Infrared Physics & Technology 116 (2021) 103740

Contents lists available at ScienceDirect

Infrared Physics and Technology

journal homepage: www.elsevier.com/locate/infrared

Precise head pose estimation on HPD5A database for attention recognition based on convolutional neural network in human-computer interaction

Hai Liu, Duantengchuan Li [*], Xiang Wang, Leyuan Liu, Zhaoli Zhang, Sriram Subramanian

National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China

ARTICLE INFO

Keywords:
Head pose estimation
Infrared imaging
Convolutional neural network
Database
Human-computer interaction
Behavioural analysis

ABSTRACT

Head pose estimation (HPE) under infrared imaging has become more and more common in the human-computer interaction. In this paper, we proposed a novel HPE with convolutional neural network and established a precise head pose database under 5° angle (HPD5A) for human attention recognition. Specially, the HPD5A database includes 729 infrared head pose images from different subjects with and without glasses, which corresponds to drivers who wear glasses or do not wear glasses. To verify the availability and usability of the HPD5A database, the benchmark evaluations are performed on our database using traditional standard HPE classification methods with and without principal component analysis. The methods include linear discriminant analysis, *K*-nearest neighbor, random forest and Naïve Bayes classifiers. We also design and implement a convolutional neural network architecture as one of elementary assessments. All the results are provided for future reference. The developed deep learning technique could obtain the state-of-the-art performance on the HPD5A database. This database will certainly help in the development of model for infrared HPE and be beneficial to the attention recognition in human-computer interaction system.

1. Introduction

Head posed estimation (HPE) under infrared (IR) imaging is leveraged to automatically estimate human head pose from an image or a video and predict the directions at which human are focusing on, which is beneficial for many practical applications [1–3], such as online learning [4–6], human-computer interaction [1,7], and assisted driving [8]. According to the latest statistics provided by the world health organization, about 1.3 million people annually die in traffic accidents around the world. Human factors, including driver distraction, fatigue driving, are main factors leading to traffic accidents [9,10]. To ensure the safety of drivers and passengers, many researchers focus on intelligent driving assistance systems [11]. However, most of them did not take into account the night driving environment, which will lead to traffic accidents increasing. Furthermore, thanks to the popularity and low cost of night infrared cameras, car manufacturers are committed to developing better advanced assistance systems. Undeniably, head pose signal can provide rich information about human motivation, intention and attention (Fig. 1). Assistance systems can estimate drivers' condition such as fatigue, distraction by tracking the head of drivers.

HPE has become a key focus area of intelligent driving assistance systems. Unfortunately, there are few nights infrared head pose database. This paper proposes and establishes the start-up assistance driving infrared head pose (HPD5A) database for human-computer interactive system. First, we will describe the design, collection, and annotation of the HPD5A database. Then, we carry out the benchmark evaluations on the proposed infrared head pose database through using not only several facial features-based methods, including linear discriminant analysis (LDA) [12], *K*-nearest neighbor (KNN) [13], random forest (RF) [14] and Naïve Bayes classifiers (NB) [15], but also an emerging approach, namely convolutional neural network (CNN). To decrease the computational complexity, we will apply a dimensionality reduction technique before classification as comparative experiments, namely principal component analysis (PCA) [16]. The performance of the proposed database can be proved by the evaluation result in the HPE task.

Over the past decade years, there are some existing databases for the HPE task. The database includes the Bosphorus [17], AFLW [18], MALF [19], BIWI Kinect [20], 300W-LP [21], AFLW2000-3D [21], MTFL [22], WFLW [23], Pointing'04 [24], CAS-PEAL-RI [25], and CMU Multi-PIE [26]. In fact, it is very difficult to achieve the high-quality HPE databases. The major reason can be summed as two aspects. Firstly, the artificial database often contains several angles, which cannot describe

[*] Corresponding author.
E-mail address: dtclee1222@gmail.com (D. Li).

https://doi.org/10.1016/j.infrared.2021.103740
Received 26 January 2020; Received in revised form 1 January 2021; Accepted 30 March 2021
Available online 6 April 2021

all the head pose in the human-computer interaction. Secondly, the time-consuming manual labeling of spontaneous expressions and the difficulty of setting up the scene often discourage the researchers who plan to establish the HPE database. Thus, we usually produce the image label manually for the existing HPE databases. In Table 1, we sum the head posed database in the past decades, including the type, released time, pose description, condition, and samples.

The infrared head pose databases do not exist, since the research is focus in visible-light imagery for the HPE task. From Table 1, we can observe that the images of these databases are visible images that can prove the above argument. Since the widespread use of the infrared cameras and the lack of infrared head pose databases, we plan to establish a precise IR HPE database, which can mitigate the issue and enrich the existing database in the assisted driving systems.

The rest of this paper is structured as follows. Section 2 introduces the setup for infrared imaging capturing. Section 3 presents the details about the data acquisition processing and environment. The benchmark evaluations and analysis are presented in Section 4. Section 5 concludes this article.

2. Setup for HPE image capturing

To set up the head pose dataset, we construct a photographic room in our laboratory. Since we prepare to mark on the tubes and the ceiling, a spacious attic is chosen. It is only two meters high which makes it very convenient for us to mark. It includes four parts in the head pose recording system, such as a chair lift in a fixed position, tagged scenes, a network-linked person computer and an IR camera system. In Fig. 2, we have demonstrated the image acquisition system in details.

2.1. Scene layout

As shown in Fig. 2, 27 markers are set at yaw angle direction with 5° intervals increasing. The yaw angles appear as a sector form. The considered yaw angles are between -65° and 65°. In each yaw angle direction, we label all the pitch angles. In other words, we marked it on the benchmarks from the ground to the ceiling. All the pitch angles are ranged in $[-65^\circ, -60^\circ, -55^\circ, -50^\circ, -45^\circ, -40^\circ, -35^\circ, -30^\circ, -25^\circ, -20^\circ, -15^\circ, -10^\circ, -5^\circ, 0^\circ, 5^\circ, 10^\circ, 15^\circ, 20^\circ, 25^\circ, 30^\circ, 35^\circ, 40^\circ, 45^\circ, 50^\circ, 55^\circ, 60^\circ, 65^\circ]$. To ensure that all the subject eyes can stare on the markers, we set a lifted chair to adjust the height of the subjects. The pitch angle is initialized as 0° in the sector center. To produce a simple background, we place a folding screen behind every subject. To acquire an accurate marked position, we calculated it through the geometric relationship, precise measuring tools and strict mathematical computation.

2.2. Setup of infrared camera

To capture the infrared images, a DS-IPC-S12A-IWT (4 mm) infrared camera is introduced. The image resolution is 1080 × 1920. An infrared-cut removable (ICR) filter is leveraged for day and night conversion modes. To achieve the accurate groundtruth label of each HPE image, the experimenters are required to stare on each marker through adjusting their head gestures instead of turning their eyeballs. However, most person are habituated to stare on the markers by rolling their eyes when it comes to taking exaggerated poses. For instance, the experimenter will look at the markers with oculogyration involuntarily rather than head rotation when pitch angle is at 65° and the yaw angle is at 65°. To suppress the errors caused by artificial factors, we arrange a volunteer to assist the experiment in finishing the specified head pose.

