CS 221: Final Project Report

Detecting Horizontal Gaze Nystagmus for Impairment Detection

Sean Konz (swkonz), Contributor: Ben Backus (mbackus)

December 9, 2018

Introduction:

Impairment detection is a task often carried out by law enforcement agencies in order to assess an individual's intoxication level due to use of drugs or alcohol. This process is often carried out through the observation of horizontal gaze nystagmus (HGN). HGN is an inherent biological process in which an individual's pupils may shake or generate saccades as a result of the consumption of certain substances such as alcohol, depressants, or certain stimulants. This process is currently carried out by law enforcement officers during traffic stops, however, due to human error, the validity of this process is often doubted as a source of viable evidence since there are many variables that can contribute to the failure of this test. We propose that the application of machine learning in controlled testing environments will provide for a more accurate and consistent HGN detection methodology.

The application of machine learning has been attempted for this problem in the past [3]. The approach was limited to pupil detection and thresholding on specific features from pupil locations however, and had limited accuracy. Additionally, there have been countless instances of medical diagnosis using machine learning techniques and images of cells [4, 5]. Our work here differs from these past works in that we apply machine learning to videos of pupil movements, and thus attempt an application of activity recognition with a very limited number of salient features to utilize for detection.

The HGN detection process details a number of particular testing processes to score the degree of a person's nystagmus. First, the smoothness of object tracking is measured by instructing the individual to use only their eyes to follow a visual prompt passing through their field of view. The second is the onset angle of nystagmus while tracking a moving object. This is the approximate angle from center of the individuals head (0 degrees) that nystagmus begins to

be observed. The third is the degree of saccades observed in the individual's pupils when the pupils are in their maximum deviation position (55 degrees). The maximum deviation position of the pupils is when they are located at the farthest position from center. These three observable states contribute to the assessment of a person's impairment based on HGN. We define features of HGN, construct feature vectors for each video test instance, and classify the videos using various feature vector based models and a Recurrent Neural Network.

Dataset:

Our dataset consisted of 190 HGN testing videos in total. The breakdown of these videos was 136 intoxicated, and 54 sober. For model testing and training, we used a 80/20 split to define our training and test sets. Since our dataset was unbalanced, we confirmed we weren't overfitting the impaired data by reviewing model performance on the test set by observing where the model misclassified data, and confirming that the model misclassified both sober and impaired datapoints. In the case where the model was overfitting impaired data, we randomly selected videos from the dataset to be removed for the training of a particular model.

Pupil Representation:

As described above, HGN can be detected by observing pupil saccades at certain stages of an HGN test. The common feature across all of these test instances is that the individual's pupils "shake" in the x dimension predominantly. With this in mind, we construct feature vectors focused on detecting when the individual's pupils are moving an excessive amount in the X direction. In order to extract these features, we first detect the individual's pupils in each frame of a video recording of the individual conducting the HGN test.

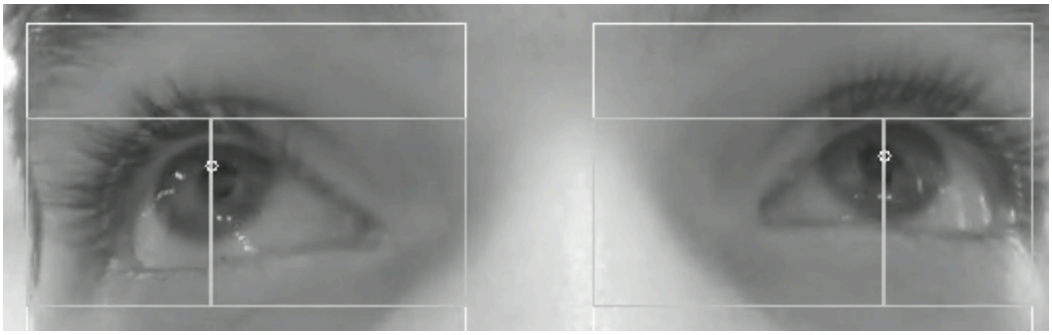


Figure 1: Example of pupil detection algorithms applied to HGN test videos.

