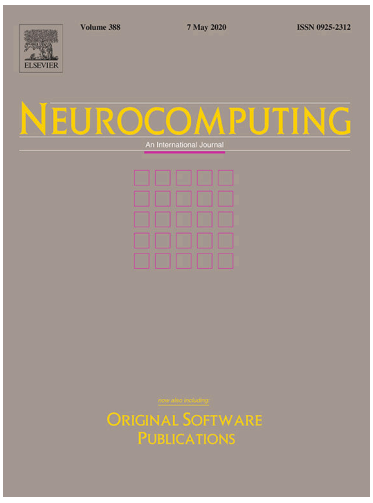
Journal Pre-proofs



Anisotropic angle distribution learning for head pose estimation

Hai Liu, Hanwen Nie, Zhaoli Zhang, You-Fu Li

PII: S0925-2312(20)31599-X
DOI: https://doi.org/10.1016/j.neucom.2020.09.068
Reference: NEUCOM 22905

To appear in: *Neurocomputing*

Received Date: 23 April 2020
Revised Date: 7 July 2020
Accepted Date: 24 September 2020

Please cite this article as: H. Liu, H. Nie, Z. Zhang, Y-F. Li, Anisotropic angle distribution learning for head pose estimation, *Neurocomputing* (2020), doi: https://doi.org/10.1016/j.neucom.2020.09.068

Anisotropic angle distribution learning for head pose estimation

Hai Liu, Hanwen Nie*, Zhaoli Zhang, You-Fu Li*

Department of Mechanical Engineering, City University of Hong Kong, Hong Kong.

E-mail: hanwennie@gmail.com and meyfli@cityu.edu.hk

Abstract

Head pose estimation is an important way to understand human attention. In this paper, we propose a novel anisotropic angle distribution learning (AADL) network for head pose estimation task. Firstly, two key findings are revealed as following: 1) Head pose image variations are different at the yaw and pitch directions with the same pose angle increasing on a fixed central pose; 2) With the fixed angle interval increasing, the image variations increase firstly and then decrease in yaw angle direction. Then, the *maximum a posterior* technology is employed to construct the head pose estimation network, which includes three parts, such as convolutional layer, covariance pooling layer and output layer. In the output layer, the labels are constructed as the anisotropic angle distributions on the basis of two key findings. And the anisotropic angle distributions are fitted by the 2D Gaussian-like distributions (groundtruth labels). Furthermore, the Kullback-Leibler divergence is selected to measure the predication label and the groundtruth one. The features of head pose images are perceived at the AADL-based convolutional neural network in an end-to-end manner. Experimental results demonstrate that the developed AADL-based labels have several advantages, such as robustness for head pose image missing, insensitivity for the motion blur. Moreover, the proposed method has achieved good performance compared to several state-of-the-art methods on the Pointing'04 and CAS_PEAL_R1 databases.

Keywords: Head pose estimation, Anisotropic angle distribution, Convolutional neural network, Regularization, Optimization

1. Introduction

Head pose estimation (HPE) is an important topic in computer vision [1-4], which can be used for human attention interpreting. More than 30% of traffic accidents are ascribed to fatigue driving and inattention. In order to avoid traffic accidents and protect personnel safety and health, car manufacturers are committed to develop lots of driving assistant system, as shown in Fig. 1. Human head pose can provide a key cue in analyzing the human attention, intention, and motivation, etc. Thus, the driver gaze zone can be estimated by tracking the driver head. It has proved that head postures can be an excellent substitute for driver attention [4]. Furthermore, it is well known that HPE is widely utilized in many fields, such as pedestrian position prediction [5], social communication skills improvement [6], and gesture recognition [7]. However, estimating accuracy head pose directions is very difficult in practice due to the multiformity of the appearance caused by the pose changes and motion, such as illumination, occlusion, facial texture, etc. To tackle those problems, numerous algorithms have been proposed over the past decades. In sum, the existing HPE methods can be briefly divided into two streams: facial feature-based HPE method and deep learning (DL)-based HPE method.

1.1 Related work

Depending on the feature extraction principles, facial feature-based HPE methods can be classified into the three categories: template matching methods (TM-HPE),

Fig. 1. Head pose estimation by the proposed method. (a) Normal driving. (b) Talking in driving. (c) Occlusion by hand in driving. (d) Bumps in fast driving. Three color lines denotes the three angle directions of human head pose.

model-based methods (MB-HPE) and feature regression methods (FR-HPE). The basic idea of TM-HPE methods is to match the head pose directions of input images with the discrete standard models, which represent the ground-truth head pose labels [8]. The input head pose images are classified into the specific angle categories by evaluating the image matching degree with the exemplar set. These HPE methods are implemented easily and work well on discrete head pose images. However, it fails for the continuous angles head pose images, since continuous head pose templates are lacked in the image dataset. The aim of MB-HPE methods is to characterize the face with several landmarks [9, 10], and then locate the landmarks on real faces by the trained appearance models. The most common approaches can estimate the distance from a

reference coordinate system via coplanar facial landmarks. But the performance of MB-HPE methods is determined by the precise degree of facial landmark detection. In the real scenarios, facial landmarks are easily obscured, which limit the accuracy of prediction in HPE. The FR-HPE methods map the images to a head pose space by a trained regression function. Different from the regression tools, such as random forest [11, 12], support vector machines and cascade [13], the FR-HPE methods are more accuracy than the aforementioned methods and achieve better real-time performance. These algorithms leverage hand-crafted features to extract information from the head pose images. While crucial improvements have been made with hand-crafted feature technologies, these hand-crafted features are not suitable for HPE task. Furthermore, it is not conducive to extract features on the large-scale datasets due to the time-consuming problem.

Recently, the adoption of deep learning (DL) technologies facilitates to overcome these issues. With the great success in speech, computer vision [14] and natural language processing [15] fields, DL has played a very important role in many areas [16, 17]. The convolutional neural network (CNN) can be trained automatically on large-scale datasets in an end-to-end manner. Patacchiola *et al.* [4] applied CNN technology in HPE field with dropout and adaptive gradient method for the first time. Over the past couple decades, various DL-based algorithms have been developed, such as multi-loss CNN [18], ordinal regression network [19], and attention network [20]. Some researchers have also attempted to improve the accuracy of head pose estimation from a multitasking perspective, such as [21, 22], and these impressive works have made significant progress. Compared with traditional algorithms, few DL-based HPE methods reveal the inherent image characteristics in HPE task. On the one hand, labels of head pose are ambiguity as human head pose is difficult to describe with an exact number. On the other hand, labels of head pose are collected at a small interval. It is difficult to exactly predict the poses if they are not included in training set.

1.2 Contributions of this paper

In this paper, two key finding are first revealed, which are called as the anisotropic property and unsmooth variation property. Based on the two properties, we propose an anisotropic angle distribution learning (AADL) model for head posed estimation. The model is learned via an end-to-end CNN which utilizes the covariance pooling layer to capture the second-order image features. The major contributions of this work can be summarized as three aspects.