2.3. Glass in the infrared imaging

When performing HPE experiments, the important information can be disguised by wearing the glasses. We requested subjects to take two sets of photos, one with glasses, one without glasses. It can make the HPD5A database effective in scenario cases. Their own glasses and the provided one can be selected in the capturing process.

3. Data acquisition for HPE database

3.1. Subjects

In our experiments, we recruit 40 healthy volunteers as subjects, including 18 females and 22 males. All the subjects are asked to sit straightly just like the assisted driving poses. They could finish the specified head poses. Each is informed that the HPE images are only used in the scientific research before the experiment. All the subjects will receive compensation for participating after finishing the experiments. In this paper, the total of 27 × 27 marked positions for each subject (729 markers). Two groups of HPE images (without and with glasses) are captured for each subject. Each group includes about 729 images.

3.2. Procedure

During our experiments, we record two groups head poses for every subject. The procedure is provided in detail as follows. (1) We introduce

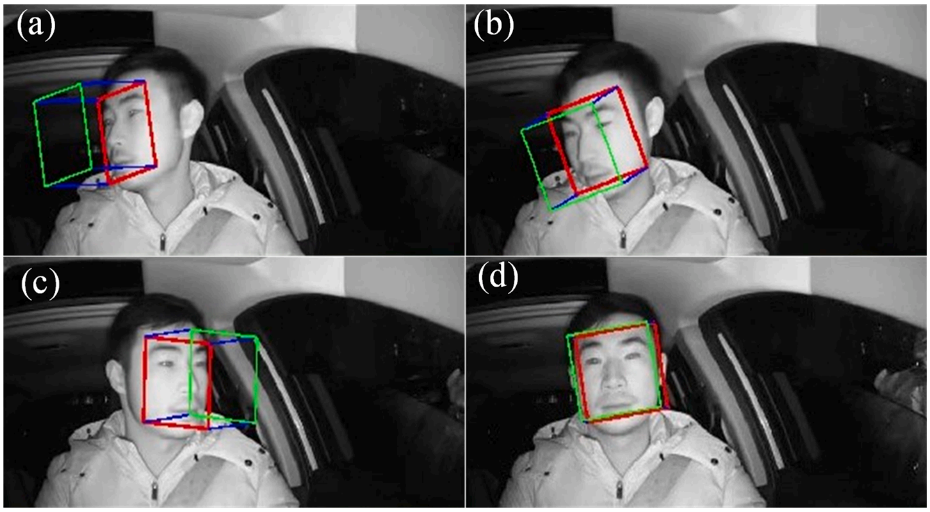


Fig. 1. Infrared HPE for aided driving at night with few ambient light sources.

Table 1
Comparison of head pose databases.

Databases	Type	Released time	Pose description	Condition	Samples
BIWI	Visible	2013	Roll: from −50° to 50°. Pitch: from −60° to 60°. Yaw: from −75° to + 75°.	Lab	15 K images
Multi-PIE	Visible	2009	Yaw: [−90°, −75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°, 90°].	Lab	More than 750,000 images
Bosphorus	Visible	2009	Pitch: Strong upwards, Slight upwards, Slight downwards, Strong upwards. Yaw: [−90°, −45°, 10°, 20°, 30°, 45°, 90°]. Cross Rotations: [(45° yaw, 20° pitch), (45° yaw, −20° pitch)].	Lab	4666 images
CAS-PEAL-RI	Visible	2008	Pitch: [−45°, −30°, −15°, 0°, 15°, 30°,45°] Yaw: [−30°, 0°, 30°].	Lab	30,863 images
Pointing'04	Visible	2004	Pitch: [−90°, −60, −30°, −15°, 0°, 15°, 30°, 60°, 90°]. Yaw: [−90°, −75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°, 90°].	Lab	2790 images
AFLW2000-3D	Visible	2016	Annotated by algorithm	Web	2000 images
300W-LP	Visible	2016	Annotated by algorithm	Web	122,450 images
MTFL	Visible	2014	Yaw: [−60°, −30°, 0°, 30°, 60°]	Web	12,995 images
AFLW	Visible	2011	Annotated by algorithm.	Web	25 K images

the experimental procedure for the subjects, including how to roll their head, the meaning of valence and arousal. (2) The chair is moved to the sector center as the initialization. The subjects are asked for sitting upright on the side opposite the tube where the yaw angle was 0°. We raised or lowered the chair to make sure that the marker and their eyes are at the same horizontal line. Then the subject can gaze on the markers where its pitch was at 0°. (3) The IR camera is placed between the tube and the subject. It is rather remarkable that the IR camera must be pointed directly at the subject's eyes. The IR camera is connected with the personal computer when it starts. The captured images can be transferred to the personal computer. (4) To locate the subjects and label all the pose images correctly, we required the subjects to finish each posture in an orderly manner. Namely, the subjects turn around their hand starting from yaw angle at −65° and pitch angle at −65°. Then we capture one image with the head rotation 5° angle. For the yaw direction, there are 27 head pose images to be captured. The head pose images are stored in our personal computer one by one.

3.3. *Design of database*

After recording all the data, we clip all the images and unified them into 224 × 224 sizes as our database for formal use. In total, the HPD5A database contains 58,320 images from 40 different subjects. Each subject displays 729 specified head pose. Fig. 3 shows example images of a subject. Indeed, each subject has to accomplish the same head pose two times. In Fig. 4, we show one group IR head pose images without and with the glasses.

4. Benchmark evaluations and experiments

In this section, we presented the procedures to achieve the benchmark evaluation results on the proposed HPD5A database by using standard HPE methods. Those elementary assessments provide an insight to the usability and effectiveness of database. In general, the existing approaches could be briefly classified into two streams: (1) traditional methods, *i.e.*, facial features-based HPE; (2) state-of-the-art approaches, *i.e.*, deep learning-based HPE. The distinction of them is that the patterns of features extraction are unlike. For an image, the former, depending on the corresponding descriptor, are invariant and the necessary element for classifying, while the latter are updated iteratively to achieve the most incarnated features in the training process. We conducted HPE experiments using different types of methods belong to two streams. Details and results of baseline evaluation are provided in the following sections.

4.1. *Facial features-based methods*

4.1.1. *Feature extraction*

The head pose feature extraction can determine the accuracy of the HPE task. Three different traditional feature extraction techniques are selected, such as grayscale intensities, Gabor wavelets and histogram of oriented gradients.

Grayscale intensities (GI): Among the conventional HPE, most of them are GI-based methods [27–29]. The kind of methods do not depend on

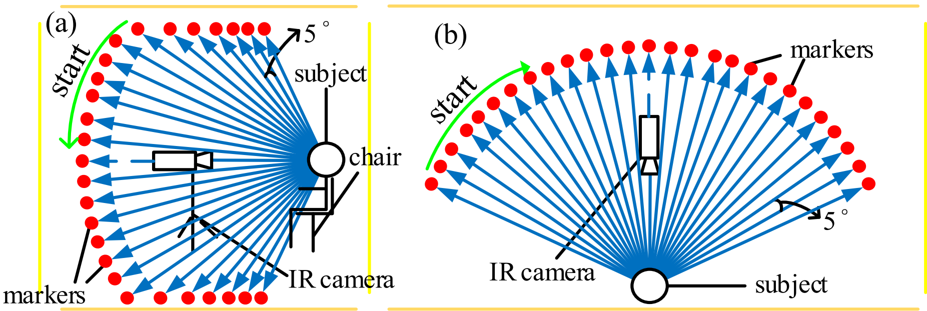


Fig. 2. 2D plan of the shooting environment. (a) Side view of the image capturing. The red dot denotes pitch angles increasing the interval 5° from −65° to +65°. (b) Top view of the image capturing. The red dot means a yaw angle from −65° to +65°.

