Neurocomputing 411 (2020) 510–520

Contents lists available at ScienceDirect

Neurocomputing

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

Infrared head pose estimation with multi-scales feature fusion on the IRHP database for human attention recognition

Hai Liu [a,c], Xiang Wang [a,b], Wei Zhang [b,*], Zhaoli Zhang [a], You-Fu Li [c]

[a] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
[b] National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan 430079, China
[c] Department of Mechanical Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 5 April 2020
Revised 18 May 2020
Accepted 16 June 2020
Available online 23 June 2020
Communicated by Ma Jiayi

Keywords:
Head pose estimation
Convolutional neural network
Feature fusion
Attention recognition
Infrared image

ABSTRACT

Head pose estimation (HPE) has been widely applied in human attention recognition, robot vision and assistant driving. Infrared (IR) images bear unique advantages of being still effective under visible scenarios, which are resistance to illumination changing and strong penetration. However, the lack of public IR database hinders the research progress in the low illumination environment. In this paper, we establish a first-of-its-kind infrared head pose (IRHP) database and propose a novel convolutional neural network architecture IRHP-Net on the IRHP database. The IRHP database contains 145 kinds of IR head pose images of subjects, and benchmark evaluations are conducted on our database by the facial features-based standard HPE classification methods to prove the usability and effectiveness of IRHP database. To extract the adaptive features for the IR images, a novel multi-scale feature fusion descriptor is developed in the proposed IRHP-Net model. Quantitative assessments of the proposed method on the IRHP images demonstrate the significant improvements over the traditional methods. The new proposed IRHP-Net model can be utilized in human attention recognition and intelligent driving assistant system.

1. Introduction

Head pose estimation (HPE) is not a new field, but it is still a valuable research topic due to its diverse applications, such as intelligent assisted driving [1–4], human attention recognition [5,6], and so on. HPE plays an important role in analyzing their attention, intention and motivation. Nevertheless, few of researchers consider the night environment, in which the probability of traffic accidents may increase due to the inability to track drivers head postures and further analyze their attention. In this case, the IR sensor technology [7–9] utilizes the special IR light [10] to artificially generate radiation. The IR light can be invisible to human eyes due to its 800–1100 nm band. And it can be captured by charge-coupled device or complementary metal oxide semiconductor to radiate "illumination" scenery not impacted by environment. Based on this, the IR sensor technology [8,11–14] is more suitable for night vision. However, HPE in IR domain has received few attentions compared with estimation in visible-light images due to the lack of IR head pose database. IR face images [10] are generated based on the temperature or radiancy while visible face images are generated in accordance with the reflectivity. This essential difference causes their different texture and edge features [15], so that existing algorithms based on visible head pose database cannot be applied directly. All these meet imperious and necessary demands to create an IR database for various head pose in the nighttime or dim light environment, which can enrich the existing database and alleviate this urgent situation for HPE in the night environment.

Although creation of an IR database is a difficult and time-consuming task, it is a vital step and advance-phrase preparation in the development of assistance systems that can track the human attention. Well-labeled head pose images are of crucial importance, which are essential for training, testing and validation of algorithms for the development of robust systems. Inspired by this, an attempt is made to construct an infrared head pose (IRHP) database and propose a novel IRHP-Net structure. The details of the experiment design for precisely tagging and capturing of genuine head pose, its protocol, and annotation of head pose in this database are described in this paper. Strategies used for addressing the challenges for database creation are provided explicitly. The major novelty of the proposed method is that IRHP-Net can extract the various features and then combine them via three feature fusion descriptors. Two contributions of this paper can be summarized as follows:

* Corresponding author.
E-mail address: zwccnu@mail.ccnu.edu.cn (W. Zhang).

https://doi.org/10.1016/j.neucom.2020.06.066
0925-2312/

(1) A first-of-its-kind IR head pose database is proposed and established. And extensive facial features-based HPE methods on this database are conducted as benchmark evaluations to verify the effectiveness of our database;
(2) A novel convolutional neural network (CNN) algorithm IRHP-Net is proposed with multi-scales feature fusion, which outperforms facial features-based HPE methods;

The reminder of this paper is organized as follows. In Section 2, we briefly review the existing head pose databases and head pose estimation methods. The new IRHP database construction is described in detail in Section 3. The IRHP-Net model is proposed and optimized by mini-batch gradient descent algorithm in Section 4. Experiment results and analysis on IRHP database are provided in Section 5, and Section 6 concludes this paper.

2. Related work

2.1. Review of HPE databases

During past couple decades, several existing databases are built for HPE task. The most popular HPE databases are summarized in Table 1. And the type of all the images are visible in these databases, together with details of size, condition, pose description and released time, etc. On the one hand, the difficulty of setting up the scene and the time-consuming manual labeling of head pose discourage researchers who plan to produce databases. On the other hand, artificial databases only contain certain fewer discrete angles that cannot fully express all the gestures of head on account of the continuity of head pose. Based on these two points, the labels in some of existing head pose databases are not artificially produced but obtained by standard HPE algorithms which may lead to inaccurate labels.

To compare these databases expediently, the HPE databases can be divided into two categories according to the condition of images: laboratory (lab)-based HPE databases and web-based HPE databases. And each category is sorted on the basis of chronological order.

For the lab-based HPE databases, Pointing'04 [16] is one of the most famous, which captures the head pose images by utilizing a single camera. To obtain different head poses, they put markers in the whole laboratory and each marker corresponds to a head pose. This makes it easier to obtain different poses. Pointing'04 database contains a total of 93 discrete head pose and 2790 images. Unlike Pointing'04, CAS-PEAL-RI [17] is a subset of the entire CAS-PEAL [17] face database. It contains 30,863 images and 27 different head poses of one subject who is asked to peer upward (about 30°), peer straight ahead and peer downward (about 30°) with nine cameras in one shot that are mounted on the horizontal semicircular arm in each facing direction. Bosphorus [18] contains 13 systematic head poses and 4,666 images in total acquired by using 3D structured-light system. In addition to pose annotation, this database also provides rich repertoire of expressions and varieties of face occlusions: hair, hand and eyeglasses. Similar to CAS-PEAL-RI, CMU Multi-PIE [19] systematically captures varying head poses by installing 15 cameras and 18 flashes, which has the largest number of images (more than 750,000 images). However, the poses of CMU Multi-PIE are a bit monotonous with only 13 discrete yaw angles, which is far from reflecting the value of HPE. Compared with these databases aforementioned, BIWI Kinect [20] contains 15 k images with depth and RGB information for the first time. And it includes continuous head orientation in the range of around ± 75° for yaw, ±60° for pitch and ± 50° for roll. Ground-truth poses are provided in the form of the 3D location of the head and its rotation. For the web-based HPE databases, AFLW [21] is released at first, which provides a large-scale collection of annotated face images (about 25 k images in total) gathered from Flicker net. The pose labels are extracted by POSIT algorithm [22] and can only be utilized for coarse head pose estimation. MTFL [23] contains five kinds of head pose as well, such as left profile, left, frontal, right and right profile. Except for head pose, 12,995 face images of this database are annotated with attributes of gender, smiling and wearing glasses to optimize facial landmark detection with heterogeneous but subtly correlated tasks. To achieve the 3D head poses, 300 W-LP [24] and AFLW2000-3D [24] databases are created and released at the same time and contain 122,450 images and 2,000 images respectively in the wild. Each image is annotated with Euler angle obtained by algorithms and 68 landmarks for 3D face alignment.

2.2. Head pose estimation algorithms

Based on the HPE databases, many HPE algorithms are proposed to estimate the human pose angle. Generally, they can be classified into two groups: facial features-based HPE method and convolutional neural network (CNN)-based HPE method. The classification is determined by the difference of feature extraction manners. On the basic of different feature extraction operators, the features extracted by the former are invariant and the essential preparation for classification, while the features extracted by the latter are adaptively changing as the training process to obtain the most appropriate features.

