Development of Deep-Learning-based Driver Monitoring Algorithms

Ruide Li[1†] and Hiromatsu Aoki[2]

[1]Department of HMI Sensing, SenseTime Japan, Kyoto, Japan
(E-mail: liruide@sensetime.jp)
[2]Department of HMI Sensing, SenseTime Japan, Kyoto, Japan
(E-mail: aoki@sensetime.jp)

Abstract: Traffic accidents cause millions of deaths worldwide each year, it is a huge challenge to improve the traffic environment. Recently, more and more automotive manufacturers are considering deploy Driver Monitoring Systems (DMS) on their future products. In this paper, we first describe what functions are required for a DMS, then explain what kind of efforts have SenseTime made in DMS, and finally discuss the further possibility for DMS in the near future.

Keywords: Driver Monitoring System (DMS), Deep Learning, Convolutional Neural Network (CNN), Computer Vision (CV)

1. INTRODUCTION

According to the report from Tokyo Metropolitan Police Department[1], over two-thirds of the causes of traffic fatal accidents currently occurring in Japan are "violations of safe driving obligations." Among them, almost 90% of the accidents are caused by the carelessness of the driver, such as "inattentive driving", "aimless driving", "inadequate operation", and "insufficient safety confirmation". Therefore, as an approach to improve traffic safety, many Driver Monitoring Systems (DMS) are developed, so that the system can warn the driver to focus on driving.

Meanwhile, as the rapid development of auto-driving research, vehicles with auto-driving-like system are increasingly manufactured. However, in SAE (J3016) Automation Levels, L1 and L2 automation require human driver performs the main aspects of the dynamic driving task; even L3 requires human driver to respond appropriately to a request to intervene. In order to avoid abuses of L1 $\sim$ L3 automation systems, DMS will play an important role before L4 automation technology getting mature.

By installing DMS in to a vehicle, we introduce an in-vehicle camera. As a result, besides key features of DMS, people may also desire other computer vision functions such as identification with face recognition to unlock the car, more interactive infotainment applications or a adjustable HUD (Head-Up Display) which can fit the drivers position.

For the structure of this paper, we first describe what functions are required for a DMS, then explain what kind of efforts has SenseTime made in DMS, and finally discuss the further possibility for DMS in the near future.

2. FUNCTIONS IN DMS

Usually, DMS is understood as the whole system which may consist of camera, SoC, operating system, software, etc., while in this paper, we only focus on the image recognition process, where the input is a video stream and the output is the status of the driver. Camera specification and how to make use of the driver status (warning lights, warning sound, taking control of the vehicle, etc.) is out of the scope for this paper.

The two main functions required in DMS are detection of drowsiness and distraction of the driver, which may cause fatal accidents. In drowsiness detection, a general approach is to detect whether the driver is feeling drowsy or dozing off. As for distraction detection, the system should tell whether the driver is focusing on driving by detecting where are the driver looking at.

Although driving safety is one of the main motivations for DMS, there are still many other in-vehicle scenarios, in which people can make benefit of computer vision algorithms. Other functions in DMS may include face recognition (identity verification), face expression recognition, action detection (such as smoking, drinking and calling on the phone), hand gesture detection or body posture detection. These functions may be not directly related to driving safety, but they may improve in-vehicle user experience a lot. Here are some examples. With face recognition technology, a hands-free unlock function can be applied; with hand gestures detection technology, the driver will be able to control audio volume or air conditioner by hand gestures, without looking at the control panel; with gaze direction estimation technology, HUD can adjust its position by tracking gaze direction of the driver.

Besides functional requirements, robustness is vital for DMS. Driving scenes can be unexpected ranged, from pitch dark scenes during the night to dazzling backlighting scenes at noon, and algorithms should be able handle all these high dynamic ranged scenes. Moreover, running computationally expensive deep learning algorithms at a real-time rate on relatively less powerful automotive SoCs is also a challenge.