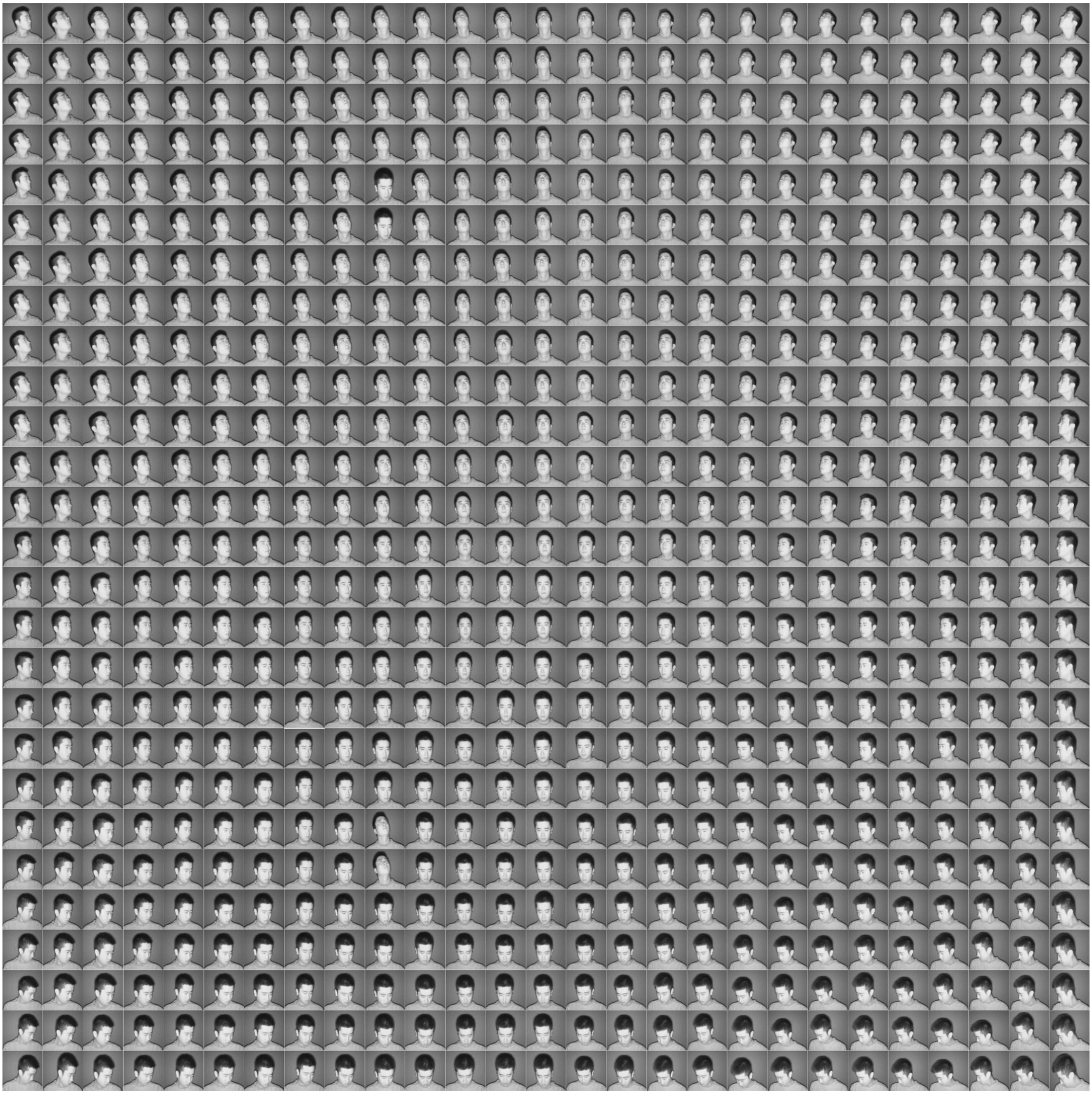


Fig. 3. Head pose images in the proposed HPD5A dataset. From left to right (from bottom to top), the yaw (pitch) angles are increased from −65° to +65°, respectively.

any special hardware, only a webcam can work normally. Thus, it is easy to promote and use, and the related research achievement yield the most. It is worth noting that facial images need to be preprocessed since the pixel intensities vary from 0 to 255, *i.e.*, they were normalized as a Gaussian distribution with zero mean value. Whereafter, the pixel intensities of the preprocessed image were used as feature vectors to classify head pose.

Gabor wavelets (GW): GW transform has good time-frequency localization characteristics. That is, it is very easy to adjust the fundamental frequency bandwidth, the directions of Gabor filter, and center frequency so as to trade off the resolution of the signal in the space-time and frequency domains. GW transform has a multi-resolution characteristic, namely a zooming capability. That is to say, using a multi-channel filtering technique, a set of GW with different time-frequency domain characteristics is introduced to the image transformation. The local features of input images can be extracted by each channel. Thus, the images can be analyzed at different coarse and fine granularities as needed. In [30], it found that Gabor filters are well suitable for texture separation and expression. A 2D Gabor filter in the spatial domain is a Gaussian type function, which is modulated by a sinusoidal plane wave. In the field of image processing, Gabor function is a linear filter for edge feature extraction. The orientation and frequency of the Gabor filters are very similar to those of the human visual system. In consequence, many image processing researchers [31–33] are keen on it for capture discriminative features. In our implementation, we opted to use four orientations and six scales; thus twenty-four Gabor filters banks described in Fig. 5.

Histogram of oriented gradients (HOG):The optical and geometric

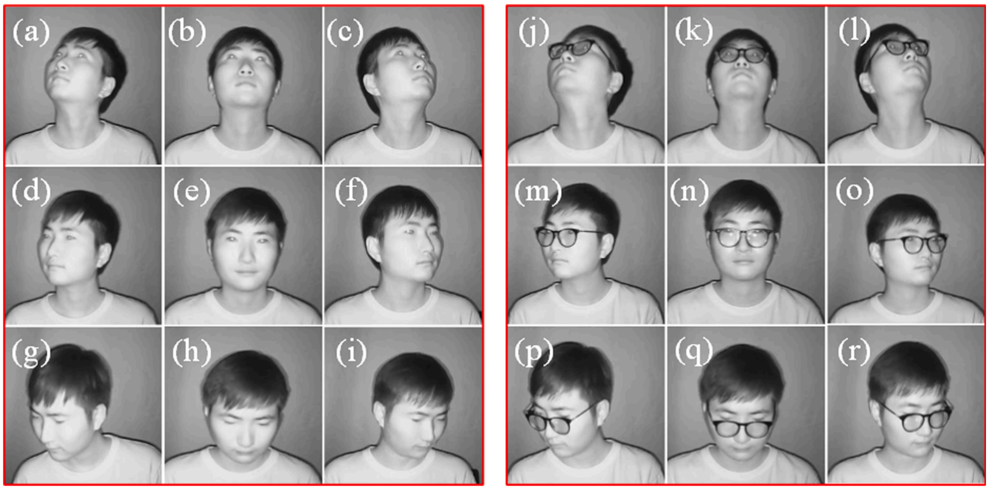


Fig. 4. One group data of a subject with and without glasses for nine different head pose, *i.e,* (−45°,45°), (0°,+45°), (+45°,+45°), (−45°,0°), (0°,0°), (+45°,0°), (−45°, −45°), (0°,−45°), and (+45°,−45°).

