Hindawi
Computational Intelligence and Neuroscience
Volume 2020, Article ID 9606908, 10 pages
https://doi.org/10.1155/2020/9606908

Research Article

Driver Distraction Detection Method Based on Continuous Head Pose Estimation

Zuopeng Zhao,[1,2] Sili Xia,[1,2] Xinzheng Xu,[1,2] Lan Zhang,[1,2] Hualin Yan,[1,2] Yi Xu,[1,2] and Zhongxin Zhang[1,2]

[1]*School of Computer Science and Technology & Mine Digitization Engineering Research Center of Ministry of Education of the People's Republic of China, China University of Mining and Technology, Xuzhou 221116, China*
[2]*School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China*

Correspondence should be addressed to Sili Xia; xiadeemail@163.com

Received 16 January 2020; Revised 12 October 2020; Accepted 18 November 2020; Published 29 November 2020

Academic Editor: Fabio Solari

In view of the fact that the detection of driver's distraction is a burning issue, this study chooses the driver's head pose as the evaluation parameter for driving distraction and proposes a driver distraction method based on the head pose. The effects of single regression and classification combined with regression are compared in terms of accuracy, and four kinds of classical networks are improved and trained using 300W-LP and AFLW datasets. The HPE_Resnet50 with the best accuracy is selected as the head pose estimator and applied to the ten-category distracted driving dataset SF3D to obtain 20,000 sets of head pose data. The differences between classes are discussed qualitatively and quantitatively. The analysis of variance shows that there is a statistically significant difference in head posture between safe driving and all kinds of distracted driving at 95% and 90% confidence levels, and the postures of all kinds of driving movements are distributed in a specific Euler angle range, which provides a characteristic basis for the design of subsequent recognition methods. In addition, according to the continuity of human movement, this paper also selects 90 drivers' videos to analyze the difference in head pose between safe driving and distracted driving frame by frame. By calculating the spatial distance and sample statistics, the results provide the reference point, spatial range, and threshold of safe driving under this driving condition. Experimental results show that the average error of HPE_Resnet50 in AFLW2000 is 6.17° and that there is an average difference of 12.4° to 54.9° in the Euler angle between safe driving and nine kinds of distracted driving on SF3D.

1. Introduction

The World Report on Road Traffic Injury Prevention points out that many factors have an impact on traffic safety, such as the mental state of drivers, the degree of fatigue, and whether the driver is drunk or distracted. According to Volvo, over 80% of road accidents are caused by distracted drivers. Compared with other dangerous driving modes, distracted driving is transient and frequent [1]. When drivers are distracted, for example, when adjusting onboard equipment, using the mobile phone, or involuntarily bowing the head due to fatigue, their head pose changes in varying degrees compared with the normal situation, Therefore, the analysis and comparison of the driver's head pose can provide a basis for judging whether the driver is in a distracted driving state.

Distraction detection is one of the application directions of head pose estimation [2]. In the field of computer vision, by inputting the head image containing the target user into the computer and combining it with image processing technology, the pose parameters of the head in space are determined based on calculation and prediction. There are two ways of expressing this pose parameter: face orientation and Euler rotation angles [3]. Compared with the expression based on face orientation, the Euler rotation angle is more accurate and comprehensive. As shown in Figure 1, the Euler rotation angle refers to a group of angular parameters in three-dimensional space: yaw, pitch, and roll. In this paper,

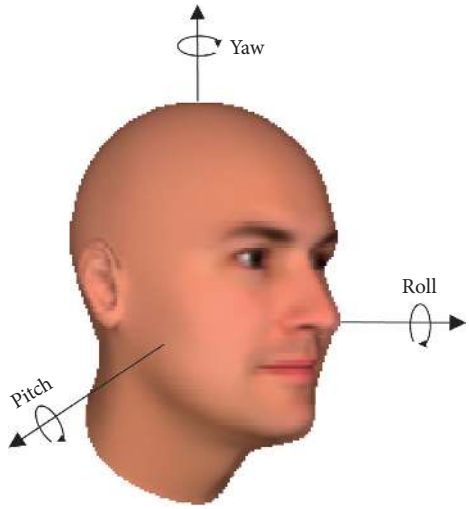


Figure 1: Example of expression of Euler rotation angle.

the Euler rotation angle is chosen as the expression of the head pose. Figure 2 shows the differences between the Euler rotation angle and face orientation expression.

Distraction detection needs to be judged according to the location of the driver's gaze. At present, there are some methods to deduce the driver's cognitive state by considering the characteristics of eye movement (e.g., gaze, saccade, and smooth tracking), such as driver distraction [4]. Other methods apply classification to eye images related to different gaze regions to detect the location of the driver's gaze while driving [5]. There are also some methods that can track facial features and 3D head posture and predict the direction of gaze relative to the position of the car to detect whether the driver's eyes are fixed on the road [6]. Other methods study the driver's gaze behavior (e.g., scan frequency and scan time) to evaluate the driver's driving performance when interacting with other devices, such as portable navigation systems while driving [7]. Finally, there are some studies on the functional relationship between the driver's head posture and gaze behavior, in order to predict the fixation position according to the driver's head position and direction [8], so as to classify different types of driver behavior while driving [9]. Considering the stringent requirements of the gaze point estimation for the data set and after studying the relationship between the driver's head posture and gaze behavior, it is found that using head posture estimation to detect the driver's distraction behavior is feasible and accurate.

In previous work on driver distraction detection, Hoang et al. [10] used faster R-convolutional neural network (CNN) to establish whether the driver's hands are on the steering wheel and whether the driver has a mobile phone as a basis for distraction detection. Robinson et al. [11] designed a distraction detection system that uses the direction of movement of the driver's eyes, mouth, and head in the vehicle image as recognition parameters in order to judge the driver's distraction. In a limited environment, the global detection accuracy is as high as 99%, while, in the practical application, the accuracy is 86%, and the average response time is 30 ms. In 2016, State Farm held a distracted driving identification competition in ten categories on Kaggle. Toshi-k uses the detector to detect the driver's body outline, cuts the picture according to the detected body outline, and then uses the depth convolution network to identify it, which gives good results. Taking advantage of the good performance of the convolutional neural network (CNN) in the field of computer vision, Baheti et al. [12] proposed a detection system based on CNN that not only identifies whether the driver is distracted or not but also identifies the types of distraction, modifies the structure of VGG16, and uses various regularization techniques to improve recognition accuracy.