For the facial features-based HPE methods, the accuracy of HPE depends upon the selection of suitable facial features to represent the head pose. The grayscale intensity (GI) of the facial images can be extracted according to the pixel values. The GI-based methods

Table 1
Comparison of head pose databases during past couple decades.

Databases	Samples	Condition	Pose Description	Released Time	IR/ Visible
Pointing'04 [16]	2,790 images	Lab	Yaw: [−90°, −75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°, 90°]; Pitch: [−90°, −60, −30°, −15°, 0°, 15°, 30°, 60°, 90°].	2004	Visible
CAS-PEAL-RI [17]	30,863 images	Lab	Yaw: [−30°, 0°, 30°]; Pitch:[−45°, −30°, −15°, 0°, 15°, 30°,45°]	2008	Visible
Bosphorus [18]	4,666 images	Lab	Yaw: [−90°, −45°, 10°, 20°, 30°, 45°, 90°]. Pitch: Strong upwards, Slight upwards, Slight downwards, Strong upwards. Cross rotations: [(45° yaw, 20° pitch), (45° yaw, −20° pitch)].	2009	Visible
CMU Multi-PIE [19]	More than 750,000 images	Lab	Yaw: [−90°, −75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°, 90°].	2009	Visible
BIWI Kinect [20]	15 K images	Lab	Yaw: from −75° to +75°; Pitch: from −60° to 60°; Roll: from −50° to 50°.	2013	Visible
AFLW [21]	25 K images	Web	Annotated by algorithm.	2011	Visible
MTFL [23]	12,995 images	Web	Yaw: [−60°, −30°, 0°, 30°, 60°]	2014	Visible
300 W-LP [24]	122,450 images	Web	Annotated by algorithm	2016	Visible
AFLW2000-3D [24]	2,000 images	Web	Annotated by algorithm	2016	Visible

[25,26] can work well without any special hardware except only a webcam. Thus, these methods become one of the most classical technologies for HPE task. However, GI cannot represent the facial structure and texture features well. To extract the orientation features, Wei et al. [27] propose a Gabor eigenspace-based model for HPE. Gabor-wavelets (GW) transform is applied to enhance pose feature information and capture discriminative features. Not only that, histogram of oriented gradients (HOG) [28] is one of the best features for describing head pose boundaries and geometry information and is robust to illumination and small offsets. Local binary pattern (LBP) [29] features are utilized to describe the local texture features of human face.

For the CNN-based HPE methods, CNN has shown its powerful performance. Since Patacchiola *et al.* [30] deeply study the application of CNN with dropout and adaptive gradient method in HPE for the first time and obtain ground-breaking improvements. Then, various CNN-based HPE methods have been sprung up in succession. For instance, Hopenet method [31] uses three separate losses for three Euler angles combined with two components: a binned pose classification and a regression component. QuatNet [32] establishes a CNN model and directly predicts the quaternion of each input head pose image. It has achieved impressing results. KEPLER [33] is a multi-task learning network [34] which captures structured global and local features and thus facilitating accurate key-points detection. As a by-product, KEPLER also provides the predicted 3D pose of the face accurately. Recently, FSA-Net [35], an attention network, propose a compact model for HPE from a single image using direct regression without landmarks.

3. New IRHP database for IRHP-Net model

3.1. Scene layout for IR images capture

To establish IRHP database, an appropriate photographic scene is built in our lab. A two-meter-high spacious attic is selected for marking on the tubes and the ceiling conveniently. The entire head pose recording system includes an IR camera system, tagged scenes, a chair lift in a fixed position and a network-linked laptop that is used to record the head pose. In Fig. 1, the image capture system is depicted in detail.

The key gaze zones focused by human include the strong left, slight left, right ahead, strong right, slight right, strong upwards, slight upwards, slight downwards, strong upwards and so on. Those zones have the dis-continuous property, which can be represented by the dis- continuous head poses. The yaw angle directions are set at intervals of 15° in the form of a semicircle. Many experiments show that the angle intervals of 15° can meet the time requirement in the real life assists system [3,4]. In Fig. 1, we have labeled pitch angle in each yaw angle direction, namely mark on the tubes, the ceiling and the ground. Similarly, the range of considered pitch angles are [−90°, −75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°, 90°]. Then, a chair lift, which can adjust the height of subjects to make sure their eyes are flush with the point where the pitch angle is 0°, is placed in the center of the semicircle. And a folding screen is installed behind each participant to produce a simple background which can be easily filtered through algorithms and improve performance. In the image capture scene, the accurate positions (red points in Fig. 1) are calculated through strict mathematical and geometric formulas, and then marked by the precise measuring tools.

3.2. IR camera setup

A DS-IPC-S12A-IWT (4 mm) infrared camera (850 nm band) is employed to record IR head pose images with resolution

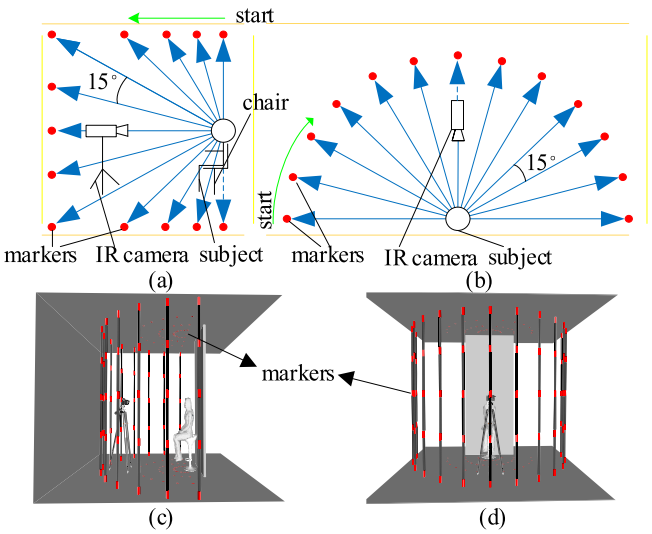

Fig. 1. IR head pose images capturation in our designed photographic scene. (a) Side view of the 2D shooting environment, where each red dot represents a pitch angle, from −90° to +90°, with an interval of 15°. (b) Top view of the 2D shooting environment, where each red dot represents a yaw angle, also from −90° to +90°, with an interval of 15°. (c) and (d) 3D display based on realistic shooting environment simulation.

1920 × 1080. The ICR infrared filter in IR camera can be utilized for day and night conversion modes. The camera is installed at a 1-meter-high tripod on the ground and 0.75 m in front of subject.

The subjects are asked to stare at each marker through rolling his/her head pose instead of turning his/her eyeball to obtain the accurate ground-truth, that is, the labels of each images are as precise as possible. One problem here is that most people are accustomed to gaze at the markers by means of turning their eyes when it comes to taking exaggerated poses. For example, subjects would peer at the marker through oculogyration unconsciously rather than head rotation when the yaw angle is 75° and the pitch angle is 75°, which will cause serious data error problems and invalidate the entire database. Therefore, a volunteer is arranged to assist subjects in completing the specified posture, which aims to reduce the error caused by artificial factors.

3.3. Subjects

Total of 40 healthy subjects (24 males and 16 females) are selected in our experiments. All people's cervical vertebrae are normal and they can complete all the specified action. Each one is informed that the experiment will be utilized for the scientific research and not for any commercial purposes. In return, they would receive compensation for participating after completing the experiment if they volunteer. There are a total of 13 × 13 marked positions (13 pitch angles per yaw angle, i.e., 169 markers) for each subject. However, the subjects could not display all the postures because of physical limitations of the human head. Accordingly, we stipulate that the pitch angle is only 0° when the yaw angle is 90° or −90°. Ultimately, for each subject, two sets of photographs (with and without glasses) are captured, each consisting of 145 images.

3.4. Glasses

The problem of occlusions is often existed in human face, such as glasses, hair. These shelters would disguise the important features when performing HPE experiments. To raise the robustness of the HPE algorithm, the subjects are asked to take two sets of

images to make the IRHP database more practical, one with glasses, and another without glasses. Subjects could choose to wear their glasses or the pairs we provide.

