# Intoxication detection in a video

Vivek Verma
2016csb1064
Vineet
2016csb1063

CSL 462 Computer Vision
Department of Computer Science and Engineering
Indian Institute of Technology Ropar

## 1 Abstract

Drunk driving is one of the major cause of accident on road. Thus, drunk driving detection system is necessary for early detection and prevention of accidents. Most existing systems require special equipment such as Infrared cameras or Breathalyzers. So, based on the limitations of these methods, we have proposed an idea of detecting drunk person by analyzing the video.

## 2 Introduction

Alcohol impaired driving poses a serious threat to the driver as well as pedestrians. Development of automated drunk detection systems is necessary to reduce traffic accidents and the related financial costs. Intoxication detection systems can be divided into following categories:

1. **Direct detection** - Measuring Blood Alcohol Content (BAC) directly through breath analysis.
2. **Biosignal based detection** - Using Electrocardiogram signals [13] or face thermal images [8] to detect intoxication.
3. **Behaviour based detection** - Detecting characteristic changes in behaviour due to alcohol consumption. This may include changes in speech, gait, or facial expressions.

Direct detection is often done manually by law enforcement officers using Breathalyzers. Biosignal based detection also requires specialized equipment to measure signals. Behaviour based detection can be performed passively by recording speech or video of the subject and analyzing it to detect intoxication.From the blog - How to Recognize the Signs of Intoxication[9] by Harrison Lewis , we concluded these three physical and behaviour aspects in our work:-

1. **Eye Fatigue** - A person eyes can tell a lot about them and their mental state in a particular moment. Normal blinking of eye in not observed in drunk people.
2. **Emotion Transition** - Drunk person show extreme level of emotion changes. They are sometime sad, sometime happy. So, we will be focusing on analyzing emotion changes in intoxicated person.
3. **Video Engagement** - Intoxicated people can not perform normal tasks as easily as they can when they are sober. They are not able to focus in the video which can easily be tracked by eye movement.

We will be focusing on above aspects for intoxication detection, specifically using video/images of subject.

## 3 Prior Work

In the recent work by D.P.Yadav and A.Dhall [14] on the dataset and experiments on videos related to intoxicated people , deep learning techniques are used. No other existing work in literature addresses the problem of detecting intoxication using RGB videos. However, several other techniques have been proposed on the three behaviour and physical signs explained in the Introduction.

### 3.1 Eye Fatique Detection

The driver fatigue problem has become an important factor of causing traffic accidents. Fatigue has high correlation with drunk person. It is hard to recognize a driver whether he is dozing because he is tired or he is drunk. In the past 10 years many researchers have worked on driver fatigue problem [6],[3]. Inspired by their work we thought of using fatigue behaviour in detecting intoxication state of a person. It is easy to detect whether the eyes are open or closed in the video, we assumed that when eyes are close over 5 consecutive frames, then the driver is regarded as dozing which can be due to alcohol or actual tiredness.

### 3.2 Emotion Recognition and Changes Detection

To recognize emotion at different time in video and observe the emotion changes. At first, we need to recognize emotion for which a model of the facial muscle motion corresponding to different expressions has to be found. The best known such model is given in the study by Ekman and Friesen [4], known as the Facial Action Coding System (FACS). Ekman has since argued that emotions are linked directly to the facial expressions and that there are six basic ”universal facial expression” corresponding to happiness, surprise, sadness, fear, anger, and disgust. The FACS codes the facial expressions as a combination of facial movements known as action units (AUs). The AUs have some relation to facial muscular motion and were defined based on anatomical knowledge and by studying videotapes of how the face changes its appearance. Ekman defined 46 such action units to correspond to each independent motion of the face.Tao and Huang [12] used a simplifies model which uses an explicit 3D wireframe model of the face. The face model consists of 16 surface patches embedded in Bezier volumes.
These earlier methods are not automated and need user to mark the patches manually.Recent method for automated Emotion recognition using PHOG and LPQ features[2] by A. Dhall is based on deep learning techniques but the feature extraction by PHOG and LPQ is considered in this work to recognize emotion.

