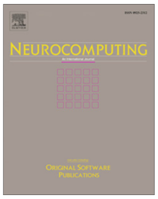
Neurocomputing 436 (2021) 210–220

Contents lists available at ScienceDirect

Neurocomputing

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



NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom

Tingting Liu[a,b], Jixin Wang[b], Bing Yang[a,*], Xuan Wang[c]

[a] School of Education, Hubei University, 368 Youyi Road, Wuhan 430062, Hubei, China
[b] Collaborative Innovation Centre for Information Technology and Balanced Development of K-12 Education, Central China Normal University, 152 Luoyu Road, Wuhan, Hubei, China
[c] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 22 August 2020
Revised 25 November 2020
Accepted 19 December 2020
Available online 5 January 2021
Communicated by Steven Hoi

Keywords:
Head pose estimation
Deep learning
On-task behavior understanding
Infrared imaging
Soft label
Human–computer interaction
K-12 education

ABSTRACT

Head pose estimation (HPE) under active infrared (IR) illumination has attracted much attention in the fields of computer vision and machine learning. However, IRHPE often suffers from the problems of low-quality IR images and ambiguous head pose. To tackle these issues, we propose a novel nonuniform Gaussian-label distribution learning network (NGDNet) for the HPE task. First, we reveal the essential properties from two different perspectives: 1) two head pose images change differently in pitch and yaw directions with the same angle increasing on the central pose; 2) the IR head pose variation first increases and then decreases in the pitch direction. Subsequently, the first property indicates the pose image label as a nonuniform label distribution (Gaussian function) with different long and short axes. The second property is leveraged to determine the distribution size in accordance with the similarities of adjacent hand poses. Lastly, the proposed NGDNet is verified on a new IRHPE dataset, which is built by our research group. Experimental results on several datasets demonstrate the effectiveness of the proposed model. Compared with conventional algorithms, our NGDNet model achieves state-of-the-art performance with 77.39% on IRHPE, 99.08% on CAS-PEAL-R1, and 87.41% on Pointing'04. Our code is publicly available at https://github.com/TingtingSL/NGDNet.

1. Introduction

Infrared head pose estimation (IRHPE) is a fundamental problem in a weak illuminated environment for computer vision. IRHPE can be widely applied in many fields, such as human–machine interaction [1,2], safe driving [3,4], and activity understanding [5–7]. Human head pose can provide a key cue in analyzing human attention, intention, and motivation. In practice, estimating the head pose of a person in deficient illumination is more challenging than estimating the head pose in an environment with sufficient light. Over the past decade, researchers have made remarkable efforts on head pose estimation (HPE) with various techniques [8–12]. To this end, we develop a new i nfrared (IR) head pose dataset to reveal the essential attributes in IRHPE. Unlike visible images, IR images often suffer the problems of random noise and being textureless and blurry. Traditional HPE methods are powerless on the developed IR dataset. This limitation motivates us to propose a novel HPE network to adapt to noisy and blurry IR images.

IR head pose images are difficult to classify with high accuracy due to image degradation and ambiguous problems. In Fig. 1, we demonstrate low-quality, noisy, and blurry IR images. The degradation usually leads to incorrect or inconsistent labeling, which will hinder the HPE process, especially for the data-driven deep learning (DL)-based HPE task. Three types of problems will be caused by a low-quality IRHPE. First, useful head pose features cannot be learned directly by the proposed network. Second, an incorrect label may result in disconvergence in the iteration optimization process.

To tackle the problems, we present a simply yet efficient network, called nonuniform Gaussian-label distribution learning network (NGDNet), to suppress the low quality for IRHPE. In this paper, two key findings are revealed, which are the difference and nonuniformity properties. On the basis of the two properties, we propose an NGDNet model for HPE in the human–computer interaction. The model is learned via an end-to-end CNN, which

[*] Corresponding author.
E-mail address: yangbing.cn88@gmail.com (B. Yang).

https://doi.org/10.1016/j.neucom.2020.12.090
0925-2312/© 2021 Elsevier B.V. All rights reserved.

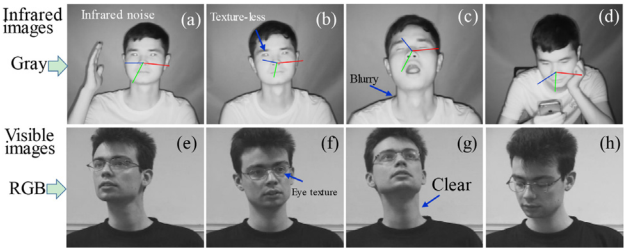


Fig. 1. Texture-reduced and low-quality IR head pose images compared with visible images. IRHPE datasets are extremely difficult to annotate because of ambiguous head poses and low-quality images.

utilizes covariance pooling to capture second-order features. Overall, the contributions can be summarized into three aspects:

1) Two important essential properties are revealed: i) if we increase the same angle on the central pose in pitch and yaw directions, two head pose images change differently; ii) the similarity value of adjacent IR head poses first increases and then decreases in the pitch direction. The two properties convert the original head pose annotations into a nonuniform Gaussian distribution.
2) A Gaussian-like multilabel distribution is designed to reduce the effects of low-quality IR images. The method is robust and effective for lost pose in head pose images and extremely flexible for arbitrary inputs.
3) We extensively validate our NGDNet on the IRHPE dataset and several public datasets. The proposed method achieves state-of-the-art performance with accuracy of 77.39% on IRHPE, 99.08% on CAS-PEAL-R1, and 87.41% on Pointing'04 and reaches new records on them.

The rest of this article is organized as follows. In Section 2, we briefly review the related works on the head pose estimation. Section 3 describes the Gaussian-like label construction method for head pose estimation. Then, the proposed NGDNet model is illustrated in section 4. Section 5 reports the experimental results and discussion. Finally, Section 6 gives the conclusion of this study.

2. Related work

2.1. Problem formulation

IRHPE predicts the head pose of a person in an IR image/video, which is beneficial to understand human actions and intentions [13–16]. IR illumination is used in active near-IR (NIR) (780–1100 nm) imaging. The head pose is generally considered as a rigid body transformation relative to the camera. The estimation targets are 2D Euler angles, which consist of pitch and yaw angles. Given an input IR head image I and a head pose angle l, HPE aims to construct an accurate mapping relation to predict exact labels l from IR image I.

2.2. Feature learning for IRHPE

Many HPE approaches have recently been proposed [11,17–24]. Existing methods can generally be divided into two groups: model-driven HPE (MD-HPE) and data-driven HPE (DD-HPE) methods. The aim of MD-HPE methods is to characterize faces with several landmarks [25,26] and then locate the landmarks on real faces by using trained appearance models. They are the most common approaches, which estimate the distance from a reference coordinate system via coplanar facial landmarks. The limitation of model-based methods is the accuracy of landmark detection. In a real scenario, facial landmarks are proned to be obscure, which results in a great influence on prediction accuracy. Feature regression methods map images to a head pose space by using a trained regression function. Unlike regression tools, such as random forest [8], support vector machine [27], and cascade [28], the feature regression method is more robust and achieves better real-time performance. These feature regression-based algorithms leverage handcrafted features to extract information from head pose images. Crucial improvements have been made with handcrafted feature techniques, but they are unbeneficial to HPE. The extraction of features on a large-scale dataset is also unconducive due to the time-consuming problem.

DD-HPE methods usually include two major stages: head feature extraction and head pose classification. In the head feature extraction, numerous methods have been proposed to obtain appearance features and face geometry caused by head rotation. Over the past couple of decades, various DL-based algorithms, such as attention network [29], ordinal regression network [30], multitask learning [31,32], and multiloss CNN [33], have been proposed. However, the IR imaging condition often leads to low-quality head pose images. Thus, previous networks and algorithms are unsuitable for the IR imaging case. This limitation motivates us to overcome the low-quality problem and suppress the uncertainties.