3. SENSETIME DMS

In this section, we will discuss the mechanism of SenseTime DMS. The processing pipeline is shown as in Figure 1. We combine deep learning based models and

[†] Ruide Li is the presenter of this paper.

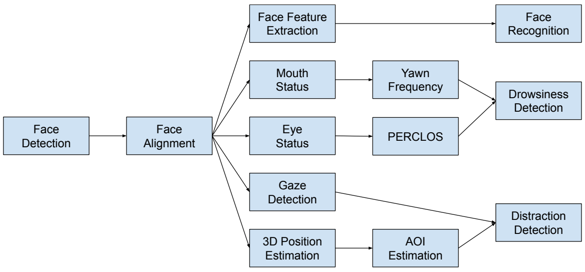

Fig. 1 SenseTime DMS Pipeline

logic based process. Face detection, face alignment, feature extraction, eye/mouth status recognition, gaze detection and 3D head position estimation modules are based on deep learning models, while yawn frequency, PERCLOS (Percentage of Eye Closure, [2]), AOI (Area of Interest) estimation, face recognition, drowsiness detection and distraction detection are based on rules and calculation.

For one frame of input video stream, we first apply face detection to locate faces in the frame, then face alignment will provide detailed information of the face. Detailed face information with cropped face image will be fed into further CNN models, which provide face feature, gaze vector, eye/mouth status, and 3D head position for further inference.

We fine-tune our algorithms on IR image dataset with ranged scenarios, so that they are robust enough to handle the high dynamic input range. Also, we specially designed algorithms in order to provide high performance on low-spec SoCs.

3.1. Face Detection/Alignment

SenseTime owns a huge annotated dataset for face detection/alignment training, also we fine-tune the model on dataset taken by IR cameras. However, besides accuracy, speed performance is a challenge for DMS application, since automotive SoCs are usually not that powerful comparing to CPUs and GPUs for PC. Face detection model is able to provide accurate results of face location, and face alignment model provide detailed face information like facial landmarks given the bounding box of faces. Yet the problem is that face detection required much longer time due to large computation, as a result, it is nearly impossible to run face detection model on automotive SoCs with a real-time frame rate. We have designed a asynchronized mechanism to combine face detection and alignment, in which face detection runs sparsely to update face location, while between the interval of face detection, face alignment is applied to track the face location.

3.2. Drowsiness Detection

After face alignment, we feed detailed face information to mouth/eye model in order to obtain the open/close status. Then we use a customizable time window to calculate PERCLOS and yawn. Finally, based on duration of eye closure and yawn frequency with a customizable threshold, a drowsiness level will be returned. In evaluation, our algorithm is able to reach over 98% TAR@FAR=0.1% (true alarm rate at false alarm rate less than 0.1%).

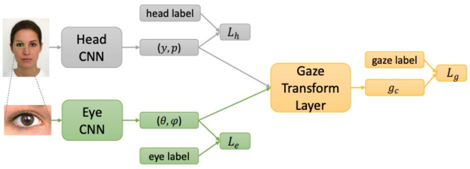

Fig. 2 Full Face Gaze Estimation Model

3.3. Full-Face-Based Gaze Detection

In gaze detection, the purpose is to estimate gaze vector(s) in camera coordinate system, which represents the direction the subject is looking in. Traditional gaze estimation works (such as [3]) use eye landmarks to calculate gaze direction in head coordinate system, and then combine head pose (head angle) information to get the gaze angle in camera coordinate system. We also used this kind of algorithm on our early DMS product.

However, our researchers have found out that simply combining gaze direction and head pose somehow cannot accurately represent the geometric relationship between them. In SenseTime's work [4], instead of only using eye information, we also feed the full face into the CNN model as a branch (as in Figure 2), so that the model can learn the geometric relationship. Other works using full face approach such as [5] use attention mechanism to constrain weights which do not relate to gaze estimation. Our recent DMS product is deployed with similar full face approach in order to provide more accurate gaze estimation.