1) Two key findings are revealed in this paper. Firstly, head pose image variations are different at the yaw and pitch directions with the same pose angle increasing on a fixed central pose. Secondly, With the fixed angle interval increasing, the image variations increase firstly and then decrease in yaw angle direction. Based on the two findings, original head pose annotations are converted into the anisotropic angle distribution labels.
2) A novel end-to-end HPE framework with the AADL-based CNN is proposed. The method adopts a CNN, which can leverage covariance pooling layer to capture second-order image features. The proposed model is trained with the RGB face images. To the best of our knowledge, this is the first work for HPE with an anisotropic angle distribution learning.
3) Experimental results on several public datasets indicate that the proposed method achieves the state-of-the-art performance on both prediction accuracy and robustness. Furthermore, the AADL method can work well for motion blur in head pose images, even for the occlusion image overwhelmed with some pose images missing.

1.3 Organization of this paper

The reminder of this paper is organized as follows. Two key findings are revealed in detail and constructed as the AADL labels in Section 2. Section 3 illustrates the concrete architecture of the AADL-based neural network model and its adversarial extension. Then, it is optimized by stochastic gradient descent algorithm. Section 4 demonstrates the evaluation results of experiments performed by the anisotropic angle label. Finally, we conclude this paper in Section 5.

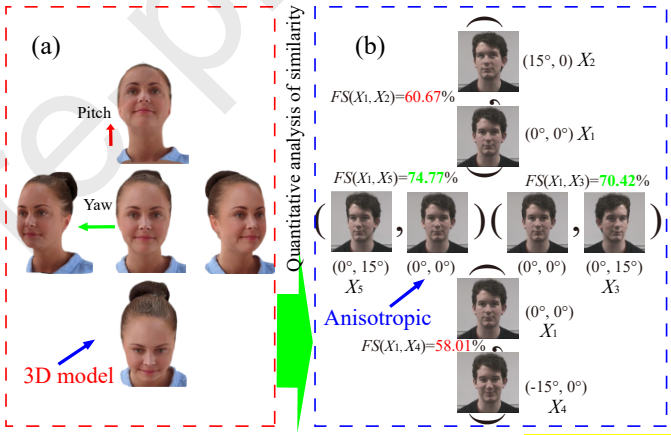


Fig. 2. Key finding 1: anisotropic property for HPE task. Head pose image variations are different at the yaw and pitch directions with the same pose angle increasing on a fixed central pose. For the central pose (0°, 0°), head pose image variations are different with increasing the same pose angle (15°) in yaw and pitch directions. The similarity of head pose in yaw direction (74.77%) is larger than that in pitch direction (60.67%).

2. Characteristics analysis for head pose estimation

2.1 Head pose estimation

Head pose estimation refers that the computer determines the position and attitude parameters of the human head in three-dimensional space by analyzing and predicting the input images or video sequence. The head pose is generally considered as a rigid body transform and the space is relative to the camera. HPE aims at estimating the two-dimensional Euler angles, which includes the pitch and yaw angles. Given an input face image X and a head pose angle Y, the task of HPE network is to construct a function to predict the exact label Y from image X.

2.2 Anisotropic angle head pose

It can be observed that human head is a non-spherical symmetry rigid body, which means that the rotation of head

varies in different directions. We observe that the rotation of human head in the pitch direction (red arrow) is more obvious than that in the yaw direction (green arrow). This property is shown in Fig. 2(a). In other words, the similarity between a certain pose and its yaw angle adjacent poses are different from that with its pitch angle adjacent poses.

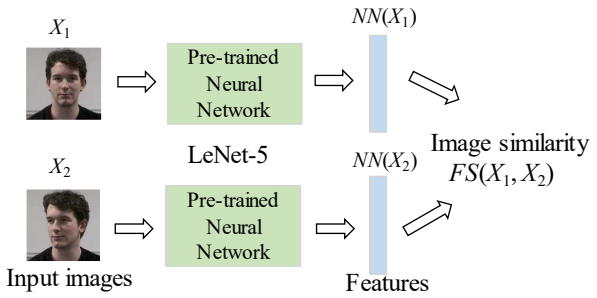


Fig. 3. Feature similarity of two face images is computed by the feature vectors which are extracted from a pre-trained neural network LetNet-5 [23].

To quantitatively formulate this observation, the cosine similarity function is employed to calculate the feature similarity (FS) of two head pose images. The pre-trained LeNet-5 [23] neural network is introduced to extract the image features and calculate the feature similarity of two head pose images. This network has six layers with two convolutional layers, two pooling layers, and two fully connected (FC) layers. The last FC layer of neural network (*NN*) contains the most representative features of one image. Two vectors extracted from the last FC layers is used to compute the cosine similarity. Given two images X_1 and X_2, *NN* is regarded as a function that outputs a feature vector. The detail process of feature similarity is illustrated in Fig. 3. The formula is defined as,

$$FS(X_1, X_2) = Similarity(F(X_1), F(X_2)) = \frac{NN(X_1) \cdot NN(X_2)}{\| NN(X_1) \| \times \| NN(X_2) \|}, \quad (1)$$

where $F(X)$ denotes the feature similarity of image X, and $NN(X)$ is the activation of the FC layer.

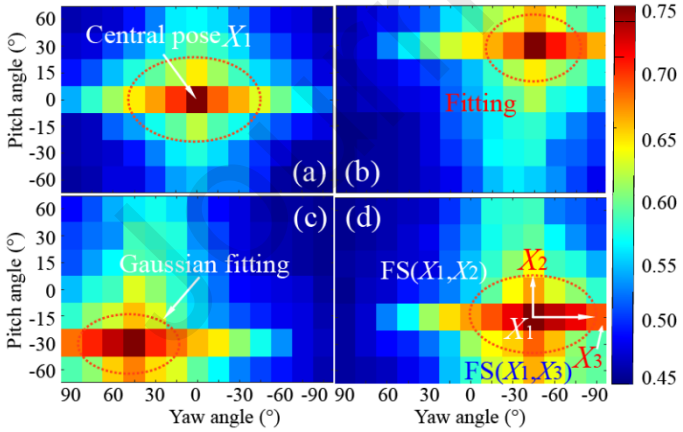


Fig. 4. Anisotropic label construction according to the key finding 1. We compute the all similarity between central pose X_1 and other poses X_2, X_3, X_4 and X_5. The shape of similarity matrix is fitted by a two-dimensional Gaussian distribution.

The formula (1) is utilized to measure the similarity of the head pose images. The feature similarities are calculated between the central pose X_1, and its neighboring poses X_2, X_3, X_3 and X_5, respectively. In Fig. 2(b), it can be observed that the feature similarities (74.77%, 70.42%) between the image of pose (0°, 0°) and pose (0°, ±15°) are significantly greater than the feature similarities (60.67%, 58.01%) between the face image of pose (0°, 0°) and pose (±15°, 0°). In Fig. 4, we plot all the similarity matrixes, which can be fitted by a two-dimensional Gaussian distribution (Figs. 4(b) and 4(c)). The ratio map can be achieved after calculating all the similarity matrixes.