2.3. Soft-label learning

The label distribution learning [34] technique can construct a single label as a fixed distribution. It considers that the distribution label can describe an instance clearly. Thus, this technique is widely used in image classification tasks, such as practical age estimation [35], facial expression recognition [19], and video parsing [17]. Zhang *et al.* [36] analyzed the characteristics of different NIR facial expressions and fitted seven basic expression probabilities as a Gaussian-based distribution. The constructed soft label was leveraged to train the network, and impressive classification results were achieved. In [35], the soft label distribution learning was extended to the age estimation community. In this work, only reasonable adjacent ages were covered in the soft labels; therefore, the age estimation performance was superior to that of other existing methods. Furthermore, the label distribution learning was utilized to count the crowd number in an intelligent video system [17]. The conventional single label was reconstructed as a multilabel distribution, which can also learn a real class label and its neighboring class labels effectively. Those works inspire us to construct a soft distribution label in accordance with IR head pose images. Then, we reveal the essential properties of the pose images under active IR illumination. Unlike traditional label distribution learning methods, the proposed nonuniform Gaussian-distributed labels can suppress the uncertainties caused by low-quality head images and ambiguous head pose angles.

3. Gaussian-like label construction for HPE

3.1. Definition of head pose similarity (HPS)

To measure the similarities of adjacent head poses, we introduce the cosine function to compute the HPS of two head image features. The most representative features of one head image are extracted using the last fully-connected layer of a pretrained network. In Fig. 2(a), the central image is the (0°, 0°) angle pose image I_0. With the increase of 15° in the pitch direction, one of the (0°, 15°) angle pose in the pitch direction is shown in image I_1 as in Fig. 2(b). For the two IR images I_0 and I_1, the pretrained network

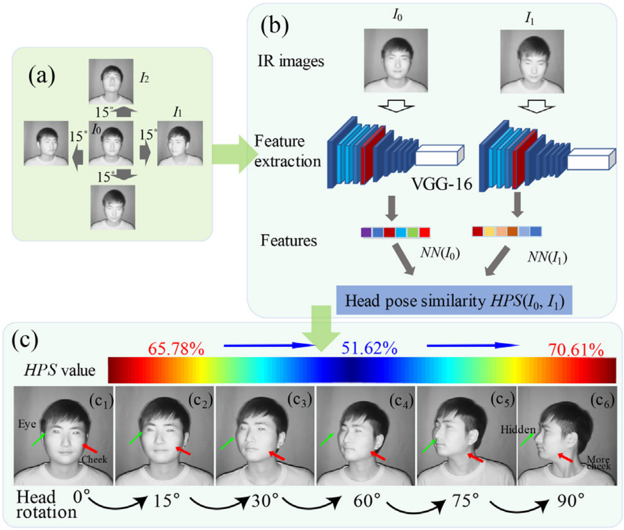


Fig. 2. Nonuniform property of HPE images in the yaw direction. (a) Different changes onhead pose for the image in center with the increase of the same angle degree in pitch and yaw directions. (b) HPS computation by using a pretrained VGG16 network. (c) Nonuniform similarity in one pose direction (yaw). The HPS value of the adjacent images first decreases to 51.62% and then increases to 70.61%. The color bar represents the trend of image similarities of adjacent head poses.

can output the feature vectors by utilizing its last layer. The expression of HPS is formulated as

$$HPS(I_0, I_1) = \frac{FV(I_0) \cdot FV(I_1)}{||FV(I_0)|| \times ||FV(I_1)||} \tag{1}$$

where $FV(\cdot)$ denotes the feature vector of the fully connected layer. Then, the Eq. (1) is used to measure the similarity of the images with head poses in the same yaw or pitch direction as well as in different directions. The essential attributes of IRHPE will be revealed by the variation in HPS values.

3.2. Nonuniform similarity in one pose direction

In Fig. 2(c), we calculate the five HPS values of the six images in the yaw direction. The variation in five HPS values can reflect the feature similarity with the angle degree increases. Considering that the label (15°) of the fixed head pose is extremely closed to its adjacent label (30°), we plan to convert the hard label into a Gaussian distribution (soft label). The nonuniform property reflects the degree of correlation between the current and adjacent poses. This condition is important for Gaussian distribution label construction.

From 0° to 30°, the HPS values decrease with the angle degree increase between the two images with adjoining poses. At those angles, we can observe the eyes and cheeks clearly. The HPS value reaches to 65.48%. Within the range of 30°–60°, the HPS value (51.62%) becomes lower. The green arrow indicates the right eye of the human, which is gradually hidden with the head rotating. If the angle reaches 60° and 90°, then more left facial profile (shown by the red arrow) can be observed, with the HPS value (70.61%) being larger than those at all other angles.

3.3. Difference property in two pose directions

To investigate the differences in two directions (Fig. 2(a)), the feature similarity is calculated between the central pose I_0 and other neighboring poses I_1 and I_2. The HPS value between the images of poses (0°, 0°) and (±15°, 0°) is considerably smaller than that between the head images of poses (0°, 0°) and (0°, ±15°). Thus, we argue that the rotation of human head in the pitch direction is more obvious than that in the yaw direction. The head rotation varies in two pose directions because the human head is a nonspherical symmetry rigid shape. Then, we compute all the neighboring pose images around the central pose (0°, 0°), as shown in Fig. 2 (a). The results are shown in Fig. 3(a). Generally, the similarity between a certain pose image and its pose-variant images in yaw directions differs from that with pose images with same angle degree in pitch direction. In this paper, this finding is defined as the difference property of IRHPE.

Furthermore, we define a criterion to describe the difference quantitatively. Given three images I_0, I_1, and I_2, the ratio of HPS is formulated as

$$Ratio(I_0, I_1, I_2) = \frac{HPS(I_0, I_1)}{HPS(I_0, I_2)} \tag{2}$$

which is utilized to calculate the ratio values among the head pose images. The values of $Ratio(I_{center}, I_{pitch}, I_{yaw})$ belong to the range [0.6, 1]. In Fig. 3, all HPS values are plotted in a matrix, which can be fitted by using a 2D Gaussian-like distribution. It can be used to construct ground-truth IRHPE labels.

3.4. Nonuniform Gaussian distribution label

Given a set of label $L = \{l_{xy}|x = 1, \ldots, k_1, y = 1, \ldots, k_2\}$, the angle distribution l can be expressed using a $k_1 \times k_2$ matrix [8]. The

