Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Unseen head pose prediction using dense multivariate label distribution*

Gao-li SANG[1,2], Hu CHEN[1], Ge HUANG[1], Qi-jun ZHAO‡[1]

([1]*State Key Laboratory of Fundamental Science on Synthetic Vision, College of Computer Science, Sichuan University, Chengdu 610064, China*)

([2]*College of Mathematics and Information Engineering, Jiaxing University, Jiaxing 314001, China*)

E-mail: g.sang@foxmail.com; huchen@scu.edu.cn; 26434368@qq.com; qjzhao@scu.edu.cn

Received July 23, 2015; Revision accepted Feb. 16, 2016; Crosschecked May 6, 2016

Abstract: Accurate head poses are useful for many face-related tasks such as face recognition, gaze estimation, and emotion analysis. Most existing methods estimate head poses that are included in the training data (i.e., previously seen head poses). To predict head poses that are not seen in the training data, some regression-based methods have been proposed. However, they focus on estimating continuous head pose angles, and thus do not systematically evaluate the performance on predicting unseen head poses. In this paper, we use a dense multivariate label distribution (MLD) to represent the pose angle of a face image. By incorporating both seen and unseen pose angles into MLD, the head pose predictor can estimate unseen head poses with an accuracy comparable to that of estimating seen head poses. On the Pointing'04 database, the mean absolute errors of results for yaw and pitch are 4.01° and 2.13°, respectively. In addition, experiments on the CAS-PEAL and CMU Multi-PIE databases show that the proposed dense MLD-based head pose estimation method can obtain the state-of-the-art performance when compared to some existing methods.

Key words: Head pose estimation, Dense multivariate label distribution, Sampling intervals, Inconsistent labels
http://dx.doi.org/10.1631/FITEE.1500235 **CLC number:** TP391.4

1 Introduction

A head pose refers to the direction of a human head with respect to his/her intrinsic coordinate system in the three-dimensional space. The origin is usually assumed to be at the head centroid or at the nose tip, x and y axes are corresponding to the horizontal and vertical directions, respectively, and z axis is perpendicular to the x-y plane. A head pose provides useful information in many computer vision related applications, such as gaze estimation (Lu *et al.*, 2012; 2014), fatigue driving detection, behavior analysis, and human-computer interfaces. In face recognition (Ma XH *et al.*, 2013), however, pose variations are a well-known challenge (Bowyer *et al.*, 2006). To better recognize non-frontal faces, various pose normalization, pose correction, and pose-adaptive methods have been proposed. Most of these methods require estimating head pose angles before extracting and matching facial features. Yet, accurately estimating pose angles from a single two-dimensional (2D) face image is still an open issue.

Existing head pose estimation methods can be roughly divided into two categories according to whether or not facial landmarks are used. The methods in the first category mainly explore the

‡ Corresponding author

* Project supported by the National Key Scientific Instrument and Equipment Development Project of China (No. 2013YQ49087903) and the National Natural Science Foundation of China (No. 61202160)

ORCID: Gao-li SANG, http://orcid.org/0000-0002-6567-1652

correlation between facial geometric features and head pose angles (Brunelli, 1997; Fitzpatrick, 2000; Wu and Trivedi, 2008; Aghajanian and Prince, 2009; Zhu and Ramanan, 2012; Cai *et al.*, 2015). They first locate a set of facial landmarks on the input face image, and then estimate head poses based on the topology of these landmarks. These methods have achieved impressive results especially on the faces with relatively small pose angles. However, their performance depends highly on the accuracy of landmark localization, which itself is a challenging task. Besides, the assumption underlying these methods is arguable, because the difference between the facial landmark configurations might be due either to different head poses or to different faces (i.e., identities). Some researchers thus proposed to learn the correlation between facial texture features and head pose angles directly from a set of training data, which leads to the methods in the second category (Krüger and Sommer, 2002; Fenzi *et al.*, 2013; Geng and Xia, 2014; Hu *et al.*, 2014; Ma *et al.*, 2015). The methods in the second category extract certain texture features (e.g., Gabor filter responses, local binary patterns (LBPs), histograms of oriented gradients (HOG), and other new user-defined features) from the cropped face images, and train a head pose predictor with a set of face images whose pose angles are known. Given a new input face image, they use the obtained predictor to estimate the head pose based on the extracted texture features. Compared with the landmark-based methods, these training-based methods do not need to detect facial landmarks, but directly extract appearance features from the face images. They are thus easier to apply, but depend on the used training data. Details about existing head pose estimation methods can be found in Murphy-Chutorian and Trivedi (2009) and Tang *et al.* (2014).

A major limitation of training-based methods is that they rely heavily on the training data. The face images in the training data usually contain only a finite number of head pose angles. Take the CMU-PIE (Sim *et al.*, 2002) and CAS-PEAL (Gao *et al.*, 2008) face image databases, two of the most widely used databases in cross pose face recognition, as examples. The yaw angles (i.e., rotation angles with respect to the y axis) covered by them are $\{-90°, -67.5°, -45°, -22.5°, 0°, 22.5°, 45°, 67.5°, 90°\}$ and $\{-45°, -30°, -15°, 0°, 15°, 30°, 45°\}$, respectively. In the Pointing'04 database (Gourier and Letessier, 2004), a database most widely used in head pose estimation, the range of considered pitch angles (i.e., rotation angles with respect to the x axis) is $\{-90°, -60°, -30°, -15°, 0°, 15°, 30°, 60°, 90°\}$. However, arbitrary head pose angles can be observed in real-world applications, such as face recognition for video surveillance. In other words, the available training data with discrete head pose angles are merely a very sparse sampling of the full head pose space, which is intrinsically continuous (Fig. 1).

