Driver Drowsiness Detection via a Hierarchical Temporal Deep Belief Network

Ching-Hua Weng, Ying-Hsiu Lai, and Shang-Hong Lai(✉)

Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
lai@cs.nthu.edu.tw

Abstract. Drowsy driver alert systems have been developed to minimize and prevent car accidents. Existing vision-based systems are usually restricted to using visual cues, depend on tedious parameter tuning, or cannot work under general conditions. One additional crucial issue is the lack of public datasets that can be used to evaluate the performance of different methods. In this paper, we introduce a novel hierarchical temporal Deep Belief Network (HTDBN) method for drowsy detection. Our scheme first extracts high-level facial and head feature representations and then use them to recognize drowsiness-related symptoms. Two continuous-hidden Markov models are constructed on top of the DBNs. These are used to model and capture the interactive relations between eyes, mouth and head motions. We also collect a large comprehensive dataset containing various ethnicities, genders, lighting conditions and driving scenarios in pursuit of wide variations of driver videos. Experimental results demonstrate the feasibility of the proposed HTDBN framework in detecting drowsiness based on different visual cues.

1 Introduction

Recent reports have suggested that drowsy driving is one of the main factors in fatal motor vehicle crashes each year [1–3]. In 2014, the National Sleep Foundation (NSF) pledged an initiative that seeks to raise public awareness on drowsy driving and asked legislators to have law enforcement, regulations and recommendations on drowsy driving and distraction prevention [4]. Therefore, developing active monitoring systems that help drivers avoid accidents in a timely manner is of utmost importance [5,6].

In recent drowsy driver detection systems, most of the work focus on using limited visual cues (often just one) [7]. However, human drowsiness is a complicated mechanism. If various cues are combined dynamically [8], results can be improved. Furthermore, drowsiness has an accumulative property and the decision cannot usually be made in a short period of time, *i.e.* drowsy status at a previous time point is a factor for the drowsy status at the current time point and the duration depends on the behavior of individuals. For example, frequent yawning is an important behavioral feature, but it does not always occur before the driver goes into a drowsy state. It should be used as a preemptive measure

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 117–133, 2017.
DOI: 10.1007/978-3-319-54526-4_9

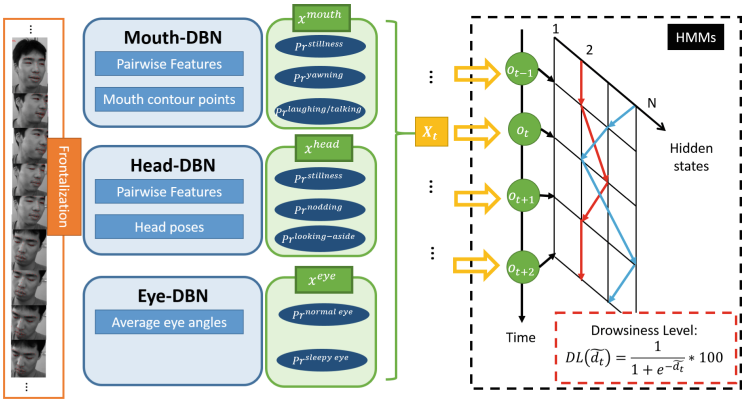


Fig. 1. A Hierarchical Temporal Deep Belief Network (HTDBN) for drowsy driver detection: inputs from the face or head after frontalization and pairwise feature extraction are first fed to deep belief networks to extract high level features and output each motion probabilities to form an observable vector X_t at each time stamp t. The deep neural nets are first pre-trained and then fine-tuned by the target drowsiness-related class. Then X_t is regarded as observation vector o_t for HMM. Two HMMs are learned on top of the deep neural nets and employed to analyze the likelihoods of the observable vector sequence within a fixed length duration. Finally, the driver's drowsiness level is predicted by inverse logit transform.

and memorized until other symptoms are captured. Otherwise, the probability of drowsiness will drop as nothing is detected for some period of time.

To resolve these issues, we introduce a novel and unified Hierarchical Temporal Deep Belief Network (HTDBN) for detecting drowsiness. The overall architecture is shown in Fig. 1. Through the proposed HTDBN framework, a set of high-level facial landmark features can first be extracted and used to learn representations for classification of several drowsiness-related symptoms. We demonstrate that drowsiness-related symptoms can be well classified using facial landmark points and head posture and the results are further composed to form observation vector sequences. For modeling temporal information, we resort to a continuous-Hidden Markov Model (HMM) which can be extended to long-term temporal information. We train two sets of parameters, drowsiness-HMM and non-drowsiness-HMM, covering many possible kinds of driving scenarios using the Baum-Welch algorithm. Finally, by using the forward-backward algorithm, the maximum likelihoods can be calculated and their differences are accumulated over a predetermined period of time.

Despite the importance of research in a practical drowsy driver detection system, most research have used relatively limited datasets. The generalization of different approaches to drowsy driver detection analysis remains unknown. In the absence of performance evaluation on a common public dataset, the comparative strength and weakness of different approaches is difficult to determine. In the field of facial expression and action recognition, comparative performance

evaluations have proven valuable [9,10], and similar benefits should be gained in the field of driver drowsiness detection. Therefore, we also describe in this paper a dataset that we specifically designed and collected for drowsy driver detection. This dataset will be soon made publicly available to researchers in the field. The dataset contains a wide variety of human subjects with various races, ethnicities and genders. The data was also collected at different situations like wearing glasses, sunglasses and various lighting conditions. We expect this dataset to be a representative test-bed for drowsy driver detection approaches.