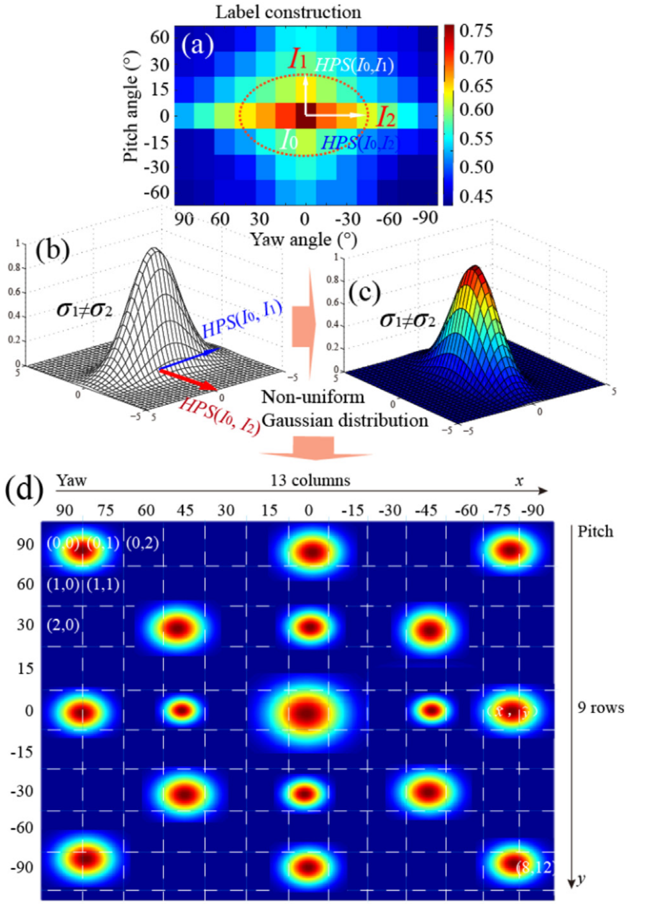


Fig. 3. Nonuniform Gaussian distribution construction on the IRHPE dataset. (a) Regions with a high degree of similarity indicated as nonuniform. (b)-(c) *HPS* values in 3(a) fitted by using a 2D nonuniform Gaussian function. The nonuniform 2D Gaussian model has a long *x* axis and a short *y* axis. (d) Ground-truth label for IRHPE.

matrix in Fig. 3(a) can be fitted by using a Gaussian-like distribution,

$$g(l_{xy}) = \frac{1}{2\pi\sqrt{|Q|}}\exp\left(-\frac{1}{2}(l_{xy}-\mu)^{\mathrm{T}}Q^{-1}(l_{xy}-\mu)\right) \quad (3)$$

where the covariance matrix Q is set as, $\begin{pmatrix} \sigma^2 & 0 \\ 0 & \eta\cdot\sigma^2 \end{pmatrix}$The lower-right element value is smaller than the upper-left element value due to the anisotropic property of the head pose. The value of η is supposed to be in the range $\eta\in(0.7, 1]$ in accordance with the quantitative analysis. Then, the angle distribution $\hat{y}$ is determined through the normalization operation, as shown as follows:

$$\hat{g} = \frac{g(l_{xy})}{\sum_x\sum_y g(l_{xy})} \quad (4)$$

In Fig. 3(d), we show the map of angle distribution when the head pose pitch and yaw angles are equal to 0°.

4. Proposed NGDNet model

4.1. NGDNet model

The pipeline of the proposed NGDNet network is shown in Fig. 4. The pooling layer and convolutional layer in the traditional network can extract the first-order statistics information (maximum or mean). In fact, the regional descriptors can be selected as the second-order statistics such as covariance, which is better than the first-order one. The distortions of head pose image key points are very important for the IRHPE task. And the distortions property can be captured by the second-order descriptors. Thus, following the last convolutional layer, the second-order covariance pooling is introduced, which builds the covariance matrix to describe pose image globally. The covariance pooling layer includes the nonlinear function, which is hard to implement the backpropagation operation. Thus, the gradients calculation method [37] is introduced in our covariance pooling layer.

4.2. Covariance pooling layer

Let $W \in \mathrm{R}^{d\times N}$ be a matrix in the last convolutional layer. Its columns includes a sample of N features at dimension d. Its covariance matrix C can be calculated as

$$C = W \quad \bar{I} \quad W^T \quad (5)$$

in which the unit matrix can be written as,

$$\bar{\boldsymbol{I}} = \frac{1}{M}(I - \frac{1}{M}11^T).$$

I is the identity matrix, and **1**=[1, … ,1]^T represents a vector. Its dimension is denoted as the symbol M. The eigenvalue decomposition is provided by

$$\mathrm{O} = \mathrm{U} \quad \Lambda \quad \mathrm{U}^T \quad (6)$$

where ^= diag (μ_1,. . .,μ_d) is a diagonal matrix whose eigenvalues μ_i is arranged in the descending order. U = [u_1, . . . ,u_d] is an orthogonal matrix and its column u_i is the eigenvector corresponding to μ_i. Matrix power can be converted to the power of eigenvalues by EVD. Consequently, we obtain

$$\Omega = \mathrm{O}^\delta = \mathrm{U}\ \mathrm{F}(\ \Lambda)\ \ \mathrm{U}^T \quad (7)$$

in which the symbol δ means a positive real number. Denote F(^) = diag(f(μ_1),. . .,f(μ_d)), and it can be computed as

$$f(\mu_i) = \mu_i^\delta \quad (8)$$

Then, $\boldsymbol{\Omega}$ is input to the subsequent, top fully connected layer. Then 90% parameters in the fully connected layers is removed. The label distributions of samples are learned after the softmax function.

4.3. Loss function of NGDNet for infrared HPE

Given one group of infrared head pose images I with the constructed ground-truth angle distribution g, the aim of training is to find the best θ by maximizing the posterior probability p ($\theta|I$, g). The equation could be formulated as,

$$\theta^* = \arg \max_\theta \ p(\theta|I, \quad g) \quad (9)$$

Based on Bayes rule $p(\theta|I,g) \propto p(I, \quad g|\theta)p(\theta)$, Eq. (9) can be rewritten by introducing the logarithm function,

$$\theta^* = \arg\ \min \quad (\ \text{-}\ \log \quad p(I,g|\theta)\ \text{-}\ \log\ p(\theta)) \quad (10)$$

It can be seen that two probability density functions need to be constructed.

Likelihood p (I, $g|\theta$). The first term is the likelihood density function, which represents a measure of the conformance of the predicted labels to the ground-truth ones. Since the HPE labels are defined as the Gaussian distributions, the Kullback-Leibler (KL) divergence is introduced to measure the distance between prediction and ground-truth labels. Thus, the likelihood probability is formulated as,

$$p(I, \quad g|\theta) \propto KL(g,\hat{g}) = \sum_t g_t \ln\frac{g_t}{\hat{g}_t} \quad (11)$$

where $\hat{g}$ means the predicted distribution label.

Prior $p(\theta)$. In this article, the smoothness constraint is added to the hyper-parameters θ. The error in the hyper-parameters θ is assumed to fit the Gaussian distribution. Then, the prior probability $p(\theta)$ could be rewritten as,

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\theta^2/2\sigma^2) \quad (12)$$

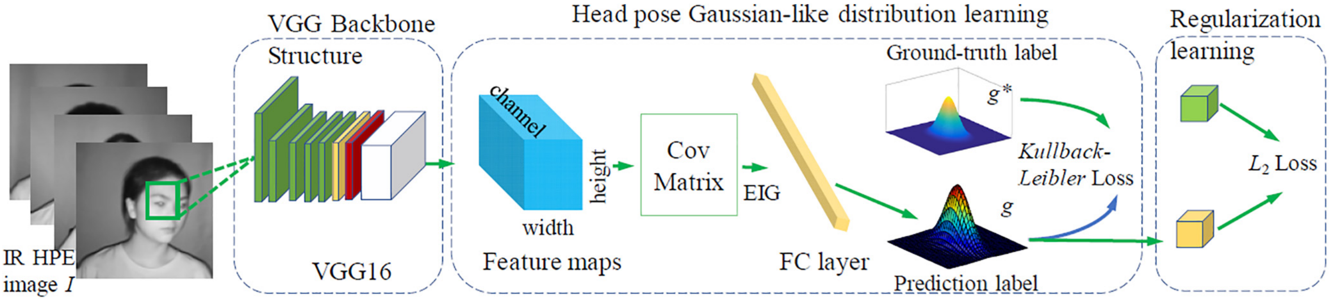


Fig. 4. Pipeline of the proposed NGDNet model.

