Neurocomputing 115 (2013) 1–10

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

A novel feature descriptor based on biologically inspired feature for head pose estimation

Bingpeng Ma [a,*], Xiujuan Chai [b], Tianjiang Wang [c]

[a] University of Chinese Academy of Sciences, Beijing, China
[b] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[c] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 24 July 2012
Received in revised form
23 October 2012
Accepted 1 November 2012
Communicated by Dr. Qingshan Liu
Available online 11 February 2013

Keywords:
Head pose estimation
Biologically inspired features
Local binary pattern

ABSTRACT

This paper proposes a novel method to improve the accuracy of head pose estimation. Since biologically inspired features (BIF) have been demonstrated to be both effective and efficient for many visual tasks, we argue that BIF can be applied to the problem of head pose estimation. By combining the BIF with the well-known local binary pattern (LBP) feature, we propose a novel feature descriptor named "local biologically inspired features" (LBIF). Considering that LBIF is extrinsically very high dimensional, ensemble-based supervised methods are applied to reduce the dimension while at the same time improving its discriminative ability. Results obtained from the evaluation on two different databases show that the proposed LBIF feature achieves significant improvements over the state-of-the-art methods of head pose estimation.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

As one of the most active research topics, automatic face recognition has received significant attention in computer vision and pattern recognition. After more than thirty years of research, face recognition systems have finally achieved very high performance under controlled conditions. However, when variations due to extrinsic factors like pose, illumination and expression changes are present, performance degrades dramatically [1]. Pose change is one of the most important and difficult issues for face recognition. To achieve the expected robustness to pose variation, one may expect to process face images differently according to their pose parameters. In this case, the pose of the input faces should be estimated as a prerequisite for subsequent processes.

This paper focuses on the problem of estimating the head pose using 2D images. Pose estimation essentially means the computation of three types of rotation of a head: yaw (looking left or right), pitch (looking up or down) and roll (tilting left or right) [2]. Due to many important applications, estimation of the yaw rotation attracts more attention than the estimation of the other two rotations [3].

Generally, the methods of head pose estimation can be categorized into two main groups [4]: model-based methods [5–8] and appearance-based methods [9–12]. Recently, Murphy-Chutorian et al. conducted a survey of the existing head pose estimation schemes [2]. Model-based methods can be divided into flexible models and geometric methods: flexible models fit a non-rigid model to the facial structure of each individual in the image plane; head pose can then be estimated from the feature-level comparisons or from the instantiation of the model parameters. Flexible models include methods, such as active shape models (ASM) [5], active appearance models (AAM) [6] and elastic graph matching (EGM) [7]. On the other hand, geometric methods use the location of fiducial points, such as the eyes, mouth, and nose tip to determine pose from their relative configuration. Using the five fiducial points (the outside corners of each eye, the outside corners of the mouth, and the tip of the nose), Gee et al. proposed to find the facial symmetry axis by connecting a line between the midpoint of the eyes and the midpoint of the mouth [13]. Nikolaidis et al. proposed a head pose estimation method based on the distortion of the isosceles triangle formed by the two eyes and the mouth [8]. In general, the performance of model-based methods relies on the accuracy of fiducial points localization. Since robust fiducial points localization is still an open problem, model-based approaches are limited in practice.

Appearance-based methods typically assume that there exists a certain relationship between the 3D face pose and some properties of the 2D facial image. Under the framework of statistical learning techniques, a large number of training images are used to infer the relationship. Darrell et al. proposed to use the separated eigenspace for head pose estimation [9]. In their method, the pose of the input image is determined by projecting it onto each eigenspace and selecting the one with the lowest residual error. Gong and his colleagues studied the trajectories of multi-view faces in the linear principal component analysis (PCA) feature space and use kernel support vector machines (KSVM) for pose estimation [10,11]. Li et al.

* Corresponding author.
E-mail addresses: mabingpeng@163.com (B. Ma), chaixiujuan@ict.ac.cn (X. Chai), tjwang@hust.edu.cn (T. Wang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.005

exploited independent component analysis (ICA) and its variants for pose estimation [14]. Chen et al. proposed the kernel-based method to deal with the non-linearity of head pose estimation [3]. They chose the face images of two specific head pose angles and utilized the classification-based non-linear interpolation to estimate the head poses between the two angles. Wei et al. proposed to select the optimal orientation of the Gabor filters for each pose to enhance pose information and eliminate other distractive information like variable facial appearance or changing environmental illumination [15]. Ma et al. proposed the method of Gabor–Fourier Fisher features (GFFF) which is based on the assumption that the asymmetry in the intensities of each row of the face image is closely relevant to the yaw rotation of head [16]. Since the set of all facial images with various poses is intrinsically a 3D manifold in image space, manifold learning [17–19] for head pose estimation has recently attracted great interests [12,20,21]. By thinking globally and fitting locally, Fu et al. proposed using the graph embedded analysis method for head pose estimation [12].