Pupils are detected using a simple gradient algorithm proposed originally by Timm and Barth in 2011 [1]. The algorithm uses image gradients to define an objective function, the maximum of which corresponds to the location where most gradient vectors intersect, which is the eye's center. We implement this algorithm by first detecting eye regions on the individual's face, cropping the image to these regions, and running the pupil detection on these cropped images.

$$c = \max_c \frac{1}{N} \sum_{i=1}^{N} (d_i^T g_i)^2$$

Equation 1: Objective function defining approximate pupil center. The variables d_i is the displacement vector of a pixel location and g_i is the gradient vector of a pixel location

The performance of this algorithm was computed using the BioID Dataset and as advertised in the initial presentation of this algorithm [1]. We use detected pupil locations in (x, y) form in order to generate representative feature vectors for each video in our dataset.

Feature Extraction:

Using our extracted pupil locations, we attempted to construct representative feature vectors for each dataset video. Our initial features were defined as follows:

1. Number of changes in X velocity – This feature counts the number of instances that the user's pupil changes direction throughout the course of the video. When an individual is intoxicated their eyes change direction more often due to the inherent HGN.
2. Average distance from pupil center a change in velocity occurs – We hypothesize that there is a correlation between the distance from center that a saccade occurs, and whether or not an individual is intoxicated. We make this hypothesis based on the understanding that saccades occur more often as the pupil moves farther from the eye center.
3. Average X velocity of eyes – The speed of motion for the HGN test is constant throughout the test. Since nystagmus produces eye saccades which are rapid movements, we predict the average velocity of eye movements will be higher in intoxicated individuals.

4. Average X acceleration – Similar to feature number two, when the pupils move rapidly, the acceleration will spike. We predict average acceleration to be higher in intoxicated individuals.
5. Average distance from center of nystagmus – This is intended to capture the approximate pixel distance from the center of the eye that nystagmus occurs.

We predicted that these features would provide additional information to threshold off of in order to provide additional context from the video for classification.

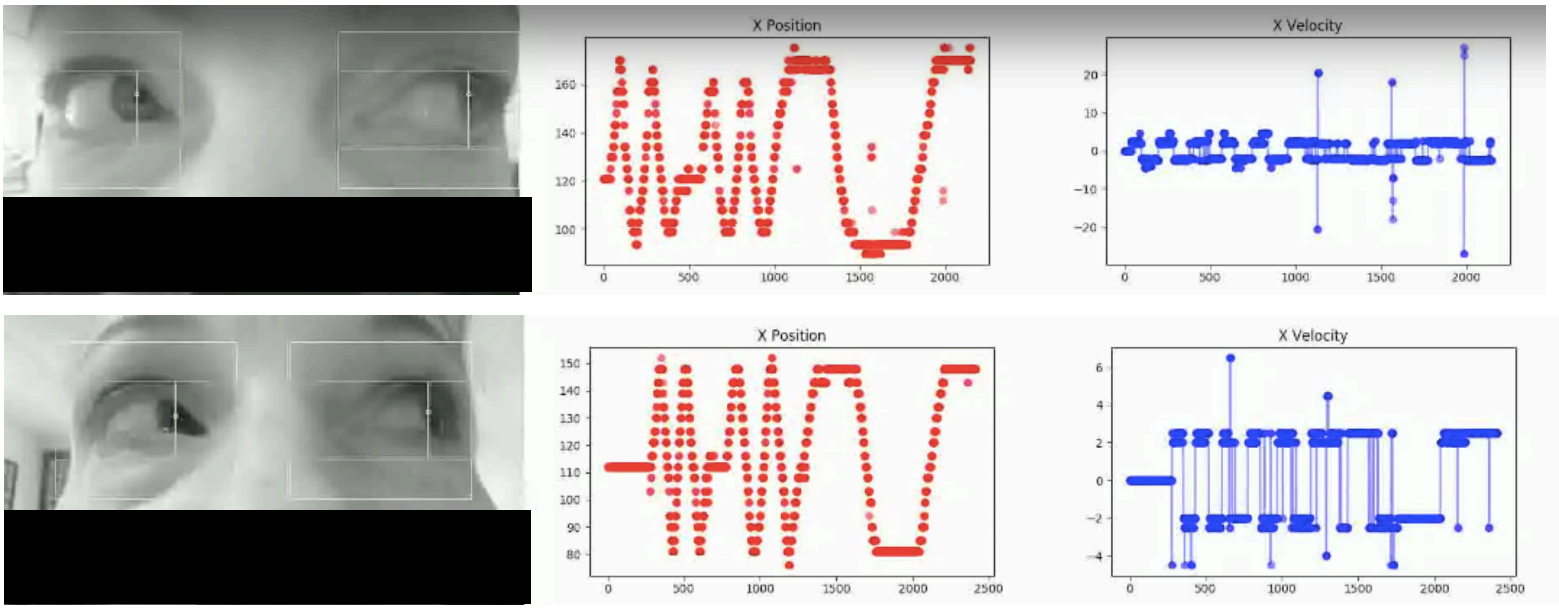


Figure 2: Plots of X position and X velocity of sober and intoxicated individuals. Note the plot y axis range and the difference in clean X position points. Top is Intoxicated. Bottom is sober.