In summary, we make the following contributions in this paper: (1) To the best of our knowledge, we are the first to combine DBN with HMM for drowsy driver detection. (2) The proposed framework captures the temporal information as well as the interactive relation among eyes, mouth and head. (3) We provide a dataset that contains drowsiness and non-drowsiness videos captured under various kinds of driving circumstances.

2 Related Work

Drowsiness-related symptom measurement methods can be generally grouped into two categories [5]: Physiological and Physical. Physiological methods offer an objective and precise way to measure sleepiness. They are based upon the fact that physiological signals start to change in earlier stages of drowsiness [11]. Despite of their reliability, the intrusive nature of measuring physiological signals remains an issue that makes them unacceptable for real-world applications. Physical methods are based upon non-invasive observation of a driver's external state. A typical focus is on facial expressions that might express some characteristics, such as eyelid movement, head movement, gaze, and facial expression [12]. The research in this area can be classified into four groups [13]:

Threshold-Based Approach. The simplest method to predict a driver's drowsiness level is to set a threshold on extracted drowsiness-related symptoms. In the system presented in [14], the percentage of eyelid closure (PERCLOS) in a time window has shown to provide meaningful message of drowsiness. Teyeb *et al.* [7] further showed that when the head inclination angle exceeds a certain value and duration, the level of alertness of the driver is lowered. In [15], yawning is detected based on the rate of change of the mouth contour and is determined as the only sign of drowsiness. This approach may encounter false-alarms when the required visual cues cannot be distinguished from the similar motions, e.g. talking or laughing.

Knowledge-Based Approach. In the knowledge-based approaches, decision of driver's drowsiness is made based on knowledge of an expert. In this approach, knowledge usually appears to be evaluated according to *if-then* rules. Rezaei and Klette [16] implemented a fuzzy control fusion system to prevent road crush. In [17], a Finite State Machine (FSM) was used for hypo-vigilance detection. However, it is difficult to provide an accurate definition of driver drowsiness

with some rules, since rules defined in the system do not provide sufficient expressive power to accommodate the large variations and uncertainties in driver videos.

Probability Theory Based Approach. Ji *et al.* [8], proposed a Dynamic Bayesian Network (DBN) system to determine the level of driver's drowsiness. In [12], they included frequent yawning, nodding, gaze distribution and eyelid movement as observation nodes and many other subjective factors such as sleep quality, sleeping time and driving environment as contextual nodes in DBN. Although DBN has the ability to represent the spatio-temporal characteristics for determining drowsiness, it also has a large computational complexity. In addition, the statistical analysis of large-scale subjective training data is difficult to obtain.

Statistical Approach. Support Vector Machine (SVM) and Neural network (NN) are the main methods in statistical pattern recognition. Jin *et al.* [7] utilized SVM and Eskandarian and Sayed [18] applied NN on distinguishing driver's drowsiness both based on the combination of driver behavioral measures and driving performance measures. Neither SVM nor NN was able to model temporal information. Thus, it is unrealistic to port such methods into real-world applications.

In summary, most of the approaches have drawbacks due to impractical reasons mentioned above or do not provide sufficient discrimination to capture the uncertainties. Moreover, most of the existing methods do not evaluate the robustness of their system against subjects from different ethnicities, races, genders, various illumination conditions and partial occlusion (e.g. glasses, sun-glasses and facial hair). In contrast, our work aims to systematically integrate spatio-temporal information using pure visual cues representing the driver's behaviors. Our algorithm is evaluated on a driver video dataset we collected while covering a wide variety of scenerios. All the knowledge needed in the detection system is learned from the training data itself without subjective and sophisticated parameter tuning.

3 Proposed HTDBN Algorithm

3.1 System Overview

The proposed model is inspired by the framework successfully applied to the speech and gesture recognition [13,19]. But instead of limiting the learning ability to Restricted Boltzmann Machine (RBM), the proposed HTDBN based method utilizes the nice property of DBN [20], i.e. modeling high-order dependencies.

As shown in Fig. 1, the learned DBNs are used to extract drowsiness-related symptoms after frontalization along with pairwise feature extractions. On top of the DBNs, two continuous-HMMs are adopted for modeling higher level temporal relationship among drowsiness and non-drowsiness using the observation vectors

obtained from the probabilities of mouth-, eye- and head-motions. To evaluate driver's drowsiness level, the accumulated differences of both HMM maximum likelihoods collected from previous time stamps to current time stamp is passed to the inverse logit transform.

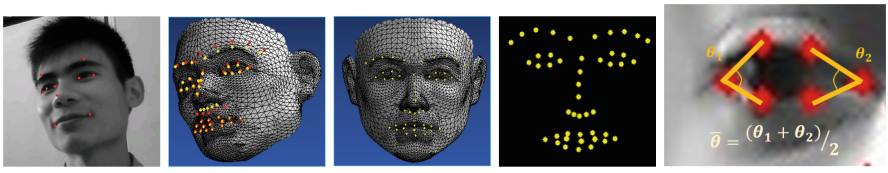
3.2 Frontalization

Specifically with faces, the success of the learned network in capturing facial appearance is highly dependent on a rapid 3D normalization step. Faces captured by the camera are considered unconstrained (due to the non-planarity of the face) and non-rigid expressions. Rezaei and Klette [16] focused on head-pose estimation, but ignored the importance of pose normalization on facial cues. Similar to the recent literature [21], our alignment is based on using facial landmark point detectors to direct the normalization process but is simplified by discarding pixel-wise transformation. In our framework, the recent algorithm of [22], based on the supervised descent concept, is to detect and track 49 facial landmarks in a video sequence.