### 3.3 Video engagement

In the recent work by A. Dhall and D.P. Yadav [14] on the dataset and experiment on videos related to intoxicated people - features like eye gaze, eye pose,eye landmark , face landmark are considered. As most of the features for the video engagement are considered in the emotion recognition. We will be considering eye gaze and eye pose for video engagement.

## 4 Dataset

We used a dataset created by D.P.Yadav and A. Dhall [14]. We have taken a subset of it 1070 videos of drunk subjects and 484 videos of sober subjects.These videos are 15 second in length each having 360 frames of dimension 224X224.

## 5 System

Our approach began with extracting useful information from the video. We extracted the features from the video using constraint local models (CLM), we used OpenFace toolkit[5] to extract video features. Then, Kmean clustering algorithm is applied to extract useful frames more detailed explanation is given later. Further, PHOG and LPQ features are extracted for these frames. Facial Action Units were also used as features. These, features were trained using machine learning model. Further, detail of each step is given below:

1. **Frames selection based on K-Mean clustering**
   The video has lot of redundant data. So, we decided to extract useful information and only those frames which are relevant. The video had the length of 15 seconds having 280 frames. Before, using landmarks as a feature vector for K-Mean, we selected frames if any one of Facial Action Unit like 1, 2, 4, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45 was detected. This lowered the

number of frames. Now, we created the feature vector X of facial landmarks obtained in CLM.

$$P = [eye\_lmk\_x\_0...eye\_lmk\_x\_n; eye\_lmk\_y\_0...eye\_lmk\_y\_n]$$

The length of **P** vector was 112. This vector was standard normalized with zero mean and unit variance. Likewise, for each frame, normalized feature vector was created. These normalized vectors were feed to K-Mean clustering algorithm with 25 cluster centers. Now, taking one cluster center we took its euclidean distance from every normalized frame vector. Frame vector with minimum euclidean distance from cluster center is chosen as key frame. Similarly, all 25 key frames are chosen. Now, all features for a video are calculated on these 25 frames.

2. **Eye gaze and Facial Action Unit as features** As discussed eye gaze can prove to be a important feature for intoxication detection. We will be using Open Face library [5] to get eye gaze.Features like Eye gaze direction vector in world coordinates, eye gaze direction in radians in world coordinates averaged for both eyes and facial action units like AU's 1, 2, 4, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, 28, and 45 are extracted for every frame. Larger vector is formed by stacking these features of all 25 frames. And thus vector of length 475 is generated. This vector act a feature vector for the given video.

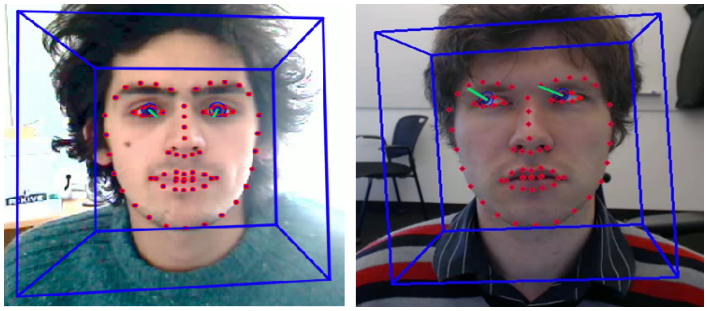

Figure 1: Eye Gaze

3. **Fatigue Feature** After collecting Video frames from K-mean clustering, we cropped the eyes from the frames based on the facial landmarks. These facial landmarks are obtained using Openface [5]. Openface toolkit give coordinates of the above landmark. Thus, box is obtained around the eyes. Then we do further analysis.

   The box filter is applied over the cropped image followed by histogram equalization. Then gaussian filter was applied to removed further noise. Then, the RGB image is convert to YUV colorspace.