The existing research on the identification of driver distraction can achieve remarkable results in specific environments and datasets. These data come either from a public driver distraction dataset or from a laboratory driving simulator, but there are few studies that analyze the distraction state based on actual driving images. For example, in State Farm Distracted Driver Detection (SF3D), the number of participants in the training set is 26, the driver image is not obscured, the light conditions in the experiment are good, and the change of light and shade is not obvious. There are some differences between this and the actual driving conditions. Such are the bumps in the driving process that easily cause the deviation of the imaging angle. In addition to other factors, the weather and sun exposure angle easily cause greater light and dark changes, as well as the large base of drivers, less participants maybe not able to accurately fit, and other related differences. In addition, a single picture can only capture an instantaneous state, while continuous human action is difficult to judge. It cannot be established with certainty whether the driver is distracted or not from a single picture. Therefore, in view of the related issues presented above, this paper makes the following contribution to the field:

(1) It proposes a head pose estimation method based on deep learning without preprocessing input images. The paper compares the results between the improved Alexnet and three kinds of Resnet based on this method.

(2) On the common distraction dataset SF3D, the head pose difference between safe and distracted driving is analyzed by (1) as the basis for distraction detection.

(3) Several actual driving videos are analyzed frame by frame on the basis of (2) to obtain the safe driving head pose range and threshold under the corresponding imaging angle.

2. Methods

2.1. Head Pose Estimation

2.1.1. Deep Learning Method. Traditional head pose estimation methods can be divided into shape-based [13, 14] and face-based key point set relationships [15–17]. They both have some shortcomings: the method based on shape template has a large error and can only render limited

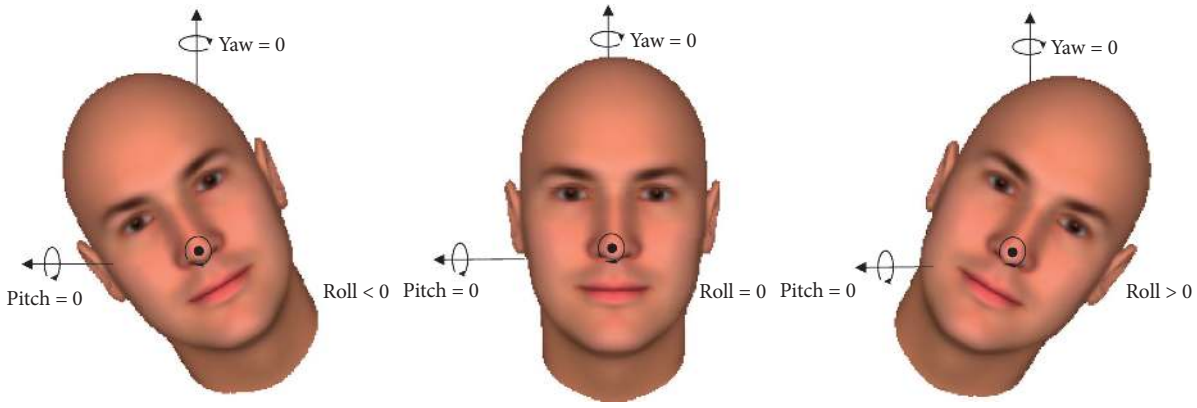


Figure 2: Frontal face image with face orientation (1, 0, 0).

discrete attitude values. On the other hand, in the method based on facial key points, the marking and positioning accuracy of facial key points directly affects the results. Compared with other methods, the deep learning method [12, 18] has a strong generalization ability and can fit different face poses so as to ensure higher accuracy.

Head pose estimation is essentially a regression problem [2], that is, mapping high-dimensional image space to low-dimensional head pose space. The solution can be divided into two parts: dimensionality reduction and regression. The mapping from image to pose space can be effectively completed by a CNN. Different from the fully connected network, the core of the CNN is the convolution layer that completes image feature extraction through convolution operation [19]. Owing to the local connection and weight sharing, the number of parameters in the network is significantly lower, which makes the network model easier to train and reduces overfitting. At the same time, the downsampling of the pool layer further reduces the number of parameters, so the generalization of the model is improved.

2.1.2. Network Structure. Based on deep learning, the head pose estimator (HPE_Resnet) used in this paper is improved on the classical residual network. The core idea of Resnet [20] is to no longer learn a complete output but only learn the difference between output and input. This idea realizes the jump transmission of information, maintains the completion of feature information, and solves the problem that the accuracy of traditional CNN decreases with the increase of depth. In addition, its time cost does not increase significantly with the deepening of Resnet layers.

HPE_Resnet does not do any preprocessing of images at input and gives them directly to the model for training. This is different from some network models which need to preprocess pictures, such as cropping, graying, equalization, and normalization, while the use of native images is closer to the actual use. HPE_Resnet retains the convolution layer and pooling layer in Resnet, uses the first half of the convolution layer as the feature extractor, uses 3-3 convolution in the two-layer residual block, and uses one convolution and one convolution in the three-layer residual block to realize the continuous operation of dimensionality reduction, convolution, and dimension enhancement. Figure 3 shows two-layer residual block used on HPE_Resnet18 and 34 and three-layer residual block used on HPE_Resnet50.

Nowadays, head pose estimation based on deep learning is mainly done in the following three ways:

(i) Based on the routine, the head pose estimation is regarded as a typical regression problem, while network parameters are continuously optimized by the loss function to approach the label value.