During the past decades, a number of regression-based methods have been developed to estimate continuous head pose (Aghajanian and Prince, 2009; Huang *et al.*, 2011; Zhu *et al.*, 2013). Zhu *et al.* (2013)

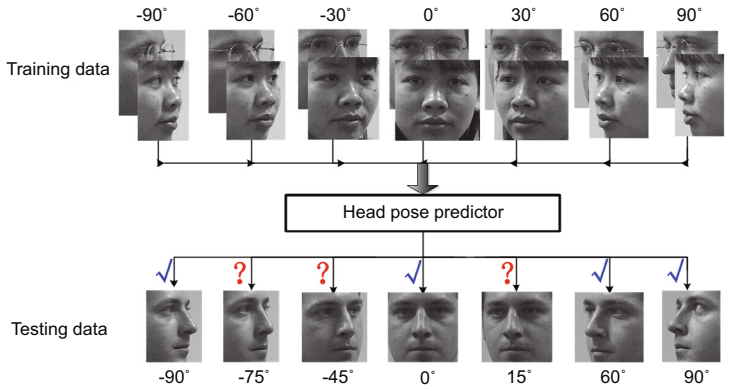

Fig. 1 Illustration of problem solving through our proposed method. The head pose predictor trained on a database which contains a finite number of pose angles could be problematic in estimating arbitrary head pose angles, especially the pose angles not included in the training data. Our proposed dense multivariate label distribution based method is able to well predict head pose angles both in and out of the training data

learned random regression forests to estimate continuous poses by assigning the face to the corresponding exact pose label. Aghajanian and Prince (2009) proposed a probabilistic framework that represents a face with a non-overlapping grid of patches. The experimental results have demonstrated good performance on face images taken under uncontrolled conditions. Assuming that one local tangent subspace can be reconstructed by its neighboring subspaces, Huang *et al.* (2011) divided the input feature space into a number of local tangent subspaces, each corresponding to a part of the output head pose space. Then the reconstructed full feature space could be predicted over pose angles even when the training set is not uniformly sampled. One limitation of these methods is that, despite their good performance, it is still not clear how these methods work on pose angles that are unseen in the training data. In other words, although these methods can estimate continuous head poses, they have not been systematically evaluated for predicting unseen head poses.

In this study, we aim at predicting unseen pose angles and further improving the head pose estimation accuracy by taking into account the limitation of the training data. Specifically, we treat the labeled head pose angles as soft labels. Each face image in the training dataset is associated with a dense multivariate label distribution (MLD), which indicates the relevance of different pose angles to the face image. Given a training face image, not only its labeled head pose but also the poses around it are set with non-zero relevance to the face image in its dense MLD. The parameters involved in computing the dense MLD of a face image from the features extracted from it are learned from the training dataset via some optimization process. For a new face image, its dense MLD can be then directly computed from its features, and its head pose is supposed to be the one with the maximum relevance in the dense MLD. According to our experimental results, our proposed method can successfully predict unseen head poses. Moreover and surprisingly, it improves the estimation accuracy for the head poses in the training data.

2 Dense multivariate label distribution

In this study, we consider two degrees of freedom in head rotation (i.e., pitch and yaw). Instead of assigning an exact head pose angle to a face image, we assume that a face image can be associated with a variety of pose angles. This is represented by an MLD. To be precise, we define the MLD of a face image as a 2D matrix $\boldsymbol{L}\in\mathbb{R}^{M\times N}$, whose rows correspond to M possible pitch angles and columns to N possible yaw angles. The entry l_{ij} in $\boldsymbol{L}$ indicates the relevance of the input face image to the head pose of the ith pitch angle (θ_i^x) and the jth yaw angle (θ_j^y). Geng and Xia (2014) constructed an MLD according to the head pose angles available in the used training database (i.e., Pointing'04). To predict pose angles which are unseen in the training data, in our method, not only the pose angles that are included in the training data but also the pose angles that are not included in the training data are considered when constructing the MLD. The basic idea is to use a smaller pose angle sampling interval through interpolation when constructing the MLD (e.g., one degree sampling interval). In this way, not only the pose angles that are included in the training data but also the pose angles close to the existing poses that are not included in the training data have a non-zero relevance to the face image in its MLD. As a result, pose angles close to the existing poses can be predicted by the non-zero relevance to the pose. Theoretically, using small enough intervals to construct an MLD can predict any pose within the scope of a full pose space. However, density of the training sampled poses matters. Too sparse sampling intervals will not help predict arbitrary pose face images, whereas too dense sampling intervals will lead to high computational cost. Fig. 2 shows an example of a dense MLD.