deformation of the face images can be well extracted by the HOG operator. It is also one of the best features for representing the object boundary and geometric information. It consists of calculating and statistic the gradient orientation histogram of the local region of the image to form the feature vector. The specific implementation method is illustrated as follows. Firstly, the facial image is divided into several small related regions. They are called as the cell units. Then, we can obtain the edge direction and gradient histogram of each pixel in the cell units. Finally, a feature descriptor can be constructed by these histograms. In some studies [38], HOG descriptors have been employed as HPE and perform much better than other feature filters. As a result, we selected the HOG feature as one of the benchmark features of our experiments.

4.1.2. *Head pose classification*

HPE can be considered as a standard pattern classification problem which commonly adopts various machine learning methods. After extracting facial feature described in Section 4.1.1, several classification models are implemented in our experiments. Dimensionality reduction of input eigenvectors is required, which is a vital step before introducing classification techniques. The input eigenvectors are achieved by the feature extraction operators, such as HOG, GI, and GW. And the reason for that is, the dimensionality of the input data is overabundant which results in great computational complexity. In our experiment, PCA will be chosen as a dimensionality reduction tool.

We conduct PCA + KNN, PCA + LDA and PCA + NB classifiers for extracting features in head pose classification. In order to compare with the performances with and without applying PCA, we also perform solely those classification techniques. In addition to the above these individual classifiers, an ensemble learning model will also be implemented in our experiments. Random forest (RF) is constructed and combined through decision trees. It is deemed to the typical representative the ensemble learning and achieves higher robustness.

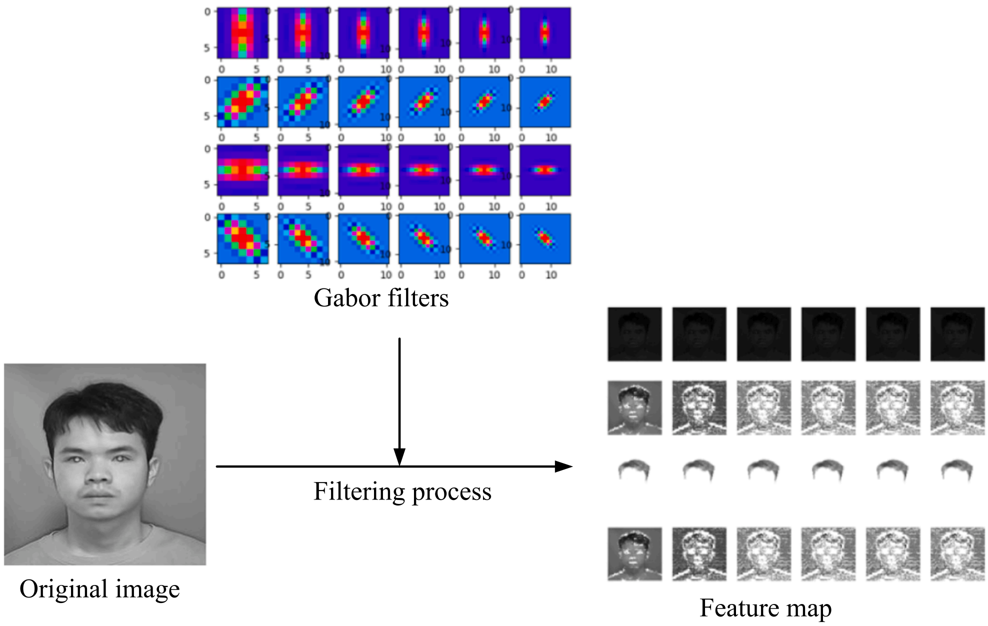


Fig. 5. Gabor features extraction. In this paper, we opted to use four orientations and six scales (Gabor orientation: 0°, 45°, 90°, 145°; scales: 7 × 7, 9 × 9, 11 × 11, 13 × 13, 15 × 15, 17 × 17).

4.2. Deep learning-based methods

Apart from the implementation of our experiments with the above traditional methods, convolutional neural network (CNN), as the most common technology for HPE nowadays, is proposed for baseline evaluation as well. Since Massimiliano *et al.* [35] deeply studied CNN with dropout and adaptive gradient method, then they introduced it into HPE for the first time and obtained good performance. In the following two years, multifarious CNN-based HPE methods had sprung up like mushrooms. For instance, FAN [36] built a multi-dimensional detector using CNN and constructed abundant facial landmark datasets. Multi-loss CNN [34] adopted a coarse-to-fine strategy by using three separate classification and regression losses to constrain head pose. QuatNet [1] established a multi-regression loss CNN model to directly predict the angles of head pose represented by unit quaternion instead of Euler angle. To facilitate accurate key-points detection, KEPLER [37] proposed a multi-task network for capturing structured features from local to global. 3D pose of the face prediction was also provided as a by-product. In [6], the authors exploited the soft stage-wise architecture and feature aggregation to develop an attention network for HPE. Therefore, CNN has already yielded unusually brilliant results in the field of HPE which is the primary cause of applying CNN-based method as our benchmark evaluation.

In this work, we developed a lightweight CNN architecture consisting of three convolutional layers and several fully connected layers. The last layer is the softmax layer, which returns the predicted classification result of input images. Details are shown in Fig. 6, the input of our CNN architecture is a fixed-size 224 × 224 image because of the specific size of our proposed database. Convolutional layers use a convolution kernel with a smaller receive domain of 3 × 3 in this architecture. Convolutional layers are followed by a max pooling layer, which is major utilized for reducing the feature dimension and over-fitting problems, compressing the number of data and raising the robustness of model. A most special noteworthiness is, as shown in Fig. 6, there may be a normalization layer between the pooling layer and the convolution layer, which can effectively prevent gradient dispersion and accelerate network training. We implemented experiments with and without the normalization layer separately. In the end, the output size of the last full connected layer is twenty-seven, which represents the number of head pose of yaw angle or pitch angle.

In the proposed architecture, the first three-convolutional layer acts as the backbone net to extract common features shared by different databases. The corresponding parameters are initialized by pre-trained models. Moreover, it follows the fundamental that the front convolutional layers share features and the later fully connected layers learn the features representation for specific tasks.

In pre-training step, the ground-truth labels and the predicted classification probabilities are transmitted into the corresponding loss functions. Then the classification scores transmit through softmax activation function. It obtains the corresponding predicted probabilities in inference step. This compact modification makes the architecture easy to be trained and achieves comparable performance on our database for HPE with fewer model parameters. We choose it to pre-train our model on the 300w-lp HPE database. It is a large HPE database which contains not only 122,450 facial images but also a great deal of persons. Based on this, the pre-trained model can extract the general features of head pose to improve its generalization capability. However, there is one detail that merits attention. We need to preprocess the images and make them greyscale images of 224 × 224 × 1 in order to ensure the same input format as our database.