(ii) The range of head pose parameters is divided into several equal intervals, and the regression problem is transformed into a classification problem [21].

(iii) Combined with the idea of classification and regression, the range of parameters is classified, and regression prediction is then carried out.

It should be pointed out that (ii) is only suitable for some specific head positions, such as head up, head down, and left and right rotation. In essence, the continuous head pose is mapped into a discrete specific interval, although it can overcome some interference, but the actual division range is relatively large (about 15°) and cannot accurately express the head pose and meet the needs of detection. In this paper, (i) and (iii) are verified on the same network at the same time. For (i), only regression training is carried out, the loss of the whole Euler angle is calculated, and a single mean square error is used as the loss function:

$$\text{loss}_1 : \ \text{MSE}(y, y') = \frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n}. \tag{1}$$

For (iii), yaw, pitch, and roll are trained separately. The range of each angle parameter is divided into several categories, and regression training is then carried out. The cross-entropy loss function is used in the classification prediction stage, and the mean square error loss function is used in the regression phase:

$$\text{loss}_2 : \ H(p, q) + \text{MSE}(y, y') = -\sum_{x} p(x) \log q(x) + \frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n}. \tag{2}$$

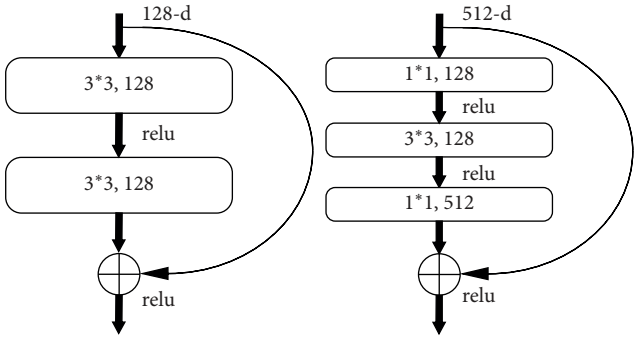


Figure 3: Two-layer residual block and three-layer residual block.

After conv1, Max pool, conv2, conv3, conv4, conv5, and Average pool processing, the image is transmitted to the four branches of the full connection layer (FC), respectively. In the FC of HPE_Resnet18 and HPE_Resnet34, the original number of output channels is unchanged, that is, it is 512. Due to the use of three-layer residual blocks in HPE_Resnet50, the output channel needs to be expanded, that is, it needs to be 512*4. The Softmax activation function is used in FC_Yaw, FC_Pitch, and FC_Roll. The sum of classification and regression errors of each Euler angle is calculated separately. In FC_All, only the overall regression error is calculated. The detailed architecture is shown in Figure 4.

2.2. Distraction Detection

2.2.1. Head Pose in SF3D. During driving, head pose is one of the most abundant parameters. On SF3D, the head poses of nine types of distracted driving are intuitively different from safe driving. In order to quantify specific differences and reduce the impact of individual extreme values, HPE_Resnet50 was used to test SF3D with a total of 20,000 images. The results show that distracted driving differs from safe driving in the head pose.

In Table 1, the mean, standard deviation, and quartiles (Q3–Q1) of the Euler angular components are calculated separately for each category, where the standard deviation reflects the degree of dispersion of the sample and the quartiles (Q3–Q1) mainly reflect the spatially approximate distribution of $c0$ to $c9$ over the angular components. The results show that the yaw component as a whole has a higher volatility and distribution range than the pitch and roll components, that is, more active offset activity in the yaw direction in real driving.

The SF3D variability at the 95% and 90% confidence levels was then verified, respectively, and the results showed that safe driving in the head position was significantly different from various types of distracted driving.

Combining Tables 1 and 2 reveals that the angular components corresponding to each type of driving action are distributed over a specific interval and that there are varying degrees of spatial distance differences between the mean points. Starting with these two characteristics can provide a statistical basis for the design of distracted driving identification methods.

Table 1 only shows that there is a certain difference in head pose between safe and distracted driving that cannot be directly used as a basis for discrimination. Due to the continuity of distraction, such as adjusting vehicle equipment, using a mobile phone, and turning head for a long time, duration is uncertain, and continuous information cannot be expressed in a single frame. For this reason, it may be difficult to define whether driving is distracted or not, which results in a high rate of misjudgment. Therefore, in order to obtain the parameter range and driving characteristics of safe driving, the driving behavior should be analyzed over a period of time, and the driver's head pose parameters should be read frame by frame.

2.2.2. Distraction Detection Based on Video Frame. It should be pointed out that head pose is closely related to the imaging angle, and it is only meaningful to discuss head pose at the same imaging angle. As shown in Figure 2, the front face of the camera is yaw = 0, pitch = 0, roll = 0. If the imaging angle is changed at this point, the head does not rotate relative to the human body, and the result will still change. Therefore, when the range of head pose parameters and threshold of safe driving are given, the corresponding imaging angle should be indicated. The specific steps for solving the safety range and threshold issue are as follows:

(1) Calculate the Euler angle of a single person with a total of n frames of a safe driving continuous video frame by HPE_Resnet50, which is recorded as $(\text{yaw}_i, \text{pitch}_i, \text{roll}_i)$.

  (i) Calculate the average value of n Euler angles in this group $(\overline{\text{yaw}}, t\overline{\text{pitch}}n, q\,h\overline{\text{roll}})_i$, and take it as the base point $b_j\,(j = 0, \ldots, m)$.

(2) Repeat Step (1) m times to get a total of m person times safe driving base point set $B(b_1, b_2, \ldots, b_m)$.

  (i) Calculate the set B mean value, which is recorded as the safe driving base point $\beta(\overline{\text{yaw}}, \overline{\text{pitch}}, \overline{\text{roll}})$.
  (ii) Based on the statistical analysis of m*n points, extreme points are removed and expanded into specific space Ω, that is, the safe driving range.