We start our alignment process by extracting 9 2D facial landmark points as temporary anchor points. The points are corners of both eyes and mouth, center of the brows and tip of the nose as illustrated in Fig. 2a. They are used to approximate the pose matrix P given by

$$P = \left[\begin{array}{c|c} R_{3\times 3} & \mathbf{t}_{3\times 1} \\ \hline \mathbf{0}_{1\times 3} & 1 \end{array}\right] \tag{1}$$

by applying POSIT algorithm [23] in which R is a rotation matrix, $\mathbf{t}$ is a translation vector. After obtaining these parameters, the reference 3D face model is then rotated and translated to align with the 2D facial image plane. Four affine transformations are respectively applied on each eye, nose and mouth regions to warp 2D anchor points to the **xy**-image plane of the 3D reference face (Fig. 2b).



(a) (b) (c) (d) (e)

Fig. 2. (a) The detected face with 9 initial facial landmark anchor points. (b) Four affine transformations are respectively applied on each eye, nose and mouth to warp 2D anchors points to the **xy**-image plane of the 3D reference model. Red dots are the original points on the 2D detected face, yellow dots are the results after warping. (c) Frontalization of the 3D shape model along with the aligned points. (d) Project the 3D frontalized points to 2D image plane. (e) The points used and the equation to calculate the average eye angle $\bar{\theta}$ for feature extraction. (Color figure online)

This generates a 3D-aligned version of the 2D facial contour and the corresponding depth information of each facial point can be estimated [24].

Since texture of the face is not considered in our framework, the fitted P can be directly used without concerning about the corruption between pixel-wise warping. Finally, the frontalization is achieved by using transpose of the rotation matrix $R^{\top}$ on the 3D-aligned facial points, as illustrated in Fig. 2d.

3.3 Pairwise Feature Extraction

Once the system is activated, the system will collect frontalized facial landmark points to form the pairwise features and then the features are passed to mouth-DBN, head-DBN or eye-DBN. The 2D coordinates of facial landmark points of current frame-c are given as: $F_c = \{f_1^c, f_2^c, \ldots, f_{\mathbb{N}}^c\}$, where $\mathbb{N}$ is the number of points used. We deploy 2D pairwise differences of points for input in the first layer of DBN. The idea comes from the usage of joints in action recognition [25], because we believe that facial landmark points are similar to 3D joints in a way that they can both use position differences to characterize motion information; however, we introduce innovations on some expects. The pairwise differences of points capture posture features, motion features, and max-pool motion features by directed concatenation: $\mathcal{A} = [f_{cc}, f_{cp}, f_{cd}]$ in which:

$$\begin{aligned} f_{cc} &= \{f_i^c - f_j^c | i, j = 1, 2, \ldots, \mathbb{N}; i \neq j\} \\ f_{cp} &= \{f_i^c - f_i^p | f_i^c \in F_c; f_i^p \in F_p\} \\ f_{cd} &= \{f_i^c - f_{i,w}^d | f_i^c \in F_c; f_{i,w}^d \in F_d; w = 10, 20, \ldots, 60\} \end{aligned} \tag{2}$$

where f^p denotes the landmark points extracted from preceding frame-p and f^d denotes the landmark points in the frame-d whose sum of differences of points to current frame is maximum in a window size w. Since our system cannot discover where the initial frame of motion is, to characterize more motion information on continuous frames, we replace the offset features f_{ci} in [25] with max-pool motion features f_{cd}. The illustration is demonstrated in Fig. 3.

Specifically, we consider that a motion can be captured by computing the differences of points between current frame and dozens of proceeding frames (previous 2, 3 seconds). However, the more the number of proceeding frames be used for features, the more dimension of features would be, which makes the following DBNs on drowsiness-related symptoms cost too much time on computation. Therefore, we design an innovative way just like max-pool layer in convolutional neural network [26] to discard less important information in time order, *i.e.* to find frames which has maximum sum of differences on certain window size, and then calculate their differences of points as features.

Dimension of $\mathcal{A}$ results in $N_{\mathcal{A}} = (\mathbb{N} \times (\mathbb{N} - 1)/2 + \mathbb{N} + \mathbb{N} \times 6) \times 2$. Ten facial landmarks ($\mathbb{N} = 10$) at both mouth corners, the middle of uppers and lower lips are considered for yawning and talking/laughing detection. All the points are normalized by the scheme mentioned in Sect. 3.2 before extraction. As for head features, they are extracted in the same fashion as facial landmark points whilst

