A Driver State Detection Framework Based on Temporal Facial Action Information Postprint

Authors: Cui Ziyan, Wang Jianming, Jin Guanghao

Date: 2018-09-12T00:00:00+00:00

Abstract

In the domain of safe driving, the driver' s physical and mental state is of paramount importance to traffic safety. Capturing driver facial video via network cameras and feeding it into networks for detection constitutes an effective approach for identifying abnormal driving behaviors such as fatigue. Previous methods primarily determined fatigue by analyzing facial expressions including mouth shapes to detect yawning, consequently misclassifying many similar states such as speaking as fatigue. To address these issues, we propose a detection framework based on temporal facial action information for driver state monitoring, which enhances detection accuracy while reducing false positive rates. This framework comprises two key components: a) extracting multiple facial features by detecting facial contours in videos to form facial action units; b) training a corresponding LSTM network to construct temporal facial action units, fusing multiple action units based on their correlations to determine the final driver state. Experimental results on the public YAW-DD dataset demonstrate that, compared to existing methods, the accuracy improves to 93.1% while substantially reducing the false detection rate for fatigue states.

Full Text

Preamble

Title: Driver State Detection Framework Based on Temporal Facial Action Information

Authors: Cui Ziyan[1], Wang Jianming[1,2], Jin Guanghao[1]†

Affiliation: 1. School of Computer Science & Software, Tianjin Polytechnic University, Tianjin 300380, China 2. School of Electronics & Information Engineering, Tianjin Polytechnic University, Tianjin 300380, China

Abstract: In the field of safe driving, a driver' s physical and mental state is critical to traffic safety. Detecting abnormal driving conditions such as fatigue through network camera input of driver facial videos has proven to be an effective approach. Previous methods primarily analyzed facial expressions such as mouth shape to detect yawning and thereby determine fatigue driving. However, this approach misclassifies many similar states, such as speaking, as fatigue. To address this issue, we propose a detection framework based on temporal facial action information to detect driver states, thereby improving detection accuracy and reducing false positive rates. The framework consists of two key components: (a) extracting multiple facial features by detecting facial contours in videos to form facial action units, and (b) training corresponding LSTM networks to create temporal facial action units that fuse multiple action units based on their correlations to detect the final driver state. Experimental results on the public YAW-DD dataset demonstrate that our method achieves 93.1% accuracy while substantially reducing the false detection rate of fatigue states compared to existing approaches.

Keywords: abnormal driving; temporal information; facial action coding system

0 Introduction

Traffic accidents represent one of the leading causes of fatalities worldwide, surpassed only by natural disasters, and result in significant property damage. Research indicates that driver state, vehicle condition, weather, and road conditions contribute to accident causation, with driver state accounting for 95% of incidents. Consequently, assessing and evaluating driver state is of paramount importance. Generally, technologies for detecting abnormal driving states can be categorized into two classes: environmental factors related to roads or vehicles, and driver behavior. This research focuses on detecting driver states to determine whether they are abnormal.

Abnormal driver states primarily include fatigue driving, drunk driving, angry driving, and distracted driving. Numerous vision-based driver fatigue detection systems have been developed, typically employing dashboard-mounted cameras that capture facial and eye features directly to discriminate driver states. Monitoring devices that acquire driving videos combined with computer vision methods offer a cost-effective and efficient solution. In previous work, researchers proposed a method for monitoring driver states during driving based on eye state and head pose, demonstrating that combining eye and head information enables more effective driver state detection. McCall et al. employed Driver Intent Inference (DII) methods instead of trajectory prediction (TF) techniques, inferring whether drivers intentionally change lanes and predicting whether vehicles will cross lane boundaries regardless of driver state.

While these approaches have achieved more effective driver state detection in

many respects, video-based fatigue detection remains challenging due to variations in lighting conditions, head pose changes, and temporal dependencies. Particularly, large variations in head pose cause severe deformation of facial shapes in videos, making it difficult for conventional methods to extract effective data. Although methods based on aligned facial points better represent fatigue features, ignoring temporal relationships in videos means they cannot effectively distinguish normal blinking from fatigue-induced blinking. Therefore, recent work has incorporated temporal information to differentiate states with long-term temporal dependencies, such as yawning and speaking.