Given a face image with a labeled 'ground truth' head pose (θ_i^x,θ_j^y), its dense MLD, $\boldsymbol{L}$, is constructed as follows. First, all entries in $\boldsymbol{L}$ are initialized as zeros. Second, the entries around l_{ij} (i.e., the one corresponds to the labeled head pose) are set according to

$$l_{mn}=\frac{1}{Z}\exp\left(-\frac{1}{2}\begin{pmatrix}\theta_m^x-\theta_i^x\\ \theta_n^y-\theta_j^y\end{pmatrix}^{\mathrm{T}}\cdot\boldsymbol{\Sigma}^{-1}\begin{pmatrix}\theta_m^x-\theta_i^x\\ \theta_n^y-\theta_j^y\end{pmatrix}\right),\tag{1}$$

where $\boldsymbol{\Sigma}=\begin{bmatrix}\tau^2 & 0\\ 0 & \tau^2\end{bmatrix}$ is a 2×2 covariance matrix (τ

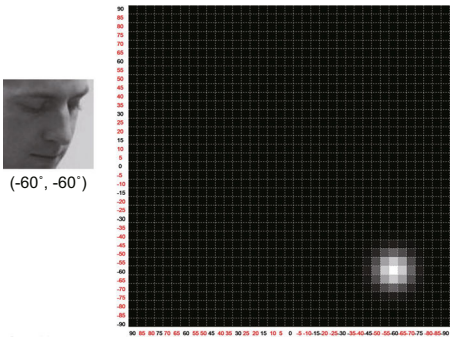

Fig. 2 A face image and its dense multivariate label distribution whose rows correspond to 37 possible pitch angles and columns to 37 possible yaw angles. The pose angles printed in red color are not included in the training data, i.e., the Pointing'04 database in this paper. References to color refer to the online version of this figure

is the finest granularity of the pose angles), and Z is a normalization factor to ensure $\sum_{mn} l_{mn} = 1$. This essentially defines a discretized bivariate Gaussian distribution centered at the labeled head pose. While the MLD has a maximum relevance at the labeled head pose (θ_i^x, θ_j^y), the relevance decreases gradually as the pose angle deviates from (θ_i^x, θ_j^y).

To estimate the head pose in a new input face image, what we have to do is to compute its MLD. Once the MLD is obtained, we can claim that the input face image has a head pose at which the relevance in the MLD is the maximum. Fig. 3 gives the flowchart of our proposed dense MLD-based head pose estimation method. In the next section, we will show how to compute the MLD of a face image based on the texture features extracted from it.

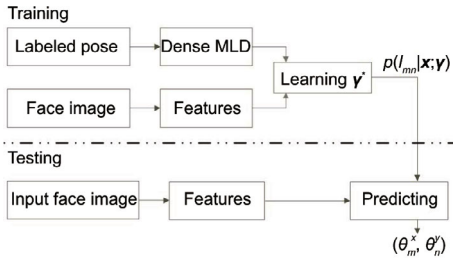

Fig. 3 Block diagram of the proposed dense multivariate label distribution (MLD) based head pose estimation method

3 Estimating MLD from facial features

As in Geng and Xia (2014), we assume that the MLD of a face image is determined by the texture features (particularly HOG) extracted from the image via a maximum entropy model (Berger *et al.*, 1996):

$$p(l_{mn}|\boldsymbol{x};\boldsymbol{\gamma}) = \frac{1}{C}\exp\left(\sum_r \gamma_{mnr}\cdot x_r\right), \qquad (2)$$

where $\boldsymbol{x}$ is the feature vector of the face image, x_r is the rth component of the feature vector, $\boldsymbol{\gamma}$ denotes the involved parameters with γ_{mnr} corresponding to the mth pitch angle, nth yaw angle, and rth feature component, and $C = \sum_{mn}\exp(\sum_r \gamma_{mnr}\cdot x_r)$ is the normalization factor.

To extract the feature vector from a face image, the face region cropped from the image is first converted into a normalized gray-scale image (32×32 pixels). Then the normalized face image is divided into a number of overlapping cells, from each of which a histogram is extracted counting the occurrences of pixels with specific gradient orientations. In our experiments, the cell size is 3×3. The gradient orientations within $[0, 2\pi)$ are evenly divided into nine histogram bins. The neighboring 3×3 cells are grouped into a block, in which the histogram bin values of these cells are combined as a vector with its magnitude normalized to one. Finally, the histograms in all the blocks are concatenated into a high-dimensional HOG feature vector.

Given a set of S training face images, their HOG facial features $\{\boldsymbol{x}^s|s=1,2,\ldots,S\}$ are extracted, and MLDs $\{\boldsymbol{L}^s|s=1,2,\ldots,S\}$ are constructed. To learn the parameters $\boldsymbol{\gamma}$, Geng and Xia (2014) used the weighted Jeffrey divergence to measure the distance between two MLDs. However, it is time-consuming in terms of computing weights between two poses. In this study, Kullback-Leibler (K-L) divergence (Do, 2003) is employed to measure the distance between two MLDs $\boldsymbol{L}^s$ and $\boldsymbol{L}^q$:

$$D_{\text{K-L}}(\boldsymbol{L}^s\|\boldsymbol{L}^q) = \sum_{i=0}^{M}\sum_{j=0}^{N}\frac{(l_{ij}^s - l_{ij}^q)^2}{l_{ij}^s + l_{ij}^q}. \qquad (3)$$

The best parameters $\boldsymbol{\gamma}^*$ can then be determined by minimizing the following objective function:

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}}\sum_{s=1}^{S} D_{\text{K-L}}\left(\boldsymbol{L}^s\|p(\boldsymbol{L}|\boldsymbol{x}^s;\boldsymbol{\gamma})\right). \qquad (4)$$

According to Eqs. (2) and (3), the above

objective function can be rewritten as

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}} \sum_{s,i,j} \frac{\left(l_{ij}^s - \frac{1}{C^s}\exp\left(\sum_r \gamma_{mnr}\cdot x_r^s\right)\right)^2}{l_{ij}^s + \frac{1}{C^s}\exp\left(\sum_r \gamma_{mnr}\cdot x_r^s\right)}, \tag{5}$$

where $C^s = \sum_{mn}\exp(\sum_r \gamma_{mnr}\cdot x_r^s)$. This minimization problem can be solved by using the limited-memory quasi-Newton method L-BFGS described in Liu and Nocedal (1989). With the obtained parameters $\boldsymbol{\gamma}^*$, we can easily estimate the head pose of a new input face image as follows. First, HOG features are extracted from the face region in the input face image. Then its MLD is computed based on the HOG features and Eq. (2). Finally, its head pose is estimated as the one corresponding to the maximum in the MLD.