To quantitatively reveal the anisotropic property, the ratio of feature similarity is defined at the yaw direction and pitch direction. Given three images X_1, X_2 and X_3, the definition is formulated as,

$$Ratio(X_1, X_2, X_3) = \frac{FS(X_1, X_2)}{FS(X_1, X_3)}. \quad (2)$$

Based on this equation, we calculate the ratio values for the pose (0°, 0°) in Fig. 2(b), namely, $Ratio_1=FS_1/FS_2=0.81$, and $Ratio_2=FS_4/FS_2=0.78$. Then, we set all other head poses as the center pose, and calculate the ratio similarity between the center head pose image and its pitch/yaw neighbor image. It finds that the ratio values are ranged as [0.6, 1]. Thus, we argue that the rotation of human head in the pitch direction is more obvious than that in the yaw direction. We define the finding as the anisotropic property for HPE in this paper.

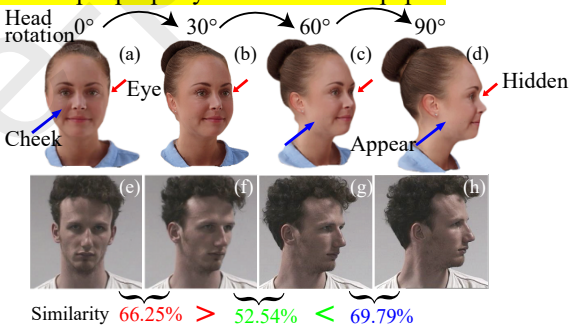


Fig. 5. Key finding 2: unsmooth variation property. With the fixed angle interval increasing, the image variations increase firstly and then decrease in yaw angle direction.

2.3 *Unsmooth variation in yaw direction*

Usually, the shape of human head can be considered as a polyhedron. And the shape of human face skull is flat since the neurocranium appears as an oval ball. Moreover, the face image captured by camera only shows a two-dimensional surface of the face, but not the three-dimensional shape. Consequently, the variations of face image are nonuniform while the head rotates the same angle in the yaw direction. We take negative yaw angle as an example, as is shown in Figs. 5(a)-5(d). From 0° (Fig. 5(a)) to 30° (Fig. 5(b)), two cheeks and both eyes can be observed clearly. However, from 30° (Fig. 5(b)) to 60° (Fig. 5(c)), the right facial profile (blue arrows) appears and left eye (red arrows) is hidden gradually. From 60° (Fig. 5(c)) to 90° (Fig. 5(d)), the variation is hard to observe between two images.

This observation is also quantitatively studied in this paper. In Figs. 5(e)-5(g), the feature similarity between images of 0° and 30° is 66.25%, 30° and 60° is 52.54%, and 60° and 90° is 69.79%. As can be seen, with the fixed angle interval increasing, the image variations increase firstly and then

decrease in yaw angle direction.

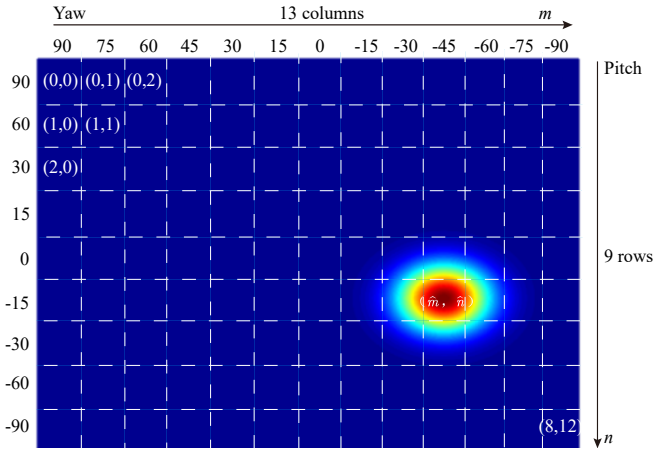


Fig. 6 Head poses $y_{\hat{m}\hat{n}} = (\hat{m}, \hat{n})$ from the Pointing'04 database together with their Gaussian distribution generated by Eqs. (3) and (4).

2.4 Anisotropic angle distribution

According to the analysis in Fig. 4, the region of feature similarity of images can be fitted by a two-dimensional Gaussian distribution, whose diagonal elements in the covariance matrix are different. The soft label is aimed to augment data on the label space by utilizing the correlate knowledge between each category [24]. Inspired by this, an attempt is made to convert the annotated head pose labels to the angle distributions. Taking Pointing'04 dataset as an example, the yaw and pitch angles are combined in the HPE. In Fig. 6, all the head poses can be plotted in a matrix, which has 13 rows and 9 columns. Given a head pose image X, its central pose angle is defined as $y_{\hat{m}\hat{n}} = (\hat{m}, \hat{n})$, where $\hat{m}$ and $\hat{n}$ are the row number and column number of pose image, respectively. For instance, y_{00}=(0, 0) denotes the head pose (pitch=90°, yaw=90°) in Fig. 6. The angle distribution $\hat{y}$ is defined as,

$$\hat{y} = \frac{g(y_{\hat{m}\hat{n}})}{\sum_{m}\sum_{n} g(y_{\hat{m}\hat{n}})}, \tag{3}$$

and

$$g(y_{\hat{m}\hat{n}}) = \frac{1}{2\pi\sqrt{|\Omega|}} \exp\left(-\frac{1}{2}((m-\hat{m})^2 + (n-\hat{n})^2)\Omega^{-1}\right), \tag{4}$$

where m and n are row number and column number in the matrix. The function of Eq. (3) is to normalize the sum of probability values as 1. And the covariance matrix Ω is set as $\begin{pmatrix} \sigma^2 & 0 \\ 0 & \eta\sigma^2 \end{pmatrix}$. In this matrix Ω, if the diagonal elements are equal, the Gaussian distribution in Eq. (4) will be isotropic shape. If the diagonal elements are not equal, the distribution will be anisotropic shape. The variable η is set to achieve the anisotropic 2D Gaussian distribution, which can represent the anisotropic property (key finding 1) for HPE task. On the basis of the quantitative calculation in Fig. 6, the η values belong to the range $\eta \in (0.6, 1)$.

In Fig. 7, the unsmooth variation property (key finding 2) is converted into the different values of the standard deviation σ in matrix Ω. Figure 7(a) shows the angle distribution when head pose (pitch = 0°, yaw = 0°). Figure 7(b) shows the angle distribution at head pose (pitch = 0°, yaw = -45°). It can be found that $\sigma_1 > \sigma_3 > \sigma_2$. Based on the Eq. (3) and Eq. (4), finding 1 and finding 2 can be expressed.

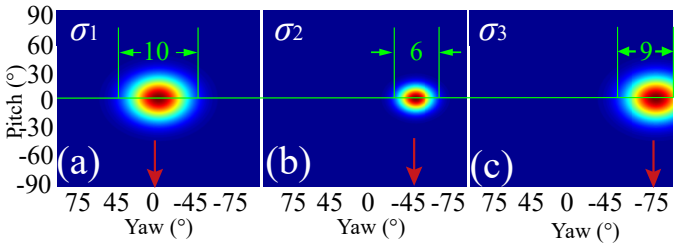


Fig. 7. Unsmooth variation of head pose angle distribution. (a) (0°, 0°), (b) (0°, -45°), and (c) (0°, -75°). The angle distribution of -45° is smaller than that of 0° and -75°. With the fixed angle interval increasing, the image variations increase firstly and then decrease in yaw angle direction.