Intuitively, appearance-based methods can naturally avoid the above-mentioned drawbacks of the model-based methods. Therefore, they have attracted more and more attentions. Appearance-based methods have been supported by machine learning techniques that use statistical and probabilistic methods. However, it is becoming increasingly clear that the development of new machine learning algorithms alone might not be the best approach to solve the recognition problems.

Recently, the human visual system is demonstrated about the ability to perform identification tasks with enough accuracy for most applications. More and more researchers have focused their attentions on how it represents visual data in order to derive features that will be useful in computer vision systems. Based on the human visual system, biologically inspired features (BIF) [22] have been proposed and shown excellent performance in some computer vision tasks, such as object category recognition [23], face recognition [24], age estimation [25] and scene classification [26]. The motivation of BIF is that the primate visual system uses a strategy of shared general early level processing that branches off into more specific higher level representations [24]. Most visual information that reaches higher levels of the cortex first passes through center-surround processing in the retina and lateral geniculate nucleus (LGN), as well as through an early localized edge/spatial frequency representation in the primary visual cortex (V1). Later processing, for example recognizing objects, is performed in different brain regions.

Different researchers proposed the different BIF model. Riesenhuber et al. [22] proposed the hierarchical "HMAX" model, which consistent with physiological data from inferotemporal cortex that accounts for this complex visual task and makes testable predictions. Their model is based on a MAX-like operation while contains alternating layers called Simple (S1) and Complex (C1) cell units. In the S1 layer, the input image is convolved with the Gabor filters. Two adjacent scales in the same orientation are then grouped together to form "bands". In the C1 layer, the maximum value within the adjacent scale with the same orientation are computed. The MAX operation locally and automatically select a relevant subsets of inputs which seems biologically plausible, and increase the tolerate to the small shifts and scale changes within a small range of position and scale.

Motivated by a quantitative model of visual cortex, Serre et al. [23] extended the "HMAX" model to include two higher level layers, called the S2 layer and the C2 layer. In the S2 layer, the response of S2 unit depends in a Gaussian-like way on the Euclidean distance between a new input patch and the stored prototype. This is consistent with well-known neuron response properties in primate inferotemporal cortex and seems to be the key property for learning to generalize in the visual and motor systems. Each stored S2 unit is convolved over an entire image at all scales. And then, C2 unit are assigned the maximum response value on S2. Built on Serre's work for object category recognition, Mutch et al. [27] proposed some improvements, such as sparsifying S2 inputs, which suppressed S1 and C1 outputs, and feature selection.

In order to create a new set of features useful for both face identification and expression recognition, Meyers and Wolf [24] modified the "HMAX" model by adding center-surround processing to handle illumination changes. To handle high dimensional data, they proposed a new method of combining lower level features based on a kernelized and regularized version of the relevant component analysis transformation.

Guo et al. [25] investigated how to use BIF to estimate the human age from the input face images. They set the size of Gabor filters as small as possible and suggested to determine the number of bands and orientations in a problem-specific manner, rather than using a predefined number. More importantly, they proposed a non-linear operator "STD" on each scale band (two scales) of S1 units after they are merged into one maximum map using the "MAX" operation. Dimension reduction is also performed on the C1 features to make them more efficient.

Given the superior performance of human vision on general object recognition, it is reasonable to explore for inspiration to improve the performance of head pose estimation. In this paper, we explore the relationship between the human vision and head pose estimation. The first contribution of this paper is the introduction of BIF to the problem of head pose estimation for the first time ever. The second contribution is to further improve the discriminant ability of BIF, based on which a novel feature extraction method named local biologically inspired features (LBIF) is proposed in this paper. LBIF is the combination of BIF and the traditional local binary pattern (LBP) feature. As a texture descriptor, LBP has gained great successes in many areas. So, in LBIF, we consider LBP as the complement of BIF. By combining of the human vision and LBP, LBIF can improve the representation ability of BIF features. To show the effectiveness of BIF and LBIF in head pose estimation, we design the experiments on the CAS-PEAL database and the Multi-PIE database. The obtained accuracies show that BIF can be applied in the problem of head pose estimation, and LBIF can improve the accuracy of BIF further while at the same time outperforming all the current state-of-the-art methods.