3.5. Protocol

Two sets of head pose images are recorded in our experiments. Details of this procedure are given as follow.

First, subjects are given an introduction to the experimental procedure, the meaning of arousal and valence, and how to roll their head.

Second, move the chair to the center of the semicircle, and then subjects should sit upright on the side opposite the tube where yaw angle is 0°. We raise or lower the chair to ensure the same horizontal line between their eyes and the marker to exactly face the marker where pitch is 0°.

Third, place a camera between the subject and the tube. It is noticeable that the camera must be level with the marker where pitch angle is 0° and subjects eyes. After installing the camera, we need to connect it on a laptop to acquire images.

Fourth, subjects are required to complete each posture in an orderly manner for the convenience of manually labelling the images and to ensure correct head location of the subjects. In other words, subjects will start to turn their head to face yaw angle of −90° and lower steadily their head to the pitch angle of −90°. Then they would raise their head in 15° intervals in the same vertical direction. After all pitch angles, the subjects restart at −60° yaw angles and so on until accomplish all head poses. The remaining two pose will be collected separately. Meanwhile, all images would be recorded and rechristened to format *No.-Yaw-Pitch* as annotation in the laptop, where *No.* is the serial number of subjects, *Yaw* and *Pitch* are corresponding head pose separately.

3.6. Design of database

After capturing all the images, each one is clipped and normalized the same size to 224 × 224 in our database. In sum, IRHP database contains 11,600 images from 40 different subjects. Each subject display 145 specified head poses. Fig. 2 shows the example images of a subject. Each subject need to accomplish the same head pose twice. Images of a subject with and without glasses are shown in Fig. 3. We randomly divide our database into three parts: training, validation and testing sets at 0.8: 0.1: 0.1 to conduct experimental comparisons fairly and meaningfully.

4. Proposed IRHP-Net method

4.1. Problem formulation

For the problem of HPE, a training set $D = \{(x_i, y_i)|i = 1, ..., N\}$ is given, where x_i represents each face image with ground-truth angle y_i and N is the number of training set. The learning task is to find a predicted function F so that it predicts $\hat{y}$=$F(x;\ \theta)$ that matches ground-truth angle y for the given image as much as possible. Thus, the final goal is to minimize the loss function $J(\theta)$,

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| = \frac{1}{N}\sum_{i=1}^{N}|F(x_i;\theta) - y_i|, \quad (1)$$

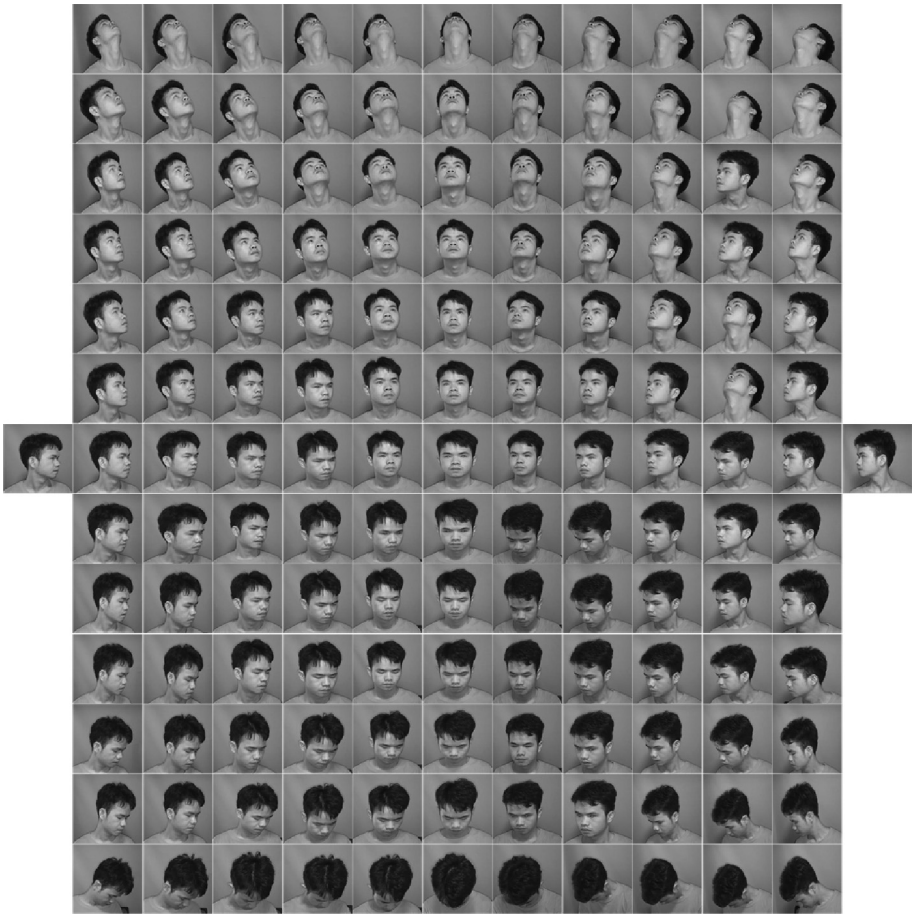

Fig. 2. IR head pose images of a subject in IRHP database. The images from left to right represent the yaw angle from −90° to +90°. The images from bottom to top represent the pitch angle from −90° to +90°.

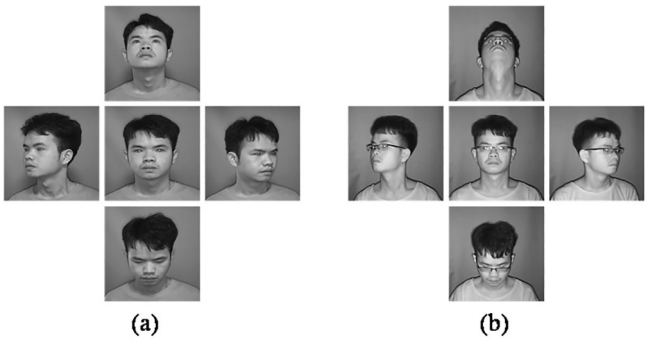

Fig. 3. Effect of the occlusions on five different head poses, i.e., (0°,0°), (0°,−45°), (0°,+45°), (−45°,0°) and (+45°,0°). (a) With glasses, (b) Without glasses.

where $\hat{y}_i = F(x_i; \theta)$ is the predicted angle for each training image x_i, θ represents the parameter of the corresponding model. However, this is an idea for address regression problems, which is inappropriate in this situation due to the discrete ground-truth, that is, types of yaw and pitch angle are all thirteen. Thus, the modified predicted function can be presented as

$$F(x_i;\theta) = \begin{bmatrix} p(\hat{y}_i = 1|x_i;\theta) \\ p(\hat{y}_i = 2|x_i;\theta) \\ \vdots \\ p(\hat{y}_i = k|x_i;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{f_j(x_i;\theta)}} \begin{bmatrix} e^{f_1(x_i;\theta)} \\ e^{f_2(x_i;\theta)} \\ \vdots \\ e^{f_k(x_i;\theta)} \end{bmatrix}, \tag{2}$$

where k is set as thirteen, $p(\hat{y}=j|x_i;\theta)$ is the probability for the j-th category and $f_j(x;\theta)$ is the corresponding value calculated through predicted function F.