After we have gaze vector, we can directly use it for distraction judgement by specifying angle range threshold (max/min pitch, yaw), or first define 3D regions as AOIs, such as front windshield, side mirrors, NAVI display, then see whether the driver's gaze vector intersect any AOI to inference where the driver is looking at.

3.4. 3D Position Estimation with Monocular Camera

Many 3D head/eye position estimation solution are using stereo camera or Kinect (from Microsoft). Extra devices not only raise the cost, but also required much more space in the vehicle. In our system, it does not require more devices than the IR camera which is the same one as is used in face detection. 3D position estimation can be used with HUD. Combining with gaze direction, head/eye position information will be provided so that the HUD can adjust its size and showing position for better user experience.

3.5. Face Recognition

In face recognition, we use deep CNN model to extract face feature (hign-dimentional vector) to represent each face, then compare the similarity of different faces to judge whether they are the same person. Similar to face detection/alignment, we also fine-tune the feature extrac-

5 seconds	95.67%
10 seconds	99.07%
30 seconds	99.67%

Table 1 Recognition accuracy in N seconds after the driver closing the door (on 10000 faces database)

tion model on IR camera data. Our analysis of face recognition accuracy on real-car scenarios is shown in Table 1. We evaluated the recognition accuracy on 10000 distractor faces after the driver get into the vehicle and close the door. Our algorithm can reach 95.67% accuracy within 5 seconds, or 99.67% accuracy within 30 seconds.

3.6. Other Functions

There are several other functions in our DMS product, such as face expression recognition, face mask/glasses/sunglasses detection, hand gestures detection and body posture detection. Although some of these functions may be not directly related driver monitoring, we provide them as an option of computer vision solution.

4. FURTHER ISSUES

Needs of DMS (or more generalized, in-vehicle computer vision application) is still far from resolved. Here we give two examples of open issues.

4.1. Combination with ADAS

Working with ADAS (Advanced Driver-Assistance Systems) with front outside camera(s), it is also possible to combine gaze estimation and road information. For instance, when ADAS detect pedestrian crossing the road, DMS is able to tell whether the driver is paying sufficient attention to the pedestrian and will send alarm if not.

Such application requires high accuracy of gaze estimation, however, in general (person independent) gaze estimation, even state-of-the-art algorithms can only achieve an error of $4^\circ \sim 5^\circ$, which is far from enough for out-of-vehicle scenarios. A possible reason is that there is a person dependent difference in structure inside the eyeball [6]. For distraction detection, this may be a sufficient accuracy, but for HUD application, much more accurate gaze estimation is required to precisely localize the point where the driver is looking at. To reach more accuracy, usually personal calibration is required. Applying personal gaze calibration in DMS remains to be a further issue.

4.2. Passenger Monitoring

Our DMS product so far is designed mainly for driver monitoring, while there are actual needs for passenger monitoring in the market. For example, whether the driver is leaving their children or pets in the car without air conditioner running.

Yet the problem is that for back seat passenger, there are only few options of camera position and even though face detection coverage is still limited due to occlusion by seat-backs. Since additional camera for back seat passenger will highly expand the complexity and narrow the flexibility of the whole system, a mature solution for back seat passenger monitoring without increasing cameras remains to be a challenge.

REFERENCES

[1] Japanese Government Statistics Website, https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00130002&tstat=000001027458&cycle=7&year=20180&month=0

[2] U. Trutschel, et al., "PERCLOS: An Alertness Measure of the Past", *Driving Assessment Conference*, 2011

[3] X. Zhang, et al., "Appearance-based gaze estimation in the wild", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015

[4] H. Deng, et al., "Monocular Free-head 3D Gaze Tracking with Deep Learning and Geometry Constraints", *The IEEE International Conference on Computer Vision (ICCV)*, 2017

[5] X. Zhang, et al., "It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation", *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016

[6] K. A. Funes Mora, et al., "Geometric Generative Gaze Estimation (G3E) for Remote RGB-D Cameras", *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014