(3) Calculate the spatial distance $d_i\,(i = 1, 2, \ldots, n)$ of m group n points from β, as shown in equation (3), and remove the partial maximum to make the distance between most points at least 90% and $\beta \le d$, where d is the recommended safety threshold:

$$d_i = \sqrt{\left(\overline{\text{Yaw}} - \text{yaw}_i\right)^2 + \left(\overline{\text{Pitch}} - \text{pitch}_i\right)^2 + \left(\overline{\text{Roll}} - \text{roll}_i\right)^2}. \tag{3}$$

(4) Test k continuous video frames, compare with space Ω, and calculate the space distance $d_i\,(i = 1, 2, \ldots, k)$ between each frame and β. In equations (4), (5), and (6), OT represents the number of all k values that exceed the threshold, MCOT represents the maximum number of k values that continuously exceed the threshold, which can reflect the driver's

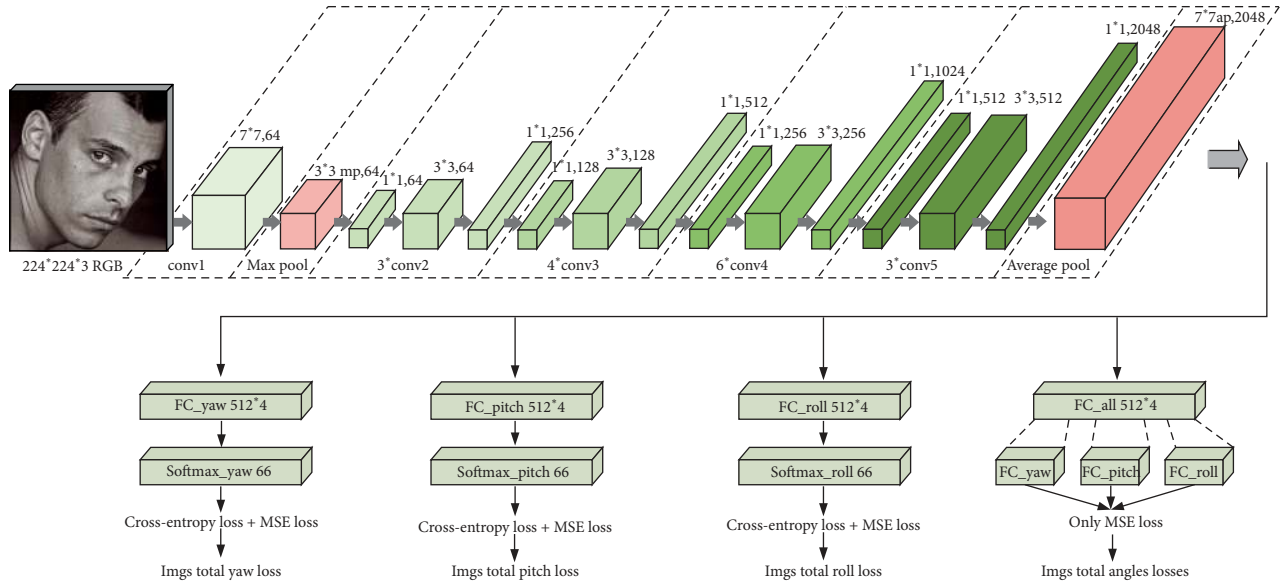


Figure 4: Architecture of HPE_Resnet50: 3*conv1 represents the three overlays of the network layer and (7*7, 64) indicates the convolution core size of 7*7, with a total of 64 layers.

Table 1: Head pose related data on SF3D through HPE_Resnet50.

SF3D	Yaw			Pitch			Roll		
	$\overline{x}$	σ	Q3–Q1	$\overline{x}$	σ	Q3–Q1	$\overline{x}$	σ	Q3–Q1
$c0$	−59.3	12.7	16.7	10.1	8.3	11.8	−21.0	6.4	8.6
$c1$	−48.0	12.6	15.0	−4.6	9.2	14.1	−26.1	7.9	9.6
$c2$	−37.7	12.7	14.1	−10.5	5.5	7.1	−16.4	10.0	16.6
$c3$	−71.6	14.3	15.9	1.3	10.1	15.5	−15.9	5.8	7.1
$c4$	−68.0	12.2	16.0	2.6	7.3	9.4	−16.2	6.1	8.1
$c5$	−40.2	6.5	7.5	−4.5	8.4	12.2	−30.0	5.9	8.0
$c6$	−43.2	13.7	16.0	−3.5	8.4	11.1	−28.6	9.5	12.1
$c7$	−13.8	21.1	28.3	−19.4	11.0	15.9	−29.3	16.2	28.9
$c8$	−38.1	17.7	21.6	−8.4	7.1	8.6	−16.5	12.0	19.9
$c9$	−16.8	9.3	14.1	−16.1	5.9	8.8	−30.6	6.2	7.6

Table 2: Analysis of head pose differences.

	Source of variation	SS	Df	F value	P value	Statistical significance	
						$\alpha = 0.05$	$\alpha = 0.1$
Yaw	SSA	$6.67*10^6$	9	3,882.04	0.00	$P < 0.01$	$P < 0.01$
	SSE	$3.81*10^6$	19,990	—	—		
	SST	$1.05*10^7$	19,999	—	—		
Pitch	SSA	$1.40*10^6$	9	2,271.66	0.00	$P < 0.01$	$P < 0.01$
	SSE	$1.37*10^6$	19,990	—	—		
	SST	$2.77*10^6$	19,999	—	—		
Roll	SSA	$7.49*10^5$	9	984.96	0.00	$P < 0.01$	$P < 0.01$
	SSE	$1.69*10^6$	19,990	—	—		
	SST	$2.44*10^6$	19,999	—	—		

distraction over a period of time, and N N indicates the number of thresholds exceeded:

$$\text{all} = \frac{\text{OT}}{k} * 100\%, \tag{4}$$

$$\text{cont} = \frac{\text{MCOT}}{k} * 100\%, \tag{5}$$

$$\text{count} = N. \tag{6}$$

3. Experiment and Result

3.1. Experiment Datasets. The datasets used in this study are 300W-LP [22], AFLW [23], AFLW2000, SF3D, and a

collection of actual driving images (Driver_Imgs). 300W-LP and AFLW are used as training sets and AFLW2000 as a test set. SF3D is used to verify the safety and head pose differences in all kinds of instances of distracted driving, while Driver_Imgs is used to test the effect of actual driving.

