Real-time, Calibration-free, Gaze Tracking for Human-Computer Interactive Systems

Avishek Bose
Architech Labs
Email: jbose@architech.ca

Valentin Tsatskin
Systems Design Engineering
University of Waterloo
Architech Labs
Email: vtsatskin@uwaterloo.ca

***Abstract*—Enhancing Human Computer Interaction through Gaze Tracking is a well studied problem with varying applications in areas such as interest detection, marketing analytics and interactive displays. However, past works rely on either intrusive apparatuses that are expensive and not conducive to natural head motions or difficult calibration schemes which need long setup time or require large data sets for training. In this work, an architecture is presented that performs real-time calibration-free gaze tracking at the cost of precision of the eye gaze point which is mapped to one of four quadrants of the display. To this end, smart feature selection is employed in order to optimize for real-time performance and completely eliminate the need for calibration, thereby increasing the usability of the proposed system in real-world environments.**

I. Introduction

Human computer interaction (HCI) systems have been an important component of modern society [1]. Human eye-gaze provides important information regarding a person's emotional state and cognitive processes [2] as it is one of the most important ways through which one perceives the world. From Literature, eye gaze is usually parametrized by head pose and eye locations [3]. The estimation of these parameters often requires the use of expensive and cumbersome hardware that is restrictive to natural head motions. Consequently, a variety of gaze tracking techniques exist in the literature and can be either classified into either feature based approaches that require intrusive apparatuses for eye tracking [4], [5] and head pose modelling or learning based methods that utilize computer vision and image processing techniques. The focus of this work is on creating a gaze tracking system that is real-time, calibration-free and agnostic to the monocular camera being used.

The underlying guiding principle is to promote responsive human interaction such that users may enter and exit the scene freely without compromising the integrity of the detection. To this end, granularity of the gaze tracking and detect eye gaze are sacrificed for the four quadrants of a display screen. To achieve this heuristics are used to create correspondences between 2D facial features observed through video input to 3D coordinates that in turn help estimate 3D head pose. Furthermore, state-of-the-art iris centre detection [6] is used which in conjunction with 3D head pose estimation are used as smart features for the gaze mapping algorithm. The contributions for this work include insights into the selection of smart features and their approximation that allow for real-time calibration-free gaze tracking along with robustness to new users.

II. Related Work

While there are various methods for tracking gaze, most approaches fall under two categories namely 1) *feature based gaze estimation* and 2) *appearance based models*. Feature based methods generally seek to identify key features in the human eye such as eye contours and corneal reflections that are then used in gaze estimation. Feature based methods can be further subdivided into model based approaches where an explicit geometric model of the eye is used to estimate the 3D gaze direction vector and shape based methods where the gaze direction is inferred from features in the image of the eye [4], [5].

The prevailing approach in model based approaches involves the reconstruction of the optical axis of the human eye in 3D. The visual axis, defined to be the vector from optical axis to the iris centre, is then used to intersect the scene (i.e. a screen). Shape based methods assume the mapping from image features to gaze co-ordinates have a particular parametric form such as a polynomial or other nonparametric forms. Appearance-based gaze estimation methods directly use eye images for mapping to screen coordinates and as a result do not rely on camera calibration and geometry of the eye and can potentially work with low-resolution eye images [7] but with an underlying assumption that there are head pose variations. This work may be classified as an appearance based method with an important distinction in that the assumption of a fixed head pose is relaxed as it is approximated thus allowing for natural viewing motions.

III. Experimental Setup

The goal of collecting data was to simulate an environment where participants were actively engaging with a stimulating display. An LG 55LS5700-UA was used to show participants points to look at. The display measured 55 inches diagonally across the panel. A Logitech HD Pro Webcam C920 was mounted on the top of the display in the centre to record the participants.

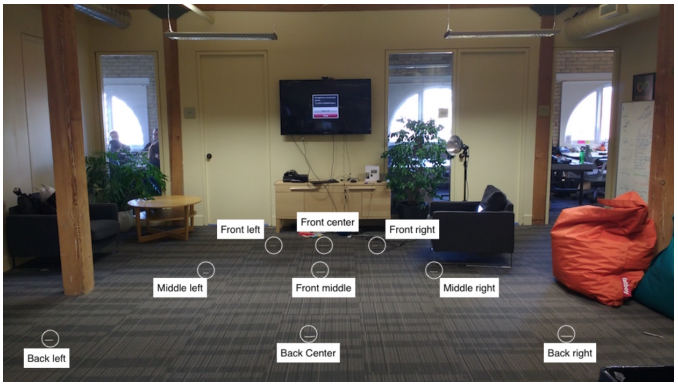


Fig. 1. Experiment Setup

Participants were guided through the process of looking at the various points of the screen by an experimenter, in various head poses as shown in Figure 3. 23 people were asked to perform the task at each of the 9 locations in Figure 1. The total cleaned data set consists of 17004 frames where the participants were volunteers selected from a public advertisement and consisted of 15 males and 8 females. For each point on the screen, participants were asked to perform two tasks: 1) head orientation fixed in the centre of the screen and only using pupils to look at a point 2) head was free to look at the point "naturally".