Employing the maximum likelihood estimation theory, the posterior probability expression of each category can be obtained,

$$p(\hat{y}_i|x_i;\theta) = \prod_{j=1}^{k} \left\{ \frac{e^{f_j(x_i;\theta)}}{\sum_{l=1}^{k} e^{f_l(x_i;\theta)}} \right\}^{\mathbf{1}(y_i=j)}, \tag{3}$$

where $\mathbf{1}(y_j = j)$ denotes the indicator function. Then, likelihood function can be written as

$$L(\theta) = p(Y|X;\Theta) = \prod_{i=1}^{N} p(\hat{y}_i|x_i;\theta) = \prod_{i=1}^{N}\prod_{j=1}^{k} \left\{ \frac{e^{f_j(x_i;\theta)}}{\sum_{l=1}^{k} e^{f_l(x_i;\theta)}} \right\}^{\mathbf{1}(y_i=j)}, \tag{4}$$

where X is the training set face images with the corresponding ground-truth angles Y. Then, logarithmic likelihood function to facilitate easy calculation is as follow:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{N}\sum_{j=1}^{k} \mathbf{1}(y_i = j)\log \frac{e^{f_j(x_i;\theta)}}{\sum_{l}^{k} e^{f_l(x_i;\theta)}}. \tag{5}$$

The maximum value of function (5) is needed to optimize. Hence, loss function can be rewritten as:

$$J(\theta) = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{k} 1(y_i = j)\log \frac{e^{f_j(x_i;\theta)}}{\sum_{l}^{k} e^{f_l(x_i;\theta)}}\right], \tag{6}$$

The L_2-norm regularization term is added in (6) to prevent overfitting during the training phase. Lastly, the loss function can be represented as

$$J(\theta) = -\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{k} 1(y_i = j)\log \frac{e^{f_j(x_i;\theta)}}{\sum_{l}^{k} e^{f_l(x_i;\theta)}}\right] + \lambda\|\theta\|_2^2 \tag{7}$$

where the symbol λ denotes the regularization coefficient. Thus, the process of minimize the loss function can be written as,

$$\begin{aligned} \theta^* &= \operatorname{argmin}(J(\theta)) \\ &= \operatorname{argmin}\left(-\frac{1}{N}\left[\sum_{i=1}^{N}\sum_{j=1}^{k} 1(y_i = j)\log \frac{e^{f_j(x_i;\theta)}}{\sum_{l}^{k} e^{f_l(x_i;\theta)}}\right] + \lambda\theta_2^2\right) \end{aligned} \tag{8}$$

where θ^* is the optimization parameter of the solution that determines our model.

4.2. IRHP-Net structure

As the analysis, HPE task is formulated as a standard multi-classification pattern recognition problem. The IRHP-Net structure we propose can be illustrated in Fig. 4, where Conv2D (3×3, c) is 2D convolution and c is a channel parameter. BN denotes the batch normalization. ReLU is activation function; $1 \times 1 \times$ c, MP, GAP and FC mean 1×1 *convolution layer*, max *pooling layer*, *global average pooling layer* and *fully connected layer*, respectively. It should be pointed out that the activation functions after $1 \times 1 \times$ c layer are omitted in Fig. 4 to simplify the illustration.

Compared with traditional convolutional neural networks, different scales feature (Conv2d-FP1, Conv25-FP2 and Conv2d-FP3) within IRHP-Net features hierarchies are explored in this paper.

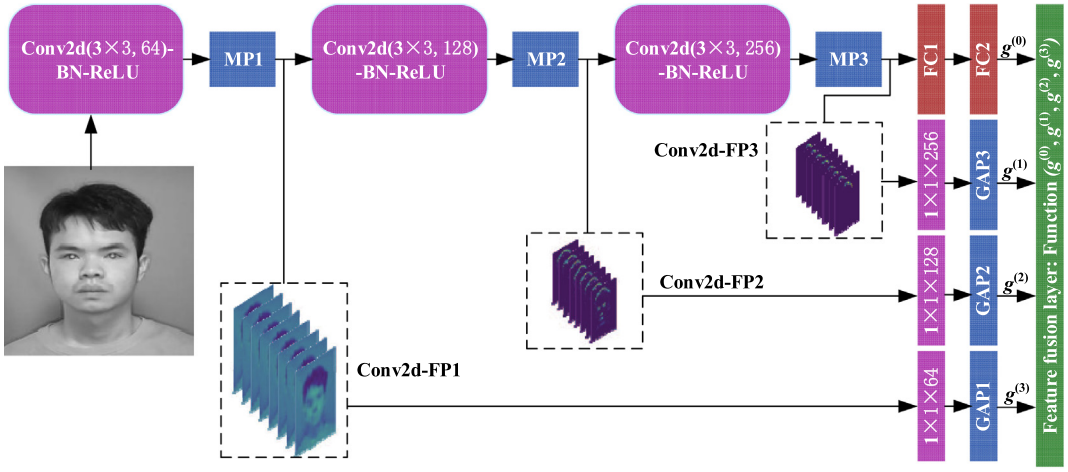

Fig. 4. Overview of the proposed IRHP-Net. The modules of Conv2d (3×3, c) integrate the convolutional layer, batch normalization layer and ReLU activation function. The modules of $1 \times 1 \times$ c are 1×1 convolutional layers. MP1, MP2 and MP3 are the max pooling layers. GAP1, GAP2 and GAP3 are the global average pooling layers. FC1 and FC2 are the fully connected layers. Conv2d-FP1, Conv2d-FP2 and Conv2d-FP3 are different scale feature maps extracted from the modules of Conv2d (3×3, c). Green box denotes the feature fusion layer.

Three feature vectors ($g^{(0)}$, $g^{(1)}$, $g^{(2)}$ and $g^{(3)}$) are combined between higher layer abstract texture information and lower layer rich and powerful spatial information. The core component of IRHP-Net is features fusion process, which determines whether representative features can be incorporated. Three feature fusion descriptors are explored: (i) Connect fusion (CF); (ii) Weighted fusion (WF); (iii) Complex vector fusion (CVF). Suppose $G^{(\zeta)}$ is the ζ-th scale feature space (ζ = 0,1,2,3) observed from Fig. 4 that defined on pattern sample space Ω. And let $g^{(\zeta)}$=(g(ζ) 1, g(ζ) 2,. . .,g(ζ) n_ζ) $\in G^{(\zeta)}$ where n_ζ is the dimension of $g^{(\zeta)}$, then for an arbitrary sample $\xi \in \Omega$, we have three feature fusion descriptors mentioned above to define ξ.

For the first descriptor, the fusion feature of ξ is obtained by connecting $g^{(0)}$, $g^{(1)}$, $g^{(2)}$ and $g^{(3)}$, that is ξ=(g(0) 1,. . ., g(0) n_0, g(1) 1,. . ., g(1) n_1, g(2) 1,. . ., g(2) n_2, g(3) 1,. . ., g(3) n_3)^T. And then the fusion feature ξ is n_ξ-dimensional vector where $n_\xi = \sum_{\zeta=0}^{3} n_\zeta$.

For the second descriptor, normalization pretreatment for $g^{(\zeta)}$ is conducted before the weighted fusion calculation, which eliminates the adverse effects caused by singular sample data. The $g^{(\zeta)}$ is defined in detail as,

$$g^{(\zeta)} = Norm\left(g^{(\zeta)}\right) = \frac{1}{\sum_{i=1}^{n_\zeta} e^{g_i^\zeta}}\left(e^{g_1^\zeta}, e^{g_2^\zeta}, \ldots, e^{g_{n_\zeta}^\zeta}\right). \tag{9}$$

Then, $\sum_{\zeta=0}^{3} \rho_\zeta g^{(\zeta)}$ is utilized to represent the weighted feature vector ξ, where $\rho_\zeta \in (0, 1)$ is a hyper-parameter and $\sum_{\zeta=0}^{3} \rho_\zeta = 1$. It is worth noting that the particular case is the most commonly popular feature extracted by CNN while ρ_1, ρ_2 and ρ_3 are all equal to 0.

For the last descriptor, its calculation procedure is similar with the WF feature vector. The complex feature vector can be written as $\xi = g^{(0)} + i\sum_{\zeta=1}^{3} g^{(\zeta)}$, where i is an imaginary unit and $g^{(\zeta)}$ (ζ = 0,1,2,3) is the vector after normalization preprocessing by function (9).

