Literature review- Intoxication detection using face video

Vineet
2016csb1063
Vivek Verma
2016csb1064

CSL 462 Computer Vision
Department of Computer Science 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 [16] or face thermal images [10] 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[11] 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 A. Dhall and D.P. Yadav [17] 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 [8],[4]. 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 [5], 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 [15] 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[3] 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 [17] 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 will using a dataset created by D.P.Yadav and A. Dhall [17]. We have extended this dataset including more videos of drunk and sober people. Our work consists of collection, processing and analysis of a new dataset of drunk and sober face videos. We extended DIF (Dataset of Intoxicated Faces) for drunk person identification. By collecting this dataset, we aim to analyze the difference in facial features of a drunk and sober person. Hence, we try to search for online videos of people exhibiting facial movements and expressions in drunk and sober conditions. We use search queries such as 'drunk reactions', 'drunk review', 'drunk challenge' etc. on YouTube (www.youtube.com), Periscope (www.pscp.tv) and Twitch (www.twitch.tv) to obtain videos of drunk people. Similarly, for the sober category, we collect several reaction videos from YouTube. Since the videos were recorded in unconstrained real world conditions, our dataset represents drunk and sober people 'in the wild'. We use the video title and caption given on the website to assign class labels. In some cases, the subject labeled as 'drunk' might only be slightly intoxicated and not exhibit significant changes in behavior. In these cases, the drunk class labels are considered as weak labels. There is no such ambiguity in the sober class labels. We collect 48 videos in the sober category. The drunk category contains 41 videos . In sober category 20 videos are of different female and rest 28 videos are of different males. In drunk category 13 videos of different females and 28 videos of different males.We process these videos using the pyannote-video library [6]. First, we perform shot detection on the video and process each shot separately in subsequent

stages. Then, we perform face detection and tracking to extract bounding boxes for faces present in each frame of the video. We also get a unique id for each tracked face. Using these ids and bounding boxes, we crop the tracked face from the video. Hence, we obtain a set of face videos where each video contains a drunk or sober person exhibiting facial motion. These face videos are of 224x224 size, maintaining the aspect ratio of the original videos. We also perform face alignment on each frame of the video using the OpenFace toolkit [1]. Each aligned face videos is split into non-overlapping sequences of 75 frames, which corresponds to 5 seconds of video at 15 frames per second. While extracting these sequences, we reject those in which a face was not detected accurately. We also apply a threshold based on the variance of facial landmark points to remove sequences having low facial movement. Our final dataset consists of 1294(old) + 3230(new) sequences for the drunk category and 1443(old) + 1220(new) sequences for the sober category.

5 System

Our system starts with face tracking using constraint local models (CLM) [14]. The shape vectors of the face in an image sequence are then normalised. The K-means clustering algorithm is applied to the normalised shape vectors and images having face shape vectors closest to the cluster centres are chosen for further processing. Further, the Viola-Jones [28] face detector is applied to the chosen images to compute the PHOG and LPQ features on the cropped faces. The CNN models in the OpenFace toolkit[6], will be used to extract eye gaze and face pose on the choosen images for a video. These features will be extracted for all the videos in the training set and will be trained by different machine learning techniques.

1. **Face tracking using CLM and Clustering based sequence quantisation**
   As the amount of motion in two consecutive frames is very sparse, we wish to remove the redundant frames so that the features are extracted on the frames which efficiently describe the temporal dynamics of the expression. We considered the technique used in Emotion recognition using PHOG and LPQ features by A. Dhall [3] in which face tracking is done by CLM[14] in the image sequences. It is based on fitting a parameterised shape model to the location landmark points of the face. It predicts locations of the model's landmarks and a row vector P containing location of each of the n landmark points is taken.

   $$P = [x_1; y_1 x_n; y_n]$$

   P is then normalised by taking the horizontal Euclidean distance between the outer eye corners on the left and right side. The vertical normalisation distance is the Euclidean distance between the tip of the nose and the midpoint between the eyebrows. We denote the normalised shape vector as $\mathbf{P^n}$.$\mathbf{P^n}$ are calculated for all the images in an sequence.For selecting key frames, K-means clustering algorithm is used on the normalised shape vectors. All the features will be extracted from these selected frames.