Hence, the loss function of the proposed NGDNet model can be constructed as,

$$E(\theta) = \sum_t g_t \ln \frac{g_t}{\hat{g}} + \alpha||\theta||_2^2 \tag{13}$$

where α means the regularization parameter. The L_2-norm is utilized in the hidden layer to avoid the substantial growth of parameters in the training phase. The promised presentation can be achieved while α is set as 0.001.

4.4. Optimization

The matrix backpropagation is used to calculated the variation of objective function $E(\theta)$. The optimization formulation can be written as,

$$\left(\frac{\partial E}{\partial \Omega}\right)^T \mathrm{d}\Omega = \left(\frac{\partial E}{\partial \mathrm{U}}\right)^T \mathrm{dU} + \left(\frac{\partial E}{\partial \Lambda}\right)^T \mathrm{d}\Lambda \tag{14}$$

Based on this equation, we can achieve,

$$\begin{cases} \frac{\partial E}{\partial \mathrm{U}} = \left(\frac{\partial E}{\partial \Omega}\right)^T + \left(\frac{\partial E}{\partial \Omega}\right) \\ \frac{\partial E}{\partial \Lambda} = M\left(\mathrm{diag}(\mu_1^{\delta-1}, \cdots, \mu_d^{\delta-1})\mathrm{U}^T \frac{\partial E}{\partial \Omega}\mathrm{U}\right)_{\mathrm{diag}} \end{cases} \tag{15}$$

in which matrix M_{diag} denotes the operation that keeping the diagonal entries of M while setting all non-diagonal entries as zeros. Then, the variation of object funtion can be calculated by,

$$\left(\frac{\partial E}{\partial \mathrm{O}}\right)^T \mathrm{dO} = \left(\frac{\partial E}{\partial \mathrm{U}}\right)^T \mathrm{dU} + \left(\frac{\partial E}{\partial \Lambda}\right)^T \mathrm{d}\Lambda \tag{16}$$

in which U means the orthogonal constraint. Consequently, the gradient of E with respect to the W could be rewritten as,

$$\frac{\partial E}{\partial W} = \bar{\boldsymbol{I}}\mathrm{W}\left(\frac{\partial E}{\partial \mathrm{O}} + \frac{\partial E}{\partial \Lambda}\right)^T \tag{17}$$

Finally, the optimization of NGDNet model is finished.

5. Experimental results and discussion

5.1. Experimental setup

*1) **Datasets**:* Three public HPE datasets are selected from different fields to evaluate the proposed model.

a) IRHPE dataset: It is captured by our research group. This dataset includes 40 subjects 11,600 images in the IR imaging environment. The total of 145 head poses is captured from each subject. The capturing environment is shown in Fig. 5.

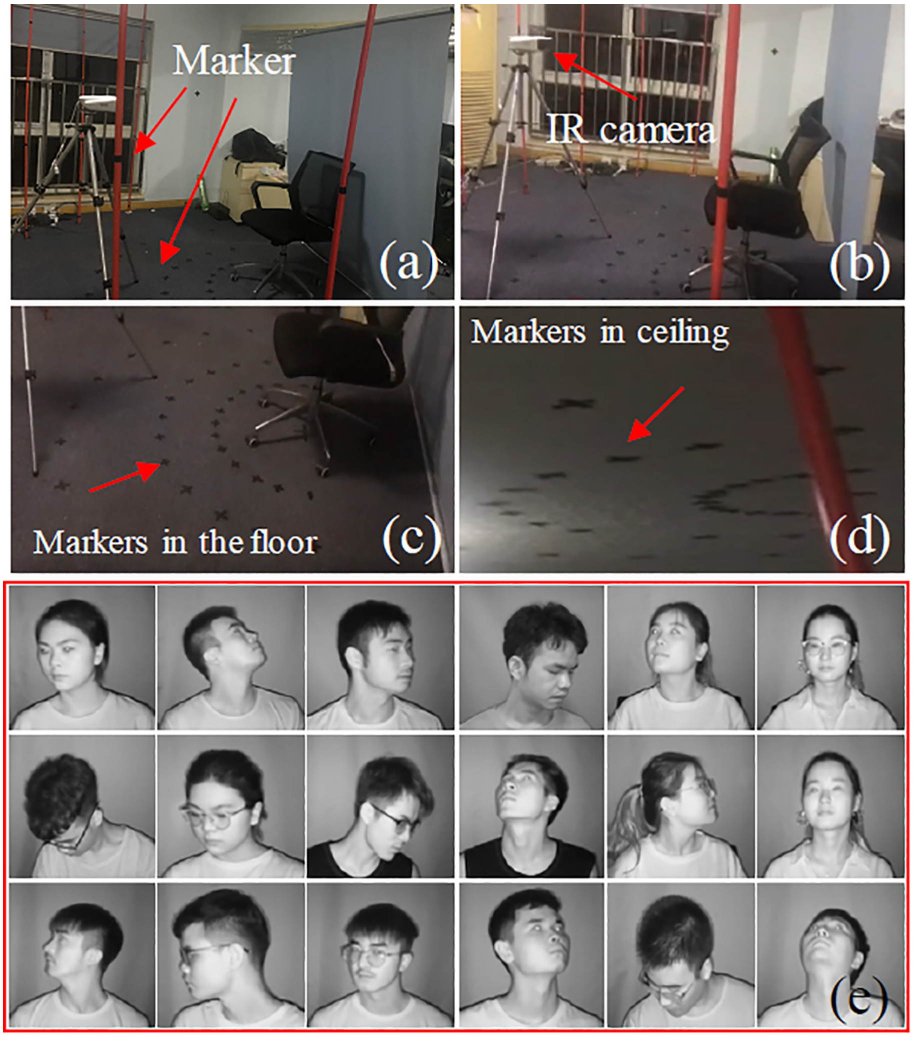


Fig. 5. Infrared head pose dataset, captured by our research group. (a)-(b) Capturing environment in our lab. (c) Markers in the floor. (d) Markers in ceiling. (e) Infrared head pose images in IRHPE dataset.

b) *CAS-PEAL-R1* dataset [38]: It includes 30,900 images from 1040 subjects. . The yaw and pitch angles of the head pose images are in range of [−30°, 0°, 30°] and [−45°, −30°, −15°, 0°, 15°, 30°, 45°], respectively.

c) *Pointint'04* dataset [39]: This database consists of 15 subjects, whose ages, genders, hairstyles are different, with a total of 2,790 head pose images. All the images are captured under the visible light conditions. The yaw and pitch angles of the captured images are in the range of [−90°, 90°]. Each person includes total 93 discrete poses, since the yaw angle equals 0° whilethe pitch angle is a±90°.

2) ***Evaluation metrics****:* Two common metrics are introduced to evaluate the performance of the comparing methods, such as Mean Absolute error (*MAE*) and Mean Accuracy (*MA*). The *MAE* is defined as,

$$MAE = \frac{1}{M}\sum_{i=1}^{M}\left(|\hat{\omega}_i - \omega_i| + |\hat{\xi}_i - \xi_i|\right) \tag{18}$$

in which ω_i and ξ_i denote the ground-truth labels in two different directions. The small *MAE* value means the high accuracy in the comparing methods. The *MA* can be formulated as follows:

$$MA = \frac{1}{M}\sum_{i=1}^{M} acc_i \tag{19}$$

in which the symbol *M* means the cross validation numbers, and acc_i denotes the accuracy in *i*-th validation.

3) Tested models. To compare with the proposed method, we selected six famous recommendedalgorithms as follows:

a) MLD [13]: The labels of the head pose image are constructed as the soft distribution, which is useful on the Pointing'04 dataset.

b) GLM [2]: High dimension probability is introduced to measure the ground-truth label of each head pose angle.

c) SIFT-RP [40]: A novel descriptor with the random projected dense technique is proposed. It can extract the facial texture feature in the head pose images.

d) CASR [41]: The cumulative attribute space regression is introduced to estimate the Euler angle of head pose images. It can classify the Pointing'04 dataset very well.

e) DLD [24]: Suppressing the label ambiguity with the deep label distribution learning, this method is extended to many image classification tasks (HPE).

f) NGDNet: To adapt the infrared imaging feature, the multi-label distributions are constructed as the Gaussian-like type labels. Two important essential properties are revealed, namely, the non-uniform similarity at one pose direction and difference property at two pose directions.