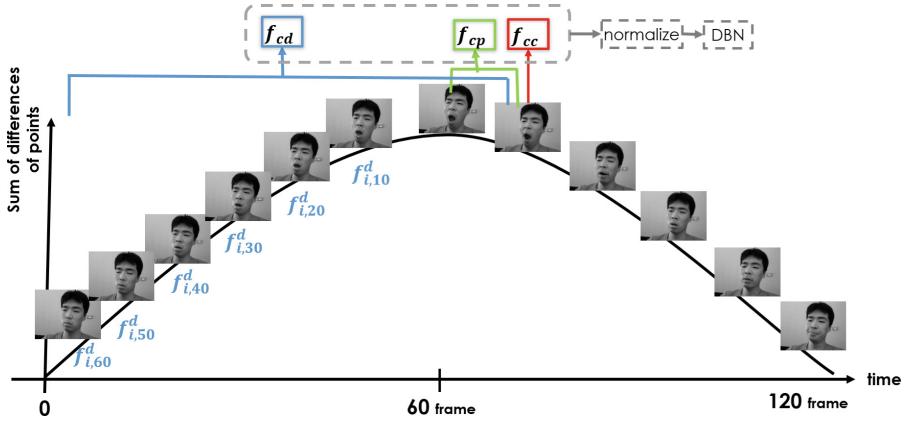


Fig. 3. Raw feature formation: three feature channels f_{cc}, f_{cp}, f_{cd} for capturing information of posture, motion, max-pool motion.

the 2D positional points are replaced by yaw, pitch and roll angles decomposed from the rotation matrix R.

Nevertheless, as for eye features, we consider the average eye angles $\overline{\theta}_{left}, \overline{\theta}_{right}$ at both eyes (Fig. 2e) in previous 10 seconds as features, because eyes have small variation that only can be identified to 3 states, opened, half-closed and closed eyes, which may make extracting eye pairwise features be less meaningful as mouth or head motions.

3.4 Learning the Higher Level Representation

RBMs were originally developed using binary stochastic units for both the visible and hidden units, but logistic units are a very poor representation for real-world data such as pixel intensities in natural images [27]. One solution is to replace the binary visible units by linear units with independent Gaussian noise [28], so called Gaussian RBM (GRBM). The original energy function can be replaced by:

$$E(\mathbf{v}, \mathbf{h}) = \sum_{i \in vis} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in hid} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} \tag{3}$$

where w_{ij} denotes the matrix of connection between visible unit i and hidden unit j with their bias terms a_i and b_j and σ_i is the standard deviation of the Gaussian noise for visible unit i. Because our pairwise features are continuous features, we use the above GRBM in the same fashion of [10] to model the energy term of the first visible layer. It is possible to learn the variance of the noise for each visible unit but it is much easier to normalize the data (mean subtraction and standard deviation division) to have zero mean and unit variance (e.g. $\sigma_i^2 = 1$) in the preprocessing phase.

Specifically, the DBN in our system consists of one visual layer (*i.e.* the lowest layer in GRBM) with continuous pairwise features and four hidden layers

to learn a hierarchical feature representation given training data extracted from mass pool of facial landmark points and head posture data. Given the nodes at the third hidden layer and the classification labels (*i.e.* stillness, yawning or laughing/talking; stillness, nodding or looking aside; normal eye or sleepy eye), the output of the DBN (*i.e.* last hidden layer) can classify by comparing the values of these nodes. The DBNs in HTDBN are activated at all times to detect mouth-motion, head-motion and eye-motion. For mouth-DBN, the model aims to discriminate yawning from mouth stillness and laughing/talking, and for head-DBN, it is used to classified head stillness, nodding and head looking aside. In addition, eye-DBN can differentiate sleepy eye from normal eye.

3.5 Continuous-HMM for Drowsy Driver Detection

A hidden Markov model describes the statistical behavior of a process in time. At each time step t, we have one 8-dimensional feature vector X_t taking values of each motion class probabilities obtained from DBNs, *i.e.* $X_t = \{x^{eye}, x^{mouth}, x^{head}\}$, where $x^{eye} = \{p^{normal}, p^{sleepy}\}$, $x^{mouth} = \{p^{stillness}, p^{yawning}, p^{laughing/talking}\}$, $x^{head} = \{p^{stillness}, p^{nodding}, p^{looking-aside}\}$. The intuition behind this is that a motion in drowsiness-related symptoms consists of a sequence of state transition, e.g. yawning is a long-term motion of mouth that has its fullest open state in between stillness at onset and offset. Therefore, the variation of motion probabilities in a sequence of feature vectors can describe a motion or a transition from different motions.

Here, X_t not only serves as an observation vector O_t for HMM but also describes the relation between eyes, mouth and head. Existing work tend to neglect such relations when detecting drowsiness-related symptoms, e.g. eyes might be closed when yawning, but only mouth features are considered.

Assume a HMM has J (unobserved) states $\{s_1, s_2, \ldots, s_J\}$ and K observation vectors $\{o_1, o_2, \ldots, o_K\}$. At time t, HMM occupies a state s_i and may undergo a state transition from the $s_i = i$ to a state $s_{i+1} = j$ at time $t+1$ with the state transition probability $a_{ij} = P(s_{t+1} = j | s_t = i)$. Associated with each state is a set of observation vectors o_t with their respective observation probability densities, Gaussian M-component mixture densities, $b_i(o) = \sum_{k=1}^{M} c_{ik} \mathcal{N}[o, \mu_{ik}, \mathbb{U}_{ik}]$, where c_{jk} is the mixture weight, $\mathcal{N}$ is the normal density and μ_{ik} and $\mathbb{U}_{ik}$ are the mean vector and co-variance matrix associated with state i, mixture k. Starting from an initial state $s_1 = i$ with probability $\pi = P(s_1 = i)$, the process undergoes a sequence of state transitions over a time duration T and generates an observation vector sequence $O = \{o_1, o_2, \ldots, o_T\}$ with a certain probability. In a sense, the HMM is specified by a triplet $\lambda = (\mathit{\Pi}, A, B)$, where $A = \{a_{ij}\}$ denotes transition probabilities, $B = \{b_j(o)\}$ denotes observation symbol probabilities, and $\mathit{\Pi} = \{\pi_i\}$.