3. Methodology and optimization

3.1 MAP-based HPE model

The *maximum a posterior* (MAP) estimation method, which inherently includes *a priori* constraints in the form of *prior* probability density functions, has been wildly used in a mass of applications. In this paper, we illustrate that the head pose estimation problem can be converted as the maximizing the posterior probability problem. Given a set of face images X with the constructed ground-truth angle distribution $\hat{Y}$, the aim of training is to find the best θ estimation by maximizing the posterior probability $p(\theta|X, \hat{Y})$. The θ denotes the all the neural network parameters in the proposed AADL model. It can be written as,

$$\theta^* = \arg\max_{\theta} p(X, \hat{Y} \mid \theta). \tag{5}$$

According to Bayes rule, Eq. (5) becomes,

$$\theta^* = \arg\max \frac{p(X, \hat{Y} \mid \theta)p(\theta)}{p(X, \hat{Y})}. \tag{6}$$

Since $p(\theta|X, \hat{Y})$ is independent of $p(X, \hat{Y})$, $p(X, \hat{Y})$ can be considered as a constant. Hence, Eq. (6) can be rewritten as,

$$\theta^* = \arg\max\, p(X, \hat{Y} \mid \theta)p(\theta). \tag{7}$$

Employing the monotonic logarithm function, Eq. (7) can be rewritten as

$$\theta^* = \arg\max \left(\log p(X, \hat{Y} \mid \theta) + \log p(\theta)\right). \tag{8}$$

There are two probability density functions need to be defined. The likelihood probability $p(X, \hat{Y}|\theta)$ represents the distance between the predicted distribution and the ground-truth distribution. Kullback-Leibler (KL) divergence is selected to measure the distance. Consequently, the likelihood probability can be presented as

$$p(X, \hat{Y} \mid \theta) = \sum_{t} \hat{y}_t \ln \frac{\hat{y}_t}{\hat{y}_t^*}, \tag{9}$$

where $\hat{y}^*$ denotes the prediction label.

To reduce the overfitting problem, a smooth constraint is regularized on the parameters θ of CNN. In this paper, the Gaussian distribution is introduced to suppress the noise error in parameters θ. Namely, the *priori* probability $p(\theta)$ in Eq. (8) can be formulated as,

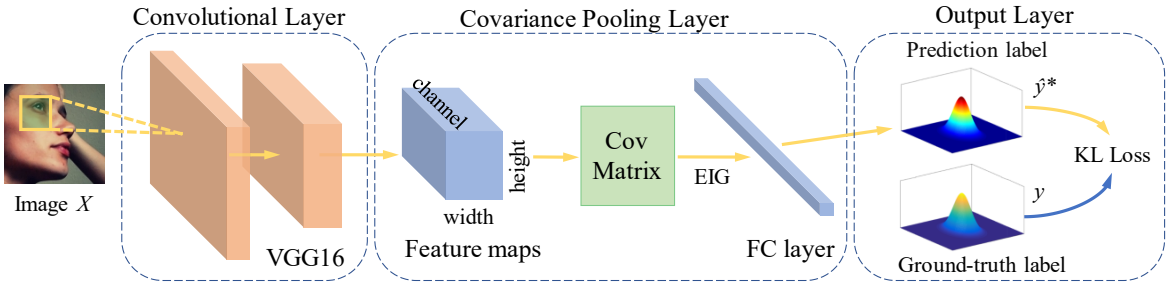


Fig. 8. Concrete architecture of the proposed AADL neural network.

$$p(\theta)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\theta^2}{2\sigma^2}\right). \tag{10}$$

Then, substituting Eqs. (9) and (10) in Eq. (8), our MAP problem is transformed to an object minimization that minimizes the negative logarithm of the probability. Hence, Eq. (8) can be rewritten as

$$\theta^*=\arg\min\left(\sum_t \hat{y}_t \ln\frac{\hat{y}_t}{\hat{y}_t^*}+\lambda\|\theta\|_2^2\right). \tag{11}$$

Finally, the loss function of AADL model is proposed as:

$$L(\theta)=\sum_t \hat{y}_t \ln\frac{\hat{y}_t}{\hat{y}_t^*}+\lambda\|\theta\|_2^2, \tag{12}$$

where the symbol λ denotes the regularization coefficient. 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 λ is set as 1×10^{-4}. It is worth noting that the proposed model (12) is an instance of the likelihood probability $p(X, \hat{Y}|\theta)$ and *prior* probability $p(\theta)$ in the MAP framework (8).

3.2 Network architecture

In Fig. 8, the proposed AADL network includes three parts: convolutional layer, covariance pooling layer and output layer. For convolutional layer, backbone of the VGG16 [25] is utilized to extract the features of the input images. The size of the image size is 224×224×3 (height, width and channel). And the feature matrix size of output of the final convolutional layer (input of the covariance pooling layer) is 7×7×512. Traditional CNNs are designed with convolutional layers, pooling layers and FC layers to capture only first-order statistics such as mean or maximum. It is considered that the second-order statistics such as covariance are deemed to be better regional descriptors than first-order statistics [26]. HPE task is more directly bound up with how facial key points are distorted, rather than presence or absence of specific key points. Consequently, second-order statistics are more suitable to capture such distortions than first-order statistics. We introduce covariance pooling rather than average or maximum pooling after the last convolutional layer, and build covariance matrices as global image representations. The backpropagation is not easy due to the nonlinear functions involved by covariance pooling. We refer the methodology of calculating gradients in [27] for end-to-end learning.

3.3 Covariance pooling layer

Let $E\in\mathrm{R}^{d\times N}$ be a matrix produced by the last convolutional layer. Its columns consist of a sample of N features of dimension d. The covariance matrix C of X is computed as

$$\mathrm{C}=\mathrm{E}\bar{\mathrm{I}}\mathrm{E}^T, \tag{13}$$

where $\bar{\mathbf{I}}=\frac{1}{N}(I-\frac{1}{N}11^T)$, $\mathbf{I}$ denotes the identity matrix and $\mathbf{1}=[1,\ldots,1]^T$ is a vector whose dimensional is N. Hence, a symmetric positive semi-definite of covariance matrix C is obtained. Its eigenvalue decomposition (EVD) is given by

$$\mathrm{C}=\mathrm{U}\Lambda\mathrm{U}^T, \tag{14}$$

where $\Lambda=\mathrm{diag}(\mu_1,\ldots,\mu_d)$ is a diagonal matrix whose eigenvalues μ_i are arranged in non-increasing order. $\mathrm{U}=[\mathrm{u}_1,\ldots,\mathrm{u}_d]$ is an orthogonal matrix and its column u_i is the eigenvector corresponding to μ_i. Matrix power is converted to the power of eigenvalues by EVD. Thus, we have

$$\mathrm{D}=\mathrm{C}^\delta=\mathrm{U}\mathrm{F}(\Lambda)\mathrm{U}^T, \tag{15}$$

where δ is a positive real number. Let $\mathrm{F}(\Lambda)=\mathrm{diag}(f(\mu_1),\ldots,f(\mu_d))$, which is given by

$$f(\mu_i)=\mu_i^\delta. \tag{16}$$

Finally, matrix D is input to the FC layer. We remove most of FC layers which roughly contain 90% parameters of the whole model instead of a covariance pooling. The label distribution of sample is learned after softmax, as shown in Fig. 8.