4) Implementation details. The learning rate is decayed by an attenuation coefficient of 0.1 every 30 epochs with initial value 0.0001. Each model is trained 200 epochs by using the batched of 128. For objective and unbiased, the five-fold cross-validation technique and 80%–20% train-test settings are employed.

5.2. Results and discussion

In this section, several experiments will be carried out on three public datasets to evaluate the developed NGDNet model. Three aspects objectives are summed as follows,

ii) Investigate the two important properties of the NGDNet method;

ii) Compare the performance of the proposed NGDNet method with the state-of-the-art methods under the active infrared illumination. Two metrics will be used to measure the difference among the methods;

iii) Extend the proposed NGDNet model to other HPE datasets, such as the CAS-PEAL-R1 and Pointing'04. It can reflect the generalization ability of the NGDNet model.

5.2.1. Accuracy analysis of NGDNet

To verify the performance, the experiments are carried out on IRHPE dataset under active infrared imaging environment. In Fig. 6(a) and 6(b), we show the iteration curves of loss values of test and training process. These curves can reflect the learning ability of NGDNet method. It can be observed that the loss curves can converge at the early steps. Furthermore, we demonstrate the high-level features of the infrared pose images, which are visualized from the IRHPE in the weak illuminated environment. The visualized results can reveal the feature extraction ability of the proposed NGDNet model. To this end, we introduce the Grad-CAM as the visualization tool. In Fig. 7, the different class-discriminative areas of the head pose images are demonstrated. With the same subject, the pose angles are from −90° to +90°. As the result, the discriminative features are extracted by the NGDNet method in the different pose images.

Furthermore, the class-discriminative areas for (15°, −15°) and (−75°, 15°) are illustrated in Fig. 8. Experiment results show that the same areas of the head pose images are activated by the class-discriminative features. This phenomenon can be observed in all the subject in the IRHPE dataset. For the (15°, −15°) angle, the class-discriminative features are presented in the mouth and cheek regions (see Fig. 8(a)–(d)). While for (−75°, 15°) angle, the region is around the eyes (see Fig. 8(i)–(l)).

5.2.2. Results on the IRHPE infrared dataset

The experimental results are carried out on three datasets by comparing with five state-of-the-arts HPE methods. The IRHPE dataset is captured by our research team. To achieve the best performance, the parameters in the comparing methods are carefully

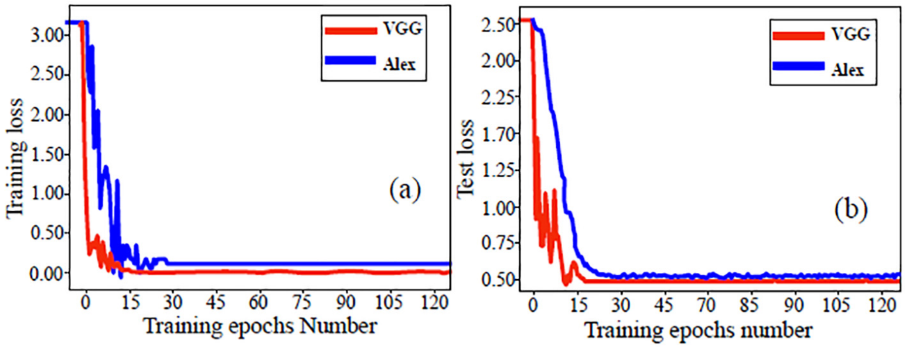


Fig. 6. Iteration curves of loss function with the epoch number increasing. (a) Training stage (b) Test stage.

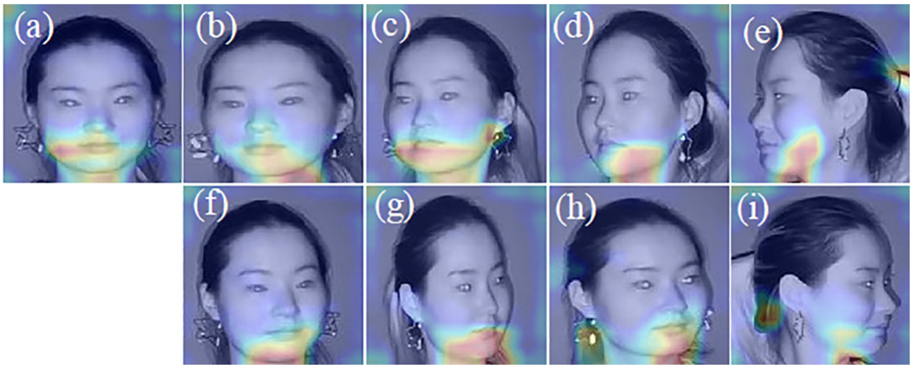


Fig. 7. Feature visualization in NGDNet model for HPE task. The red regions correspond to high scores for classes.

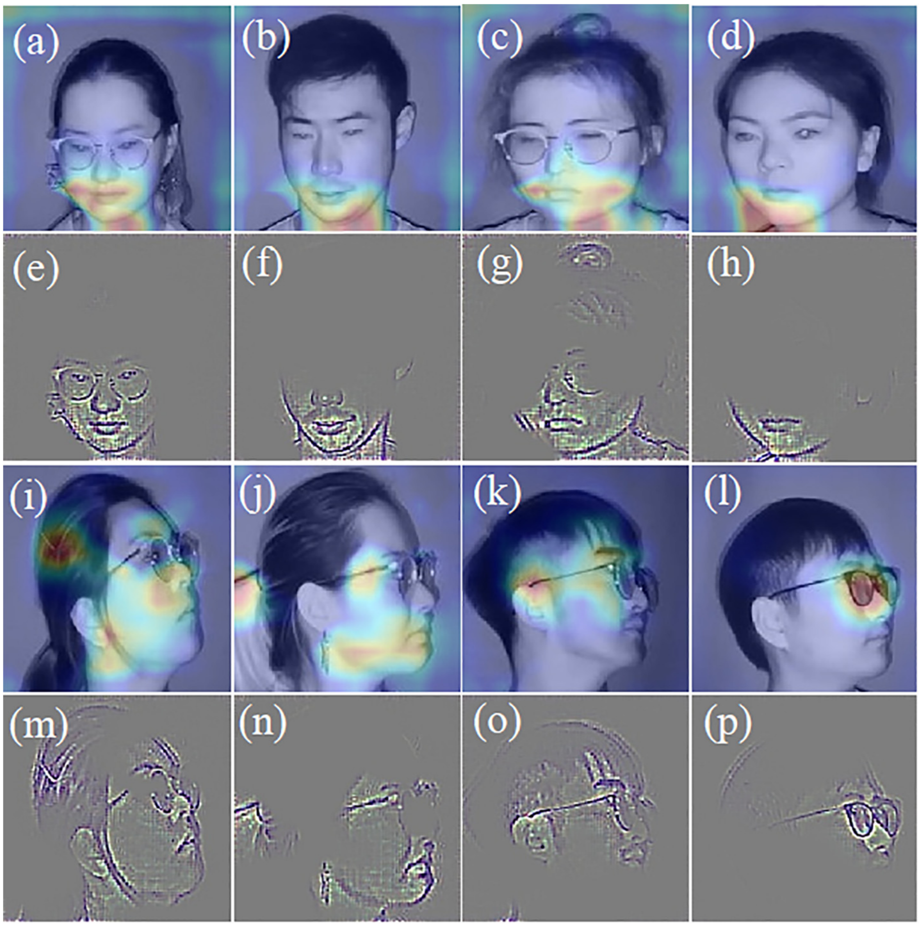


Fig. 8. The feature visualization by the proposed NGDNet model, which are on the same head pose angle of different subjects.