Note that for the descriptors WF and CVF, padding the lower dimension ones with zeros until the dimensions of $g^{(\zeta)}$ are equal while their dimensions are not equal. Consequently the dimension of ξ is max(n_ζ | ζ = 0, 1, 2, 3). For example, if $g^{(0)} = (g_1^{(0)}, g_2^{(0)})^T$, $g^{(1)} = (g_1^{(1)}, g_2^{(1)}, g_3^{(1)})^T$, $g^{(2)} = (g_1^{(2)}, g_2^{(2)}, g_3^{(2)}, g_4^{(2)})^T$ and $g^{(3)} = (g_1^{(3)}, g_2^{(3)}, g_3^{(3)})^T$, $g^{(0)}$ and $g^{(3)}$ are first turned into $(g_1^{(0)}, g_2^{(0)}, 0, 0)^T$ and $(g_1^{(3)}, g_2^{(3)}, g_3^{(3)}, 0)^T$, respectively. Then the feature fusion vector ξ can be calculated by descriptors (ii) and (iii) mentioned above. Finally, the fusion feature can be obtained to calculate the predicted classification result according the descriptor (ii) or (iii).

4.3. Training strategy

A total of 9,280 random face image of IRHP database are fed into the proposed IRHP-Net for training. To address the large-scale image representation problem learning, mini-batch gradient descent (MBGD) algorithm is adopted to optimize the objective function in (8). According to the MBGD strategy, θ is updated by the rule as follows,

$$\arg\min_\theta(J(\theta)) \Rightarrow \begin{cases} \theta^* = \theta - \eta\frac{\partial J(\theta)}{\partial \theta} \\ = \theta + \eta\frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{k}\mathbf{1}(y_i = j)\left\{\frac{\partial f_j(x_i;\theta)}{\partial \theta}\right\} \\ -\eta\frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{k}\mathbf{1}(y_i = j)\left\{\frac{\sum_l^k e^{f_l(x_i;\theta)}\frac{\partial f_l(x_i;\theta)}{\partial \theta}}{\sum_l^k e^{f_l(x_i;\theta)}}\right\} - 2\eta\lambda\theta \end{cases}, \tag{10}$$

where η is the learning rate in gradient-based optimization algorithm. Then, the IRHP database and source code will be available upon request. The algorithm described in this section can be summarized as follows.

Algorithm 1: Training strategy for the proposed IRHP-Net method.

Input: the training set of IRHP database
Set: iteration number: max-*Iter*; batch size *b*; ρ_ζ; λ; η.
1: Initialize IRHP-net parameters θ via random number or Gaussian distribution.
2: IRHP-Net forward propagation
3: **while** epoch <= max-*Iter* do:
Sample a mini batch with size s from IRHP database.
Adopt loss function (10) to optimize parameters θ, i.e., update θ with Adam optimizer.
end while
Output: optimization parameter θ^*

5. Experiment and analysis

In this section, we first verify the effectiveness of the new IRHP database by the serval traditional machine learning methods. Then, the proposed IRHP-Net is executed on the IRHP database. Our experiments are implemented on a workstation with an NVIDIA TITAN RTX, 16 Intel (R) Core (TM) i9-9900 K (3.60 GHz) CPUs and 64 GB RAM. TensorFlow, an open-source platform developed by Google is utilized to execute our proposed IRHP-Net model. 1000 epochs are used to train the network with Adam optimizer with the initial learning rate of 10^{-3}, β_1 = 0.9, β_2 = 0.999 and $\varepsilon = 10^{-8}$.

5.1. Metrics

To compare the performance of different methods on HPE, two metrics (accuracy (*ACC*) and mean absolute error (*MAE*)) are adopted. *ACC* is formulated as follows,

$$ACC = \frac{1}{k}\sum_{i=1}^{k} acc_i = \frac{1}{k}\sum_{i=1}^{k}\frac{N_{i-ture}}{N}, \tag{11}$$

where k is the number of cross-validations, acc_i is the accuracy in the *i*-th validation, $N_{i\text{-}true}$ denotes the number of correct predictions in the *i*-th validation, and *N* denotes the total number of validation samples. *MAE* is defined as follows,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(|\hat{y}_i - y_i|), \tag{12}$$

where y_i represents the ground-truth of yaw or pitch angles of the *i*-th sample, $\hat{y}_i$ is the corresponding prediction. The smaller the *MAE* value is, the higher the *ACC* achieves.

5.2. Compared methods

To verify the usability and effectiveness of IRHP database, serval facial features-based methods are introduced on the new database. GI, GW, HOG, LBP feature extraction techniques are conducted as benchmark evaluation in our experiments. After extracting the face feature, several classification techniques are carried out.

Before introducing classification techniques, by reason of the overabundance of the dimensionality of the input data obtained by the feature extraction techniques such as GI, GW, HOG and LBP, it will be at the expense of great computational complexity. Hence, dimensionality reduction is a vital step before feeding the

data to a classifier. Principal component analysis (PCA) [36], as a dimensionality reduction tool, can well reduce the dimension of the features with minimal information loss and is widely utilized.

To classify head pose, three classifiers PCA + KNN (K nearest neighbor) [37], PCA + LDA (linear discriminant analysis) [38] and PCA + NB (Naïve Bayes) [39] are conducted to extract the IR image features. These classification techniques without PCA are also carried out to compare and analyze their performances. Apart from the above individual classifiers, random forest (RF) [40,41] algorithm is conducted in our experiments. As a typical representative of ensemble learning, it implements classification tasks by constructing and combining decision trees and has better robustness.

After verifying the usability and effectiveness of IRHP database, our algorithm is compared with the following state-of-the-art methods for HPE. CNNs [42] addresses the problem of HPE with two degrees of freedom (pitch and yaw) using a low-resolution image. MLD [43], firstly introduced by Geng *et al.*, utilizes a Gaussian distribution function to construct a label distribution for each image. DLDL [44] propose a deep convolutional neural network method based MLD and obtain promising performance. Chen *et al.* [45] focus on cumulative attribute space regression for HPE and propose two methods: indepCA and CartCA/MvCA, and achieve competitive results.

5.3. Benchmark evaluation on IRHP database

Feature selection is an important step before feeding the data to a classifier. The involved four facial feature extractor and four classification techniques are carried out and compared on the IRHP database. And a dimensionality reduction tool PCA is utilized to reduce the dimension of the facial features for reducing computation time complexity. As observed from Tables 2–5, the time-cost of methods with and without PCA, the *ACC* and *MAE* of two angles are provided.

First, from the perspective of time-cost with and without PCA, the former takes much longer to model than the latter by using KNN, LDA and NB methods. However, two consumed almost the same time while applying RF, and even the latter would consume more time than the former. The reason may be that RF are integrated by many classifiers, namely decision trees. In our experiments, the number of decision trees are set as 400, and the

Table 2
Comparison of time-cost, *ACC* and *MAE* by different methods with grayscale intensities (*GI*).

Methods	Time-cost(s)	*ACC* (%)			*MAE* (°)		
		Yaw	Pitch	All	Yaw	Pitch	All
KNN	8.46	38.42	42.61	40.52	11.34	10.22	10.78
PCA + KNN	0.07	38.55	43.35	40.95	11.21	10.27	10.74
LDA	43.40	30.30	42.24	36.27	24.70	12.10	18.40
LDA + PCA	0.25	35.34	44.58	39.96	19.86	10.49	15.18
NB	2.29	18.60	35.96	27.28	36.43	14.80	25.62
NB + PCA	0.03	25.62	32.39	29.01	28.13	15.53	21.83
RF	108.04	54.93	48.03	51.48	8.40	8.94	8.67
RF + PCA	114.17	39.53	45.94	42.74	15.18	10.49	12.84

Table 3
Comparison of time-cost, *ACC* and *MAE* by different methods with Gabor wavelets (*GW*).

Methods	Time (s)	*ACC* (%)			*MAE* (°)		
		Yaw	Pitch	All	Yaw	Pitch	All
KNN	298.18	38.73	41.01	38.73	12.91	13.13	13.02
PCA + KNN	0.76	35.96	40.89	38.43	13.93	13.61	13.77
LDA	2546.32	43.65	47.69	45.67	13.64	10.81	12.23
LDA + PCA	7.40	49.38	52.71	51.05	11.19	7.57	9.38
NB	55.64	14.66	26.48	20.57	37.20	22.50	29.85
NB + PCA	0.28	15.63	25.30	20.47	34.79	21.18	27.99
RF	102.19	46.43	47.41	46.92	10.05	9.44	9.75
RF + PCA	166.80	36.08	39.90	37.99	17.83	12.58	15.21

Table 4
Comparison of time-cost, *ACC* and *MAE* by different methods with histogram of oriented gradients (*HOG*).