3.4 Optimization

Matrix backpropagation is utilized to compute the partial derivative of loss function (12). The expression of chain rule is given as,

$$\left(\frac{\delta L}{\delta \mathrm{D}}\right)^T\mathrm{dD}=\left(\frac{\delta L}{\delta \mathrm{U}}\right)^T\mathrm{dU}+\left(\frac{\delta L}{\delta \Lambda}\right)^T\mathrm{d}\Lambda. \tag{17}$$

According to Eq. (17), we can obtain,

$$\begin{cases}\dfrac{\delta L}{\delta \mathrm{U}}=\left(\left(\dfrac{\delta L}{\delta \mathrm{D}}\right)^T+\left(\dfrac{\delta L}{\delta \mathrm{D}}\right)^T\right)\mathrm{UF}\\ \dfrac{\delta L}{\delta \Lambda}=M\left(\mathrm{diag}(\mu_1^{\delta-1},\cdots,\mu_d^{\delta-1})\mathrm{U}^T\dfrac{\delta L}{\delta \mathrm{D}}\mathrm{U}\right)_{\mathrm{diag}}\end{cases}, \tag{18}$$

where M_{diag} represents the operation, which can keep the diagonal entries of M while all non-diagonal entries are set as 0. Furthermore, the derivative of L is computed by,

$$\left(\frac{\delta L}{\delta C}\right)^T\mathrm{d}C=\left(\frac{\delta L}{\delta \mathrm{U}}\right)^T\mathrm{dU}+\left(\frac{\delta L}{\delta \Lambda}\right)^T\mathrm{d}\Lambda, \tag{19}$$

where U is the orthogonal constraint. Thus, the gradient of L with respect to the E can be represented as following:

$$\frac{\delta L}{\delta E} = \bar{\mathbf{I}}\mathrm{E}\left(\frac{\delta L}{\delta C} + \frac{\delta L}{\delta \Lambda}\right)^{T}. \tag{20}$$

Then, the AADL algorithm can be presented as Algorithm 1. The source Python code will be available upon request.

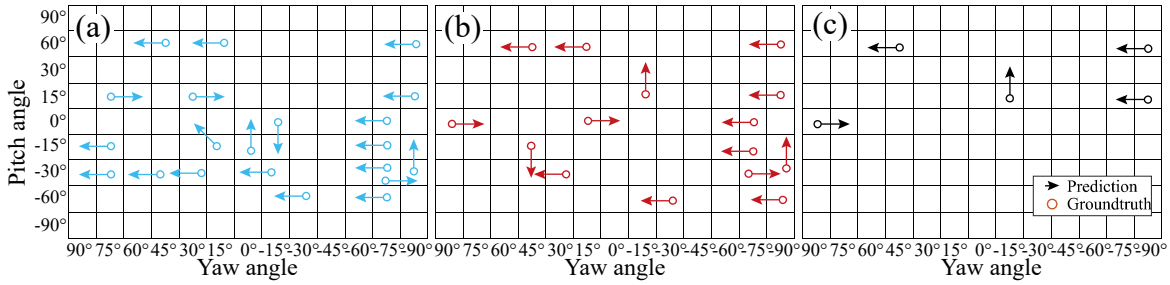


Fig. 9. Comparison of false predictions by different HPE methods. The circle represents the ground-truth poses, and the poses pointed by the arrows are the false predicted poses. (a) Hard label. (b)DLDL method [31]. (c) Proposed AADL method.

Algorithm 1. Training strategy for the proposed AADL model.

Input: Head pose image X in image set.
Set: Batch size s, η and α, *parameters* β_1 and β_2
1: Initialize network parameters θ via pre-trained model;
2: Network forward propagation;
3: **while not** loss function L is converged **do**:
sample a mini batch with size s from image X;
update θ with mini batch via stochastic gradient descent algorithm;
end while
Output: AADL model θ.

4. Experiment and discussion

4.1 General setting

1) Experiment platform: The famous Tensorflow is utilized to implement the proposed AADL network. The experiments are executed on a computer with an NVIDIA GeForce GTX TITAN V GPU and an Intel i7 CPU.

2) Training details: All images in this paper are scaled to the size of 224 × 224. All networks are optimized by the Adam [28] optimizer. The parameters are set as $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in our experiments. 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.

3) Datasets: Two public HPE datasets are utilized in this paper.

a) *Pointing'04* dataset [29]: It contains 15 subjects, whose ages, genders, hairstyles are different, with a total of 2790 human face images. Yaw and pitch angles are in the range [-90°, 90°]. Due to the yaw angle is 0 when the pitch angle is ± 90°, each subject consists of total 93 discrete poses.

b) *CAS-PEAL-R1* dataset [30]: It includes 1040 subjects 30,900 images with a total of 30,900 images. The ranges of yaw and pitch angles are in range [-30°, 0°, 30°] and [-45°, -30°, -15°, 0°, 15°, 30°, 45°].

Table 1. Comparison of *MA* and *MAE* values by five different architectures.

Network architecture	*MAE* (°)			*MA* (%)		
	Pitch	Yaw	All	Pitch	Yaw	All
AlexNet	1.60	2.62	2.17	90.24	81.33	79.25
ResNet50	1.20	2.74	1.97	90.45	83.26	80.26
ZfNet	1.17	2.47	1.82	92.36	83.78	81.45
VGG16	0.89	2.01	1.48	95.18	88.54	84.45
ADDL	**0.71**	**1.62**	**1.23**	**97.09**	**90.57**	**87.31**

Table 2. Comparison of *MA* and *MAE* by different algorithms on Pointing'04.

Methods	*MAE* (°)			*MA* (%)		
	Pitch	Yaw	All	Pitch	Yaw	All
CNNs [4]	8.06	6.93	-	73.91	66.6	-
GLLiM [27]	7.2	6.7	13.2	-	-	-
SIFT-RP [28]	5.84	6.05	-	-	-	-
MLD [26]	2.83	4.41	6.74	84.98	71.61	61.76
IndepCA(HoG) [29]	2.76	4.31	6.53	85.34	72.87	63.84
CartCA/MvCA [29]	2.04	3.25	5.01	89.21	78.96	70.93
DLDL [31]	1.69	3.16	4.64	91.65	79.57	73.15
AADL	**0.71**	**1.62**	**1.23**	**97.50**	**89.87**	**87.32**

Table 3. Comparison of *MA* and *MAE* with different algorithms on the public database CAS-PEAL-R1.