adjusted. In Table 1, we present the *MA* and *MAE* values of all the comparing methods. Obviously, the head pose recognition accuracies are 79.86% and 87.49% at the yaw and pitch directions, respectively. The *MA* values are raised by 13.55% (CASR) and 14.24% (DLD), which outperforms the comparing models. It suggests that the distinguishing features can be explored by the proposed NGDNet across different categories via distillation the latent knowledge from constructed anisotropic angle distribution and robust network architecture. MLD can estimate the head pose by training the random forest and combined linear discriminant analysis. In [42], fisher vector of local descriptors or its variant are distilled and nearest centroid classifier is employed to estimate head pose. However, all the methods cannot utilize potential contact information between head pose, and the best accuracy they reported is still very low. The proposed NGDNet is executed on the CAS-PEAL-R1 and Pointing'04 datasets. The metrics values are also highlighted.

5.2.3. Determination of learning label parameters

In this paper, there are three parameters need to be discussed, namely, one regularization parameter α and two parameters σ and η in the non-uniform Gaussian-label distribution. We conduct some experiments with different parameters on three public datasets. The regularization parameter α can tradeoff between the data

Table 1
Performance comparison in *MA* and *MAE* for each epoch on IRHPE, CAS-PEAL-R1, Pointing'04 dataset.

IRHPE dataset						
Methods	*MA* (%)			*MAE* (°)		
	Yaw	Pitch	All	Yaw	Pitch	All
GLM [2]	53.25	62.41	42.69	7.71	8.20	15.19
SIFT-RP [40]	57.36	68.19	48.21	7.15	6.76	12.56
MLD [13]	61.62	74.98	61.76	5.39	3.83	8.46
CASR [41]	62.87	75.35	63.84	4.28	3.72	7.48
DLD [24]	69.58	81.65	63.15	3.16	2.69	5.64
NGDNet	**79.86**	**87.49**	**77.39**	**2.62**	**1.71**	**2.23**
CAS-PEAL-R1 dataset						
Methods	*MA* (%)			*MAE* (°)		
	Pitch	Yaw	All	Pitch	Yaw	All
GLM [2]	87.96	88.45	87.24	1.49	1.57	1.49
SIFT-RP [40]	86.25	87.35	87.13	1.31	1.34	1.36
MLD [13]	91.43	92.41	91.15	1.05	0.97	0.99
CASR [41]	93.26	94.58	93.16	0.81	0.91	0.86
DLD [24]	94.19	95.42	94.26	0.76	0.68	0.77
NGDNet	**98.56**	**99.31**	**99.08**	**0.21**	**0.28**	**0.24**
Pointing'04 dataset						
Methods	*MA* (%)			*MAE* (°)		
	Pitch	Yaw	All	Pitch	Yaw	All
GLM [2]	63.35	72.44	52.45	6.38	7.14	13.17
SIFT-RP [40]	67.45	78.26	58.15	6.03	5.75	12.56
MLD [13]	71.58	84.01	61.81	4.39	2.28	6.69
CASR [41]	72.91	85.29	63.79	4.28	2.81	6.49
DLD [24]	79.62	91.66	73.21	3.21	1.71	4.58
NGDNet	**89.81**	**97.61**	**87.41**	**1.58**	**0.82**	**1.31**

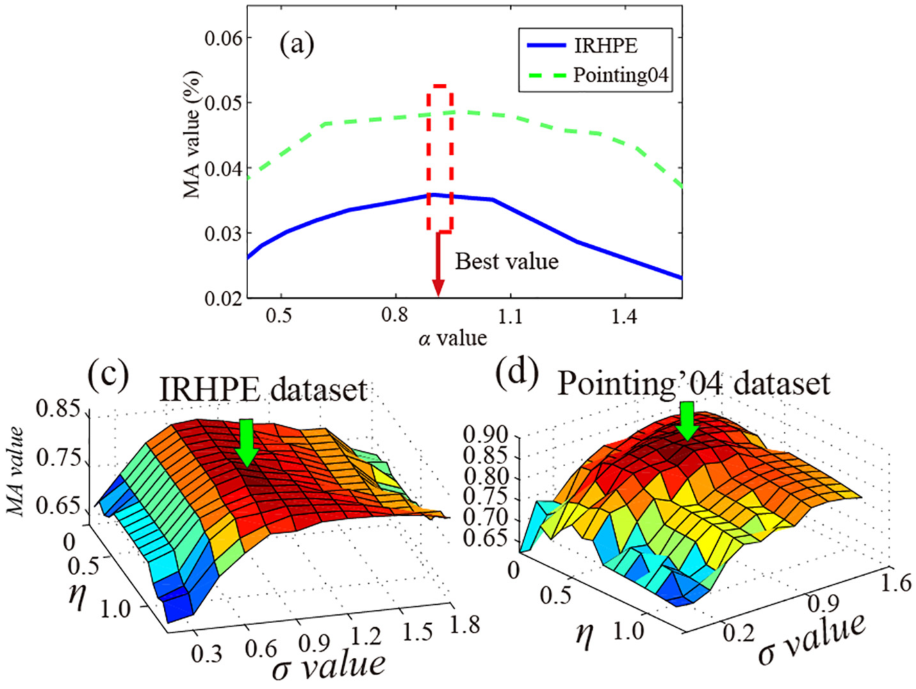


Fig. 9. Three parameters discussion on IRHPE dataset and Pointing'04 databases. (a) *MA* value changing with regularization parameter α increasing. (b) Non-uniform Gaussian distribution parameters on IRHPE dataset and (c) Pointing'04 datasets.

item and regularization item in Model (13). In Figs. 9(a), we find that setting the parameter as 0.9 obtains the highest *MA* values. Increasing the value of regularization parameter α from 0.9 to 1.7 dramatically degrades the performance which means that smooth constraint is extremely crucial.

The labels of all head pose angles are different in Fig. 3. It is necessary to discuss the effect of two key parameters. Firstly, the parameter η is fixed as 1 for each label. To be specific, the Gaussian distributions are initialized as the uniformed ones. For the parameter σ, the small value means that the label is a sharp distribution.