Methods	Time (s)	*ACC* (%)			*MAE* (°)		
		Yaw	Pitch	All	Yaw	Pitch	All
KNN	0.0080	52.09	54.24	53.17	15.68	11.24	13.46
PCA + KNN	0.0040	46.18	48.65	47.72	17.25	14.63	15.94
LDA	0.0229	22.86	26.41	24.64	28.32	25.34	26.83
LDA + PCA	0.0120	21.01	24.78	22.90	35.19	20.31	27.75
NB	0.0070	15.66	25.63	20.65	37.49	30.45	33.97
NB + PCA	0.0020	12.68	25.47	19.08	40.63	32.95	36.79
RF	10.1769	53.33	58.41	55.87	13.50	10.37	11.94
RF + PCA	6.7909	45.20	50.36	47.78	19.36	15.41	17.39

Table 5
Comparison of time-cost, *ACC* and *MAE* by different methods with local binary pattern (*LBP*).

Methods	Time (s)	*ACC* (%)			*MAE* (°)		
		Yaw	Pitch	All	Yaw	Pitch	All
KNN	8.25	32.64	34.25	33.45	23.70	21.74	22.72
PCA + KNN	0.37	23.40	26.11	24.76	28.64	24.86	26.75
LDA	40.98	33.62	42.61	38.12	18.52	12.09	15.31
LDA + PCA	6.81	42.88	45.00	43.94	16.23	13.20	14.72
NB	2.53	21.55	32.88	27.22	34.50	17.44	25.97
NB + PCA	0.15	14.84	12.93	13.89	38.95	42.12	10.54
RF	83.62	36.20	35.34	35.77	18.86	17.34	18.10
RF + PCA	94.32	33.15	33.50	33.33	21.06	22.37	21.72

maximum depth of search is set as 30. This results in the time consumed by using the abundant decision tree classifiers make the reduction time for applying the PCA technique negligible, even if it may increase the time consumption due to the alteration of extracted features.

Then, from the overall performances of the *ACC* and *MAE* of HPE, they have obvious benefits and drawbacks with respect to different feature extractors. In Tables 2 and 4, the best results for *ACC* obtained by RF classifier, using GI and HOG features respectively, are quite competitive. Moreover, RF classifier with GI features yields better *MAE*, its yaw angle about 8.40° and pitch angle about 8.94°, comparing with HOG features. Meanwhile, PCA + LDA outperforms the rest classification techniques in our experiments when using GW and LBP features in Tables 3 and 5.

It is worth noting that PCA + LDA with GW features harvests the best results with *MAE* of pitch angle about 7.57° in all facial features-based methods. However, NB and PCA + NB do not perform well whether applying the GI, GW, HOG or LBP feature extractor. The reason behind the failure of classifying correctly may be explained by the fact that NB model assumes that the attributes are independent of each other. This assumption is often unacceptable in practical applications. When the number of attributes is large or the correlation between attributes is large, the classification effect is not good. Indeed, the attributes of head images are very abundant and they have evident spatial characteristics, namely spatial correlation. Thus, NB and PCA + NB classifiers is not an appropriate tool for finding the hyperplane that minimizes the intra-class scatter while maximizing the inter-class scatter. But in real-time applications such as assistance driving system, these classifiers need to be improved to apply since their computational complexity is very low compared to other classifiers.

5.4. Sensitivity analysis of IRHP-Net structure

In this scenario, the comparison between our method and other state-of-the-art algorithms on IRHP database is shown in Table 6. As can be seen, we can clearly observe that the proposed IRHP-Net outperforms many other competitive methods. It indicates that our method can effectively explore the distinguishing features across different categories by fusing different scales feature information and distilling the latent knowledge.

For further analysis of CF, WF and CVF of IRHP-Net on IRHP database, the *ACC* and the *MAE* of each feature fusion descriptors are summarized in Table 7. It can be seen that the *ACC* and *MAE*

Table 7
Results of IRHP-Net method with three different feature fusion descriptors.

Feature fusion	*ACC* (%)			*MAE* (°)		
	Yaw	Pitch	All	Yaw	Pitch	All
CF	76.53	77.26	76.90	6.23	5.32	5.78
CVF	83.24	83.47	83.36	5.53	5.14	5.34
WF	**84.13**	**87.69**	**85.91**	**5.16**	**4.28**	**4.72**

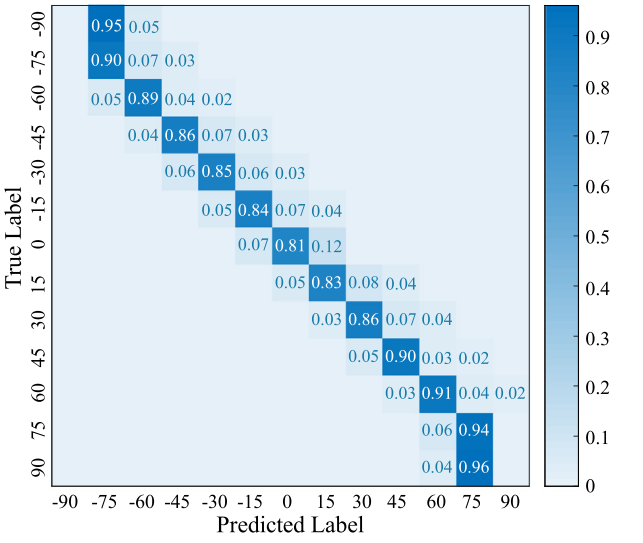

Fig. 5. Representation of the confusion matrix for each angles including Yaw and Pitch angle, which is the total accuracy for each angle. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class.

Table 6
Comparison of ACC and *MAE* on IRHP database by different algorithms.

Methods	*ACC* (%)			*MAE* (°)		
	Yaw	Pitch	All	Yaw	Pitch	All
CNNs [39]	73.78	75.90	74.84	7.55	6.13	6.84
DLDL [41]	80.26	83.61	81.94	6.18	5.87	6.03
MLD [40]	70.69	73.76	72.23	8.56	7.21	7.89
IndepCA(HOG) [42]	72.83	75.34	74.09	7.83	6.77	7.30
CartCA/MvCA [42]	75.49	77.39	76.44	7.21	6.15	6.68
Ours	**84.13**	**87.69**	**85.91**	**5.16**	**4.28**	**4.72**

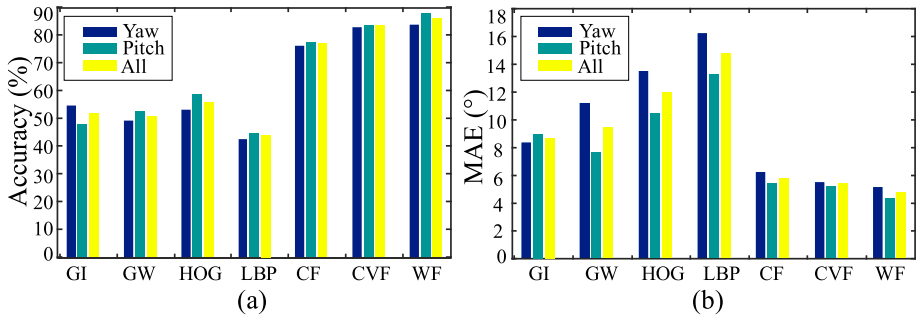

Fig. 6. Representation of the performance of different methods.

of WF are 85.91% and 4.72°, respectively, which are significantly better than the other two. To further analyze the classification details for each angle, the normalized confusion matrix for WF is reported in Fig. 5. The darker blocks are arranged on the main diagonal. It means that the prediction has a higher accuracy. The erroneous prediction results are limited to the adjacent range of the correct classification. Generally, IRHP-Net model achieves the better performance than the facial features-based methods, as manifested in Fig. 6. Three descriptors of IRHP-Net have prodigious improvement relative to the benchmark methods. Compared with best result of benchmark evaluation, WF has 30.04–41.97% raising in *ACC* and 3.95°–10° in *MAE*. Hence, the experimental results illustrate that IRHP-Net model has the best performance on the IRHP database. The reason of the excellent performance may be that the facial features extracted by IRHP-Net are of automatic adjustment to search for the most representative features along with the training process; and three different scales features are fused to represent head pose features which contain high-layer semantic information and low-layer abstract features of some edge corners as shown in Figs. 7–9.