Given the complexity of human behavior and facial expressions, this paper proposes a temporal information-based, multi-feature fusion framework for facial action recognition to comprehensively discriminate driver states. Our approach first employs the Facial Action Coding System (FACS) to evaluate driver facial muscle actions based on visual facial information. We then utilize temporal information from the data for continuous discrimination across multiple frames, enabling more accurate driver state classification. By introducing Long Short-Term Memory (LSTM) networks to address long-term dependency issues in detection, our method leverages temporal information without additional computational overhead. The framework' s contributions are threefold: (a) comprehensive utilization of multiple facial feature information for driver facial action discrimination; (b) driver state judgment through temporal information from consecutive frames; and (c) intensity-based discrimination of facial actions for more precise facial expression analysis. These improvements establish a more effective driver state detection framework that enhances detection accuracy while reducing false positive rates.

1.1 Facial Action Coding System

The Facial Action Coding System (FACS), developed based on the theories of Swedish anatomist Carl-Herman Hjortsjö, classifies human facial movements through facial appearance. Ekman, Friesen, and Joseph C. Hager published a significant FACS improvement in 2002. Using FACS, subtle and instantaneous changes in facial appearance encode individual facial muscle movements. FACS has become an automated computer system capable of detecting faces in videos, extracting geometric facial features, and generating motion characteristics for each facial movement.

FACS is arguably the most widely used method for coding facial expressions in behavioral science. The system describes facial expressions through 46 component movements, roughly corresponding to individual facial muscle actions. Figure 1 [Figure 1: see original paper] illustrates an example. FACS provides an objective and comprehensive approach to analyzing facial states or actions and has proven effective for detecting human emotional states. FACS can encode almost any anatomically possible facial expression, deconstructing it into specific Action Units (AUs) that objectively describe different facial expressions. In this paper, we employ this system to discriminate driver expressions and states by

detecting actions such as chin (AU17), nasolabial furrow depth (AU11), outer (AU2) and inner eyebrow raising (AU1) in videos, then combining these action units to determine the driver' s current state.

1.2 Long Short-Term Memory Networks

Long Short-Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) capable of learning temporal information from features. Proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997 and later improved and popularized by Alex Graves, LSTM has achieved remarkable performance in numerous temporal detection problems. The network' s ingenuity lies in adding input gates, forget gates, and output gates, making the self-recurrent weights variable. This enables dynamic integration modifications at different moments while model parameters remain fixed, thereby avoiding gradient vanishing or explosion issues in deep neural networks.

In problems exhibiting temporal dynamics, LSTM networks applied to spatial feature representations across consecutive frames can distinguish states with temporal relationships, such as yawning versus laughing, or blinking versus eye closure. Consequently, our framework incorporates LSTM networks to mine temporal cues in videos.

2 Our Method

As is well known, drowsy states as fatigue manifestations concentrate information in several primary facial regions: eyes, nose, and mouth. Speaking and anger expressions also focus on these facial areas. Therefore, our proposed method utilizes the Facial Action Coding System to analyze driver facial information in videos, then encodes individual facial muscle movements through subtle and instantaneous changes in facial appearance. By combining these action units, we comprehensively judge driver facial states in single-frame data. Subsequently, we perform temporal analysis on the acquired frame data to make more scientific judgments about driver states during that time period. The procedure is illustrated in Figure 2 [Figure 2: see original paper].

Previous methods only determined whether a driver was fatigued, often misclassifying many speaking actions as yawning due to their similarities. In practical applications, such misjudgments can affect assessments of driver states, preventing effective determinations and precautions. To reduce misclassification probability, we categorize speaking as a separate action and apply temporal learning to speaking action units, thereby reducing false fatigue detection from speaking. Our experimental section demonstrates the existence of these misclassifications and the critical role our method plays in mitigating them.