Statistics show that the range of head motion of adult males is limited. Under the condition of frontal imaging, specific angle ranges are yaw [−79.8°, 75.3°], pitch [−60.4°, 69.6°], and roll [−40.9°, 36.3°]. Combined with the range of head pose in the dataset and the actual driving situation, the Euler angles of images are limited to [−99°, 99°], as shown in Table 3.

Figure 5 shows some images from the experiment datasets. 300W-LP is a large comprehensive head pose dataset different from the previous medium and small range (±45°) datasets. By expanding the side image of the human face, it can solve the problem of landmark occlusion in a large range and provide high-precision head pose data. There is a situation when a picture contains multiple faces in AFLW. In order to meet the objective of the experiment, which is only to judge the distraction state of a single person, this kind of image is removed from the dataset, leaving only a single face in the range of [−99°, 99°], excluding the pictures in AFLW2000. Finally, a total of 16,825 pictures were used in AFLW. At the same time, the images not in the range of [−99°, 99°] in AFLW2000 were removed, and 1,968 test images were used.

SF3D based on a distracted driving identification competition on Kaggle in 2016 was divided into ten categories: drive safe, text left, talk left, text right, talk right, adjust radio, drink, reach behind, do hair and makeup, and talk to passenger. This was a total of 22,424 training pictures. These training pictures were taken by 26 participants in the experimental environment. In the experiment, only 20,000 pictures classified by Kaggle, 2,000 in each category, were selected.

Driver_Imgs comes from the "BeiDou +" vehicle video surveillance platform in Jiangsu Province. It is composed of continuous video frames, and the data are all from an actual driving video. A total of 90 drivers' driving video clips were selected. The imaging angle was fixed on the right side of the face and the face was clear and unobscured. As the driving time covers the whole day, and there are interference factors such as light and weather, the light and shade of the picture obviously change.

3.2. Results and Discussion

3.2.1. Head Pose Estimation Result. Experiment 1 aimed to train a reliable head pose estimator. In order to improve the generalization ability of the model and have a good recognition effect on the general images, 300W-LP and AFLW were used for training. First, 300W-LP was used for 10 epochs of training, and then AFLW was used for 15 epochs of training. The learning rate was 10^{-5}, and the batch size was 16. Finally, the training effect was verified in AFLW2000, and the difference between single regression and combined classification and regression thoughts was obtained.

Table 4 shows the results of each model on the verification set. After many tests, the average error is stable at 6.17°. In this training, with the deepening of the number of network layers, the error becomes smaller and smaller, and the training idea of combining classification and regression is obviously better than single regression. As far as a single Euler angle is concerned, the training of yaw under classification and regression is more successful. On the contrary, the prediction error of yaw under only regression is larger than that of pitch and roll.

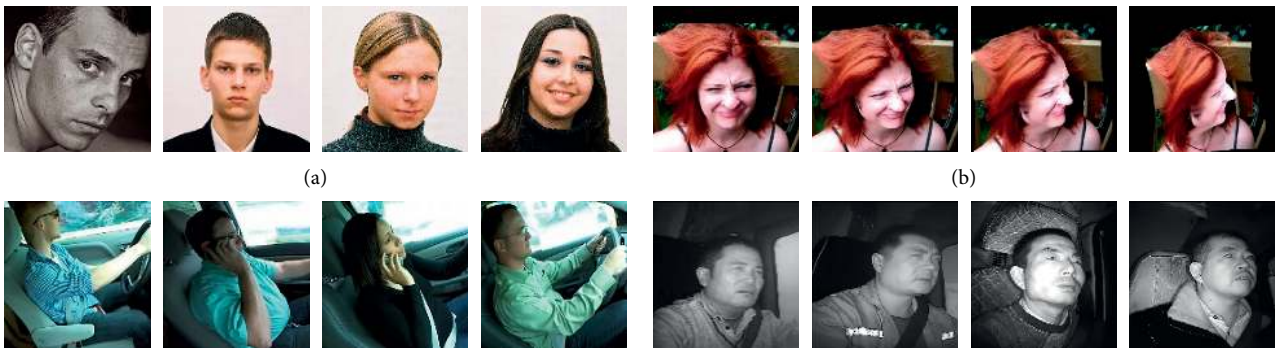
3.2.2. Distraction Detection Result. Experiment 2 was designed to verify the application of head pose estimation in the actual driving image, as shown in Figure 6. The range of safe driving head pose at this imaging angle was obtained by calculating a total of 8,000 safe driving pictures of 80 drivers, removing extremes at both ends, and selecting the range containing most of the data points (>90%), as shown in Figure 7. In Table 4, $\beta(\overline{\text{yaw}}, \overline{\text{pitch}}, \overline{\text{roll}})$ is the average point of 8,000 data sets, and D is the safe driving threshold. In Table 5, the spatial distance between each point and β was calculated, and the distance between 90% of the points and β was less than 18.6, so 18.6 was set as the safe driving threshold.

A total of 3,000 video frames, including two categories of safe driving and distracted driving, were collected from 30 drivers, broken down into 3 categories of safe driving actions and 5 categories of distracted driving actions, as shown in Table 6 and Figure 8. After obtaining the respective head posture data through the depth learning model, the corresponding distraction parameters ALL, MAX, and COUNT were calculated and the spatial distances from the base point β were plotted chronologically as fold lines. In order to ensure the stability of the recognition, the experiment stipulates that only if the threshold is exceeded continuously by 5 frames and above, the COUNT will be plus 1, that is, 0.25s continuous overthreshold at the current FPS, to prevent the phenomenon of only individual video frames exceeding the threshold due to vehicle bumps or posture calculation errors and the driver is judged to be distracted, resulting in excessive sensitivity of the method.