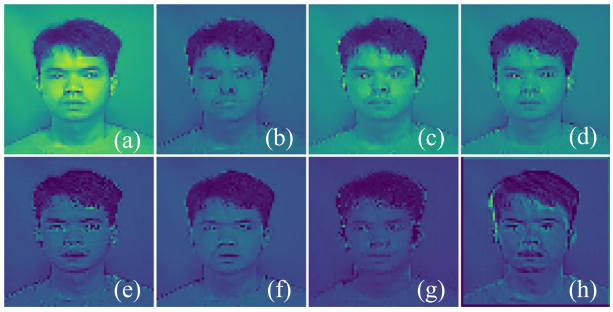

Fig. 7. Feature maps after Conv1 in the proposed IRHP-Net method.

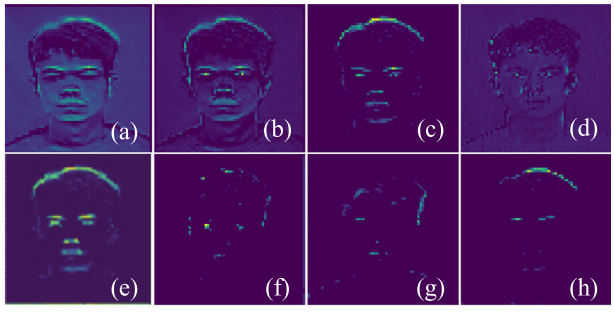

Fig. 8. Illustration of feature maps after Conv2.

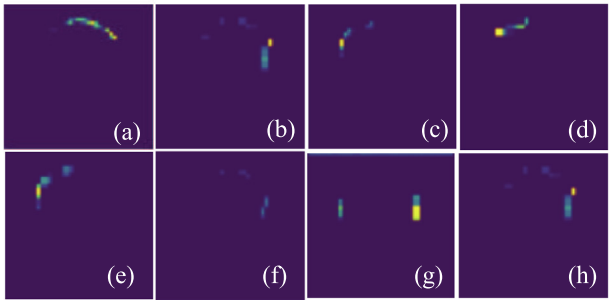

Fig. 9. Feature maps in Conv3.

The regularization technique is introduced to boost the generalization capabilities of IRHP-Net model. And to analyze the effect of the regularization coefficient λ, extensive experiments are performed on three feature fusion descriptors mentioned above with $\lambda \in (0, 1)$. The *ACC* and *MAE* values of different sampled λ are depicted in Fig. 10. It can be observed that the best result is obtained by the proposed method when $\lambda = 0.6$. And on the whole, the tendency of the *ACC* is to increase and then decrease, while *MAE* is the opposite. Consequently, the regularization coefficient is suggested as $\lambda \in (0.55, 0.65)$ in the real applications.

5.5. Application for driving assistant

To demonstrate the generalization practicability of IRHP-Net model in real-life applications, the model trained on our IRHP database is used to estimation the driver head pose in a car at night. Fig. 11 visualizes the results of different test images captured from the infrared camera. One point should be noted that the roll angle would be obtained by taking the cross product of yaw angle and pitch angle. The result indicates that the IRHP database and IRHP-Net we propose have great application prospect and reference value.

6. Conclusion

In this paper, we propose a novel convolutional neural network architecture IRHP-Net with multi-scales feature fusion for head pose estimation task. Firstly, a first-of-its-kind infrared head pose (IRHP) database is established for HPE in the night environment. Based on the IRHP database, we design three convolution layers to extract the infrared image texture information in the IRHP-Net model. Three-scales features can reveal the higher layer abstract texture information and lower layer rich and powerful spatial information. In the feature fusion layer, three feature fusion descriptors (CF, WF, CVF) are constructed to combine the multi-scale features adaptively for infrared HPE images. Then, benchmark evaluations are conducted on our database by the facial features-based standard HPE classification methods to prove the usability

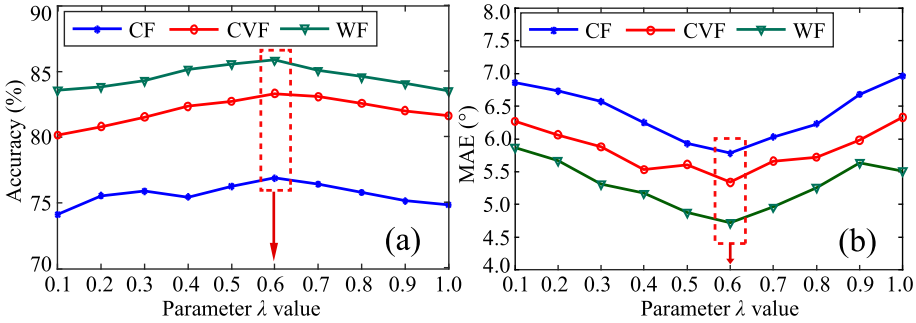

Fig. 10. Performance of IRHP-Net measured by *ACC* and *MAE* with increasing from 0.1 to 1.

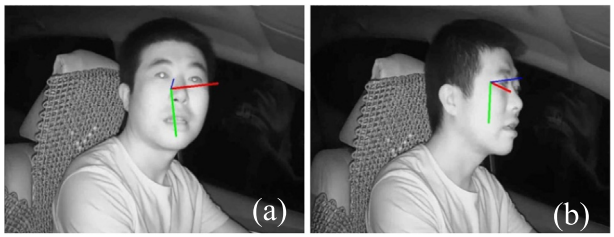

Fig. 11. HPE for assistance driving at night with few ambient light sources. The blue axis points toward the front of the face, green pointing downward, and red pointing to the side.

and effectiveness of IRHP database. Extensive experiments verify the advantages of our IRHP-Net and demonstrate state-of-the-art performances on IRHP database. In future, we will examine the infrared video scenario for HPE in the night environment.

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 under Grant 61875068, Grant 61873220, Grant 61977031 and Grant 61505064, the National Key Research and Development Program of China under Grant 2018YFB1004505, the Research Grants Council of Hong Kong under Project CityU 11205015 and Project CityU 11255716, and the Fundamental Research Funds for the Central Universities under Grant CCNU20ZT017 and Grant CCNU2020ZN008.