From these plots, we can see that there is a significant difference in range of detected X velocities, and the X position values are very noisy in the intoxicated instance. These results support our feature choices. From the acceleration plots in figure 3, we can see that in the intoxicated plot, there is far more distribution of the acceleration values in the HGN test video. Although the accelerations for both videos is centered around the same value, the fact that the intoxicated video has a larger spread supports the effectiveness of this feature.

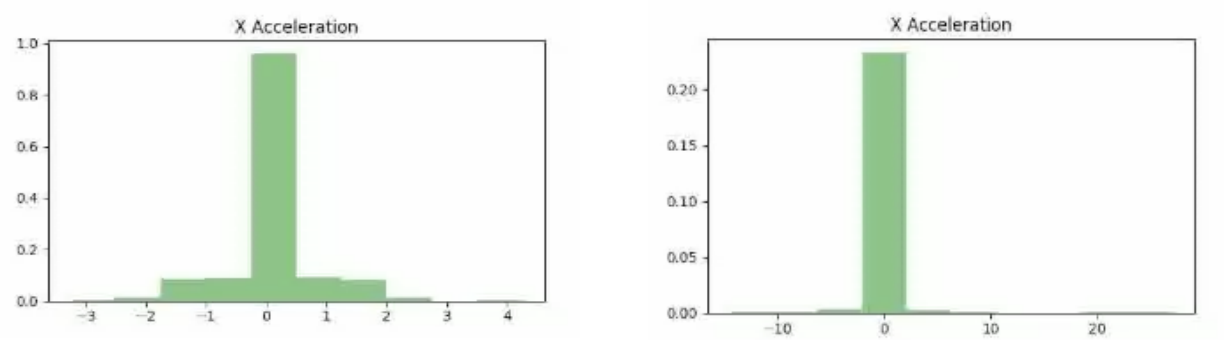
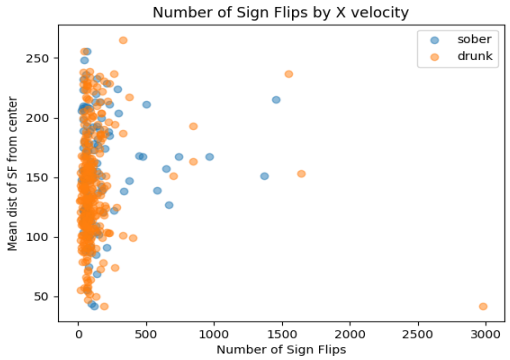


Figure 3: Acceleration plots of both sober and intoxicated individuals. Left: Sober. Right: Intoxicated

After computing our feature vectors for each video, we observed that certain features were not seperable when comparing sober and impaired data. From the plot in figure 4, we can see that the distribution of sober and impaired data has no distinct boundary when observing sign flips and sign flip distance from center. We computed statistics on the raw velocity values and found that we could capture the observed difference in velocity values in a more meaningful way by including velocity variance and skewness in our feature vectors. The table in figure 4 shows how these values might differ when compared on individual test videos.



Velocity		
	Sober	Impaired
Mean	0.0266	0.0022
Variance	4.6802	5.9038
Skewness	0.4287	1.7150

Figure 4: plot and table of extracted velocity features.

After running some models (see next section), we experimented with using alternate features which more closely represented the observed data. These features included velocity and acceleration variance, skew, and mean values for each dataset video. We found that the performance of these statistical features out performed our initial feature vectors, and instead chose to use these feature vectors in for the remainder of our tests. The final feature vector description was as follows:

1. Number of sign flips
2. Mean velocity

3. Variance in velocity values
4. Skew in velocity
5. Mean acceleration
6. Variance in acceleration values
7. Skew in acceleration

Since we found that statistical features proved more effective in representing our input videos, we hypothesized that a model operating directly on input videos or detected pupil points might prove more effective in distinguishing between sober and impaired videos.