4.3. Results and analysis

The experiments were performed with the software library tensorFlow, and which was used to implement our proposed CNN architecture. We carried out on a workstation with Intel Xeon Gold 6126 CPU and Nvidia Tesla V100 GPU. HPD5A database was divided into three components: training, validation and testing sets. The ratio is set as 0.7: 0.2: 0.1 for fair and meaningful experimental comparisons. It is noteworthy that the division is based on person independent. This is not only to consider the actual situation that the system will be tested on persons who were not labelled on the training phase, but also to prevent the problem of data leakage during the experiment. Adam optimizer is adopted to update the model parameters for 50 epochs and the batch size is 64. The learning rate is initialized to 1e-4 and reduced by a factor of 0.1 every 5 epochs.

4.3.1. Performances of different feature extractors

As observed from Tables 2-4, the results of relevant evaluation metrics are provided in our experiment. Firstly, based on the perspective of time-cost, PCA technique can reduce considerable time complexity whether utilizing KNN, LDA and NB as classification approaches. Nevertheless, there is little distinction between the two while utilizing RF. The reason behind this is that it consists of a number of decision

Table 2
Comparison of MAE, accuracy, and time-cost by different approaches with GI.

Approaches	Accuracy (%)		MAE (°)		Time-cost
	Pitch	Yaw	Pitch	Yaw	
KNN	84.94	83.69	8.10	8.56	15.24 s
PCA + KNN	86.17	89.56	7.69	7.31	1.24 s
LDA	83.46	81.11	8.79	8.92	58.35 s
PCA + LDA	85.99	85.63	7.61	7.55	5.25 s
NB	80.33	78.58	13.10	11.61	5.69 s
PCA + NB	81.49	79.46	12.33	10.77	0.78 s
RF	89.64	88.67	7.33	7.01	153.06 s
PCA + RF	**91.79**	**90.74**	**7.11**	**6.99**	**150.69 s**

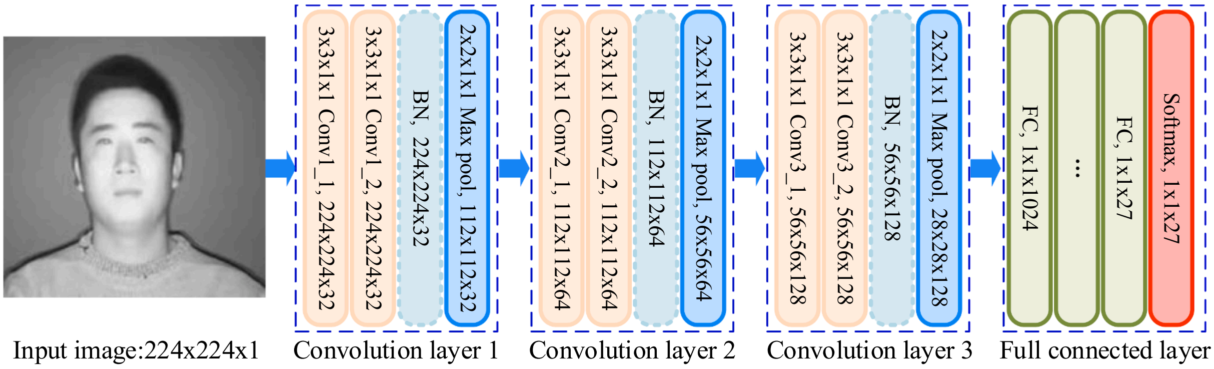

3x3x1x1 Conv1_1, 224x224x32
3x3x1x1 Conv1_2, 224x224x32
BN, 224x224x32
2x2x1x1 Max pool, 112x112x32
3x3x1x1 Conv2_1, 112x112x64
3x3x1x1 Conv2_2, 112x112x64
BN, 112x112x64
2x2x1x1 Max pool, 56x56x64
3x3x1x1 Conv3_1, 56x56x128
3x3x1x1 Conv3_2, 56x56x128
BN, 56x56x128
2x2x1x1 Max pool, 28x28x128
FC, 1x1x1024
...
FC, 1x1x27
Softmax, 1x1x27
Input image:224x224x1
Convolution layer 1
Convolution layer 2
Convolution layer 3
Full connected layer

Fig. 6. Architecture of our CNN-based HPE method. The architecture is one of our experiment based on CNN. In another experiment, it is similar to this architecture in the figure except for the non-normalization layer.

Table 3
Comparison of MAE, accuracy, and time-cost by different approaches with GW.

Approaches	Accuracy (%)		MAE (°)		Time-cost
	Pitch	Yaw	Pitch	Yaw	
KNN	88.67	87.52	8.60	9.14	394.61 s
PCA + KNN	88.11	87.94	8.32	9.01	5.36 s
LDA	86.31	85.14	8.36	8.11	3498.15 s
PCA + LDA	**89.11**	**88.60**	**6.52**	**7.04**	15.68 s
NB	79.02	76.14	14.80	10.02	47.30 s
PCA + NB	76.15	79.68	13.74	12.47	1.48 s
RF	84.02	83.63	8.62	8.14	150.33 s
PCA + RF	86.09	85.34	8.63	8.10	148.96 s

Table 4
Comparison of MAE, accuracy, and time-cost by different approaches with HOG.

Approaches	Accuracy (%)		MAE (°)		Time-cost
	Pitch	Yaw	Pitch	Yaw	
KNN	86.72	86.11	9.26	8.01	0.0103 s
PCA + KNN	87.54	86.08	8.55	7.94	0.0063 s
LDA	78.22	75.39	15.13	13.19	0.0360 s
PCA + LDA	80.96	79.63	13.99	12.77	0.0165 s
NB	79.10	78.63	14.95	14.41	0.0096 s
PCA + NB	76.73	77.90	13.33	13.19	0.0033 s
RF	89.07	88.06	7.89	7.74	15.6942 s
PCA + RF	**91.86**	**90.14**	**5.66**	**6.12**	7.6961 s

trees. The parameter is set 500 for the number of decision trees, and the maximum depth parameter for searching is 30 in our experiment. This results in a negligible reduction in time complexity even with PCA, and it may take more time to make voting decisions because of the alteration of algorithms.

Then, based on the overall performances of the accuracy and MAE, it is obvious that these different feature extractors have their strengths and weaknesses. In general, the highest accuracy for yaw and pitch angle are 90.74% and 91.86%, respectively. As Tables 2 and 4 show, when choosing GI and HOG as classification feature respectively, the corresponding highest accuracy obtained by RF classifier with PCA are almost comparable. Further, HOG features yielded better results than GI with MAE of yaw angle about 6.12° and pitch angle about 5.66°. And they are also the best results in all non-deep learning techniques. However, NB has poor performance whether utilizing GI, GW or HOG. The reason for the incorrect classification is that, NB model is based on the independence of sample attributes. This assumption only exists in the ideal and is often hard to hold in practical applications. Hence, the classification consequence may be not good because of the correlation of the attributes. Undoubtedly, the attributes of face images contain textured property which have evident spatial correlation characteristics. So with that being said, NB may be not a suitable tool for searching the model that need to minimize the intra-class distance and maximize the inter-class distinction. But it perhaps is a better choice in real-time application on account of the low computational complexity compared to other models.