4 Experiments and results

4.1 Databases and experimental setup

To evaluate the performance of the proposed method, we compare our proposed method with several state-of-the-art methods on three public databases: Pointing'04, CAS-PEAL, and CMU Multi-PIE.

The Pointing'04 database (Gourier and Letessier, 2004) contains 2D face images of 15 individuals. Each subject is captured in two series, where he/she displays a number of different head poses. A pose is represented by the combination of a yaw angle and a pitch angle. The yaw angle includes 13 values {−90°, −75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°, 90°}, and the pitch angle includes nine values {−90°, −60°, −30°,−15°, 0°, 15°, 30°, 60°, 90°}. When the pitch angle is −90° or 90°, the yaw angle is always 0° due to physical limitations of the human head. Thus, there are in total 93 (13×7+2) poses involved in the dataset.

The CAS-PEAL face database (Gao *et al.*, 2008) contains 99 594 images of 1040 individuals (595 males and 445 females). Each subject displays 21 poses combining seven yaw angles {−45°, −30°, −15°, 0°, 15°, 30°, 45°} and three pitch angles {−30°, 0°, 30°}.

The CMU Multi-PIE database (Gross *et al.*, 2010) contains head images of 336 subjects illuminated by a frontal light source under 13 viewing angles and six expressions. The considered yaw angles are {−90°, −75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°, 90°}.

For all the images in these databases, the face detector in Zhang *et al.* (2014) is applied to locate the face regions (when the method fails to detect the right face region, we crop the face region manually), and the face regions are cropped and normalized to the same size (32 × 32 pixels). Several example images of these databases are shown in Figs. 4–6. All experiments in this study were conducted using MATLAB on a 64-bit Windows workstation with Intel I3 CPU and 4 GB memory.

Fig. 4 Example face images in the Pointing'04 database

Fig. 5 Example face images in the CAS-PEAL database

Fig. 6 Example face images in the CMU Multi-PIE database

We employ the MLD-based method (denoted as MLD+BFGS) as the baseline. The proposed method is compared with the MLD-based methods, namely MLD-J and MLD-wJ (Geng and Xia, 2014), and several other state-of-the-art head pose estimation methods, including LAG (Hu *et al.*, 2014), VRF+LDA (Pang *et al.*, 2006), SL^2 (Huang *et al.*, 2011), LPLS and KPLS (Haj *et al.*, 2012), WRF (Zhu *et al.*, 2013), MGD (Jain and Crowley, 2013), CGM (Fenzi *et al.*, 2013), CovGa and kCovGa (Ma *et al.*, 2014), and kVoD (Ma *et al.*, 2015).

We measure the head pose estimation performance using the following metrics: (1) the mean absolute error (MAE) between the predicted pose and the 'ground truth', and (2) the estimation

accuracy, i.e., the percentage of the samples whose pose angles are estimated correctly. A sample's head pose is supposed to be estimated correctly if the predicted pose angle differs from the 'ground truth' within half of the pose angle sampling interval (i.e., 7.5° in the Pointing'04 database). Furthermore, the MAE of yaw+pitch is calculated by the Euclidean distance between the predicted (yaw, pitch) pair and the 'ground truth' (yaw, pitch) pair. The accuracy of yaw+pitch is calculated by regarding each (yaw, pitch) pair as a class.

4.2 Parameter selection

The covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\tau^2, \tau^2)$ in Eq. (1) indicates how the nearby poses are related to the 'ground truth'. The larger the parameter τ, the more the neighboring poses are assumed to be related to the 'ground truth'. We evaluate the performance of the proposed method by using the Pointing'04 database in terms of its MAE versus different values of τ ranging from 1 to 15. Fig. 7 shows the MAE of yaw and pitch angles on testing data when different τ's are used for dense MLD. Here, we define dense MLD as $\boldsymbol{L} \in \mathbb{R}^{37\times 37}$ to represent pitch and yaw angles (both from $-90°$ to $90°$ and with an increment step of 5°). Fig. 8 shows the required training time.

From Figs. 7 and 8, we can see that as τ increases, the MAEs of yaw and pitch decrease and then increase, while the training time is relatively stable (about 1 min). Too small or too large τ may lead to performance deterioration. In the following experiments, the finest granularity of the pose angles (τ) is thus set to five (just the same as the dense sampling interval).