Models:

The models we chose to run on our dataset were motivated by two factors: relevance to the course material, and predicted effectiveness. A summary of our model performances is presented by the table in figure 5.

Model	Accuracy
K-Means	50%
K-Nearest Neighbors	71%
Logistic Regression	73%
LSTM	76%

Figure 5: Accuracy of each model for performing on our dataset

From the table, we can see that our models which operated on representative feature vectors were outperformed by the Long-Short Term Memory (LSTM) Recurrent Neural Network (RNN), which operated directly on detected pupil points. Our K-Means model was unable to prove any classification value since the variables in our feature vectors were not linearly sperable. K-Nearest Neighbors (KNN) and Logistic regression were comparable in performance, but logistic regression saw a slight improvement in performance over KNN when we switched to using the improved feature vectors. We believe the reason for this divergence in performance is due to the fact that KNN becomes less effective as feature vector dimensions increase. Our RNN model used a simple architecture consisting of 124 LSTM nodes followed by a dense softmax layer for output prediction. Since the model was operating directly on pupil x

locations, feature selection was unnecessary, which we hypothesize contributed to the improved performance over the feature based models. The ability for the LSTM model to better extract features based on time series data most likely contributed to its heightened performance as well.

Error Analysis:

Overall, our final feature based models could not compete with the performance of models operating directly on raw extracted data from our videos. We believe this is due to the noise captured by our pupil detection algorithms. Although our pupil detection had a high degree of accuracy on the BioID dataset, we observed a significant amount of noise in our pupil detections when visualizing the videos. Most of the noise was observed when the user blinked their eyes, which caused the detected pupil location to jump a significant amount (> 200 pixels). This causes spikes in the observed velocity and acceleration, which are inherently represented in feature vectors, but might not be weighed heavily in the LSTM model. Additionally, when reviewing the videos, not all test processes were observed at the same time step in each video. What this means is that each test process might have a time differential when observing raw pupil locations. This would not be a problem for the feature vector models, but might be a problem for models that utilize time series information such as our LSTM model. This raises the question of how refined we need our dataset to be in order to be confident in our classifications. Overall, noisy pupil detections, and variable video quality, and consistency of test processes most likely contributed to our error levels.

Conclusion:

Overall, we did not beat our oracle accuracy of 90% achieved by humans in this task, but we learned a significant amount about best practices for understanding this classification task, and what components of the classification pipeline need refinement in order to achieve a greater accuracy. With a more refined and curated dataset, we are confident that we'd be able to achieve a higher accuracy. Our next steps in this project would be to focus on developing models that operated directly on the videos. This would include developing a time distributed

convolutional neural network models or a dual stream model which might utilize sub-regions of the input video in order to limit the amount of noise generated from the background regions in an input video. Additionally, a larger and a more refined dataset and dataset acquisition process will have a significant impact on the upper limit on the accuracy of our models. Generally, we feel that models operating directly on videos and raw pupil locations will most likely have the greatest success in this space looking forward.

Work Distribution:

Work on the project was distributed as follows:
Sean: Project proposal, background research, baseline model, project progress report, pupil location extraction, defining initial feature vectors, feature visualizations, analysis on initial features, refined feature vector definition, all model development, error analysis on models, experimenting with models, final poster, poster presentation, final report.
Ben: Some code for creating feature vectors.
Due to personal circumstances, Ben was unable to contribute to the project.

References:

1. BioID Face Database | Dataset for Face Detection | facedb. (n.d.). Retrieved from https://www.bioid.com/facedb/
2. Timm and Barth. Accurate eye centre localisation by means of gradients. In Proceedings of the Int. Conference on Computer Theory and Applications (VISAPP), volume 1, pages 125-130, Algarve, Portugal, 2011. INSTICC.
3. Wilson and Avakov. Image Processing Methods For Mobile Horizontal Gaze Nystagmus Sobriety Check. Stanford University Electrical Engineering. 2011.

4. Pawel Liskowski and Krzysztof Krawiec. Segmenting retinal blood vessels with pub newline deep neural networks. IEEE transactions on medical imaging, 35(11): 2369-2380, 2016.
5. Rahul Paul, Samuel H Hawkins, Lawrence O Hall, Dmitry B Goldgof, and Rober J Gillies. Combining deep neural network and traditional features to improve survival prediction accuracy for lung cancer patients from diagnostic ct. In Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on,pages 002570-002575, IEEE, 2016.