Table 7 records the distraction parameters for $t1$–$t15$ videos $t1$–$t5$ are the safe driving videos, and all of the essentially undeflected head positions are positive observation positions, which are the positions that drivers hold for the longest time in daily driving. From Table 7, it can be seen that all three distraction parameters did not reach the threshold of distraction recognition theory and had small values. To verify the method's ability to recognize other driving actions in safe driving, distraction judgments of five drivers in the $t6$–$t10$ were added. Unlike the $t1$–$t5$, the driver in the $t6$–$t10$ performed several observations of the mirrors on both sides. Looking in the rearview mirror is a normal operation during driving, more common but with greater head deflection. Distraction parameters were all increased compared to $t1$–$t5$, but neither reached the threshold for distraction recognition. Where COUNT basically reflects the number of times the rearview mirror is viewed, MAX reflects

Table 3: Experiment datasets.

Datasets	300W-LP	AFLW	AFLW2000	SF3D	Driver_Imgs	
					Safe_Imgs	Test_Imgs
Number	61,225	16,825	1,968	20,000	80*100	15*100
Angle range	±90°	±99°	±99°	—	—	—



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

Figure 5: (a), (b), (c), and (d) come from datasets AFLW, 300W-LP, SF3D, and Driver_Imgs, respectively.

Table 4: Verification results for all deep learning models on AFLW2000 with 1,968 images.

	Classification and regression				Only regression			
Model	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
Alexnet	27.16	14.95	16.28	19.47	28.22	15.60	17.14	20.32
HPE_Resnet18	4.96	8.57	9.23	7.59	26.90	14.31	15.42	18.88
HPE_Resnet34	4.80	7.01	8.00	6.60	26.04	13.62	13.40	17.69
HPE_Resnet50	4.16	6.57	7.75	**6.17**	23.05	12.75	9.29	15.03

Figure 6: Part of continuous video frames in Safe_Imgs.

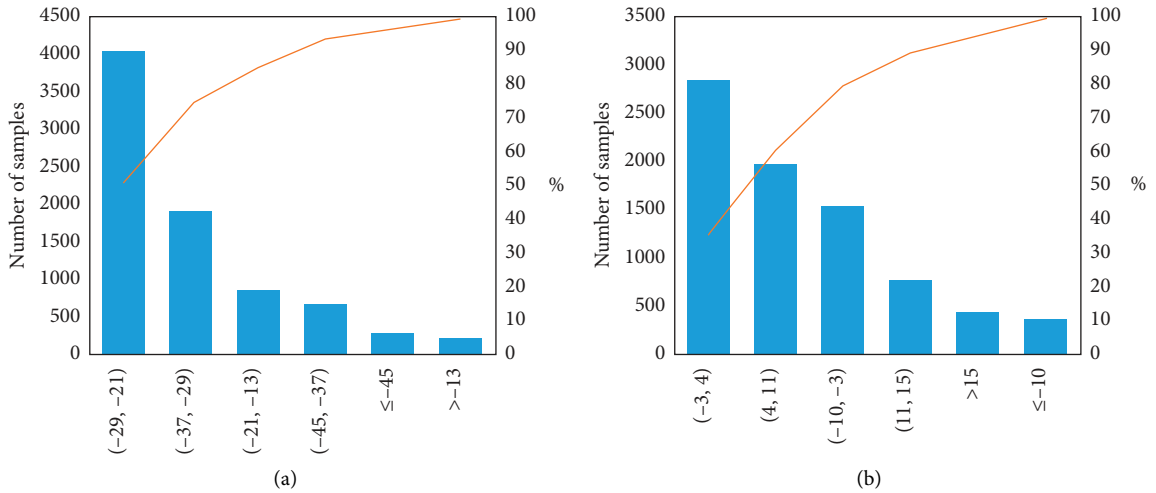


Figure 7: Continued.

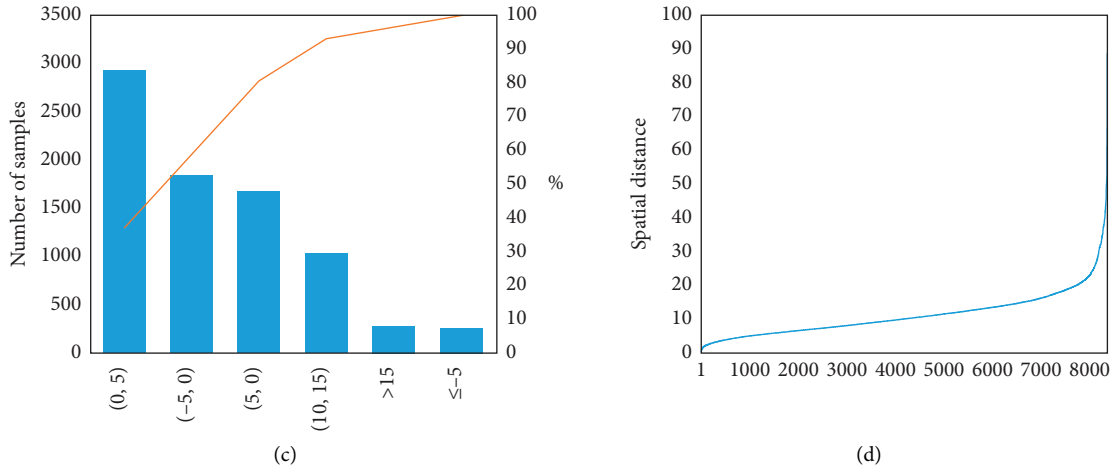


Figure 7: Head pose statistics in Safe_Imgs. (a) Yaw range. (b) Pitch range. (c) Roll range. (d) Distance range.

Table 5: Safe driving range and recommended threshold obtained from Safe_Imgs.