For each HMM, the model parameters can be trained from the selected training vector sequences by applying Baum-Welch algorithm [29]. Here we assume our HHM to be an ergodic model. A model is ergodic means its transition matrix is fully connected, which means all transitions have non-zero probabilities.

With the above definition for a model λ, a given observation symbol sequence O may be generated from one or more state sequences $S = \{s_1, s_2, \ldots, s_T\}$. The probability requires summation over all possible state sequences and an efficient way to do so is a forward-backward algorithm [30]:

$$\begin{aligned} P(O|\lambda) &= \sum_{allS} P(O|S,\lambda)P(S|\lambda) \\ &= \sum_{allS} \pi_{s_1} b_{s_1}(o_1) \prod_{t=2}^{T} a_{s_{t-1}s_t} b_{s_t}(o_t). \end{aligned} \tag{4}$$

Given a video sequence with length T, we obtain the likelihood difference $d = \mathrm{P}(\mathrm{O}|\lambda^{drowsy}) - \mathrm{P}(\mathrm{O}|\lambda^{nondrowsy})$ from every 300 frames. We then accumulate the likelihood differences within a specific among of time τ to obtain $\tilde{d}_t$. Finally, the drowsiness level (DL) is determined using the inverse logit transform (Eq. 5). While drowsiness level is more than 50%, then the drowsiness detector would consider the current state as drowsiness.

$$DL(\tilde{d}_t) = \frac{1}{1 + e^{-\tilde{d}_t}} * 100\% \tag{5}$$

4 Dataset Acquisition

Most of the previous works on drowsy driver detection attempted to recognize a small set of cases for driver drowsiness detection. Although [31] provided a freely-available dataset for yawning detection, it is still insufficient for a comprehensive drowsy driver study based on pure visual cues. Drowsiness detection from yawning alone is too restricted for use in practice. It should be combined with some additional indicators of drowsiness. Therefore, we collect a large video dataset for performance evaluation of drowsy driver detection methods.

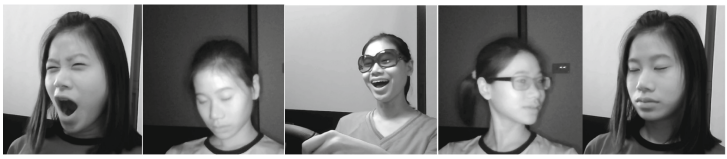
Camera Setting. To cope with the night time or poor lighting problem, we used the active infrared (IR) illumination and acquire IR videos in the dataset collection. To ensure a realistic setup, all the videos were captured by D-Link DCS-932L, a stand-alone surveillance digital camera with the resolution set at 640×480 pixels. The built-in infrared LEDs allow us to view in any light condition from daytime to nighttime. The advantage of activating IR illuminators also in daytime is because it captures occluded eyes with people wearing sunglasses better than using RBG camera (Fig. 4b). However, for the sake of completeness, 24-bit true color (RGB) Logitech C310 HD at 30 frames per second webcam was also set up simultaneously to record the data at 720p in daytime only (Fig. 4c).

Environment Setting. In the collection of our dataset, two rounds of video recordings were performed for each subject. The first round was recorded during the daytime and the second one was performed in the night. In order to

(a)

(c) Samples of participants.

(b)



(d) Different Scenarios and performances.

Fig. 4. (a) Participants were equipped with a fixed but tunable chair and a simulated driving wheel with pedals; they are instructed to perform a series of facial displays shown in Fig. 4d. A standalone IR camera and RBG webcam were placed at the left hand side of the driver while the ambient light was augmented with high-intensity lamp to simulate the condition of sunny day. (b) IR images can capture better occluded eyes for subjects wearing sunglasses. (c) Diversity of skin colors, genders and ethnicities among the participants in the dataset collection. IR and RGB videos were taken simultaneously. (d) Situations like yawning in *BareFace*, nodding when falling asleep in *Night-BareFace*, laughing in *Sunglasses*, looking aside in *Night-Glasses*, blink slowly in *BareFace* and *etc.* are considered and separately recorded.

simulate the condition of sunny day, for approximately one third of subjects, ambient room lighting augmented by a high-intensity lamp was used (Fig. 4a) for the daytime recording. Illuminance was measured by a light meter to ensure environment was well-established.

Our camera was placed on the top left hand side of the subject to emulate the position in the A-Pillar, a common used location in cars. In contrast, most datasets and the corresponding algorithms were based upon fully frontal face views, which is impractical to set up in a real cars since the camera would block the driver's view and the dashboard.

Participants. To make sure the algorithm works for various skin race and genders, 36 adults, aged from 18 to 40 years old with various ethnicities and diverse skin colors (32.5% of black or brown, 32.5% of white, and 35% of yellow) and genders (50% of female), participated in the video collection. The subjects with different hairstyles and clothing were recorded with and without glasses/sunglasses to simulate a wide variety of driving scenarios. Figure 4c and d show some samples of the participants under different conditions and driving scenarios.