The remaining part of this paper is organized as follows: Section 2 describes the proposed LBIF method in detail and analyzes its characteristics. Experiments are given in Section 3. Conclusions are drawn in Section 4 along with some discussions on the future works.

2. Local bio-inspired features

In this section, we first introduce the proposed LBIF descriptor for face representations. Then, to gain the pose of the input head image, how LBIF is combined with the classifiers, such as nearest centroids (NC) classifier or support vector machines (SVMs) [28] is introduced in details.

The flowchart of the LBIF-based framework is shown in Fig. 1. Like the traditional BIF, there are four layers in LBIF. In the S1 layer of LBIF, a pyramid of Gabor filters is applied at all positions of the input image; in the C1 layer, the non-linear "MAX" operator is applied to extract the non-linear features; in the S2 layer, local binary pattern (LBP) is improved and then applied to extract the texture features; in the C2 layer, the ensemble-based dimension reduction method is then applied to improve the discriminant-ability and reduce the dimension of features. In the follows, we introduce each layer in details.

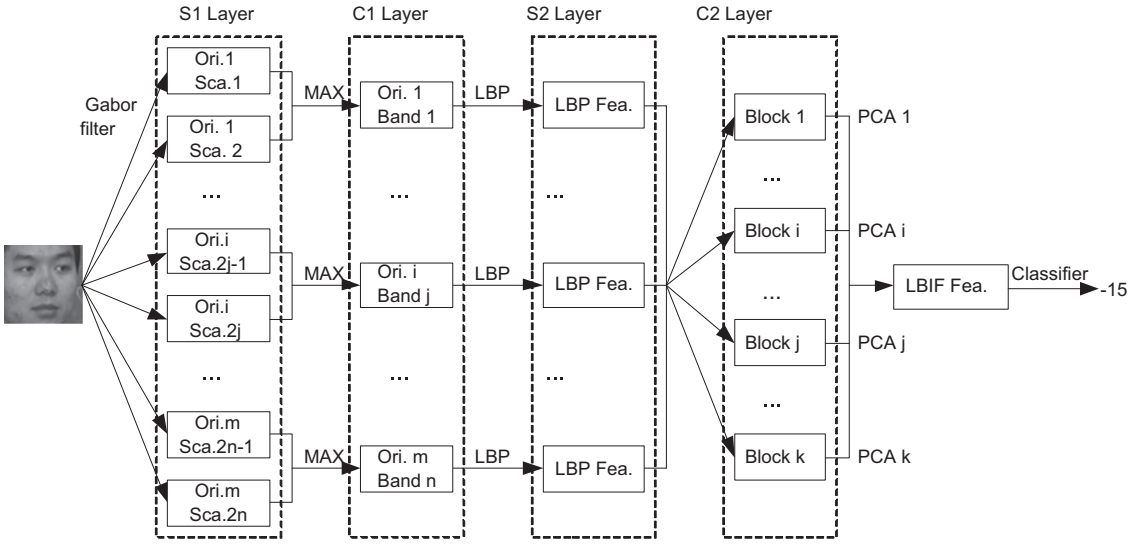

Fig. 1. The flowchart of LBIF.

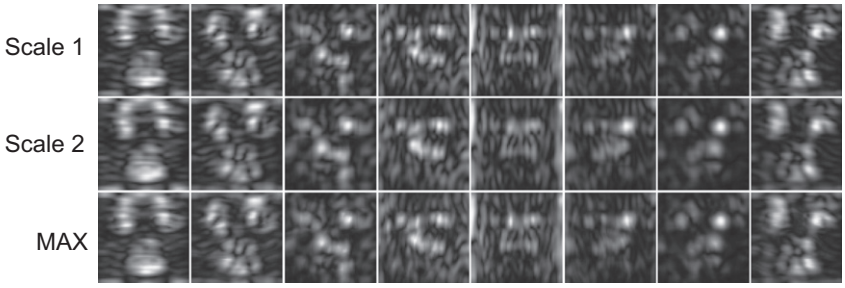

Fig. 2. The Gabor representations of a face image. The images in the first and second rows are the Gabor representations of the same image with the neighborhood scales. The images in the third row are the results of the MAX operator on the images of the first and second row.