IV. Proposed Method

An overview of the system architecture is presented in Figure 2. As input to the system a video feed from a single monocular camera positioned above the interactive display (i.e. a TV screen) is used. A standard off-the-shelf face detector from the open source DLIB library is used. Next, facial landmarks are retrieved from the detected face using the algorithm proposed by Kazemi and Sullivan that uses an ensemble of regression trees [8]. A 3D head pose is approximated by fitting a generic 3D human head model. In parallel to the head pose estimation each eye is cropped out from the input frame as given by the eye contour based on the previously detected landmarks. Iris centres are then detected from the cropped eye images using Skordra's Radial Symmetry Transform on the cropped images [6], [9]. Head pose estimation, iris centres, and normalised facial landmarks are combied into features for Random Forest Decision Trees (RF) clustering to estimate eye-gaze.

A. Face Detection and Landmarks

Facial landmarks provide critical information regarding the shape of the eye contour which is indicative of the location of the gaze i.e. looking below the optical axis without moving the head generally results in a flatter eye contour shape. The frame rate of the monocular camera used is sixty frames per second and as a result,the assumption that the human face does not move an inordinate amount in a predefined time slice is used. The proposed method uses the DLIB face detector which employs a Histogram of Oriented Gradients (HOG) feature

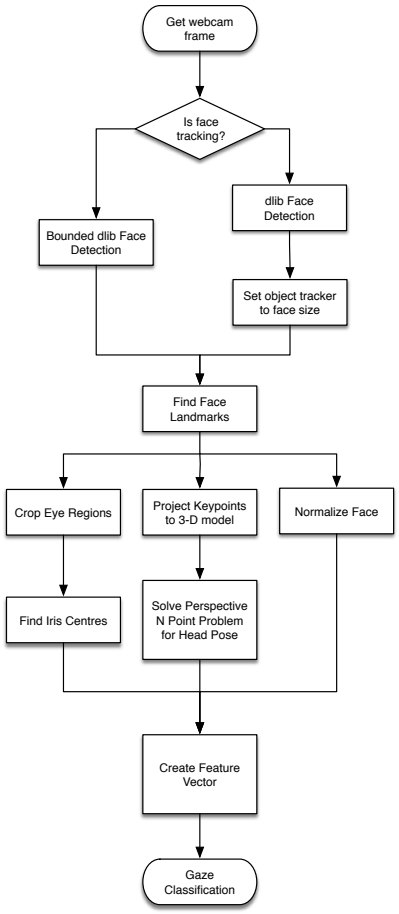


Fig. 2. Gaze Tracking System Architecture

combined with a linear classifier, an image pyramid, and sliding window detection scheme to detect faces [10]. In order to optimise the speed of detection, faces are only detected in the entire input frame once, and then an object tracker is used to bound the face search space until face detection fails. The face bounding box is then used as the input for the DLIB's facial landmark detector which returns 68 landmarks for every face [8], [10].

B. 3D Head Pose Estimation

To estimate 3D head pose a generic 3D model of a human head is used from which manually selected landmark features (nose tip, left eye corners, right eye corners, left eye centres, right eye centre, left mouth edge, right mouth edge) are used (Figure 4a). The intrinsic camera matrix is approximated using the image height and width. From the extracted feature landmarks, a Perspective N point problem is solved using an iterative solution in O(n) to create correspondences between 2D image points to 3D camera coordinates. Solving the Perspective N point problem [11] results in the rotation matrix and translation vector of the camera which is used to model the roll, pitch, and yaw needed for 3D head pose. Due to the nature of the problem, using these heuristic approaches allows

to approximate 3D head pose without compromising speed or gaze mapping accuracy of the system. The head pose angles vary as functions of distance from the interactive display due to the increase in people's field of view. Consequently, the experimental setup includes large variability of head pose as users are not constrained to a specific distance relative to the camera.

C. Iris Centre Detection

Accurate localisation of the iris centre is critical for gaze tracking. While it may seem an obvious approach to locate the centre of the iris by honing in on the centre of the eye, this assumption is fallacious and can be easily disproved by by extreme sideways glances (e.g. when a person looks to the far left or right without moving their head). The proposed solution follows the method outlined by Skodras and Fakotakis [6], [9], [12], [13]. First begin by defining a Region Of Interest (ROI). The ROI is selected based on the eye corner points given by the facial landmarks as well as the interocular distance between the two eye centres which is approximated using the centre of the eye contour for each eye. In the figure below, parameter d is the scale of the ROI and is selected based on the detection accuracy of the central and corner points.