Finally, based on local performances of the accuracy and MAE, excluding tine-cost, the performance of each classifier with and without PCA varied greatly or little. It is observed that the performances with PCA are better than without PCA. We surmise by referring to the relevant literature that the original features vector before dimension reduction contains a great deal of redundant information. It may be the cause of the classification function overfitting, thus reducing the robustness of the model.

4.3.2. Performance of deep learning-based

We show the results in terms of accuracy and MAE between deep learning-based method and facial features-based method in Table 5. The first three methods are the best in facial features-based methods from Tables 2-4. Among all the methods, deep learning -based methods, being compared with facial features-based methods, manifest better performance and present higher accuracy and lower MAE. Compared with CNN (without normalization), CNN (with normalization) has 2.59–3.49% improvement in Accuracy and 41.39–41.80% in MAE. Moreover, CNN (with normalization) attains the best performance in both accuracy and MAE. It proves that CNN (with normalization) is a high-performance framework and has the capability to extract accurate head pose features from our database. This section focuses on the performance comparison of the latter two methods.

We severally analyzed the convergence speed of accuracy and MAE, observing the value during each of the 50 epochs. We plotted the schematic diagram of convergence results for both accuracy and MAE in the Fig. 7. Whether it is accuracy or MAE, they achieved convergence after the 10th epoch iteration. The accuracy for yaw angle by using normalization method converged to 96% approximately. It was significantly higher than the accuracy without using normalization method, which converged to about 95% in Fig. 7(a). A similar experimental phenomenon also occurred at the pitch angle as shown in Fig. 7(b). On the other side, by observing from Fig. 7(c) and 7(d), we could also find that curve of MAE was gradually starting to converge after reaching the 10th epoch. MAE of yaw angle and pitch angle converged to 2.5 by deep learning-based with normalization layer. And those values were significantly lower than those without normalization. Brightening, these results were the best of all experimental methods, including the facial features-based methods mentioned above.

For our proposed network architecture, the choice of active function and the number of fully connected layers affect performance directly. Several comparative experiments were used to reveal the correlation between model performance and parameters. Three activation functions (Tanh function, Sigmoid function, and ReLU) and four different number of fully connected layers were selected for the comparison experiment on our proposed database. Fig. 8 manifested the MAE value about several parameters on histograms. For the active function, the experimental results proved that the ReLU functions achieved the smallest MAE with all fully connected layers. Furthermore, Fig. 8 also shows that ReLU achieves the highest accuracy with all fully connected layers. ReLU makes some neuron output zero which causes sparsity in the network and reduces the interdependence of parameters. Thus, the model generalization and robustness capacity are enhanced by using ReLU. Moreover, we explore the effect of different number of fully connected layers on model performance. Figs. 8 and 9 displayed that three-layer fully connected network with ReLU activate function reached the best performance. Because deeper structure does not only increase the difficulty of the gradient propagating process when backpropagation, but also makes the training process unsmooth. Therefore, the three-layer network may be suggested to avoid this problem.

In summary, we have the following three conclusions about the experiments we have done. First, RF classifier is a better choice for HPE among the facial features-based methods. Compared with the best performing LDA classifier, the results of the two are not much different even if GW feature is selected. And the results by using RF classifier are the best if we choose other features. Then, whether it is from the perspective of accuracy or MAE, CNN using normalization methods performs

Table 5
Comparison of accuracy and MAE between facial features-based method and CNN-based method.

Approaches	Accuracy (%)		MAE (°)	
	Pitch	Yaw	Pitch	Yaw
PCA + RF (GI)	91.79	90.74	7.11	6.99
PCA + LDA (GW)	89.11	88.60	6.52	7.04
PCA + RF (HOG)	91.86	90.14	5.66	6.12
CNN (without normalization)	92.23	93.92	4.45	4.18
CNN (with normalization)	**95.82**	**96.51**	**2.59**	**2.45**

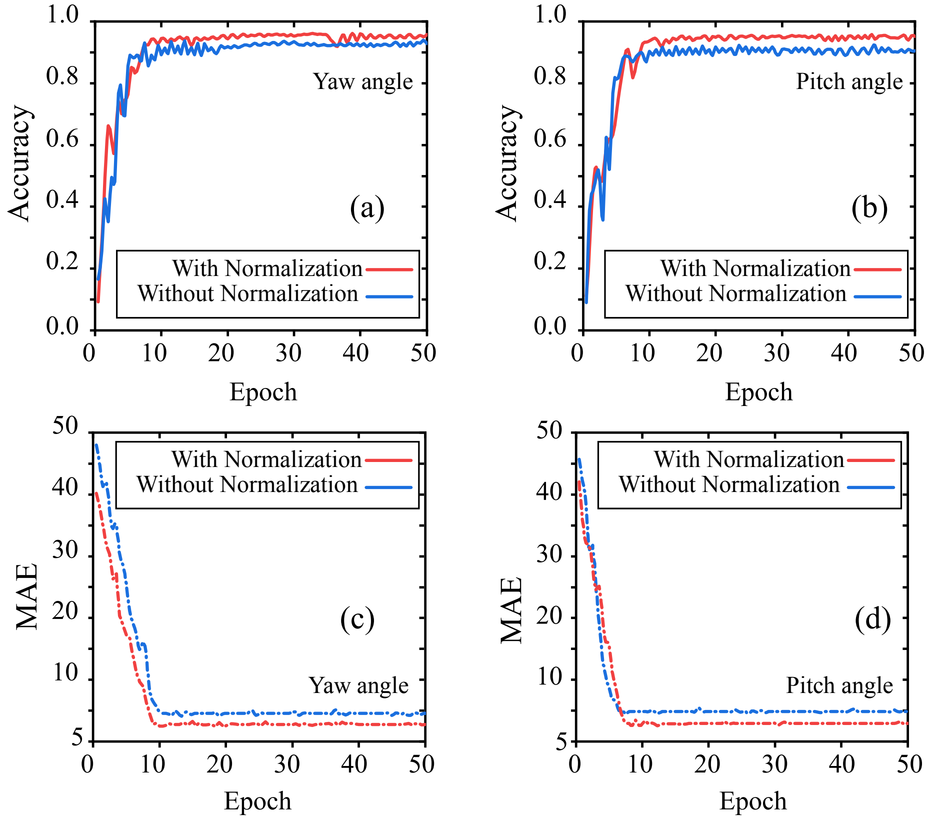


Fig. 7. Performance of two CNN architecture (with and without normalization) for the accuracy of the yaw angle (a) and the pitch angle (b) on the HPD5A database. Performance of two CNN architecture (with and without normalization) for the MAE of the yaw angle (c) and the pitch angle (d) on the HPD5A database.

