Appearance Based Gaze Tracking Using Supervised ML

Daniel Melesse'20, Mahmood Khalil'20, Elias Kagabo'20
Advisors: Dr. Huang and Dr. Ning

ABSTRACT

Applications that use gaze have become increasingly more popular in the domain of human-computer interfaces. Advances in eye gaze tracking technology over the past few decades have led to the development of promising gaze estimation techniques. However, gaze detection was previously constrained to indoor conditions, so techniques required in most cases additional hardware. Recently, Gaze detection techniques try to eliminate the use of any additional hardware that can be expensive and intrusive, while maintaining the accuracy of detection. In this paper, an in-house video camera-based tracking system has been analyzed. This process involves image acquisition using a USB web camera mounted on the user's PC or laptop at a fixed position.

In order to determine the point of gaze, Viola Jones is used to detect the face of the user from the image frame. The gaze is then calculated using image processing techniques to extract the coordinates of the pupil. Using the calculated center, a database is created which forms the training and testing data set. This database is used to train a classifier using a multi-class supervised learning method-Support Vector Machine (SVM) to distinguish point of gaze from input face image. Cross validation is used to train our model. Also, confusion matrices are used to evaluate the performance of the classification model. Accuracy is determined to assess performance. Evaluation of the proposed appearance-based technique using various kernel functions is assessed in detail.

INTRODUCTION

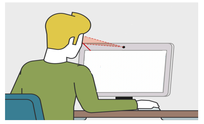
Decades of research has been put into the development of gaze tracking techniques. Although there has been significant progress in the field, gaze tracking is still an expensive commodity, thus making it inaccessible to everyone. This work aims to build on existing algorithms to develop an appearance gaze tracking technique and study its performance. Classification of gaze coordinates is done using multi-class SVM, while K-fold cross validation is used to verify robustness of classification. Different kernel functions are used to assess the separability of the data in different states. After analyzing the performance of the technique, future improvements are suggested that can help increase accuracy.

SYSTEM OVERVIEW

TECHNICAL APPROACH

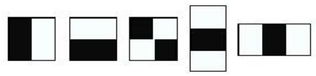
Image acquisition:

For this stage of the process, a USB 2.0 C920 Logitech Web camera is used. The web camera is declared initially in the code. Following this, a folder is created to store the acquired images that are going to be eventually processed. A for loop is also created that helps capture a predefined "N" number of images every time the code runs. A single person was used as a test subject in this work. For this, a standard LCD monitor was split into different sections: 2, 4, 6, and 9. In total, 100 data images were collected for 2 by 1 grid, 400 for 2 by 2 grid, 600 for 2 by 3 grid, and 900 for 3 by 3 grid.



Viola-Jones:

The algorithm uses Haar features as a mean to detect facial features. Haar features are made up of a range of combinations of black and white pixels. The black pixels with pixel intensity 1 and white pixels of pixel intensity 0. The features are composed of either two or three rectangles. Face candidates are scanned and searched for Haar features of the current stage.



Hough Circle Transform

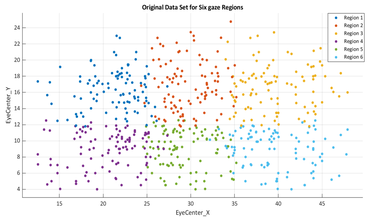
After detecting the face and the eye region from an image, HCT is used to approximate the coordinates of the pupil center by first defining the minimum and maximum values of the eye's radius. Minimizing the region scanned for circles reduces the number of detected, therefore increasing the accuracy for pupil center detection.

Classification: SVM an Cross Validation

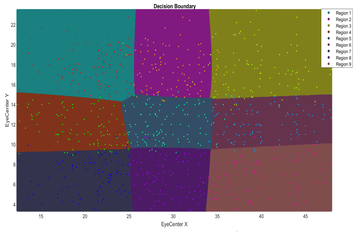
Support Vector Machine was used to classify the coordinates. It performs classification by finding the hyperplane that maximizes the margin between the two classes. In our cases, our data was linearly inseparable, thus kernel functions were introduced. Polynomial kernel of degree two and Radial basis kernel are used. In addition, SVM is inherently a binary classifier, however for our project we have a multi class problem. Therefore, one against all classification is used. Cross validation was then employed to evaluate the supervised learning model. To achieve this, 80 % of the data was used as training data to fit SVM and 20% was used as testing data set to evaluate SVM.