2.1 Facial Landmark Marking

Our driver state judgment relies on facial information, which can be represented through facial landmarks. To obtain these landmarks, we first construct a facial shape model using the Active Shape Model (ASM) function as shown below:

$$x = \bar{x} + Pb$$

where $\bar{x}$ represents the mean face shape, P is a matrix composed of principal components of shape variation, and b is the parameter vector. The coordinate vector of N feature points on an image is represented as:

$$x = [x_1, y_1, x_2, y_2, ..., x_N, y_N]^T$$

The mean face shape across all M images can be expressed as:

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i$$

Subtracting each face vector from the mean face vector yields a shape variation matrix X with zero mean:

$$X = [x_1 - \bar{x}, x_2 - \bar{x}, ..., x_M - \bar{x}]$$

It should be noted that X is a zero-mean shape variation matrix because the mean face vector is subtracted from each row vector. Principal Component Analysis (PCA) on XX^T yields decisive shape variation components—eigenvectors P_j and corresponding eigenvalues λ_j. Selecting the first K eigenvectors arranged in columns forms the shape variation matrix P. These eigenvectors essentially form the basis for all sample transformations and can represent any variation within the samples. The shape variation matrix P and parameters B can be obtained through:

$$B = P^T(x - \bar{x})$$

After constructing the shape model, we can initialize the face shape model on detected faces. The subsequent task involves finding the best matching point for each landmark in its vicinity. Long-duration facial tracking easily leads to tracking drift or face disappearance, necessitating a facial landmark confirmation step. Simple Convolutional Neural Networks (CNN) can be trained to detect faces and generate landmarks, as illustrated in Figure 3 [Figure 3: see original paper].

2.2 Facial Action Coding

Facial landmarks enable extraction of individual facial regions for facial action coding. This paper adapts Tada et al.' s method to implement the facial action coding system.

2.2.1 Face Tracking The Constrained Local Model (CLM) used can be described through parameters $p = [s, R, p, t]$, which vary to obtain various model instances: scale factor s; object rotation R (first two rows of the 3D rotation matrix); 2D translation t; and vector p describing non-rigid shape variation. The Point Distribution Model (PDM) is:

$$x = s \cdot R \cdot \Phi(\bar{x} + Pb) + t$$

2.2.2 Alignment and Marking To better analyze facial texture, we map it to a common reference frame and remove variations caused by scaling and in-plane rotation. For this purpose, we apply a similarity transformation from currently detected landmarks to neutral expression landmarks (projection of the mean shape from 3D PDM). This yields a 112$×$112 pixel facial image with 45-pixel inter-pupillary distance. Procrustes superposition, which minimizes the mean square error between aligned pixels, is used to compute the similarity transformation. We separately train eyelid and eyebrow distribution models, then combine each.

2.2.3 Appearance Features After aligning facial images to 112$×112size, weextractappearancefeatures.T$ histograms with 31 dimensions, creating descriptors for faces.

2.3 Action Unit Selection

We judge driver states by combining multiple facial action unit information. To address our specific problem, we select appropriate facial action units while removing redundant information to reduce noise and improve efficiency. Through experiments on multiple datasets, we screen facial actions and ultimately select specific units (AU2, AU4, AU7, AU9, AU26, and AU45) as primary action units for discrimination. The action unit selection process is detailed in the experimental section.

2.4 Temporal Dynamic Feature Creation

The LSTM model consists of an input gate, a forget gate, an output gate, and a memory cell. With these three gates, LSTM modules can learn long-term dependencies in sequential data, and their parameters are easier to train. The memory cell can store long-term information in its vector, which can be rewritten or manipulated in the next time step.

Since yawning and speaking are continuous actions, discrimination based solely on action unit information from single images yields significant errors. Using continuous data as judgment criteria is more reasonable—driver states should be determined by all frame data within a time period rather than a single frame' s state. Therefore, we apply LSTM to model the temporal dynamics of driver states.

3 Model Construction and Experimental Analysis

This experiment uses two datasets: a self-collected dataset gathered through driving simulation, and the public Yaw-DD dataset. We construct and validate our framework using both datasets.

3.1 Dataset

The Yaw-DD dataset is widely used for validating fatigue detection algorithms and models. It contains videos from 57 male and 50 female volunteers of different ages, ethnicities, and facial characteristics. For Yaw-DD, we first perform preliminary data segmentation and screening. In each video, test subjects perform three behaviors: natural (no action), speaking, and yawning. We label and segment these three behaviors separately. The re-generated dataset contains videos primarily recorded during daytime but includes various lighting conditions—from early morning to sunset. Additionally, weather conditions encompass sunny and rainy days, creating diverse driving conditions. Some videos include other passengers moving in the vehicle, introducing background motion. Test subjects are categorized as wearing glasses or not. Videos are captured in 640$×$480 24-bit true color (RGB) AVI format at 30 fps without audio.

