis	MAE, y axis	MAE overall
kNN	600	300	675
CNN with Image Processing	400	200	450
CNN with MLP (Estimated)	200	100	225

The first idea was to replace kNN model with multilayer perceprton, because there was observed significant deviations of output within very close input values. The results of MLP showed reduction of mean absolute error by 33%. However, that still was comparable to 20% of display size and was not accurate enough to give reasonable data set for LFC project.

Table 2: Pupil Localization methods

Methods vs MSE in pixels^2	MSE, x axis	MSE, y axis	MSE overall
Image Processing	4	4	5.66
MLP	2	1	2.24

Next, the detailed evaluation of system's software components revealed the relatively low accuracy of eye pupil localization method. Based on Image Processing techniques to preprocess eye image to remove unnecessary features, that method was using Min-Max localization to find the center of eye pupil. However, its result was dependent on lightning conditions of environment and overall results were low. As an alternative method, there was trained CNN model with similarly preprocessed images. The data set for the model was consist of 700 manually labelled eye images with shape of 50 by 35. The comparison of those two methods is shown in the table 2. The localization error was reduced by 50% on x axis and 75% on y axis. Due to time limitations, that method was not tested within overall system performance, so there is given the expected influence of it to system's intention point prediction accuracy.

References

[1] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou and Georgios Tzimiropoulos *Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression*

[2] Bulat, Adrian and Tzimiropoulos, Georgios *How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)*, International Conference on Computer Vision, 2017.

[3] Carlos H. Morimoto, Marcio R.M. Mimica *Eye gaze tracking techniques for interactive applications*. Computer Vision and Image Understanding, Volume 98, Issue 1, 4-24, 2005.

[4] Qiang Ji, Xiaojie Yang. *Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance.*. Real-Time Imaging, Volume 8, Issue 5, 357-377, 2002.

[5] R. Newman, Y. Matsumoto, S. Rougeaux, A. Zelinsky. *Real-time stereo tracking for head pose and gaze estimation*. IEEE International Conference on Automatic Face and Gesture Recognition, 2000.