RESULTS

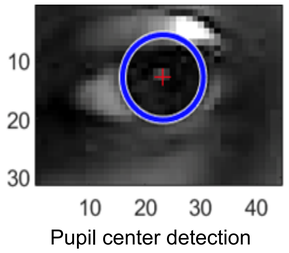
Face and eye detection



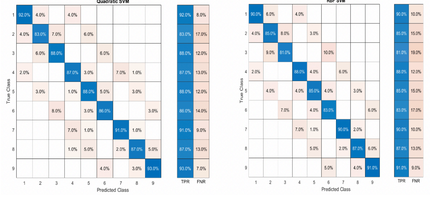
Original data for the 3x3 grid



Quadratic decision boundary for the 3x3 grid



Pupil center detection



Confusion Matrix for the 3x3 grid

Grid Type	Non-Linear Kernel Type	Accuracy (%)	Misclassification Cost	Training Time(sec)
2 by 1	Quadratic	96.7	10	0.3039
	RBF	95.7	13	0.3268
2 by 2	Quadratic	94.5	22	0.8249
	RBF	93.3	27	0.9971
2 by 3	Quadratic	89.5	63	1.2745
	RBF	89	66	1.4354
3 by 3	Quadratic	88	108	2.5168
	RBF	85.3	132	4.6094

Performance for all grid types

CONCLUSION & FUTURE DIRECTIONS

In this work, a gaze tracking technique has been introduced. Features were then introduced to a multi-class supervised learning method-Support Vector Machine (SVM) for classification. Classification was done on 4 different grid types, 2 by 1, 2 by 2, 2 by 3, and 3 by 3. Cross validation is then used to train our model using training data set and then evaluation is done using testing data set. Two nonlinear kernels, Quadratic and Radial basis are used for classification. Evaluation of the performance of the classifiers using confusion matrix shows that for such given data, quadratic kernel performs better. It was found that quadratic kernel outperformed RBF kernel in all four cases for our gaze tracking system.

A significant improvement, for future work, is the training and testing of the method using a live streaming instead of a dataset. Also, and related to the above, a time consumption evaluation must be performed to accurately guarantee on-line performance and determine if one of the previous preprocessing stages needs to be optimized or replaced in terms of the time consumption of the stage.

REFERENCES

[1] U.S. (2019). *Eye-tracking devices may help ICU patients communicate*. [online] Available at: https://www.reuters.com/article/us-health-icu-eye-tracker/eye-tracking-devices-may-help-icu-patients-communicate-idUSKCN0S324Q20151009 [Accessed 16 Dec. 2019].

[2] Viola, P. and Jones, M. (2019). *Robust Real Time Object Detection*. [online] Cs.cmu.edu. Available at: https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-IJCV-01.pdf [Accessed 16 Dec. 2019].

[3] Rafael C. Gonzalez and Richard E. Woods. Digital Image Processing. Prentice Hall, 2007. ISBN 0-13-168728-x.

[4] Daniel C. H. Schleicher and Bernhard G. Zager "Image Processing and Concentric Ellipse Fitting to Estimate the Ellipticity of Steel Coils" in Intech open Journal October 11, 2009

[5] Medium. (2019). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. [online] Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a [Accessed 16 Dec. 2019].

[6] Pdfs.semanticscholar.org. (2019). *An Approach to Face Detection and Feature Extraction using Canny Method*. [online] Available at: https://pdfs.semanticscholar.org/5393/f0adb198cf14f206eb0505cf365ef52171fe.pdf [Accessed 16 Dec. 2019]

ACKNOWLEDGEMENT

We would like to thank Prof. Huang, Prof. Ning, Prof. Mertens, Prof. Blaise, and Andrew Muslin for their assistance.