2. **Shape feature extraction using PHOG**
   For extracting shape information we use PHOG [2] 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 3×3 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].

3. **Fatigue Feature** After collecting Video frames from K-mean clustering, we cropped the eyes from the frames based on the landmarks like, left eye, right eye, nose tip and forehead. Eye cropping will be done using Open Face library [6]. It give coordinates of the above landmark. Thus, box is obtained around the eyes. Then we do further analysis. When the eyes are open, there are some eyeball pixels, as shown in Figures 1(a) and 1(b). When the eyes are closed, there are no eyeball pixels, as shown in Figures 1(c) and 1(d). By checking the eyeball pixels, it is easy to detect whether the eyes are open or closed in this system, these eye ball pixel count will be used as a feature. The similar approach was taken by W.B. Horng[8]

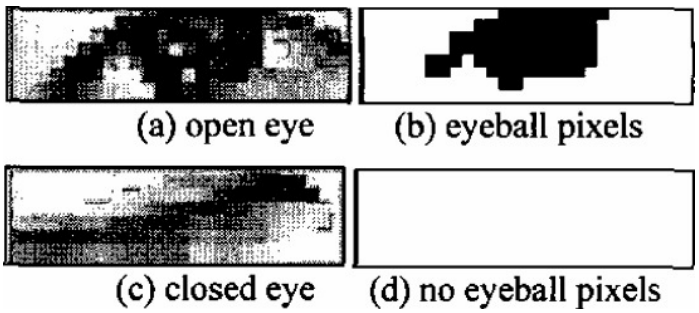

Figure 1: Eyeball Detection

4. **Eye gaze** As discussed eye gaze can prove to be a important feature for intoxication detection. We will be using Open Face library [6] to get eye gaze. Eye gaze direction vector in world coordinates are extracted from the video for every frame.This vector will serve as a feature vector for classification. Below figure[7] shows the eye gaze(green line) of people in world coordinates.

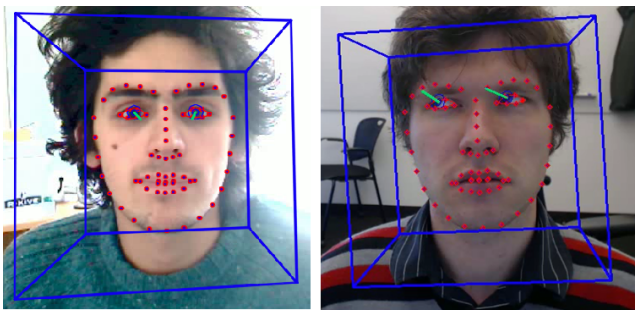

Figure 2: Eye Gaze

5. **Appearance feature extraction using LPQ** This feature taken from A.Dhall research work on emotion recognition [3] Local binary patterns (LBP) family of descriptors (LBP [12], LBP-TOP [18], LPQ [13] and LPQ-TOP [9]) have been extensively used for texture analysis, static and temporal facial expression analisys and face recognition. We use LPQ (Local Phase Quantization) appearance descriptor. Though LPQ-TOP [24] has been proposed for temporal data analayis, 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 6144 for an image sequence where the number of cluster centers m = 6.

6. **Classification** We will we using different machine learning technique to train the model. As a baseline techniques will be using Support Vector Machine(SVM) or Decision Tree for classification.

6 References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE, 2016.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] Open Face, . URL `https://github.com/TadasBaltrusaitis/OpenFace/wiki`.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Face alignment through subspace constrained mean-shifts. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1034–1041. Ieee, 2009.

[15] 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.

[16] 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.

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

[18] 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.