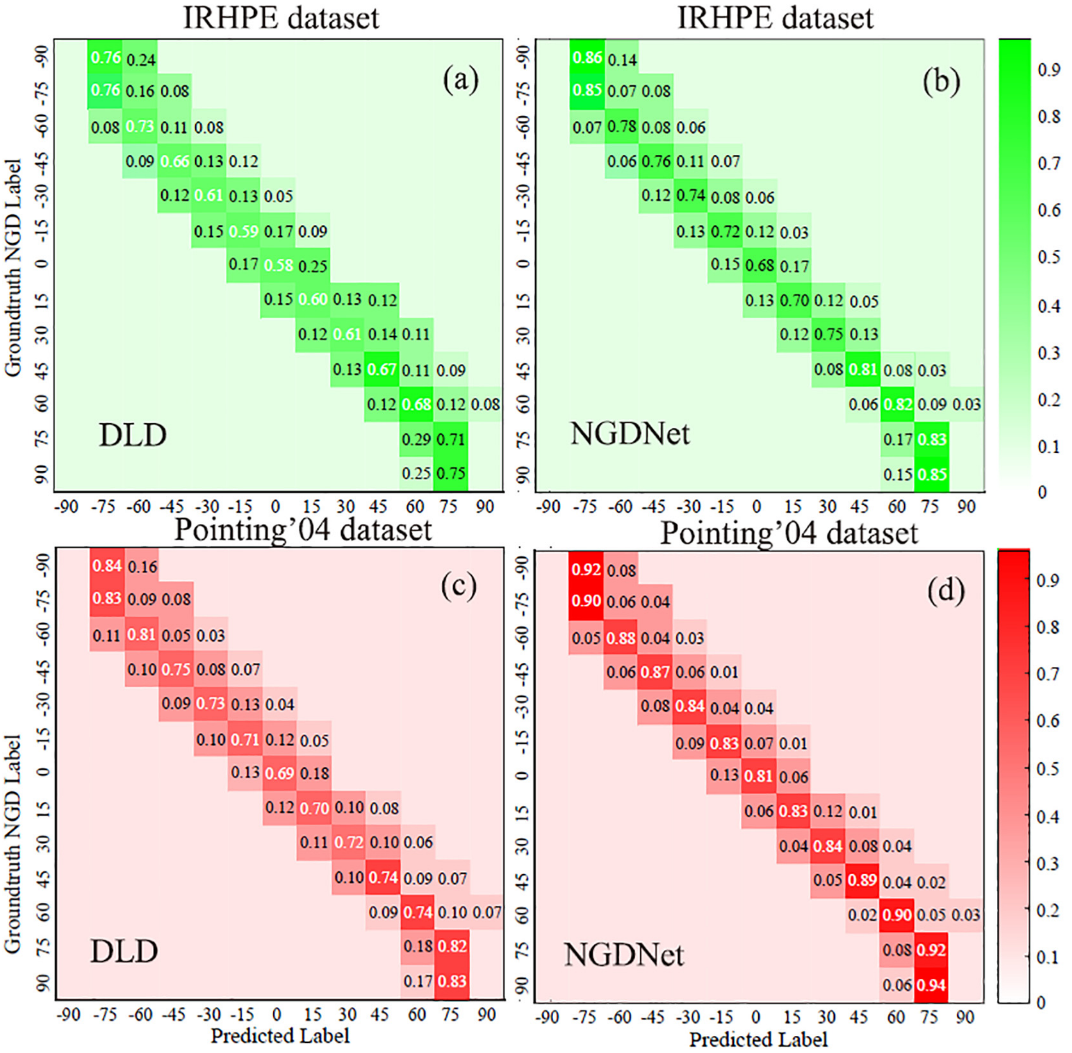


Fig. 10. Confusion matrix by the DLD method and NGDNet method. (a)-(b) IRHPE dataset. (c)-(d) Pointing'04 dataset.

On the contrary, the label distribution becomes very smooth if we increase the σ value. To analyze the effect of σ, we conduct some experiments on the IRHPE and Pointing'04 datasets. The rate value is increased from 0.6 to 1.4. In Figs. 9(b) and (c), we plot the *MA* curve with parameter σ increasing. The parameter experiment illustrates that the model achieves the best performance when $\sigma = 0.85$. Then, with the fixed σ value, the parameter η is ranged from 0 to 1.2. We show the mesh values in the Figs. 9(b) and (c). The green arrows denote the highest recognition rates in IRHPE and Pointing'04 datasets. From the experiment results presented in Fig. 10(b), we could observe that the results are better than any other cases when $\eta = 0.65$ and $\sigma = 0.85$. According to the quantitative analysis of the feature similarity, we assume the values of parameters η and σ are in the range of (0.6, 1). The experimental results show that the values of η and σ are indeed within this range. Thus, it can be concluded that the analysis aforementioned is reasonable.

5.2.4. Head pose missing in IRHPE dataset

Since the labels are constructed as the non-uniform Gaussian distributions, the NGDNet can accurately classify the similar head pose images. However, there are still some inconsecutive HPE images, which are not captured in IRHPE. The consecutive HPE images include the 7.5°, 22.5°, and 37.5° angles. To explore the ability of angle missing, NGDNet model is executed with the different missing levels. Considering two network model, one using a training dataset that contains all available yaw angles and another one using a training dataset which the angles with uniform sampling of 30°. Evaluate the accuracy of estimation with the testing set which includes the images whose yaw angles are either seen or unseen in the training dataset. The results on IRHPE and Pointing'04 datasets can be observed in Figs. 10(a)-(b) and (c)-(d), respectively.

The experiment of pitch angles is performed on the IRHPE due to the sparsity of the pitch angle. It can be observed that, (i) The increase of the *MA* is larger when the angle sampling interval in the training dataset is within 30° for both yaw and pitch angles; (ii) The *MA* of head pose increases when the angles of the testing data are not in the training set; (iii) The *MA* of HPE prediction decreases substantially when the angle sampling interval in the training dataset is beyond 30°. The results illustrate that the proposed NGDNet model can well predict the missed pose angles when the angle sampling interval in the training dataset is within 30°.

5.2.5. Comparison of time complexity in IRHPE

The computing times of all comparing methods are tested on the IRHPE dataset. The experiments are executed on a workstation with the following computer hardware configuration: one NVIDIA TITAN RTX, 16 Intel (R) Core (TM) i9-9900 K, RAM (64 GB) and

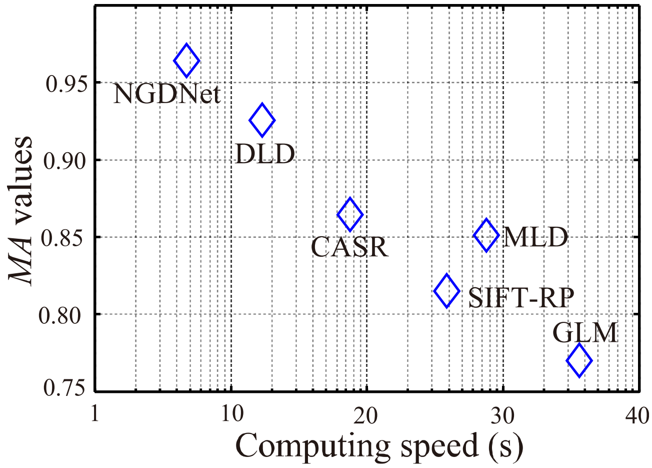


Fig. 11. Comparison of the computing time between the proposed NGDNet model and other HPE models.

CPUs (3.60 GHz). We jointly compare the head pose recognition rate and the computing times of all models, as shown in Fig. 11. The proposed NGDNet method obtains the best performance in both *MA* values and computing speed.

6. Conclusion

In this study, a nonuniform Gaussian-label learning framework is developed for IRHPE under active IR illumination. First, an IRHPE dataset is constructed for the IRHPE task. Two important properties, namely, difference and nonuniformity properties, are revealed among IR images. The cosine similarity function is employed to compute the similarities of different head pose images and then soften the label as a nonuniform Gaussian distribution. Then, the NGDNet model is proposed, and its data term is constructed as Kullback–Leibler divergence. To optimize the proposed model, the mini-batch gradient descent algorithm is introduced with good convergence performance. The experimental results on the IRHPE dataset and two public datasets illustrate state-of-the-art performance in terms of prediction accuracy and robustness. The results indicate that it is beneficial to the human–machine interaction in a weak illumination condition. In the future, IR video sequences will be examined on a real-time computing speed.

CRediT authorship contribution statement

Tingting Liu: Writing - original draft. **Jixin Wang:** Data curation. **Bing Yang:** Writing - review & editing. **Xuan Wang:** Conceptualization, Methodology.

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 article. Thank Professor Hai Liu for the helpful discussion. This work was supported in part by the National Natural Science Foundation of China under Grant 62005092, the China's ministry level project under Grant MCM20160409, and the National Science Foundation under Grant 18YJA880090.