4.3 Effectiveness in predicting unseen pose angles

It is not easy to compare our proposed method with other methods such as regression-based methods (Aghajanian and Prince, 2009; Huang *et al.*, 2011; Fenzi *et al.*, 2013; Zhu *et al.*, 2013), because these methods did not report the pose angles which are unseen in the training data. Instead, to demonstrate the effectiveness of trained pose angle predictors on face images whose poses are unseen in the training set, different biases of the head pose and

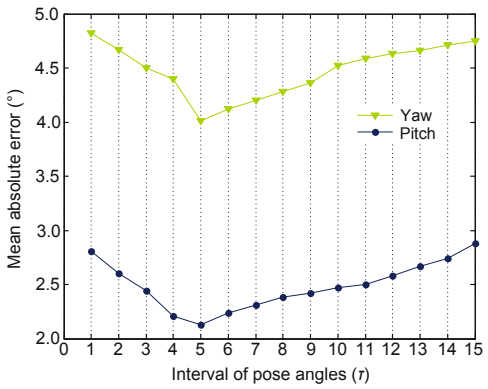

Fig. 7 Mean absolute error of yaw and pitch obtained using the proposed method with different τ's on the Pointing'04 database

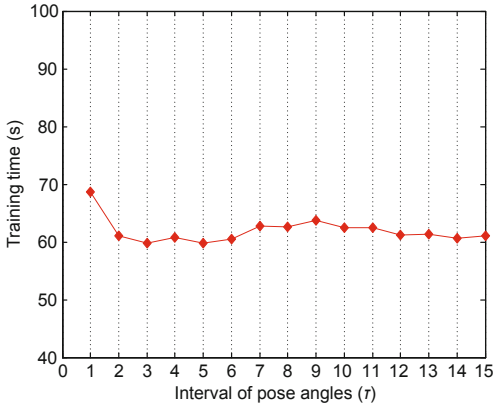

Fig. 8 Training time under different τ for dense MLD on the Pointing'04 database

different densities of training data were conducted on the Pointing'04 database. All experiments were conducted in the way of a five-fold cross-validation.

Experiment 1 Consider two head pose predictors, one using a training dataset which includes all available yaw pose angles, and the other using the same training dataset but the poses with uniform sampling of 30°. Then evaluate them with the whole testing set which contains the images whose yaw angles are either seen or unseen in the training dataset. The results are given in Table 1.

Experiment 2 Consider two head pose predictors, one using a training dataset which includes all available pitch pose angles, and the other using the same training dataset but the poses with uniform sampling of 30°. Then evaluate them with the whole testing set which contains the images whose pitch angles are either seen or unseen in the training dataset. The

results are shown in Table 2.

Experiment 3 Consider two head pose predictors, one using a training dataset which includes yaw pose angles with uniform sampling of 60°, and the other using a training dataset which includes pitch pose angles with non-uniform sampling. Then evaluate them with the whole testing set which contains the images whose yaw and pitch angles are either seen or unseen in the training dataset. The results are shown in Tables 1 and 2.

To facilitate comparison, the MAEs for yaw and pitch pose angles of the three experiments that are not included in the training dataset are shown in Tables 1 and 2. We also list the MAEs for yaw and pitch pose angles that are included in the training dataset as a baseline. From Tables 1 and 2, we can see that: (1) The head pose prediction error increases when the pose angles of the testing data are not included in the training dataset; (2) The increase of the prediction error is relatively small when the pose angle sampling interval in the training dataset is up to 30° both on yaw and pitch directions; (3) The head pose prediction error, however, increases substantially when the pose angle sampling interval in the training dataset is beyond 45°, while the head pose prediction error is irrelevant to the uniform or non-uniform sampling of the training data. These results suggest that the proposed method can predict unseen pose angles with a relatively low error when the pose angle sampling interval in the training dataset is up to 30°.

4.4 Comparison with several state-of-the-art methods

4.4.1 Results on the Pointing'04 database

For a fair comparison with MLD+BFGS, MLD-J, and MLD-wJ (Geng and Xia, 2014), we used all the 2790 images of 15 individuals to evaluate the effectiveness of the proposed method in improving head pose estimation accuracy. These images were divided into five subsets according to their subject number. Four of the subsets were used as the training set, and the remaining one as the test set. In this way, the persons for training and testing are totally different. The average results of the five-fold cross-validation are shown in Table 3.

As can be seen from Table 3, the proposed method achieves an MAE of 4.01° and 2.13° for yaw and pitch, respectively, which is superior to

Table 1 Head pose estimation performance for yaw pose angles not included in the training dataset

Experiment	−90°	−75°	−60°	−45°	−30°	−15°	0°	15°	30°	45°	60°	75°	90°
Experiment 1	√	√	√	√	√	√	√	√	√	√	√	√	√
	6.57°	5.92°	5.64°	5.38°	3.57°	2.14°	2.16°	1.86°	2.42°	4.71°	4.42°	5.01°	5.43°
	√	×	√	×	√	×	√	×	√	×	√	×	√
	4.54°	6.28°	5.87°	5.42°	4.64°	3.35°	3.33°	3.28°	1.07°	5.35°	4.50°	5.36°	4.28°
Experiment 3	√	×	×	√	×	×	√	×	×	√	×	×	√
	3.24°	9.28°	8.51°	2.42°	8.21°	7.15°	1.24°	9.25°	8.47°	2.35°	9.27°	6.07°	3.06°

√: training with the pose; ×: training without the pose. Experiment 1 contains two groups of experiments: the upper one is the baseline (all pose angles in the dataset are considered, totally 13 pose angles, with pose angle sampling interval 15°) and the lower one does not consider those pose angles with '×' (the pose angle sampling interval is 30°)

Table 2 Head pose estimation performance for pitch pose angles not included in the training dataset