Methods	*MAE* (°)			*MA* (%)		
	Pitch	Yaw	All	Pitch	Yaw	All
DLDL [31]	0.57	0.49	0.51	98.45	97.96	97.38
RF + LDA [13]	-	0.42	0.54	-	96.25	97.23
DCNN	-	-	0.60	-	-	97.17
LGBP [30]	-	-	0.65	-	-	97.14
kVoD [30]	-	1.02	-	-	94.2	-
AADL	**0.21**	**0.15**	**0.19**	**99.35**	**98.87**	**99.26**

In this paper, the fivefold cross-validation technique and 80%–20% train-test settings are employed.

4) Metrics: Two metrics are adopted to evaluate different algorithms: mean accuracy (*MA*) and mean absolute error (*MAE*). The *MA* index is defined as following:

$$MA = \frac{1}{k}\sum_{i=1}^{N} acc_i \,, \tag{25}$$

where acc_i is the accuracy in i-th validation, and k represents the number of cross validation. The *MAE* is formulated as follows,

$$MAE = \frac{1}{2N}\sum_{n=1}^{N}\left(|\hat{\theta}_n - \theta_n| + |\hat{\varphi}_n - \varphi_n|\right), \tag{26}$$

where θ_n and φ_n are the ground-truth labels in yaw and pitch directions, respectively. The smaller the *MAE* value is, the higher the accuracy achieves.

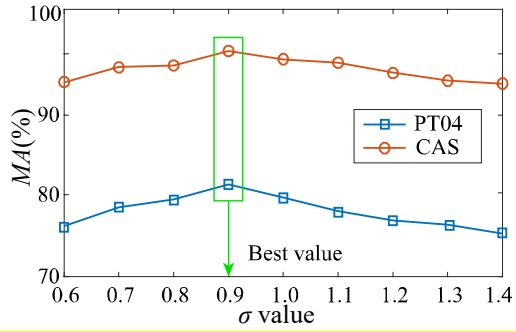


Fig. 10. *MA* values change with the parameter σ increasing on Pointing'04 and CAS-PEAL-R1 datasets. The parameter σ is robust on the different datasets.

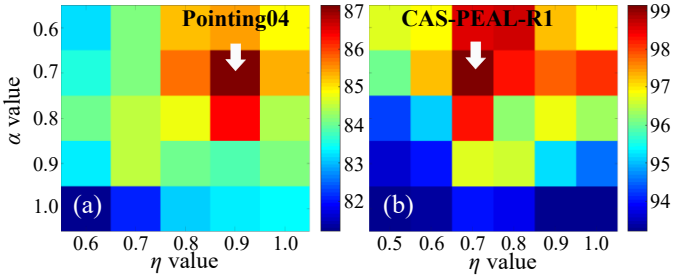


Fig. 11. Effect of the parameters η and α on the mean accuracy. (a) Pointing04 and (b) CAS-PEAL-R1.

4.2 Results and discussion

4.2.1 Accuracy analysis of different architecture

Four different advanced networks in the proposed method are compared on Pointing'04 dataset, such as AlexNet [31], ResNet50 [32], Inception-v2 [33], and VGG16. Our covariance pooling network uses the convolutional layer of VGG16 as backbone net to extract facial features.

In Table 1, we demonstrate the comparison of *MAE* and *MA* values among the five different architectures. It can be observed that the shallow architectures including the proposed network surpass the other two deep ones. One potential reason is that deep architectures are not suitable for small-scale datasets such as Pointing'04.

4.2.2 Accuracy analysis of AADL

For accuracy comparison, our algorithm is compared with several state-of-the-arts algorithms on two datasets. It includes the MLD [34], GLLiM [35], SIFT + RP [36] and IndepCA [37]. In our experiments, HoG features are used by MLD, GLLiM, SIFT-RP and IndepCA, and CNN features is utilized by other methods. The *MA* and *MAE* of these models are summed in Table 2 and Table 3. For CAS-PEAL-R1 dataset, our method achieves better *MA* and *MAE* values than any other algorithms. It indicates that the proposed method could explore the distinguishing features 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 (RF) and combined linear discriminant analysis (LDA). In [38], fisher vector of local descriptors (VoD) or its variant (kVoD) are distilled and nearest centroid (NC) classifier is employed to estimate head pose. The accuracy is still very low proposed since the proposed method can adequately utilize internal information between head pose images.

To verify the effectiveness of two findings, we record the results predicted by three compared methods, such as general classification with hard labels, DLDL [39] and our AADL. The prediction results by the three compared methods shown in Fig. 9. The circle represents the ground-truth pose, and the poses pointed by the arrow is the predicted pose. As can be seen, the pose tends to be mis-predicted in the adjacent yaw direction rather than those in pitch direction. And, false predictions are mostly concentrated on the range of +30° to -30°, 60° to 90° and -60° to -90° in yaw direction. It can be considered that since face images are more similar in these intervals, models are hard to predict accurately. However, the proposed AADL model can alleviate high prediction error rate as more reliable information is utilized to train the model.

4.2.3 Parameters discussion

The anisotropic angel labels distributions are determined by two parameters σ and η. The distribution become very sharp if the parameter σ is set as a small value. If σ is set as a large value, the distribution will become too smooth and joint less similar angle labels such as 45° to 0° in yaw angle. Thus, it is necessary to choose the value of σ cautiously. For this reason, to analyze the effect of σ, we conducted some experiments on two public datasets, which keep the rate value increasing from 0.6 to 1.4. The *MA* with different sampled σ is described in Fig. 9. From Fig. 10, it can be observed that the proposed method can achieve the best result when σ = 0.9, which is robust on Pointing'04 and CAS-PEAL-R1 datasets.

Moreover, according to the unsmooth variation property (key finding 2), we define the parameter α as the unsmooth ratio ($\alpha=\sigma_2/\sigma_1=\sigma_2/\sigma_3$) to explore the best values of the η. Then, we conducted some experiments with different parameters η and α on two public datasets. Experiment result is presented in Fig. 11(a), we can find that the result in η = 0.9 and α = 0.9 is better than any other cases on Pointing'04 dataset. Different

Table 4. *MAE* values comparison for pose angles when some poses are missed in the training process (Pointing'04).

Pose direction		-90	-75	-60	-45	-30	-15	0	15	30	45	60	75	90
Yaw (°)		√	√	√	√	√	√	√	√	√	√	√	√	√
		2.81	2.48	2.26	1.88	1.94	1.87	1.52	1.97	2.01	1.67	2.43	2.38	3.01
		√	×	√	×	√	×	√	×	√	×	√	×	√
		2.79	3.87	2.45	2.88	2.01	2.25	1.58	3.45	2.42	3.25	2.81	4.02	3.17
		√	×	×	√	×	×	√	×	×	√	×	×	√
		1.29	6.78	5.67	2.07	4.97	4.08	1.65	6.04	5.79	1.88	6.45	6.12	2.56
Pitch (°)		√	--	√	--	√	√	√	√	√	--	√	--	√
		1.57	--	0.93	--	0.62	0.84	0.67	0.97	0.58	--	0.87	--	1.25
		√	--	√	--	×	√	√	√	×	--	√	--	√
		1.79	--	1.45	--	2.11	1.39	0.98	0.95	1.62	--	1.21	--	1.27
		√	--	√	--	×	×	√	×	×	--	√	--	√
		1.12	--	0.85	--	6.51	7.39	0.68	6.95	7.62	--	0.91	--	1.07

from Pointing'04 dataset, CAS-PEAL-R1 dataset has only three angles [-30°, 0°, 30°] in the pitch direction and seven angles [-45°, -30°, -15°, 0°, 15°, 30°, 45°] in the yaw direction. Furthermore, the angle interval in the yaw direction is 15° but 45° in pitch direction. The value of σ is set as 0.9 according to the result in Fig. 10. Since the maximum yaw angle is 45°, we set $\eta = 1$ when the yaw angle equals ±45° based on the unsmooth variation property. From the experiment results presented in Fig. 11(b), it can be observed that the result is better than any other cases when $\eta = 0.7$ and $\alpha = 0.9$.