Participants number	FPS	Video	Safe range	β	D
80	20	5s/segment	Yaw (−45°, −13°) Pitch (−10°, 15°) Roll (−5°, 15°)	(−27.6, 2.5, 3.9)	18.6

Table 6: Driving action description.

Safe	Safe_l	Safe_r	Dist_h	Dist_p	Dist_t	Dist_r	Dist_d
Safe	Left rearview mirror (safe)	Right rearview mirror (safe)	Small hand movements	Phone use	Turn around	Adjust radio	Drink or eat

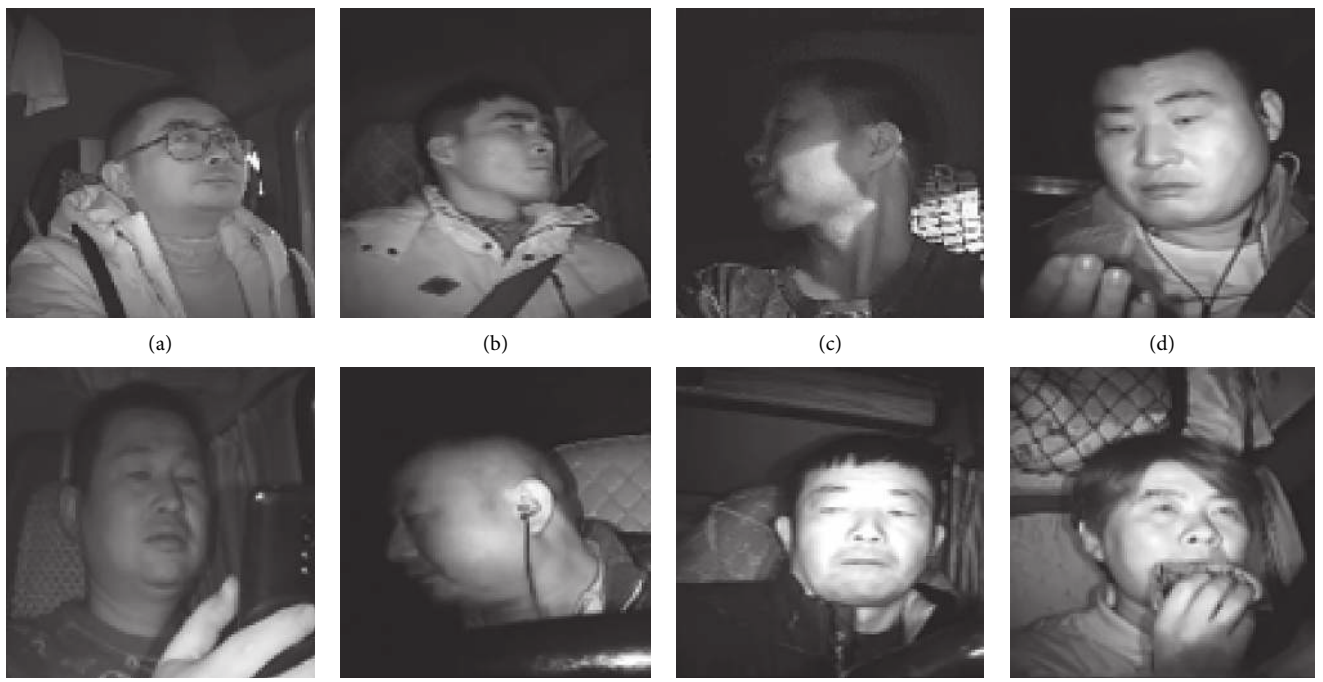


Figure 8: Some pictures of test_Imgs. (a) Safe. (b) Safe_l. (c) Safe_r. (d) Dist_h. (e) Dist_p. (f) Dist_t. (g) Dist_r. (h) Dist_d.

Table 7: Distraction parameter of t1-t15.

Test_Imgs	*t*1 (safe)	*t*2 (safe)	*t*3 (safe)	*t*4 (safe)	*t*5 (safe)
All	9%	23%	27%	14%	18%
max	4	3	4	2	3
Count	0	0	0	0	0
Test_Imgs	*t*6 (Safe_l)	*t*7 (Safe_l)	*t*8 (Safe_l)	*t*9 (Safe_r)	*t*10 (Safe_r)
All	38%	42%	46%	20%	48%
max	23	15	14	14	22
Count	2	3	4	1	3
Test_Imgs	*t*11 (Dist_h)	*t*12 (Dist_p)	*t*13 (Dist_p)	*t*14 (Dist_p)	*t*15 (Dist_t)
All	**100%**	**92%**	**66%**	**100%**	**100%**
max	**100**	33	14	**100**	**100**
Count	1	3	**5**	1	1

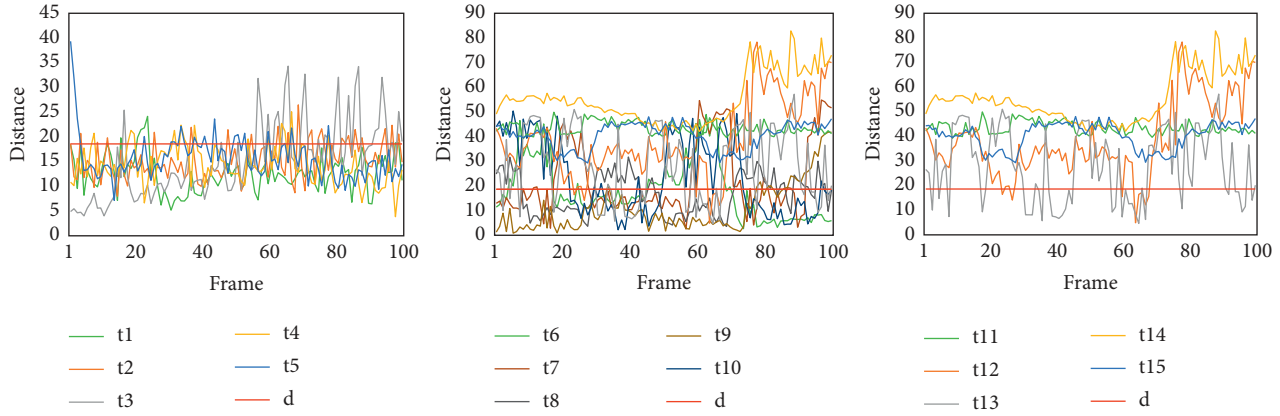


Figure 9: The chart of t11-t15.

the maximum number of hours for a single observation, and *t*11–*t*15 is the driver's true distraction during the actual driving trip. During the window of 5 seconds, some of the driver's head movements included the entire window period. Although the distraction movements varied, they all had varying degrees of loss of safe driving head position, and one or two of the distraction parameters were significantly higher than the threshold in the distraction recognition theory. Figure 9 shows the line chart of *t*1–*t*5.