Experiment	−90°	−60°	−45°	−30°	−15°	0°	15°	30°	45°	60°	90°
Experiment 2	√	√	√	√	√	√	√	√	√	√	√
	3.67°	3.15°	1.84°	1.07°	0.98°	0.80°	0.84°	0.93°	1.56°	2.57°	2.86°
	√	√	×	√	×	√	×	√	×	√	√
	3.25°	5.24°	4.50°	4.28°	2.33°	1.07°	2.12°	1.25°	3.21°	4.18°	3.11°
Experiment 3	√	×	×	√	×	√	×	√	×	×	√
	3.21°	10.24°	9.52°	3.42°	3.71°	2.01°	3.27°	3.35°	6.75°	8.35°	2.56°

√: training with the pose; ×: training without the pose. Experiment 2 contains two groups of experiments: the upper one is the baseline (all pose angles in the dataset are considered, totally 11 pose angles) and the lower one does not consider those pose angles with '×' (the pose angle sampling interval is 30°)

MLD+BFGS and other counterpart methods. Moreover, its accuracy is 74.40% for yaw and 87.23% for pitch, the best among all these methods. By comparing the results of our method and MLD+BFGS, we can see that the proposed dense MLD is effective in improving the pose estimation accuracy.

Fig. 9 shows the confusion matrices of the baseline method (MLD+BFGS) and the proposed method on the Pointing'04 database. Here, we just show the prediction results for yaw angles. Similar trends can be observed for pitch angles. By comparing the two confusion matrices, we can see that the predictions by the proposed method are much closer to the ground truth than the baseline method. The error bound of the proposed method is about 30°, whereas that of the baseline method is about 60°. These results demonstrate again the effectiveness of incorporating dense pose angle bins into MLD.

Table 3 Head pose estimation results under different methods on the Pointing'04 database

Method	Mean absolute error (°)			Accuracy (%)		
	Yaw	Pitch	Yaw+Pitch	Yaw	Pitch	Yaw+Pitch
Proposed	**4.01**	**2.13**	**6.27**	**74.40**	**87.23**	**65.77**
MLD+BFGS	5.11	4.15	8.89	65.96	80.85	53.62
MLD-wJ	4.24	2.69	6.45	73.30	86.24	64.27
MLD-J	5.02	3.54	7.94	67.96	81.51	55.66
LAG	4.60	2.40	–	72.44	85.53	–
LPLS	11.29	10.52	–	45.57	58.70	–
KPLS	6.56	6.61	–	67.36	80.36	–
WRF	7.50	7.80	–	–	–	–
VRF+LDA	–	11.05	–	–	66.95	–
MGD	6.90	8.00	–	64.51	62.72	–
CGM	5.94	6.73	–	–	–	–
kVoD	6.59	–	–	–	–	–

'–' denotes that the measure is not available in the paper where the method was proposed. For LAG, images from the first session were used for training and images from the second session for testing; For MGD, 80% of Pointing'04 images were used for training and 10% for evaluation; For kVoD, the results were obtained using three-fold cross-validation

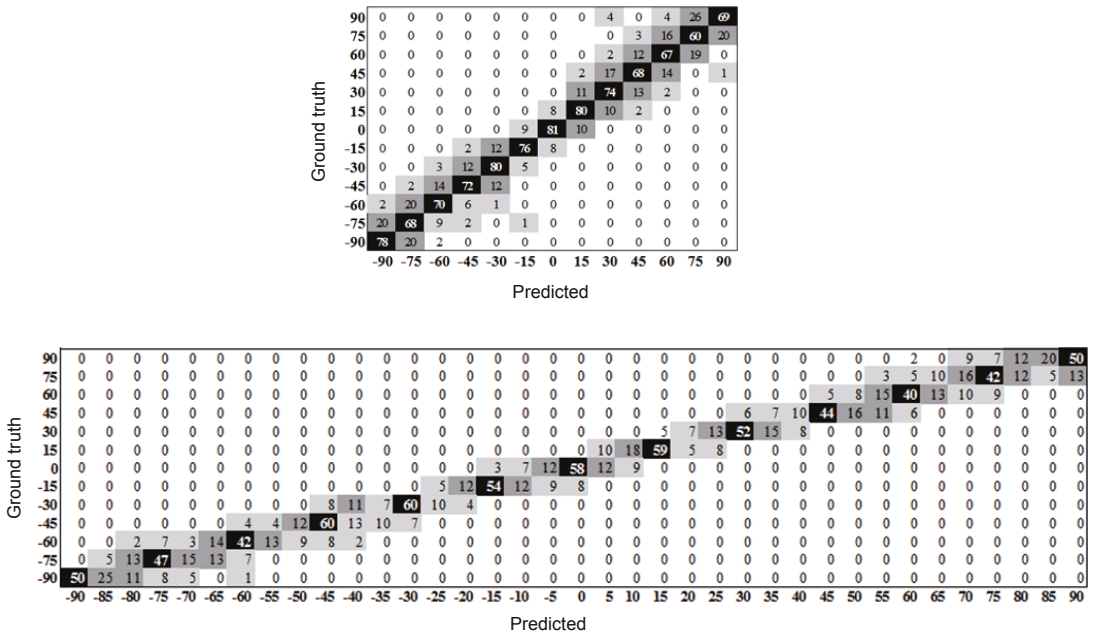

Fig. 9 Confusion matrices (in %) of the MLD+BFGS (up) and proposed (down) methods for yaw angles on the Pointing'04 database

4.4.2 Results on the CAS-PEAL database

For a fair comparison on the CAS-PEAL database with LBIF, sLBIF (Ma BP *et al.*, 2013), and kCovGa (Ma *et al.*, 2014), we used a subset containing totally 4200 images of 200 subjects whose IDs range from 401 to 600. Three-fold cross-validation was used to avoid over-training. Specifically, the images were ranked by their subject IDs and then divided into three subsets. Two subsets were taken as the training set and the other subset as the testing set.