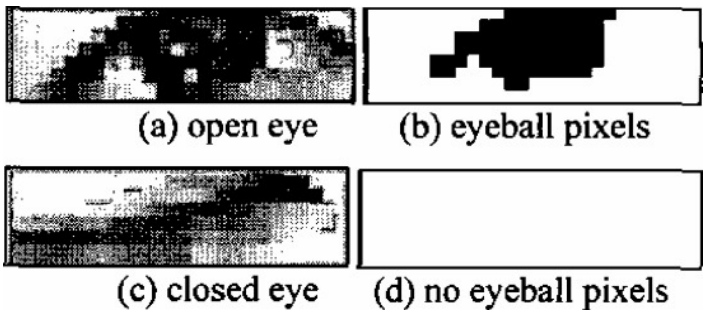


Figure 2: Eyeball Detection

4. **Shape feature extraction using PHOG**
   For extracting shape information we use PHOG [1] features. PHOG is a spatial pyramid extension of the histogram of gradients (HOG) descriptors. The HOG descriptor technique counts occurrences of gradient orientation in localized portions of an image and has been used extensively in computer vision methods. PHOG features being an extension of HOG have shown good performance , PHOG descriptors have been used for static facial expression analysis. At the start the canny edge detector is applied to the cropped face. Then the face is divided into spatial grids at all pyramid levels. After this a $3 \times 3$ Sobel mask is applied to the edge contours for calculating the orientation gradients. Then the gradients of each grid are joined together at each pyramid level. There is an option for two orientation ranges, [0-180] and [0-360].In our experiment, we will use number of pyramids L=3 the bin size N=8 and the orientation range is [0-360].

5. **Appearance feature extraction using LPQ** This feature extraction step is inspired from A.Dhall research work on emotion recognition [2] Local binary patterns (LBP) family of descriptors (LBP [10], LBP-TOP [15], LPQ [11] and LPQ-TOP [7]) have been extensively used for texture analysis, static and temporal facial expression analysis and face recognition. We use LPQ (Local Phase Quantization) appearance descriptor. Though LPQ-TOP [24] has been proposed for temporal data analysis, but as we do not have labeling of an onset, apex and offset in the database in our experiments, we use LPQ only. LPQ is based on computing short-term Fourier transform (STFT) on local image window. At each pixel the local Fourier coefficients are computed for four frequency points. Then the signs of the real and the imaginary part of the each coefficient is quantized using a binary scalar quantiser, for calculating the phase information. The resultant eight bit binary coefficients are then represented as integers using binary coding. This is step is similar to the histogram construction step in LBP. In the end we get a 256 dimensional feature vector. In our experiments we divided the cropped face of size 60x60 into four blocks. This gave us a vector dimension of 1024 for an image and 25600 for an image sequence where the number of cluster centers m = 25.

## 6 Experiments

We trained eight different support vector machine model.
In SVM, 'rbf' kernel is used and grid search is used for parameter tuning. We have taken 75% training instance and 25% test instance. Model 1 used vector having eye gaze , facial action units and eye fatigue value all three stacked together. We have named these features OPNF features(OpenFace Features). The length of this vector was 500. (475 eye gaze and AU's and 25 eye fatigue values). Model 2 used PHOG features as a feature vector(normalized) of length 17000. Model 3 used LPQ feature as a feature vector(normalized) of length 25000. Model 4 is LPQ + PHOG. Model 5 is OPNF + LPQ. Model 6 OPNF + PHOG. Model 7 is PHOG + LPQ + OPNF. Model 8 include model 1 , model 2 and model 3 . These three model are combined together and accuracy is calculated on the basis of majority prediction. The accuracy and the parameters finalized by grid search for individual model is shown below in table.