4. Conclusions

This study focused on training a stable, accurate, and available head pose estimator that does not require preprocessing for either training pictures or actual test images. After testing, the error of the pose estimator is within an acceptable range, which can be used not only to analyze the images collected in the experimental environment but also pictures from the real world.

Based on the calculation and statistics of the head pose on the common distraction dataset performed using the head pose estimator, it is concluded that there are specific differences in the head pose between safe and distracted driving. This difference can be used for providing quantitative data basis for distraction detection. Since the single-frame image can only capture an instantaneous state while human action is continuous, it is not reliable for identifying a distracted state based on a single-frame image. Therefore, continuous video frames can overcome the sudden impact caused by the bumps and light changes on the driving route, provide the safety range and threshold at a specific driving angle, and establish parameters for identifying distracted driving.

Point-of-gaze as an important indicator to judge the driver: we can use point-of-gaze estimation and head posture estimation for information fusion in the future. In more complex driving scenes, there is a certain correlation between them, which can complement each other, and in theory, the accuracy of distraction recognition can continue to be improved. At the same time, lightweight is also an important work, which makes it possible to deploy on vehicle equipment and maintain high accuracy.

Data Availability

The dataset in the paper can be obtained by contacting Sili Xia (xiadeemail@163.com).

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (nos. 51874300 and 61976217) and Xuzhou Key R&D Program (no. KC18082).

References

[1] J. D. Lee, "Fifty years of driving safety research," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 3, pp. 521–528, 2008.

[2] E. Murphy-Chutorian, A. Doshi, and M. M. Trivedi, "Head pose estimation for driver assistance systems: a robust algorithm and experimental evaluation," in *Proceedings of the 2007 IEEE Intelligent Transportation Systems Conference*, pp. 709–714, IEEE, Washington, DC, USA, October 2007.

[3] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2008.

[4] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 340–350, 2007.

[5] M.-C. Chuang, R. Bala, E. Bernal et al., "Estimating gaze direction of vehicle drivers using a smartphone camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 165–170, Boston, MA, USA, June 2014.

[6] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1–14, 2015.

[7] R. Zheng, K. Nakano, H. Ishiko, K. Hagita, M. Kihira, and T. Yokozeki, "Eye-gaze tracking analysis of driver behavior while interacting with navigation systems in an urban area," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 546–556, 2016.

[8] S. Jha and C. Busso, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *Proceedings of the IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 1–6, Rio de Janeiro, Brazil, November 2016.

[9] L. Fridman, T. Victor, J. Lee, and B. Reimer, "'Owl' and 'Lizard': patterns of head pose and eye pose in driver gaze classification," *IET Computer Vision*, vol. 10, no. 4, pp. 308–313, 2016.

[10] T. Hoang Ngan Le, Y. Zheng, C. Zhu et al., "Multiple scale faster-rcnn approach to driver's cell-phone usage and hands on steering wheel detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 46–53, Las Vegas, NV, USA, July 2016.

[11] R. Jiménez Moreno, O. Avilés Sánchez, and D. Amaya Hurtado, "Driver distraction detection using machine vision techniques," *Ingeniería Y Competitividad*, vol. 16, no. 2, pp. 55–63, 2014.

[12] B. Baheti, S. Gajre, and S. Talbar, "Detection of distracted driver using convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1032–1038, Salt Lake City, UT, USA, June 2018.

[13] M. D. Breitenstein, D. Kuettel, T. Weise et al., "Real-time face pose estimation from single range images," in *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, Anchorage, AK, USA, June 2008.

[14] P. Padeleris, X. Zabulis, and A. A. Argyros, "Head pose estimation on depth data based on particle swarm optimization," in *Proceedings of the 2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 42–49, IEEE, Washington, DC, USA, June 2012.

[15] T. Xu, C. Wang, Y. Wang et al., "Saliency model based head pose estimation by sparse optical flow," *IEEE*, in *Proceedings of the First Asian Conference on Pattern Recognition*, pp. 575–579, Beijing,China, June 2011.

[16] L.-P. Morency, J. Whitehill, and J. Movellan, "Monocular head pose estimation using generalized adaptive view-based appearance model," *Image and Vision Computing*, vol. 28, no. 5, pp. 754–761, 2010.

[17] X. Liang and W. Tong, "Face pose estimation using near-infrared images," in *Proceedings of the 2012 international conference on communication systems and network technologies*, pp. 216–220, IEEE, Rajkot, GJ, India, May 2012.

[18] S. S. Mukherjee and N. M. Robertson, "Deep head pose: gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015.

[19] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[20] K. He, X. Zhang, S. Ren et al., "Deep Residual Learning for Image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[21] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2074–2083, Salt Lake City, UT, USA, June 2018.

[22] X. Zhu, Z. Lei, X. Liu et al., "Face alignment across large poses: a 3d solution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 146–155, Las Vegas, NV, USA, June 2016.

[23] M. Koestinger, P. Wohlhart, P. M. Roth et al., "Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization," in *Proceedings of the 2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pp. 2144–2151, IEEE, Barcelona, Spain, September 2011.