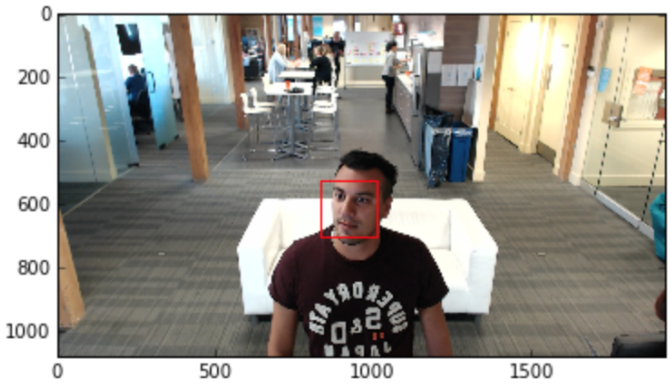
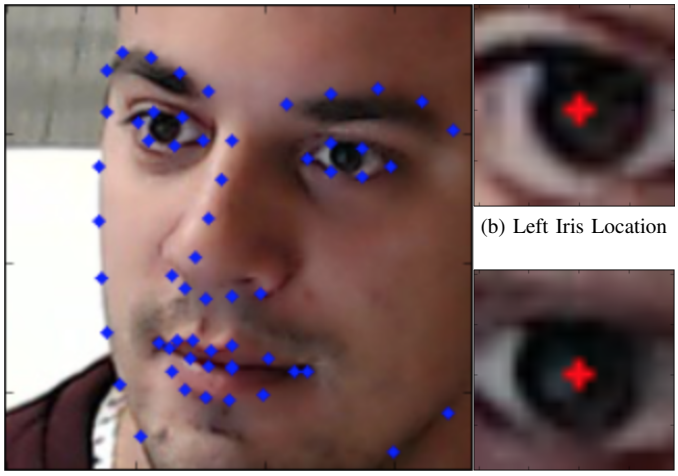


Fig. 3. Face detection

D. Gaze Mapping

The biggest challenges pertinent to gaze tracking are poor lighting conditions, variability in head pose and in pupil location due to fast eye movements, eye blinks and variability in position of the person relative to the camera. Machine learning is used for gaze mapping. Training is done with head positions and iris centre locations corresponding to specific locations on the screen. To normalise for variations in face shape and feature location, average face coordinate are found which are defined as the mean coordinate from the detected landmarks. The following features serve as input to training for the classifier:

- Roll, Pitch, Yaw
- Normalized Eye Contours (Left and Right)
- Normalised Displacements between iris centre and contour centre (Left and Right)



(a) Face Landmarks (b) Left Iris Location (c) Right Iris Location

Fig. 4. Facial features extracted

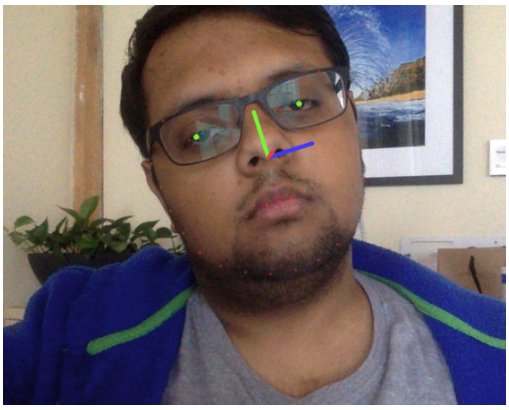


Fig. 5. 3D Head Pose and Iris Centre

- Normalized iris centres (Left and RIght)
- Average Face point, Standard deviation for Face Landmarks

The Normalisation step involves subtracting each feature's coordinates with the average face coordinate and then dividing by the standard deviation of the face. To model variations in head pose as well tricky cases i.e. Tilting the Head towards one quadrant and using side ways glances, the training procedure contains an equal number of frames for each case. The proposed method uses Random Forest Decision Trees to perform the gaze mapping after normalisation so that the model is robust to face and head pose variations. Based on the experimental results, the 2D normalized displacement vector for each eye and the three head pose angles were the most important features used by the RF algorithm.

V. Results

The proposed framework encompasses training using Support Vector Machines (SVM), Random Forest Decision Trees (RF), Gradient Boosted Trees (GBT), and AdaBoost. Following that, the accuracy of each method was evaluated by training and testing on a withheld test data set, split between

participants randomly into $^2/_3$ and $^1/_3$ respectively. In order to evaluate robustness of new users to the system. Each of these experiments were run 10 times to understand the variance of participants' data sets.

As discussed in the following *Limitations* subsection, samples from the back row of the setup were excluded during testing and training due to failures in face detection.