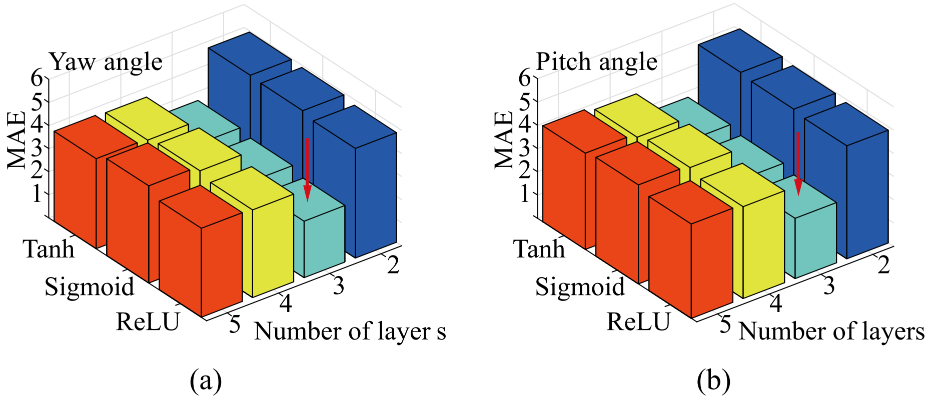


Fig. 8. Performance of CNN architecture with normalization by MAE with fully connected layers structure and different active function on HPD5A database.

significantly better than without normalization methods. Finally, the performance of CNN-based methods is better than that of facial features-based methods. Although MAE of CNN without normalization layer is sometimes lower than that obtained by using RF, we can conclude that the designed CNN has a shallower network layer and a simpler structure. It results in insufficient extracted features to obtain the best classification.

4.3.3. *Applications*

To demonstrate the applications of assisted driving in the night environment and infrared HPE, we collected and annotated thousands of images of the actual scene. Fig. 10 shows six different zones, namely, left, right, forward, center console, speedometer, and center rear-view mirror. The drivers often gaze on those zones when they are driving. Each zone is determined by a series of different head poses. Some testing images are shown in Fig. 11. The confusion matrix represents the results, as shown in Fig. 12. We observe that most gaze zones are effectively classified. Some mistakes occur between the “speedometer” and “forward” zones. Because the head poses are similar when the driver gazes at these zones. The results show that the approach is beneficial to driving assistant.

5. Conclusion

In this paper, we construct an infrared database contains 729 kinds of HPE in human computer interaction. In order to obtain accurate data, we use strict mathematical geometric formulas and precision measuring instruments in the experimental scene layout. During the data recording

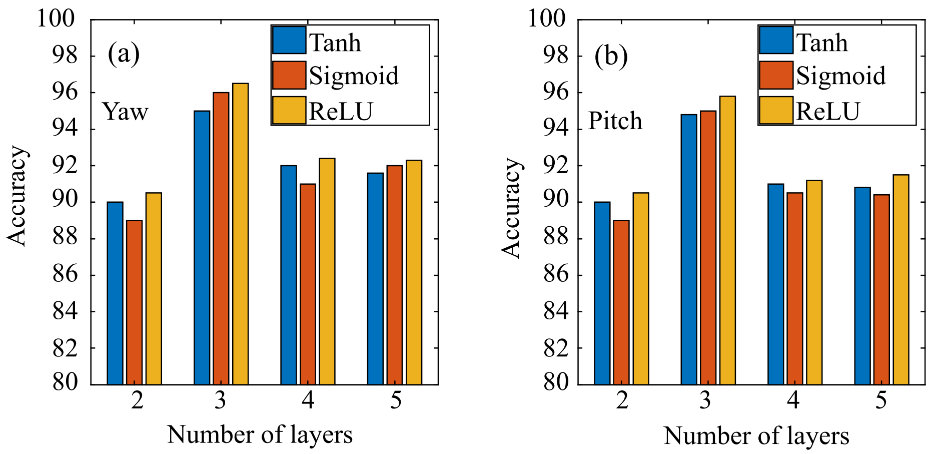


Fig. 9. Performance of CNN architecture with normalization by Accuracy with different fully connected layers and different active function on HPD5A database.

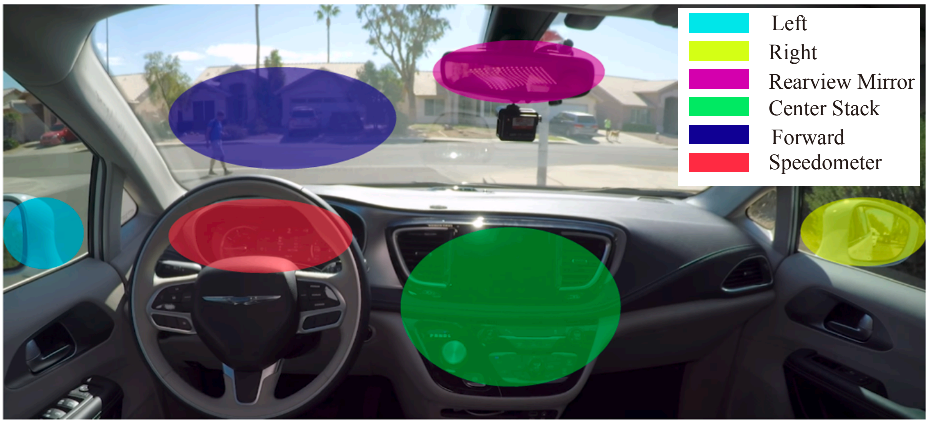


Fig. 10. Illustration of the driving assistant. Highlighted locations represent the gaze of driver.

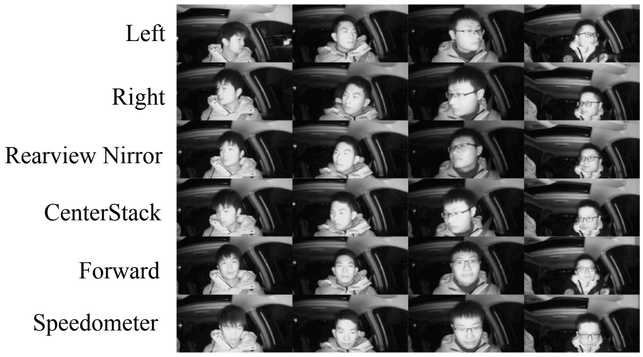
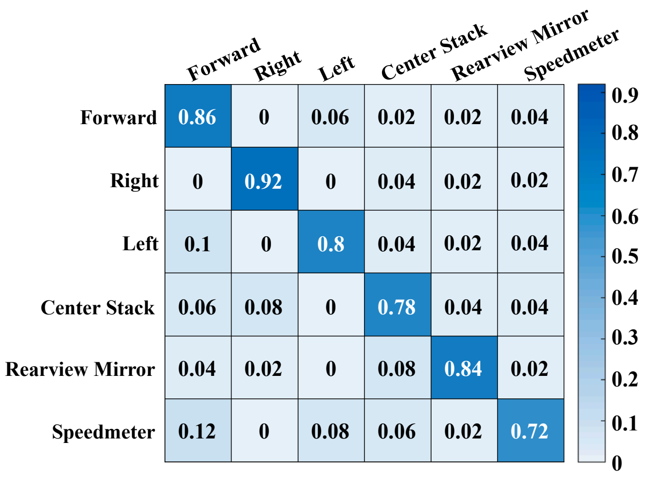


Fig. 11. Example images of several subjects for drive assistant.

phase, we arrange volunteers to assist the subjects to strictly complete the specified head pose. We implemented many baseline algorithms for infrared HPE with facial features-based methods and deep learning-based methods to verify the effectiveness and usability of the HPD5A database, and provide reference evaluation results for researchers for further improvement. We also compare the advantages and disadvantages of facial features-based methods and the performance distinction between facial features-based methods and deep learning-based methods.