Table 4 summarizes the accuracy achieved by different methods on the CAS-PEAL database. It can be seen that our proposed method achieves the best results. Fig. 10 plots the accuracy of the proposed and baseline methods with respect to the 'ground truth' pose angles. These curves show that the proposed method consistently overwhelms the baseline method at various pose angles.

Table 4 Accuracies under different methods on the CAS-PEAL database

Method	Accuracy (%)	
	Yaw	Pitch
LBIF	94.57	–
sLBIF	94.55	–
kCovGa	94.20	–
MLD+BFGS	95.21	97.08
Proposed	97.79	98.15

'–' denotes that the measure is not available in the paper where the method was proposed

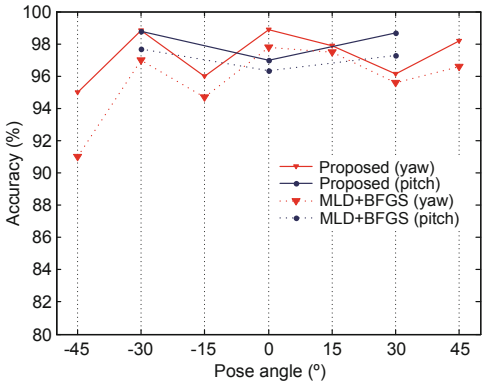

Fig. 10 Accuracies of yaw and pitch angles achieved by the proposed and MLD+BFGS methods on the CAS-PEAL database

4.4.3 Results on the CMU Multi-PIE database

There are four sessions in the CMU Multi-PIE database. In our experiment, we used only the first session (Ma *et al.*, 2014). This session contains 3735 images from different subjects. Each subject displays 13 poses from the right profile (denoted as $-90°$) to the left profile (denoted as $+90°$) with an increment of $15°$ in the yaw rotation. The images were ranked by their subject IDs and divided into three subsets. Two subsets were used for training and the remaining for testing. The experiments were done using three-fold cross-validation.

Table 5 summarizes the MAEs achieved by different methods on the CMU Multi-PIE database. Obviously, the proposed method is much better than the other methods. Detailed results of the proposed and baseline methods are shown in Fig. 11. Again, we can see that the proposed method is consistently better than MLD+BFGS at different poses.

5 Conclusions

In this paper, a dense multivariate label distribution (MLD) method has been presented for estimating head pose angles. This is motivated by the

Table 5 Mean absolute errors of different methods on the CMU Multi-PIE database

Method	Mean absolute error (°)
SL^2	4.33
CovGa	4.50
kCovGa	3.80
MLD+BFGS	2.51
Proposed	1.95

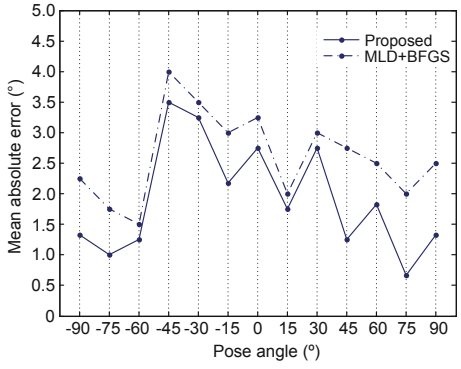

Fig. 11 Mean absolute errors of yaw angles achieved by the proposed and MLD+BFGS methods on the CMU Multi-PIE database

observation that there are usually finite and discrete poses available in the training database. The proposed method aims to estimate the pose angles that are unseen in the training dataset. To this end, it assigns each face image with a dense MLD which includes head poses not seen in the training data. Experimental results on the Pointing'04 database demonstrated that our proposed method can successfully predict unseen head poses. Besides, we have compared the proposed method with several state-of-the-art methods on the Pointing'04, CAS-PEAL, and CMU Multi-PIE databases. The results showed that the proposed method overwhelms all the other methods under consideration. In the future, we are going to further improve the accuracy of the proposed method by exploring new feature representation methods.

References

Aghajanian, J., Prince, S.J.D., 2009. Face pose estimation in uncontrolled environments. Proc. British Machine Vision Conf., p.1-11.

Berger, A.L., Pietra, V.J.D., Pietra, S.A.D., 1996. A maximum entropy approach to natural language processing. *Comput. Ling.*, **22**(1):39-71.

Bowyer, K.W., Chang, K., Flynn, P., 2006. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. *Comput. Vis. Image Understand.*, **101**(1):1-15. http://dx.doi.org/10.1016/j.cviu.2005.05.005

Brunelli, R., 1997. Estimation of pose and illuminant direction for face processing. *Image Vis. Comput.*, **15**(10):741-748. http://dx.doi.org/10.1016/S0262-8856(97)00024-3

Cai, Y., Yang, M.L., Li, Z.Q., 2015. Robust head pose estimation using a 3D morphable model. *Math. Prob. Eng.*, **2015**:678973.1-678973.10. http://dx.doi.org/10.1155/2015/678973

Do, M.N., 2003. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Process. Lett.*, **10**(4):115-118. http://dx.doi.org/10.1109/LSP.2003.809034

Fenzi, M., Leal-Taixé, L., Rosenhahn, B., *et al.*, 2013. Class generative models based on feature regression for pose estimation of object categories. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.755-762.