According to the quantitative analysis of the feature similarity, we suggest the values of parameters η and α are in the range (0.6, 1). The experiment results show that the values of η and α are indeed within this range. Thus, it can be concluded that our analysis is reasonable.

Fig. 12. Effect of the head pose missing. AADL method can well predict head pose angles if they are partly missed in the training set. The face images with red borders denote the missed pose images. The results are showed in Table 4.

4.3 Robust performance for head pose missing

To demonstrate the robustness of the angle missed, we execute the proposed AADL method with different levels of image data missed. In Fig. 12, we show one group of pose images with partly data missed in the training set. Two network models are considered in this experiment. One model is trained on all the yaw angles, and the compared one is trained on the part yaw angles. We evaluate the estimation accuracy in the testing dataset. It is worth noting that the yaw angles of pose images in the testing dataset may be removed in the training dataset. Experimental results on Pointing'04 and CAS-PEAL-R1 datasets are illustrated in Table 4 and Table 5, respectively.

Table 5. *MAE* for pose angles not included in the training dataset (CAS-PEAL-R1). (Unit: °)

Yaw (°)	-45	-30	-15	0	15	30	45
	√	√	√	√	√	√	√
MAE	0.21	0.20	0.13	0.12	0.15	0.18	0.21
	√	×	√	×	√	×	√
MAE	0.27	0.41	0.19	0.58	0.17	0.48	0.23

The experiment of pitch angles is only performed in the Pointing'04 since the pitch angles are too sparse in CAS-PEAL-R1 dataset. The result is presented in Table 5. It can be observed that: i) The *MAE* of head pose increases when the angles of the testing data are not in the training set; ii) The increment of the *MAE* is small when the angle sampling interval in the training dataset is less than 30° both on yaw and pitch angles; iii) The *MAE* of head pose prediction increases substantially when the angle sampling interval in the training dataset is larger than 30°. These results suggest that our method can well predict the missed pose angles when the angle sampling interval is less than 30° in the training dataset.

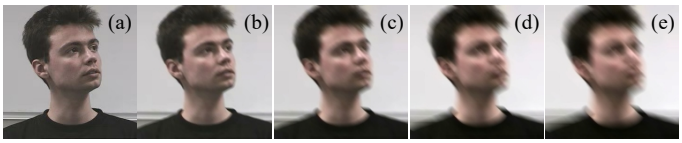


Fig. 13. From left to right, length values are 0 (original image), 5, 10, 15 and 20, respectively. With the length increasing, the motion blur is also raised correspondingly.

4.4 Effect of motion blurry for HPE

Usually, head pose images exist the motion blur problem since the car is waggling in the driving environment. To investigate the performance of the AADL network, the test image dataset is blurred with different blurry kernel. In Fig. 13, we show one group of the motion blurred images from the Pointing'04 dataset. The kernel of motion blurry is determined by two parameters, angle (or direction) and length. The angle is fixed at 0. The values of length are in range of {5, 10, 15, 20}. The larger the length value of blurry kernel, the less clear the head pose image.

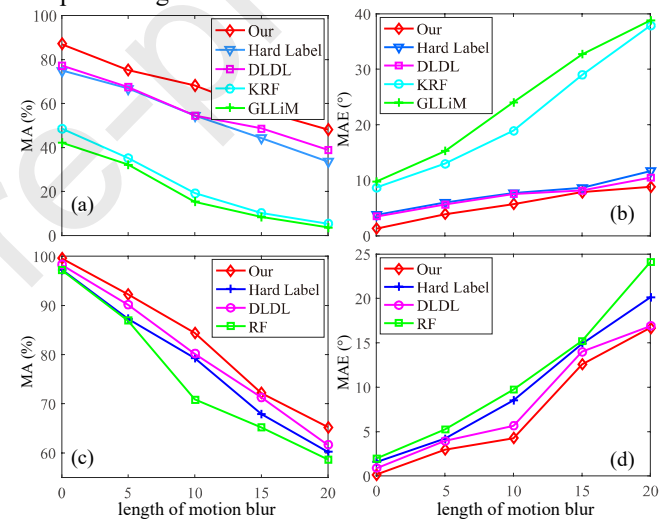


Fig. 14. Effect of motion blurred images on the ability of HPE on (a) (b) Pointing'04 and (c) (d) CAS-PEAL-R1.

Four compared methods are executed on the two public datasets, in which the images are blurred by different length blurry kernels. In Fig. 14, we plot the *MA* and *MAE* values with different length of motion blur kernel on Pointing'04 and CAS-PEAL-R1 datasets, respectively. It can be observed that *MA* values are decreased and *MAE* values are raised with the length value increasing from 0 to 20. Our method (red line with rhombus) still achieves the highest *MA* values in Fig. 14(a) and Fig. 14(c). With the motion blur increasing, the head pose image recognition rates are decreased correspondingly. In fact, the blur in head posed images leads to that it is very hard to extract the accurate image features. However, the proposed method can also work well due to the anisotropic angle distribution labels. Furthermore, the comparisons of comprehensive properties of all the HPE methods are provided in Table 6. The properties include the method type, input image type, feature extraction, computing speed, and performance. It can be observed that the proposed method achieves the

highlight performance.

Table 6. Comparison the performance between the proposed method and the state-of-the-art methods.

Methods	Type	Input	Feature	Speed	Performance
Multi-Loss [14]	Re + Cl	RGB	DL	★★★★☆	★★★★☆
FAN [40]	Re	RGB	DL	★★★★	★★★★
FSA-Net [19]	Re	RGB	DL	★★★★★	★★★★☆
KEPLER [18]	Re	RGB	DL	★★★★	★★★
MLD [27]	Cl	RGB	Hand	★★★	★★★
CartCA /MvCA [29]	Cl	RGB	Hand	★★★	★★★☆
KRF [30]	Cl	RGB	Hand	★★	★★☆
Our	Cl	RGB	DL	★★★★★	★★★★☆

5. Conclusion

In this paper, we propose a novel head pose estimation algorithm with the anisotropic angle distribution. Firstly, we analyze and reveal the two key findings in the human head pose, namely, anisotropic property and unsmooth variations property. Based on the MAP theory, the AADL model is proposed, in which the likelihood probability is constructed by KL divergence, and the priori probability is constructed as the Gaussian-based distribution. To train the model, a novel network which adopts a CNN with covariance pooling is proposed and the results shows it is superior to other well-known networks. Moreover, we investigate the effect of missing angle and motion blur in head pose estimation. Extensive experiments illustrate the advantages of our algorithm and demonstrate state-of-the-art performance at the aspect of prediction accuracy and good robustness on two public head pose datasets. In future, we will examine the 3D head pose video including roll direction for HPE task.