	Forward	Right	Left	Center Stack	Rearview Mirror	Speedmeter
Forward	0.86	0	0.06	0.02	0.02	0.04
Right	0	0.92	0	0.04	0.02	0.02
Left	0.1	0	0.8	0.04	0.02	0.04
Center Stack	0.06	0.08	0	0.78	0.04	0.04
Rearview Mirror	0.04	0.02	0	0.08	0.84	0.02
Speedmeter	0.12	0	0.08	0.06	0.02	0.72

Fig. 12. Confusion matrix for six gaze zones using HPE.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors sincerely thank anonymous reviewers for their constructive comments, which helped improve this paper. This work was supported in part by the National Natural Science Foundation of China (62011530436, 62077020, 61875068, and 61702208), the Fundamental Research Funds for the Central Universities under Grant 2020CXZZ009, Grant CCNU20ZT017, and Grant CCNU2020ZN008.

References

[1] H. Hsu, T. Wu, S. Wan, W.H. Wong, C. Lee, QuatNet: Quaternion-Based Head Pose Estimation With Multiregression Loss, IEEE Trans. Multimedia 21 (2019) 1035–1046.

[2] M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods, Pattern Recognit. 71 (2017) 132–143.

[3] Z. Zhang, H. Liu, J. Shu, H. Nie, N. Xiong, On automatic recommender algorithm with regularized convolutional neural network and IR technology in the self-regulated learning process, Infrared Phys. Technol. 105 (2020), 103211.

[4] Z. Zhang, Z. Li, H. Liu, T. Cao, S. Liu, Data-driven Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology, J. Ed. Comput. Res. 58 (2020) 63–83.

[5] T. Liu, Z. Chen, H. Liu, Z. Zhang, FTIR spectral imaging enhancement for teacher's facial expressions recognition in the intelligent learning environment, Infrared Phys. Technol. 93 (2018) 213–222.

[6] Y. C. T. Yang, Y. Lin et al., FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation from a Single Image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1087-1096.

[7] L. Xu, J. Chen, Y. Gan, Head pose estimation with soft labels using regularized convolutional neural network, Neurocomputing 337 (2019) 339–353.

[8] H. Liu, H. Nie, Z. Zhang, Y.-F. Li, Anisotropic angle distribution learning for head pose estimation, Neurocomputing (2021), https://doi.org/10.1016/j.neucom.2020.09.068.

[9] W.-H. Zhang, H.-X. Deng, X.-B. Wang, Safety factor analysis for traffic accident scene based on computer simulation. 2010 International Conference On Computer Design and Applications, 2010 pp. V5–89-V85-91.

[10] H. Liu, X. Wang, W. Zhang, Z. Zhang, Y.-F. Li, Infrared head pose estimation with multi-scales feature fusion on the IRHP database for human attention recognition, Neurocomputing 411 (2020) 510–520.

[11] J. Park, H. Son, J. Lee, J. Choi, Driving Assistant Companion With Voice Interface Using Long Short-Term Memory Networks, IEEE Trans. Ind. Inf. 15 (2019) 582–590.

[12] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, IEEE Trans. Pattern Analy. Mach. Intell. 19 (2002) 711-720.

[13] G. Guo, W. Hui, D. Bell, Y. Bi, K. Greer, KNN Model-Based Approach in Classification, 2003.

[14] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Anim. Ecol. 77 (2008) 802–813.

[15] Q. Wang, G.M. Garrity, J.M. Tiedje, J.R. Cole, Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy, Appl. Environ. Microbiol. 73 (2007) 5261.

[16] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–520.

[17] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, L. Akarun, Bosphorus Database for 3D Face Analysis, Biometrics Identity Manage. (2008).

[18] M. Köstinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization, IEEE International Conference on Computer Vision Workshops (2012).

[19] B. Yang, J. Yan, Z. Lei, S.Z. Li, Fine-grained evaluation on face detection in the wild, IEEE International Conference & Workshops on Automatic Face & Gesture Recognition (2015).

[20] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L.V. Gool, Random Forests for Real Time 3D Face Analysis, Int. J. Comput. Vision 101 (2013) 437–458.

[21] X. Zhu, L. Zhen, X. Liu, H. Shi, S.Z. Li, Face Alignment Across Large Poses: A 3D Solution, IEEE Conference on Computer Vision & Pattern Recognition (2016).

[22] Z. Zhang, L. Ping, C.L. Chen, X. Tang, Facial Landmark Detection by Deep Multi-task Learning, European Conference on Computer Vision (2014).

[23] W. Wu, Q. Chen, S. Yang, W. Quan, Z. Qiang, Look at Boundary, A Boundary-Aware Face Alignment Algorithm (2018).

[24] N. Gourier, J. Crowley, Estimating Face orientation from Robust Detection of Salient Facial Structures, FG Net Workshop on Visual Observation of Deictic Gestures (2004).

[25] G. Wen, C. Bo, S. Shan, X. Chen, D. Zhou, X. Zhang, D. Zhao, The CAS-PEAL large-scale Chinese face database and baseline evaluations, IEEE Trans Syst. Man Cybernetics 38 (2008) 149–161.

[26] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, 2010.

[27] Y. Li, S. Gong, H. Liddell, Support vector regression and classification based multi-view face detection and recognition, Proc Fourth IEEE International Conference on Automatic Face & Gesture Recognition (2000).

[28] S.G. Kong, R.O. Mbouna, Head Pose Estimation From a 2D Face Image Using 3D Face Morphing With Depth Parameters, IEEE Trans. Image Process. 24 (2015) 1801–1808.

[29] G. Fanelli, J. Gall, L.J.V. Gool, Real time head pose estimation with random regression forests, Cvpr 617 (2011) 617–624.

[30] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans Image Process A Publ IEEE Signal Process Soc 11 (2002) 467.

[31] L. Wiskott, J.M. Fellous, N. Kruger, C.V.D. Malsburg, Face recognition by elastic bunch graph matching, International Conference on Image Processing (1997).

[32] T. Randen, J.H. Husoy, Filtering for texture classification: a comparative study, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1999) 291–310.

[33] B.S. Manjunath, W.Y. Ma, Texture features for browsing and retrieval of image data, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 837–842.

[34] N. Ruiz, E. Chong, J.M. Rehg, Fine-Grained Head Pose Estimation Without Keypoints, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2018 (2018) 2155–215509.

[35] M. Patacchiola, A. Cangelosi, Head Pose Estimation in the Wild using Convolutional Neural Networks and Adaptive Gradient Methods, Pattern Recogn. 71 (2017).

[36] M. Liu, Y. Li, H. Liu, 3D Gaze Estimation for Head-Mounted Eye Tracking System With Auto-Calibration Method, IEEE Access 8 (2020) 104207–104215.

[37] A. Kumar, A. Alavi, R. Chellappa, KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), 2017, pp. 258-265.

[38] B. Wang, W. Liang, Y. Wang, Y. Liang, Head pose estimation with combined 2D SIFT and 3D HOG features, Seventh International Conference on Image and Graphics 2013 (2013) 650–655.