Driver Videos. Subjects were recorded when they sit on a chair and play a plain driving game with simulated driving wheel and pedals; meanwhile, they were instructed by an experimenter to perform a series of 8 actions under 5 kinds of scenarios: *BareFace*, *Glasses*, *Sunglasses*, *Night-BareFace* and *Night-Glasses*. The sequences recorded from each subject can be regarded as two branches: drowsiness and non-drowsiness. For drowsiness-related sequences, yawning, slow blink rate (high PERCLOS) and falling asleep (high PERCLOS followed by frequent nodding) were taken about 1 min long, and the combination of drowsiness-related symptoms sequences (yawning, high PERCLOS, frequent nodding) were recorded about 1.5 min. On the other hand, sequences of normal driving (low PERCLOS), shocked face and burst out laughing in about 1 min and the combination of non-drowsiness actions (talking, laughing, looking at both sides) recorded about 1.5 min are represented as the non-drowsiness data. Some examples can be seen in Fig. 4d. Overall, *360 videos* were taken to complete the dataset.

Moreover, to simulate more practical driving situations, 18 subjects from the proposed dataset are randomly selected yet kept the balance of various gender, skin races. Their sequences are edited and combined into a 2–10 min mixing video for each subjects under 5 kinds of scenarios which contains various situations with different number of transitions from non-drowsiness state to drowsiness state, or drowsiness state to non-drowsiness states. Overall, there are 90 mixing videos be added to the dataset for evaluation.

5 Experimental Results

5.1 Experimental Setup

We evaluate the proposed *HTDBN* framework by using the provided dataset mentioned in Sect. 4. We first train the DBNs with a four-hidden-layer structure, where the numbers of the nodes in all the layers are $[N_{\mathcal{A}}, 1000, 1000, 500, N_y]$ from the lowest layer to the highest one, respectively. The numbers of nodes in the visible layer are $N_{\mathcal{A}} = 230$, 45, and 200 for mouth-, head-, and eye-DBN, respectively. The number of outputs N_y for each DBNs is equivalent to the number of classes, *i.e.* there are 3 outputs for mouth- and head-DBN, 2 outputs for eye-DBN.

In our experiments, the dataset is divided into two parts: training, testing dataset. The subjects that have edited mixing videos are for testing, and the sequences from the other subjects are for training. In drowsiness-related symptoms detection, all the sequences from each subject in the training dataset are taken to train DBNs. For each of the sequence, the mouth-motion, head-motion, and eye-motion class probabilities are extracted at each frame and the observable vectors can be obtained. In drowsy driver detection, to maintain accurate classification capability, the selection of training observable vectors sequences is very important. This is since the chosen sequences can be used to adjust the model parameters that can also be used to recognize other sequences of observable symbols, e.g. a drowsy observable vector sequence should get higher probability of

re-generation from drowsiness HMM than non-drowsiness HMM. The data we captured for each person consists of long videos. To perform our training, two kinds of videos, the combination of drowsiness-related symptoms videos and the combination of non-drowsiness-related actions videos, are used for training data. We randomly subdivide each of the videos into various shorter overlapping and non-overlapping videos with fixed length 300 frames.

5.2 Drowsy Driver Detection Performance Evaluation

We present the performance of the proposed HTDBN for driver's drowsiness detection using the collected dataset. We use not only accuracy but also F_1-score to evaluate the performance of the proposed detection algorithm since it is a relatively fair measure for unbalanced data:

$$F_1\text{-score} = 2 * \frac{prec(\Delta) * rec(\Delta)}{prec(\Delta) + rec(\Delta)} \tag{6}$$

Δ is the length of the sequence of its likelihoods decoded from drowsiness-HMM and non-drowsiness-HMM. To ensure the stability and to fully scrutinize the strength of the proposed framework, we conduct an experiment to show the performance of the fine-tuned system on *BareFace*, *Glasses*, *Sunglasses*, *Night-BareFace* and *Night-Glasses* separate scenarios. The average scores of different scenarios for drowsiness and non-drowsiness detection are shown in Table 1.

Table 1. The 1^{st} experiment: performance of HTDBN on *separate scenarios* in the drowsy driver detection dataset

Scenario	Drowsiness F_1-score	Non-drowsiness F_1-score	Accuracy
BareFace	92.17%	92.64%	92.42%
Glasses	88.17%	85.04%	86.79%
Sunglasses	74.17%	78.59%	76.58%
Night-BareFace	92.60%	90.97%	91.87%
Night-Glasses	77.74%	73.73%	75.90%
Overall	**85.39%**	**84.19%**	**84.82%**

As shown in the table, the first three scenarios were evaluated in daytime. It can be seen that for the cases with no occlusion on face (*BareFace*) and enough ambient lighting, we can achieve about 92% accuracies on both drowsiness and non-drowsiness detections. As the detector encounters driver-wearing-glasses scenario, sometimes the reflections in glasses and the glasses frames may cause disturbance of eye openness detection and thus lose a good information from eye-motion detection.