Acknowledgements

The authors would like to thank the reviewers for their valuable and helpful suggestions. This work was supported in part by the National Natural Science Foundation of China under Grant 61875068, Grant 61873220, and Grant 61505064, 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] K. Wang, R. Zhao, Q. Ji, Human computer interaction with head pose, eye gaze and body gestures, 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 789-789.

[2] Y. Yu, K.A.F. Mora, J.-M. Odobez, HeadFusion: 360° Head Pose Tracking Combining 3D Morphable Model and 3D Reconstruction, IEEE transactions on pattern analysis and machine intelligence, 40 (2018) 2653-2667.

[3] Y. Wang, W. Liang, J. Shen, Y. Jia, L.-F. Yu, A deep Coarse-to-Fine network for head pose estimation from synthetic data, Pattern Recognition, 94 (2019) 196-206.

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

[5] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, P. Bao, Appearance based pedestrians' head pose and body orientation estimation using deep learning, Neurocomputing, 272 (2018) 647-659.

[6] A. Balderas, L. De-La-Fuente-Valentin, M. Ortega-Gomez, J.M. Dodero, D. Burgos, Learning management systems activity records for students' assessment of generic skills, IEEE Access, 6 (2018) 15958-15968.

[7] H. Tang, H. Liu, W. Xiao, N. Sebe, Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion, Neurocomputing, 331 (2019) 424-433.

[8] J. Wu, M.M. Trivedi, A two-stage head pose estimation framework and evaluation, Pattern Recognition, 41 (2008) 1138-1158.

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

[10] L. Liang, R. Xiao, F. Wen, J. Sun, Face alignment via component-based discriminative search, European conference on computer vision, (Springer2008), pp. 72-85.

[11] 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.

[12] C. Huang, X. Ding, C. Fang, Head pose estimation based on random forests for multiclass classification, 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 934-937.

[13] 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.

[14] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, Z. Zeng, CLU-CNNs: Object detection for medical images, Neurocomputing, 350 (2019) 53-59.

[15] M. Giménez, J. Palanca, V. Botti, Semantic-based padding in convolutional neural networks for improving the performance in natural language processing. A case of study in sentiment analysis, Neurocomputing, 378 (2020) 315-323.

[16] 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.

[17] 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.

[18] 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, (2017) 2074-2083.

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

[20] 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 Recognition2019), pp. 1087-1096.

[21] G. Zhang, J. Liu, H. Li, Y.Q. Chen, L.S. Davis, Joint Human Detection and Head Pose Estimation via Multistream Networks for RGB-D Videos, IEEE Signal Processing Letters, 24 (2017) 1666-1670.

[22] X. Xu, I.A. Kakadiaris, Joint head pose estimation and face alignment framework using global and local CNN features, 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 642-649.

[23] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (1998) 2278-2324.

[24] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531, (2015).

[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, (2014).

[26] O. Tuzel, F. Porikli, P. Meer, Region Covariance: A Fast Descriptor for Detection and Classification, European Conference on Computer Vision, 2006.

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

[28] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv preprint arXiv:14126980.2014, (2014).
[29] N. Gourier, D. Hall, J.L. Crowley, Estimating face orientation from robust detection of salient facial structures, FG Net workshop on visual observation of deictic gestures, (FGnet (IST–2000–26434) Cambridge, UK, 2004, pp. 7-10.
[30] 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 Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 38 (2007) 149-161.
[31] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 2012, pp. 1097-1105.
[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.
[33] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, 2015.
[34] X. Geng, Y. Xia, Head Pose Estimation Based on Multivariate Label Distribution, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1837-1842.

Abstract:

Head pose estimation is an important way to understand human attention. In this paper, we propose a novel anisotropic angle distribution learning (AADL) network for head pose estimation task. Firstly, two key findings are revealed as following: 1) Head pose image variations are different at the yaw and pitch directions with the same pose angle increasing on a fixed central pose; 2) With the fixed angle interval increasing, the image variations increase firstly and then decrease in yaw angle direction. Then, the *maximum a posterior* technology is employed to construct the head pose estimation network, which includes three parts, such as convolutional layer, covariance pooling layer and output layer. In the output layer, the labels are constructed as the anisotropic angle distributions on the basis of two key findings. And the anisotropic angle distributions are fitted by the 2D Gaussian-like distributions (groundtruth labels). Furthermore, the Kullback-Leibler divergence is selected to measure the predication label and the groundtruth one. The features of head pose images are perceived at the AADL-based convolutional neural network in an end-to-end manner. Experimental results demonstrate that the developed AADL-based labels have several advantages, such as robustness for head pose image missing, insensitivity for the motion blur. Moreover, the proposed method has achieved good performance compared to several state-of-the-art methods on the Pointing'04 and CAS_PEAL_R1 databases.

HAI LIU (S'12–M'14) 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 2020. He has authored more than 60 peer-reviewed articles in international journals from multiple domains such as pattern recognition, image processing.

His current research interests include big data processing, artificial intelligence, spectral analysis, optical data processing and pattern recognition. Dr. Liu has been frequently serving as a reviewer for more than six international journals including the *IEEE Translations on Industrial Informatics, IEEE Translations on Cybernetics, IEEE/ASME Transactions on Mechatronics,* and *IEEE Translations on Instrumentation and Measurement*. He is also a Communication Evaluation Expert for the National Natural Science Foundation of China.

[35] V. Drouard, S. Ba, G. Evangelidis, A. Deleforge, R. Horaud, Head pose estimation via probabilistic high-dimensional regression, 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 4624-4628.
[36] H.T. Ho, R. Chellappa, Automatic head pose estimation using randomly projected dense sift descriptors, 2012 19th IEEE international conference on image processing, 2012, pp. 153-156.
[37] K. Chen, K. Jia, H. Huttunen, J. Matas, J.-K. Kämäräinen, Cumulative attribute space regression for head pose estimation and color constancy, Pattern Recognition, 87 (2019) 29-37.
[38] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: A survey, IEEE transactions on pattern analysis and machine intelligence, 31 (2008) 607-626.
[39] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Transactions on Image Processing, 26 (2017) 2825-2838.
[40] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021-1030.

HANWEN NIE received the B.S. degrees from Lanzhou Jiaotong University, LanZhou, China, in 2016. 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 human–computer interaction.

ZHAOLI ZHANG (M'18) 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 (M'91–SM'01) 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.

Highlights:
1. A novel anisotropic angle distribution learning (AADL) method is proposed for head pose estimation.
2. For a central pose, head pose image variations are different even increasing the same pose angle in yaw and pitch directions.
3. A 2D Gaussian-like distribution is defined to fit the anisotropic angle labels.
4. The robustness of the proposed model is verified by extensive experiments.

Credit Author Statement

Hai Liu: Writing - review & editing.
Hanwen Nie: Writing-original draft.
Zhaoli Zhang: Data curation.
You-Fu Li: Conceptualization, Methodology.