TABLE I
MEDIAN CLASSIFICATION PERFORMANCE ON DATA SETS SPLIT BETWEEN PEOPLE

Method	Accuracy	Recall			
		Top Left	Top Right	Bottom Left	Bottom Right
AdaBoost	68.10	68.63	66.63	70.46	67.84
GBT	90.48	89.00	89.48	91.36	92.71
RF	90.76	90.56	90.81	91.24	92.34
SVM	82.71	95.32	79.56	78.10	76.35

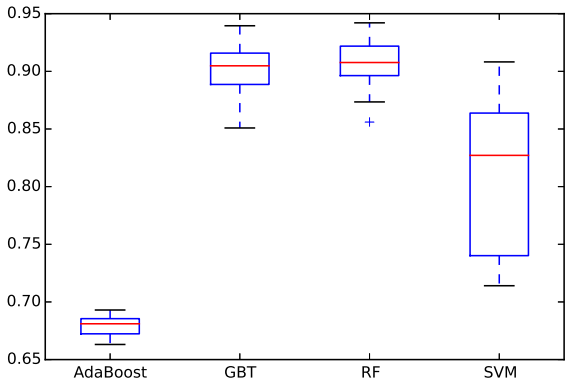


Fig. 6. Accuracy Scores over 10 trials

Based on the performance metrics, RF was selected for use in the final system. The proposed system, using RF, performs between 85.61–94.21% in accuracy when tested on unseen users located in the front and middle rows. Thus, the system is relatively robust to unseen users located within 2.5 meters of the camera and is less reliable beyond that.

A. Limitations

During feature extraction, 17.4% of the data set failed at detecting a face. Upon investigation, it was revealed that 75.5% of the cases of face detection failure were from the back row, which indicates failures begin to occur due to distance at 2.2–3.75 meters from the camera. This is due to the minimum size of face the DLIB face detector attempts to search for. Setting this parameter to search for smaller faces would increase face range at the expense of speed of face detection. The remaining 24.5% of face detection errors are speculated to be faces tilted into ranges outside of DLIB's face detection supported face angles.

From the frames which had the face detected successfully, 0.36% were not able to have their iris centres extracted.

All frames which failed on either face detection or iris centre extraction were not included in the machine learning training or testing.

VI. DISCUSSION

This paper proposed a gaze tracking framework which incorporates 3D head pose and state of the art iris detection in order to create smart features for the gaze mapping algorithm. By design, heuristics for the 3D Human face and intrinsic camera matrix are used which allow to approximate the 3D head pose. The proposed gaze mapping algorithm is also compared against several machine learning. The system is capable of performing real-time gaze tracking and does not need any prior calibration. As a result, the system sacrifices the precision of gaze mapping to four predefined locations on a display. Limitations are suspected to be overcome by using a camera with a higher resolution such that there is sufficient visual information to detect faces and iris centres. This work provides insights and evaluations on the trade offs required to achieve real-time and calibration-free gaze tracking for interactive systems.

ACKNOWLEDGMENT

The authors would like to thank Yanshuai Cao, Fartash Fagrhi, Olesya Peshko, JinSung Kang, Marc Bender, Foteini Agrafioti and Architech for their advice and support in the pursuit of this research.

REFERENCES

[1] F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab, "Human-computer interaction: Overview on state of the art," *International Journal on Smart Sensing and Intelligent Systems*, vol. vol.1 no.1, 2008.

[2] A. Matthews, E. Fox, J. Yiend, and A. Calder, "The face of fear: Effects of eye gaze and emotion on visual attention," *Visual Cognition*, vol. 10, no. 7, pp. 823–835, 2003.

[3] R. Valenti, N. Sebe, and T. Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 21, no. 2, pp. 802–815, 2012.

[4] H. Chennamma and X. Yuan, "A survey on eye gaze tracking techniques," *Indian Journal of Computer Science and Engineering*, vol. Vol. 4, No. 5, pp. 388–393, 2013.

[5] D. Witzner Hansen and Q. Ji, "In the eye of the beholder- a survey of models for eyes and gaze," *I E E E Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478 – 500, 2010.

[6] E. Skodras and N. Fakotakis, "Precise localization of eye centers in low resolution color images," *Image and Vision Computing Journal*, 2015.

[7] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Perceptual user interfaces group, 2scalable learning and perception group," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

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

[9] F. Timm and E. Barth, "Accurate eye centre localisation by means of gradients," *VISAPP*, p. 125130, 2011.

[10] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[11] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o(n) solution to the pnp problem," *International Journal of Computer Vision*, vol. vol. 81, issue 2, pp. 155–166, 2008.

[12] S. S. W. G. Peng Yang, Bo Du, "A novel pupil localization method based on gaboreye model and radial symmetry operator," *International Conference on Image Processing*, vol. vol. 1, pp. 67–70, 2004.

[13] G. Loy and A. Zelinsky, "A fast radial symmetry transform for detecting points of interest," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 25(8), p. 959973, 2003.