In addition, the performance of *Sunglasses* falls behind very much with other scenarios, although eyes occluded by sunglasses have better visibility

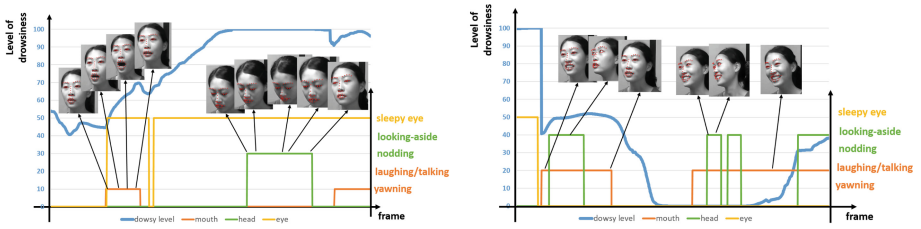


Fig. 5. Plots of the processed data aligned with level of drowsiness versus time: (a) Non-drowsy (alert) driver falling asleep example and (b) drowsy driver awake example. Blue curve represents level of drowsiness; below are orange, green and yellow curves represent predictions of mouth- and head- and eye-motions from DBNs, respectively. (Color figure online)

using IR camera compared with RGB camera, the stronger reflections and hindrance deteriorate on eye-contour-point detection. Therefore, to deal with the big challenge on driver-wearing-sunglasses scenario, firstly we design a sunglasses detector to find the region of sunglasses by the image intensity from detected eye points area, and then the sunglasses region is employed gamma correction ($\gamma = 0.4$) to adjust brightness and simple reflection removal on the reflection that would block eye contours. After these improvements, the result of *Sunglasses* is increased 9% which is about 77% accuracy.

The rest two scenarios were experimented in nighttime. The IR videos collected from the chosen camera can still capture clear faces, but facial landmark points are less accurate than in daytime resulting in lower accuracies. The accuracy of the overall scenarios is about 85%, and so do the scores for both drowsiness and non-drowsiness detection.

Figure 5a depicts an example of processing a 60-second video of an alert driver gradually falling asleep by using the proposed algorithm. Orange, green and yellow curves represent the predictions of mouth- and head- and eye-motions from DBNs, respectively, co-existing at every time stamp. As shown, when mouth-DBN first recognizes a yawning cycle the level of drowsiness goes up drastically to about 70%. As the blink rate goes slower and eye closure duration goes longer, the drowsiness percentage keeps growing up to the top 100% until the head-DBN detects a nodding cycle, a high-risk warning should be raised. In contrast, a drowsy driver waking up case illustrated in Fig. 5b shows the sustainability of the proposed system. While no sleepy eye is noticed, the level of drowsiness decrease dramatically. Mouth- and head-DBN detect several laughing/talking motions and head looking aside motions, thus the system determines the drowsiness level of the driver to be below 50%.

5.3 Comparison with Other Solutions

In this experiment, our proposed HTDBN model is compared with baseline model: support vector machine (SVM) on mixed scenarios in the Drowsy Driver Dataset. Before deep learning appears, support vector machine was the most popular technique for data classification, especially the SVM with kernel trick that can efficiently perform a non-linear classification. For a fair comparison, every solution is trained and tested same as the proposed HTDBN solution, yet the only difference is that the determination of the drowsiness state. As for SVM, the probability of drowsiness determines the final drowsiness state, while for HMMs, it is decided by the difference between drowsiness-HMM and nondrowsiness-HMM. Table 2 shows the effectiveness of applying different models (SVM or DBN, HMM) on the proposed framework.

Table 2. The 2^{nd} experiment: comparison between HTDBN and baseline solution on *mixed scenarios* in the drowsy driver detection dataset

	Drowsiness	Non-drowsiness	Accuracy
SVM+SVM	81.16%	74.99%	78.51%
SVM+HMM	81.30%	77.14%	79.43%
DBN+SVM	84.26%	79.20%	82.08%
Ours	**85.39%**	**84.19%**	**84.82%**

The accuracy is apparently lower when SVMs (SVM+SVM solution) are substituted for all detectors in the proposed algorithm. To further discuss the separate effectiveness in different part of the framework, the SVM+HMM solution only replaces DBNs with SVMs in drowsiness-related symptoms, and only HMMs are replaced with a binary-class SVM in drowsy driver detection for DBN+SVM solution. From the results, DBNs play important roles in the overall system which improves 5% accuracy, whilst the usage of HMMs increases 3%. In conclusion, our proposed HTDBN algorithm is more suitable for classification in time series.

5.4 Yawning Detection Performance on YawDD Dataset

Although *Yawning Detection Dataset (YawDD)* [31] is not a sufficient dataset for comprehensive drowsy driver detection because it is only determined by yawning detection. However, to compare with existing approaches, the yawning detector mouth-DBN in our proposed system has been evaluated on the YawDD dataset as well. According to the evaluation scheme from [32], we got 94% yawning detection accuracy with 2% false alarm rate on CASE I *(camera under the mirror)*, and 92% accuracy with 5% false alarm rate on CASE II *(camera on the dash)*. Our results outperform recent methods presented in [31–33], whose accuracies on the better case are 60%, 75%, and 92% but 13% false alarm rate, respectively.

5.5 Computational Complexity

Our driver drowsiness detection system consists of an off-line training phase and on-line detection phase. Though the learning in the network is uninterestingly long, once the model training is finished, with low inference cost, the entire system is able to perform in real-time with MATLAB implementation using a Core i5, 3.1GHz PC with 16GB RAM, at an average speed of 20 fps. More precisely, a single multi-layer feedforward neural networks incurs in linear running time $\mathcal{O}(T)$ and the forward-backward algorithm applied in HMM has time complexity $\mathcal{O}(N^2T)$, where T is the length of the sequence and N is the number of states.

6 Conclusion

In this paper, we presented a novel HTDBN that utilizes DBNs for learning contextual frame-level representations for drowsiness-related symptoms. By encoding dynamic structure of drowsiness and non-drowsiness information into HMM-based models, the results are robust and promising under different circumstances. The proposed continuous-HMM can model the interactive relations among eyes, mouth and head. Moreover, for performance evaluation, we collected a large drowsy driver detection dataset in which various skin colors, scenarios and lighting conditions are considered. Experimental results on various kinds of scenarios and fusion all together demonstrated the power of the proposed framework in estimating the driver's drowsiness level.