References

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

[2] V. Drouard, S.O. Ba, G.D. Evangelidis, A. Deleforge, R. Horaud, Head Pose Estimation via Probabilistic High-Dimensional Regression, in: international conference on image processing, 2015, pp. 4624–4628.

[3] G. Borghi, M. Venturelli, R. Vezzani, R. Cucchiara, POSEidon: Face-from-Depth for Driver Pose Estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5494–5503.

[4] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.

[5] K. Geng, G. Yin, Using deep learning in infrared images to enable human gesture recognition for autonomous vehicles, IEEE Access 8 (2020) 88227–88240.

[6] T. Liu, Y. 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. Inf. (2021), https://doi.org/10.1109/TII.2019.2930463.

[7] J. Ma, J.i. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (2019) 512–531.

[8] Y. Liu, J. Chen, Z. Su, Z. Luo, N. Luo, L. Liu, K. Zhang, Robust head pose estimation using Dirichlet-tree distribution enhanced random forests, Neurocomputing 173 (2016) 42–53.

[9] J. Ma, J. Jiang, H. Zhou, J. Zhao, X. Guo, Guided locality preserving feature matching for remote sensing image registration, IEEE Trans. Geosci. Remote Sensing 56 (2018) 4435–4447.

[10] 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.

[11] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.

[12] H. Liu, Y. Li, Z. Zhang, S. Liu, T. Liu, Blind Poissonian reconstruction algorithm via curvelet regularization for an FTIR spectrometer, Opt. Express 26 (2018) 22837–22856.

[13] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014) 1837–1842.

[14] 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.

[15] 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.

[16] 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.

[17] Z. Zhang, M. Wang, X. Geng, Crowd counting in public video surveillance by label distribution learning, Neurocomputing 166 (2015) 151–163.

[18] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1837–1842.

[19] H. Zheng, X. Geng, D. Tao, Z. Jin, A multi-task model for simultaneous face identification and facial expression recognition, Neurocomputing 171 (2016) 515–523.

[20] L. Xu, J. Chen, Y. Gan, Head pose estimation using improved label distribution learning with fewer annotations, Multimed. Tools Appl. (2019).

[21] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized Gaussian fields criterion, Pattern Recognit. 48 (2015) 772–784.

[22] H. Liu, L. Yan, Y. Chang, H. Fang, T. Zhang, Spectral deconvolution and feature extraction with robust adaptive Tikhonov regularization, IEEE Trans. Instrum. Meas. 62 (2013) 315–327.

[23] J. Ma, J. Zhao, J. Tian, A.L. Yuille, Z. Tu, Robust point matching via vector field consensus, IEEE Trans. Image Process. 23 (2014) 1706–1721.

[24] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Trans. Image Process. 26 (2017) 2825–2838.

[25] L. Liang, R. Xiao, F. Wen, J. Sun, Face alignment via component-based discriminative search, European conference on computer vision, (Springer) (2008) 72–85.

[26] Q. Liu, J. Yang, J. Deng, K. Zhang, Robust facial landmark tracking via cascade regression, Pattern Recogn. 66 (2017) 53–62.

[27] E. Murphy-Chutorian, A. Doshi, M.M. Trivedi, Head pose estimation for driver assistance systems: a robust algorithm and experimental evaluation, Intelligent Transportation Systems Conference (2013) 4624–4628.

[28] C. Gou, Y. Wu, F.Y. Wang, Q. Ji, Coupled cascade regression for simultaneous facial landmark detection and head pose estimation, IEEE International Conference on Image Processing (2018) 4624–4628.

[29] 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, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) 1087–1096.

[30] 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.

[31] R. Ranjan, V.M. Patel, R. Chellappa, Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2017) 121–135.

[32] A. Kumar, A. Alavi, R. Chellappa, KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors, 2017, pp. 258–265.

[33] N. Ruiz, E. Chong, J.M. Rehg, Fine-Grained Head Pose Estimation Without Keypoints (2017).

[34] X. Geng, Label distribution learning, IEEE Trans. Knowl. Data Eng. 28 (2016) 1734–1748.

[35] Z. Li, H. Liu, Z. Zhang, et al., Learning Knowledge Graph Embedding with Heterogeneous Relation Attention Networks, IEEE Transactions on Neural Networks and Learning Systems (2021) 1–12, https://doi.org/10.1109/TNNLS.2021.3055147.

[36] Z. Zhang, C. Lai, H. Liu, Y.-F. Li, Infrared facial expression recognition via Gaussian-based label distribution learning in the dark illumination environment for human emotion detection, Neurocomputing 409 (2020) 341–350.

[37] P. Li, J. Xie, Q. Wang, W. Zuo, Is second-order information helpful for large-scale visual recognition?, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2070–2078.

[38] W. Gao, B. Cao, 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 Cybernet.-Part A: Syst. Hum. 38 (2007) 149–161.

[39] N. Gourier, D. Hall, J.L. Crowley, Estimating face orientation from robust detection of salient facial features, in: ICPR International Workshop on Visual Observation of Deictic Gestures, (Citeseer2004).

[40] H.T. Ho, R. Chellappa, Automatic head pose estimation using randomly projected dense sift descriptors, in: 2012 19th IEEE international conference on image processing, 2012, pp. 153–156.

[41] K. Chen, K. Jia, H. Huttunen, J. Matas, J.-K. Kämäräinen, Cumulative attribute space regression for head pose estimation and color constancy, Pattern Recognit. 87 (2019) 29–37.

[42] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 31 (4) (2009) 607–626, https://doi.org/10.1109/TPAMI.2008.106.

Tingting Liu (S'18-M'20) received the M.S. degree in Natural language processing from Huazhong University of Science and Technology, in 2014, Ph.D degree in education information technology from Central China Normal University (CCNU), Wuhan, China, in 2019. During Sep.2017-Sep. 2019, she was selected as a visiting scholar in School of Computer Science, Carnegie Mellon University, Pittsburgh, USA. From 2019 to 2020, she was a member of research staff with the Collaborative Innovation Centre for Information Technology and Balanced Development of K-12 Education, Faculty of Artificial Intelligence in Education, in CCNU, where she was hosted by the Professor Jixin Wang. She joined Hubei University, Wuhan, in 2020, and is currently a lecturer in School of Education. Her current research interests include Learning behavior analysis, head pose estimation, label distribution learning and deep learning. Dr. Liu has been frequently serving as a Reviewer for several international journals including the IEEE Transactions on Industrial Informatics, IEEE/ASME Transactions on Mechatronics, Neurocomputing, and Pattern Recognition Letters. She is also a Communication Evaluation Expert for the National Natural Science Foundation of China from 2020.

Jixin Wang received the B.S. degree from Central China Normal University (CCNU) in 1984, and worked in the educational technology center in CCNU in the same year, then served as the deputy director of the center; In 1999, he worked in the Department of Information Technology. He is now the executive Deputy director of the Information Office of CCNU, the director of the Digital Education Resource Center, the executive director of the Collaborative Innovation Center for the Balanced Development of Information and K-12 Education, the professor and doctoral supervisor of the School of Education Information Technology. His current research interests include head pose estimation, learning behavior analysis and deep learning.

Bing Yang is currently a full professor of computing with the School of Education, Hubei University, Wuhan, China. Professor Yang obtained his PhD from Huazhong University of Science and Technology, Wuhan, Hubei, China. Professor Yang research involves computer networking, educational technology, e-commerce and data mining. He has published extensive research in computer science journal, Chinese journal of computer and other international conferences and workshops.

Xuan Wang is a PhD student in the National Engineering Research Center for E-Learning at Central China Normal University. Her research interests include educational communication, instruction and learning behavior analysis and digital learning. (Xuan Wang' photograph not available at the time of publication.)