Fitzpatrick, P., 2000. Head Pose Estimation Without Manual Initialization. Report, Massachusetts Institute of Technology, Cambridge.

Gao, W., Cao, B., Shan, S.G., *et al.*, 2008. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Trans. Syst. Man Cybern. A*, **38**(1):149-161. http://dx.doi.org/10.1109/TSMCA.2007.909557

Geng, X., Xia, Y., 2014. Head pose estimation based on multivariate label distribution. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1837-1842.

Gourier, N., Hall, D., Crowley, J.L., 2004. Estimating face orientation from robust detection of salient facial features. Proc. Int. Workshop on Visual Observation of Deictic Gestures. Available from http://www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html.

Gross, R., Matthews, I., Cohn, J., *et al.*, 2010. Multi-PIE. *Image Vis. Comput.*, **28**(5):807-813. http://dx.doi.org/10.1016/j.imavis.2009.08.002

Haj, M.A., Gonzàlez, J., Davis, L.S., 2012. On partial least squares in head pose estimation: how to simultaneously deal with misalignment. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2602-2609. http://dx.doi.org/10.1109/CVPR.2012.6247979

Hu, C.L., Gong, L.Y., Wang, T.J., *et al.*, 2014. An effective head pose estimation approach using Lie algebrized Gaussians based face representation. *Multim. Tools Appl.*, **73**(3):1863-1884. http://dx.doi.org/10.1007/s11042-013-1676-5

Huang, D., Storer, M., de la Torre, F., *et al.*, 2011. Supervised local subspace learning for continuous head pose estimation. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2921-2928. http://dx.doi.org/10.1109/CVPR.2011.5995683

Jain, V., Crowley, J.L., 2013. Head pose estimation using multi-scale Gaussian derivatives. Proc. 18th Scandinavian Conf. on Image Analysis, p.319-328. http://dx.doi.org/10.1007/978-3-642-38886-6_31

Krüger, V., Sommer, G., 2002. Gabor wavelet networks for efficient head pose estimation. *Image Vis. Comput.*, **20**(9-10):665-672. http://dx.doi.org/10.1016/S0262-8856(02)00056-2

Liu, D.C., Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. *Math. Program.*, **45**(1):503-528. http://dx.doi.org/10.1007/BF01589116

Lu, F., Sugano, Y., Okabe, T., *et al.*, 2012. Head pose-free appearance-based gaze sensing via eye image synthesis. Proc. 21st Int. Conf. on Pattern Recognition, p.1008-1011.

Lu, F., Okabe, T., Sugano, Y., *et al.*, 2014. Learning gaze biases with head motion for head pose-free gaze estimation. *Image Vis. Comput.*, **32**(3):169-179. http://dx.doi.org/10.1016/j.imavis.2014.01.005

Ma, B.P., Chai, X.J., Wang, T.J., 2013. A novel feature descriptor based on biologically inspired feature for head pose estimation. *Neurocomputing*, **115**:1-10. http://dx.doi.org/10.1016/j.neucom.2012.11.005

Ma, B.P., Li, A.N., Chai, X.J., *et al.*, 2014. CovGa: a novel descriptor based on symmetry of regions for head pose estimation. *Neurocomputing*, **143**:97-108. http://dx.doi.org/10.1016/j.neucom.2014.06.014

Ma, B.P., Huang, R., Qin, L., 2015. VoD: a novel image representation for head yaw estimation. *Neurocomputing*, **148**:455-466. http://dx.doi.org/10.1016/j.neucom.2014.07.019

Ma, X.H., Tan, Y.Q., Zheng, G.M., 2013. A fast classification scheme and its application to face recognition. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **14**(7):561-572. http://dx.doi.org/10.1631/jzus.CIDE1309

Murphy-Chutorian, E., Trivedi, M.M., 2009. Head pose estimation in computer vision: a survey. *IEEE Trans. Patt. Anal. Mach. Intell.*, **31**(4):607-626. http://dx.doi.org/10.1109/TPAMI.2008.106

Pang, H., Lin, A., Holford, M., *et al.*, 2006. Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**(16):2028-2036. http://dx.doi.org/10.1093/bioinformatics/btl344

Sim, T., Baker, S., Bsat, M., 2002. The CMU pose, illumination, and expression (PIE) database. Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition, p.46-51. http://dx.doi.org/10.1109/AFGR.2002.1004130

Tang, Y.Q., Sun, Z.N., Tan, T.N., 2014. A survey on head pose estimation. *Patt. Recogn. Artif. Intell.*, **27**(3):213-225.

Wu, J.W., Trivedi, M.M., 2008. A two-stage head pose estimation framework and evaluation. *Patt. Recog.*, **41**(3):1138-1158. http://dx.doi.org/10.1016/j.patcog.2007.07.017

Zhang, Z.P., Luo, P., Loy, C.C., *et al.*, 2014. Facial landmark detection by deep multi-task learning. Proc. 13th European Conf. on Computer Vision, p.94-108. http://dx.doi.org/10.1007/978-3-319-10599-4_7

Zhu, R.H., Sang, G.L., Cai, Y., *et al.*, 2013. Head pose estimation with improved random regression forests. Proc. 8th Chinese Conf. on Biometric Recognition, p.457-465. http://dx.doi.org/10.1007/978-3-319-02961-0_57

Zhu, X.X., Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.2879-2886. http://dx.doi.org/10.1109/CVPR.2012.6248014