Acknowledgement. The authors would like to thank Qualcomm Technologies Inc. for supporting this research work.

References

1. Bergasa, L., Nuevo, J., Sotelo, M., Barea, R., Lopez, M.: Real-time system for monitoring driver vigilance. IEEE Trans. Intell. Transp. Syst. **7**, 63–77 (2006)
2. World Health Organization: Global status report on road safety 2013: supporting a decade of action: summary. World Health Organization (2013)
3. Wheaton, G., Shults, R.: Drowsy driving and risk behaviors 10 states and Puerto Rico. Online article (2014)
4. National Sleep Foundation: Drowsy driving reduction act of 2015 (2014)
5. Colic, A., Marques, O., Furht, B.: Driver Drowsiness Detection: Systems and Solutions. Springer, Heidelberg (2014)
6. Mercedes-Benz: Attention assist: drowsiness-detection system warns drivers to prevent them falling asleep momentarily. Online article (2008)
7. Teyeb, I., Jemai, O., Zaied, M., Ben Amar, C.: A drowsy driver detection system based on a new method of head posture estimation. In: Corchado, E., Lozano, J.A., Quintián, H., Yin, H. (eds.) IDEAL 2014. LNCS, vol. 8669, pp. 362–369. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10840-7_44
8. Qiang, J., Lan, P., Looney, C.: A probabilistic framework for modeling and real-time monitoring human fatigue. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **36**, 862–875 (2006)

9. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2010)
10. Wu, D., Shao, L.: Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 724–731 (2014)
11. Yang, G., Lin, Y., Bhattacharya, P.: A driver fatigue recognition model based on information fusion and dynamic Bayesian network. Inf. Sci. **180**, 1942–1954 (2010)
12. Ji, Q., Zhu, Z., Lan, P.: Real-time nonintrusive monitoring and prediction of driver fatigue. IEEE Trans. Veh. Technol. **53**, 1052–1068 (2004)
13. Mohamed, A., Dahl, G., Hinton, G.: Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. **20**, 14–22 (2012)
14. Dasgupta, A., George, A., Happy, S., Routray, A.: A vision-based system for monitoring the loss of attention in automotive drivers. IEEE Trans. Intell. Transp. Syst. **14**, 1825–1838 (2013)
15. Alioua, N., Amine, A., Rziza, M.: Drivers fatigue detection based on yawning extraction. Int. J. Veh. Technol. **2014** (2014)
16. Rezaei, M., Klette, R.: Look at the driver, look at the road: no distraction! No accident! In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 129–136 (2014)
17. Smith, P., Shah, M., da Vitoria Lobo, N.: Determining driver visual attention with one camera. IEEE Trans. Intell. Transp. Syst. **4**, 205–218 (2003)
18. Eskandarian, A., Sayed, R.: Analysis of driver impairment, fatigue, and drowsiness and an unobtrusive vehicle-based detection scheme. In: Proceeding of International Conference on Traffic Accidents (2005)
19. Taylor, G., Hinton, G., Roweis, S.: Modeling human motion using binary latent variables. In: Neural Information Processing Systems, pp. 1345–1352 (2006)
20. Hinton, G., Osindero, S.: A fast learning algorithm for deep belief nets. Neural Comput. **18** (2006)
21. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
22. Xiong, X., de la Torre, F.: Supervised descent method and its application to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 532–539 (2013)
23. DeMenthon, F., Davis, L.: Model-based object pose in 25 lines of code. Int. J. Comput. Vis. **15**, 123–141 (1995)
24. Heo, J., Savvides, M.: Gender and ethnicity specific generic elastic models from a single 2D image for novel 2D pose face synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 2341–2350 (2012)
25. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 14–19 (2012)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates Inc, Red Hook (2012)

27. Hinton, G.E.: A practical guide to training restricted Boltzmann machines. In: Montavon, G., Orr, G.B., Müller, K.-R. (eds.) Neural Networks: Tricks of the Trade. LNCS, vol. 7700, 2nd edn, pp. 599–619. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35289-8_32
28. Freund, Y., Haussler, D.: Unsupervised learning of distributions on binary vectors using two layer networks. Technical report, University of California at Santa Cruz, Santa Cruz, CA, USA (1994)
29. Yang, L., Widjaja, B., Prasad, R.: Application of hidden Markov models for signature verification. Pattern Recogn. **28**, 161–170 (1995)
30. Devijver, P.A.: Baum's forward-backward algorithm revisited. Pattern Recogn. Lett. **3**, 369–373 (1985)
31. Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., Hariri, B.: YawDD: a yawning detection dataset. In: Proceedings of the 5th ACM Multimedia Systems Conference, pp. 24–28. ACM (2014)
32. Omidyeganeh, M., Shirmohammadi, S., Abtahi, S., Khurshid, A., Farhan, M., Scharcanski, J., Hariri, B., Laroche, D., Martel, L.: Yawning detection using embedded smart cameras. IEEE Trans. Instrum. Meas. **65**, 570–582 (2016)
33. Zhang, W., Murphey, Y.L., Wang, T., Xu, Q.: Driver yawning detection based on deep convolutional neural learning and robust nose tracking. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2015)