Since each Yaw-DD video contains long periods of mixed driver states (alternating between fatigue and non-fatigue), which complicates accurate driver state determination, we further segmented and relabeled these videos for LSTM training. Videos were clipped into segments containing only speaking, fatigue, or other states, and labeled accordingly. To describe transition characteristics between these three states, we retained 10 normal frames at the beginning and end of each state clip.

To further validate framework effectiveness on the public dataset, we built a self-collected dataset for experimental verification. Real driving data is difficult to obtain safely, so we used a driving simulator. The experimental setup is shown in Figure 4 [Figure 4: see original paper]. This dataset includes 10 test subjects (6 male, 4 female) simulating three driving states (normal, fatigued, and communicating) under three traffic conditions (urban, national highway, and rural road). During simulation, test subjects maintained optimal driving states and complied with driving rules. The dataset uses the same 640$×$480 24-bit true color (RGB) AVI format at 30 fps with audio, generating 300 minutes of data.

3.2 Action Unit Selection Experiments

3.2.1 Fatigue Information Unit Screening To understand facial action units related to fatigue, our method individually trains and tests each facial action unit using CNN networks. Detection results for each unit reveal which are effective for fatigue detection. Table 1 shows the correlation between fatigue and partial facial actions.

Through these experiments, we identify the six most predictive facial actions for fatigue: AU45 (blink/eye closure), AU2 (outer eyebrow raise), AU7 (inner orbicularis oculi tightening), AU4 (frown), AU9 (nose wrinkle), and AU26 (jaw drop). Our method combines these six facial action units for fatigue detection. Video analysis also reveals that when yawning, many subjects attempt to raise their eyebrows to keep eyes open, consistent with our experimental observation of strong AU2 correlation.

3.2.2 Speaking Action Unit Screening Speaking while driving is a common behavior that easily confuses with yawning in fatigue detection. Correctly distinguishing speaking from yawning is therefore necessary. Compared to fatigue detection, speaking involves fewer facial actions, primarily concentrated in the mouth region. We similarly test the correlation between each action unit and speaking, with results shown in Table 2 .

Table 2 indicates that the action units most correlated with speaking are AU25, AU26, and AU17. Therefore, when discriminating among the three states, our method selects eight action units: AU2, AU4, AU7, AU9, AU17, AU25, AU26, and AU45.

3.3 Dynamic Feature Analysis Experiment

For the selected eight units, we compare and analyze action intensity variation patterns during temporal processes in datasets with identical formats (5 seconds/150 frames). Results are shown in Figure 5 [Figure 5: see original paper].

Based on these experimental results, we can clearly observe substantial differences in intensity variation trends among action units within the same time window across different states. Therefore, discriminating among the three states using this data is effective.

3.4 LSTM Network Training and Testing

Based on the above experimental results, we establish an LSTM network framework and conduct validation experiments. We train on selected facial action units, starting with AU45 (most effective for fatigue detection), then progressively adding the next detection feature until all features are included. Table 3 shows accuracy results for LSTM networks with different feature combinations.

3.5 Binary vs. Ternary Classification

Using the highest-accuracy feature combination (AU45+AU2+AU4+AU7+AU9+AU17+AU25+AU26), we compare binary classification (fatigue vs. natural) and ternary classification (fatigue vs. speaking vs. natural). Results are shown in Figure 6 [Figure 6: see original paper].

Figure 6 clearly demonstrates that separating speaking as an independent class substantially reduces misclassification probability while improving fatigue state detection accuracy.

3.6 Traditional Algorithms vs. Our Algorithm

We compare several typical traditional machine learning algorithms with our temporal-information-enhanced algorithm for fatigue detection accuracy and false positive rate, with results shown in Table 4 .

Experimental results demonstrate that incorporating temporal information with identical features effectively reduces false positive rates and significantly improves state judgment accuracy. We also compare our method with recent approaches [18,19,20], with Table 5 presenting the comparison. Experiments show our method achieves 93.1% accuracy in fatigue detection, surpassing previous methods.