References

[1] 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.
[2] G. Christian, J. López, A. Pinilla, Driver behavior classification model based on an intelligent driving diagnosis system, in: 2012 15th International IEEE Conference on Intelligent Transportation Systems, 2012, pp. 894–899.
[3] S. Jha, C. Busso, Analyzing the relationship between head pose and gaze to model driver visual attention, in: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 2016, pp. 2157–2162.
[4] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation and augmented reality tracking: an integrated system and evaluation for monitoring driver awareness, IEEE Trans. Intell. Transp. Syst. 11 (2010) 300–311.
[5] J. Chen, N. luo, Y. Liu, L. Liu, A hybrid intelligence-aided approach to affect-sensitive e-learning, Computing 98 (2016) 215–233.
[6] Y. Yan, E. Ricci, R. Subramanian, G. Liu, O. Lanz, N. Sebe, A multi-task learning framework for head pose estimation under target motion, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 1070–1083.
[7] C. Marshall, T. Parker, T. White, Infrared sensor technology, in: Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society, 1995, pp. 1715–1716.
[8] T. Liu, H. Liu, Z. Chen, A.M. Lesgold, Fast blind instrument function estimation method for industrial infrared spectrometers, IEEE Trans. Ind. Inf. 14 (2018) 5268–5277.
[9] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, et al., Infrared and visible image fusion via detail preserving adversarial learning, Information Fusion 54 (2020) 85–98.
[10] T. Liu, Y.F. Li, H. Liu, Z. Zhang, S. Liu, RISIR: rapid infrared spectral imaging restoration model for industrial material detection in intelligent video systems, IEEE Trans. Ind. Inform. (2019), https://doi.org/10.1109/TII.2019.2930463.
[11] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Information Fusion 45 (2019) 153–178.
[12] T. Liu, H. Liu, Y. Li, Z. Zhang, S. Liu, Efficient Blind Signal Reconstruction With Wavelet Transforms Regularization for Educational Robot Infrared Vision Sensing, IEEE/ASME Trans. Mechatron. 24 (2019) 384–394.
[13] T. Liu, H. Liu, Y. Li, Z. Chen, Z. Zhang, S. Liu, Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing, IEEE Trans. Ind. Inf. 16 (2020) 544–554.
[14] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: a generative adversarial network for infrared and visible image fusion, Inform. Fusion 48 (2019) 11–26.
[15] J. Ma, H. Xu, J. Jiang, X. Mei, X. Zhang, DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.
[16] N. Gourier, J. Crowley, "Estimating Face orientation from Robust Detection of Salient Facial Structures," FG Net Workshop on Visual Observation of Deictic Gestures, 2004.
[17] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, et al., The CAS-PEAL large-scale chinese face database and baseline evaluations, IEEE Trans. Syst., Man, Cybern. – Part A: Syst. Humans 38 (2008) 149–161.
[18] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, et al., Bosphorus database for 3D face analysis, in: European Workshop on Biometrics and Identity Management, 2008, pp. 47–56.
[19] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (2010) 807–813.
[20] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L. Van Gool, Random forests for real time 3D face analysis, Int. J. Comput. Vision 101 (2013) 02/01.
[21] 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, in: IEEE International Conference on Computer Vision Workshops, 2011, pp. 2144–2151.
[22] D.F. Dementhon, L.S. Davis, Model-based object pose in 25 lines of code, Int. J. Comput. Vision 15 (1995) 123–141.
[23] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: European conference on computer vision, 2014, pp. 94–108.
[24] X. Zhu, Z. Lei, X. Liu, H. Shi, S.Z. Li, Face alignment across large poses: a 3D solution, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 146–155.
[25] H. Liu, L. Yan, Y. Chang, et al., Spectral deconvolution and feature extraction with robust adaptive Tikhonov regularization, IEEE Trans. Instrum. Meas. 62 (2013) 315–327, https://doi.org/10.1109/tim.2012.2217636.
[26] 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.
[27] Y. Wei, L. Fradet, T. Tan, Head pose estimation using Gabor eigenspace modeling, in: Proceedings. International Conference on Image Processing, 2002, pp. 144–155.
[28] B. Wang, W. Liang, Y. Wang, Y. Liang, Head pose estimation with combined 2D SIFT and 3D HOG features, in: 2013 Seventh International Conference on Image and Graphics, 2013, pp. 650–655.

[29] H. Liu, Y. Li, Z. Zhang, et al., Blind Poissonian reconstruction algorithm via curvelet regularization for an FTIR spectrometer, in: Opt. Express, 2018, pp. 22837–22856.
[30] M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods, Pattern Recogn. 71 (2017) 132–143.
[31] N. Ruiz, E. Chong, J.M. Rehg, Fine-grained head pose estimation without keypoints, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2074–2083.
[32] 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.
[33] 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.
[34] Y. Liu, J. Zeng, S. Shan, Z. Zheng, Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 458–465.
[35] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, Y.-Y. Chuang, 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.
[36] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417.
[37] H. Liu, Z. Zhang, J. Sun, et al., Blind spectral deconvolution algorithm for Raman spectrum with Poisson noise, in: Photon. Res. 2015, pp. 168–171.
[38] Y. Ling, X. Yin, S.M. Bhandarkar, Sirface vs. Fisherface: recognition using class specific linear projection pp. III-885, Proceedings 2003 International Conference on Image Processing, 2003.
[39] K. Liu, T. Wong, Naïve Bayesian classifiers with multinomial models for rRNA taxonomic assignment, IEEE/ACM Trans. Comput. Biol. Bioinf. 10 (2013) 245–256.
[40] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J. Animal Ecol. 77 (2008) 802–813.
[41] Y. Liu, Z. Xie, X. Yuan, J. Chen, W. Song, Multi-level structured hybrid forest for joint head detection and pose estimation, Neurocomputing (2017) 206–215.
[42] S. Lee, T. Saitoh, Head pose estimation using convolutional neural network, IT Convergence and Security 2017 (2018) 164–171.
[43] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1837–1842.
[44] B.B. Gao, C. Xing, C.W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Trans. Image Process. PP (2016) 2825–2838.
[45] K. Chen, K. Jia, H. Huttunen, J. Matas, J.-K. Kämäräinen, Cumulative attribute space regression for head pose estimation and color constancy, Pattern Recogn. 87 (2019) 29–37.

Hai Liu received the M.S. degree in applied mathematics from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010, and the Ph.D. degree in pattern recognition and artificial intelligence from the same university, in 2014.

Since June 2017, he has been an Assistant Professor with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan. He was a “Hong Kong Scholar” postdoctoral fellow with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong, where he was hosted by the Professor You-Fu Li; he held the position two years till March 2019. He has authored more than 60 peer-reviewed articles in international journals from multiple domains such as patternrecognition, robot vision, and image processing. More than six papers are selected as the ESI highly cited papers.

His current research interests include big data processing, artificial intelligence, educational information technology, optical data processing and pattern recognition. Dr. Liu has been frequently serving as a reviewer for more than six international journals including the *IEEE Transations on Industrial Informatics, IEEE Transations on Cybernetics, IEEE/ASME Transactions on Mechatronics,* and *IEEE Transations on Knowledge and Data Engineering*. He is also a Communication Evaluation Expert for the National Natural Science Foundation of China. He wins the first prize of Science and Technology Progress Award by the Ministry of Education of China in 2019.

Xiang Wang received the B.S. degrees from Nanchang University, Nanchang, China, in 2018. He is currently pursuing the M.S. degree with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, under the supervision of Professor Hai Liu. His research interests include head pose estimation, pattern recognition, machine learning, and robot vision.

Wei Zhang is currently an associate professor in the National Engineering Research Center for E-Learning and National Engineering Laboratory for Educational Big Data at the Central China Normal University. He holds a Ph.D. degree from Huazhong University of Science and Technology. His research interests include computer applications, big data analysis, data mining, and application of information technology in education. He published more than 40 papers in the academic journals, including 20 papers indexed by SSCI, SCI, EI, ISTP.

Zhaoli Zhang received the M.S. degree in Computer Science from Central China Normal University, Wuhan, China, in 2004, and the Ph.D. degree in Computer Science from Huazhong University of Science and Technology in 2008. He is currently a professor in the National Engineering Research Center for E-Learning, Central China Normal University. His research interests include signal processing, knowledge services and software engineering. He is a member of IEEE and CCF (China Computer Federation).

You-Fu Li received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, and the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, Oxford, U.K., in 1993. From 1993 to 1995, he was a Research Staff in the Department of Computer Science, University of Wales, Aberystwyth, U.K. He joined the City University of Hong Kong, Hong Kong, in 1995, and is currently a Professor in the Department of Mechanical and Biomedical Engineering. His current research interests include robot sensing, robot vision, three-dimensional vision, and visual tracking. Professor Li has served as an Associate Editor of the *IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING* and is currently an Associate Editor of the *IEEE ROBOTICS AND AUTOMATION MAGAZINE*. He is an Editor of the IEEE Robotics and Automation Society Conference Editorial Board, and the IEEE Conference on Robotics and Automation.