| Model1-OPNF C:10 gamma :0.01 Kernal:rbf | | |
|---|---|---|
| | Precision | Recall |
| Drunk | 0.69 | 0.9 |
| Sober | 0.38 | 0.13 |
| Overall Accuracy | 65.46 | |

| Model2-PHOG C:1 gamma :0.001 Kernal:rbf | | |
|---|---|---|
| | Precision | Recall |
| Drunk | 0.72 | 0.98 |
| Sober | 0.14 | 0.1 |
| Overall Accuracy | 70 | |

| Model3-LPQ C:1 gamma :0.001 Kernal:rbf | | |
|---|---|---|
| | Precision | Recall |
| Drunk | 0.7 | 0.99 |
| Sober | 0.33 | 0.1 |
| Overall Accuracy | 70.3 | |

| Model4-LPQ+PHOG C:36 gamma :0.01 Kernal:rbf | | |
|---|---|---|
| | Precision | Recall |
| Drunk | 0.68 | 0.98 |
| Sober | 0.55 | 0.05 |
| Overall Accuracy | 67 | |

| Model5-OPNF + LPQ C:1 gamma :1 Kernal:rbf | | |
|---|---|---|
| | Precision | Recall |
| Drunk | 0.68 | 0.99 |
| Sober | 0.50 | 0.02 |
| Overall Accuracy | 68.29 | |

| Model6-OPNF+PHOG C:10 gamma :0.01 Kernal:rbf | | |
|---|---|---|
| | Precision | Recall |
| Drunk | 0.7 | 0.99 |
| Sober | 0.33 | 0.02 |
| Overall Accuracy | 69.07 | |

| Model7-PHOG+LPQ+OPNF C:1 gamma :0.001 Kernal:rbf | | |
|---|---|---|
| | Precision | Recall |
| Drunk | 0.72 | 0.97 |
| Sober | 0.25 | 0.03 |
| Overall Accuracy | 70.10 | |

| Model8 weighted(model1,model2,model3) | | |
|---|---|---|
| Overall Accuracy | 70.62 | |

## 7 Conclusion and Future work

We presented a method for intoxication detection.We used a discriminative classifier named SVM(Support Vector Machine) for classification.For capturing emotion features, we used PHOG , LPQ and for fatigue we use features related to eye such as gaze and action units. We trained the model on 1070 videos and used 484 videos. We achieved good accuracy on the small dataset.
For future work we would like to explore deep learning techniques for the intoxication detection.

## 8 References


[1] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.

[2] Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883. IEEE, 2011.

[3] Wenhui Dong and Xiaojuan Wu. Fatigue detection based on the distance of eyelid. In *VLSI Design and Video Technology, 2005. Proceedings of 2005 IEEE International Workshop on*, pages 365–368. IEEE, 2005.

[4] P Ekamn and W Friesen. Facial action coding system (facs): manual, 1978.

[5] Open Face. URL `https://github.com/pyannote/pyannote-video`.

[6] Wen-Bing Horng, Chih-Yuan Chen, Yi Chang, and Chun-Hai Fan. Driver fatigue detection based on eye tracking and dynamk, template matching. In *Networking, Sensing and Control, 2004 IEEE International Conference on*, volume 1, pages 7–12. IEEE, 2004.

[7] Bihan Jiang, Michel F Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE, 2011.

[8] Georgia Koukiou and Vassilis Anastassopoulos. Drunk person identification using thermal infrared images. *International journal of electronic security and digital forensics*, 4(4):229–243, 2012.

[9] Harrison Lewis. How to recognize the signs of intoxication, August 2017. URL `https://www.wikihow.com/Recognize-the-Signs-of-Intoxication`.

[10] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[11] Ville Ojansivu and Janne Heikkilä. Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer, 2008.

[12] Hai Tao and Thomas S Huang. Connected vibrations: a modal analysis approach for non-rigid motion tracking. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 735–740. IEEE, 1998.

[13] Chung Kit Wu, Kim Fung Tsang, Hao Ran Chi, and Faan Hei Hung. A precise drunk driving detection using weighted kernel based on electrocardiogram. *Sensors*, 16(5):659, 2016.

[14] Devendra Pratap Yadav and Abhinav Dhall. Dif: Dataset of intoxicated faces for drunk person identification. *arXiv preprint arXiv:1805.10030*, 2018.

[15] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29 (6):915–928, 2007.