4 Conclusion

This paper proposes a framework for automatic driver state detection through monitor-acquired videos. Our target application scenario involves using ordinary cameras placed above the dashboard to capture driver facial video data for fatigue detection. Our method encodes facial action units and obtains appropriate combinations through practical experiments. To improve driver state detection accuracy, our framework adds Long Short-Term Memory networks for temporal sequence analysis, achieving more effective driver state recognition compared to existing methods. Testing on the Yaw-DD dataset validates the method' s applicability across multiple ethnicities. Compared to previous approaches, this framework is low-cost, requires no additional devices for the driver, and thus does not affect driving behavior. Future research will focus on discriminating more complex driver states to enable more accurate and precise abnormal driving behavior detection.

References

[1] Organization, World Health. World report on road traffic injury prevention [M]// World Report On Road Traffic Injury Prevention. World Health Organization, 2004: 270-275.

[2] Al-Sultan S, Al-Bayatti A H, Zedan H. Context-aware driver behavior detection system in intelligent transportation systems [J]. IEEE Trans on Vehicular

Technology, 2013, 62 (9): 4264-4275.

[3] Wang Zhongmin, Li Zhuo, Fan Lin. Driving behavior recognition algorithm for deep belief network based on sliding window feature fusion [J]. Application Research of Computers, 2018, 35 (4): 1096-1100.

[4] Yang Qiufen, Gui Weihua, Zhou Shuren. Facial expression recognition algorithm for fatigue driving [J]. Application Research of Computers, 2008, 25 (10): 3039-3041.

[5] You Chuangwen, Lane N D, Chen Fanglin, et al. CarSafe app: alerting drowsy and distracted drivers using dual cameras on smartphones [C]// Proc of International Conference on Mobile Systems, Applications, and Services. New York: ACM Press, 2013: 13-26.

[6] Mbouna R O, Kong S G, Chun M G. Visual analysis of eye state and head pose for driver alertness monitoring [J]. IEEE Trans on Intelligent Transportation Systems, 2013, 14 (3): 1462-1469.

[7] Mccall J C, Trivedi M M, Wipf D P, et al. Lane change intent analysis using robust operators and sparse bayesian learning [J]. IEEE Trans on Intelligent Transportation Systems, 2007, 8 (3): 431-440.

[8] Nakamura T, Maejima A, Morishima S. Detection of driver' s drowsy facial expression [C]// Pattern Recognition. 2014: 749-753.

[9] Akrout B, Mahdi W. Spatio temporal features for the automatic control of driver drowsiness state and lack of concentration [J]. Machine Vision & Applications, 2015, 26 (1): 1-13.

[10] Hjortsjö C H. Man's face and mimic language. Malmo: Nordens Boktryckeri, 1970.

[11] Ekman P, Friesen W V. Facial action coding system (FACS): a technique for the measurement of facial actions [J]. Rivista Di Psichiatria, 1978, 47 (2): 333-341.

[12] Ekman P. CiNii articles [J]. Facial Action Coding System the Manual on Cd Rom, 2002.

[13] Hamm Jihun, Kohler C G, Gur R C, et al. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders [J]. Journal of Neuroscience Methods, 2011, 200 (2): 237-56.

[14] Baltrušaitis T, Mahmoud M, Robinson P. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection [C]// Proc of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. 2015: 1-6.

[15] Cootes T. An Introduction to Active Shape Models [J]. Reviews Computational Statistics, 2010, 2 (4): 503–508.

[16] Felzenszwalb P, Mcallester D, Ramanan D. A discriminatively trained, multiscale, deformable part model [J]. Cvpr, 2008, 8: 1-8.

[17] Abtahi S, Omidyeganeh M, Shirmohammadi S, et al. YawDD: a yawning detection dataset [C]// ACM Multimedia Systems. New York: ACM, 2014.

[18] Vural E, Cetin M, Ercil A, et al. Drowsy driver detection through facial movement analysis [C]// Proc of IEEE International Conference on Human-Computer Interaction. Springer-Verlag, 2007: 6-18.

[19] Shih T H, Hsu C T. MSTN: multistage spatial-temporal network for driver drowsiness detection [C]// Proc of Asian Conference on Computer Vision. Cham: Springer, 2016: 146-153.

[20] Wang Lin, Zhang Chen, Yin Xiaowei, et al. A non-contact driving fatigue detection technique based on driver' s physiological signals [J]. Automotive Engineering, 2018 (3): 333-341.

Note: Figure translations are in progress. See original paper for figures.