2.1. S1 layer

In the first step of LBIF, the input face image is analyzed by an array of simple S1 units that correspond to the classical simple cells found in the primary visual cortex (V1). Usually, Gabor filters, which provide a good model of cortical simple cell receptive fields, are applied at the S1 layer [22,24,25], motivating our choice of such features.

For the face image $I(x,y)$ with width w and height h, its convolutions with Gabor filters can be computed from the following equation [29]:

$$G(\mu,v) = I(x,y) * \psi_{\mu,v}(z) \tag{1}$$

where

$$\psi_{\mu,v}(z) = \frac{\|k_{\mu,v}\|^2}{\sigma^2} e^{(-\|k_{\mu,v}\|^2\|z\|^2/2\sigma^2)}[e^{ik_{\mu,v}z} - e^{-\sigma^2/2}] \tag{2}$$

$$k_{\mu,v} = k_v e^{i\phi_\mu}, \quad k_v = 2^{-(v+2)/2}\pi, \quad \phi_\mu = \mu\frac{\pi}{8} \tag{3}$$

where μ and v are the scales and orientations parameters, respectively. The images in the first and second rows of Fig. 2 are the Gabor representations of the same image with variations in the neighborhood scales.

In the traditional BIF-based methods, each band has a pair of adjacent filter sizes. Then, the number of scales is often fixed on 16 (8 bands) while the number of orientations is flexible and often set to 4, 8 or 16. Since Gabor filters with eight orientations are often applied in face recognition and other related areas, in LBIF, the orientation number of Gabor filters is also set to 8. In Section 3.4, we also show the accuracies of head pose estimation while the orientation number is 4, 8 and 16 show the influence of the orientation numbers.

In LBIF, like the work of Serre [23], the filters in the S1 layer are also arranged to form a pyramid of scales, spanning a range of sizes. Since the size of the face image in head pose estimation is much smaller than that in object category recognition, in LBIF, the sizes of Gabor filters start from a smaller size, 5×5, instead of 7×7 in [23]. The sizes of Gabor filters with the different bands are shown in Table 1.

From the above introduction, in LBIF, we can know that there are 128 Gabor magnitude pictures (GMPs) with eight bands and eight orientations. These GMPs can be rewritten as $\{\mathbf{G}_{ij}^q, i = 0,\ldots,m, j = 0,\ldots,n, q = 0,1\}$, where $\mathbf{G}_{ij}^0$ and $\mathbf{G}_{ij}^1$ are the filtered values at band i and orientation j, m and n are the total number of bands and orientations, respectively and q is the scale index for the specifically band. Then, for a 32×32 image, the dimension of each GMP is 1024, and the dimension of the final output of the S1 layer is 131, 072 ($=32 \times 32 \times 16 \times 8$).

2.2. C1 layer

The C1 layer corresponds to cortical complex cells that tend to have larger receptive fields [23]. The C1 layer pool over S1 units with the same orientation and the same band. Since the pooling of the maximum operation over two consecutive scales (i.e., in the same band) increases the tolerance to 2D transformations, such as scale changes [22,23], in LBIF, "MAX" is selected as the non-linear operator and the pooling filter. The maximum value $\mathbf{C}_{ij}$ of two consecutive S1 units at band i and orientation j can be computed

Table 1
The sizes of Gabor filters.

Band	1	2	3	4	5	6	7	8
Filter sizes	5×5	9×9	13×13	17×17	21×21	25×25	29×29	33×33
Filter sizes	7×7	11×11	15×15	19×19	23×23	27×27	31×31	35×35

from the following equation:

$$\mathbf{C}_{ij} = \max(\mathbf{G}_{ij}^0, \mathbf{G}_{ij}^1) \tag{4}$$

In Fig. 2, the images in the third row are the results of MAX operator between the images in the first and second rows.

Finally, after the C1 layer, the representation of a face image can be wrote as

$$\mathbf{C} = (\mathbf{C}_{11}, \mathbf{C}_{12}, \ldots, \mathbf{C}_{1n}, \mathbf{C}_{21}, \mathbf{C}_{22}, \ldots, \mathbf{C}_{mn}) \tag{5}$$

2.3. S2 layer

In the S2 layer of LBIF, an improvement based on the well-known LBP is proposed and then applied to extract the texture feature. LBP is originally introduced as a texture descriptor by Ojala [30]. It has been applied successfully in face recognition, head pose estimation, face expression recognition and other related areas.

The original LBP operator labels the pixels of an image by thresholding the pixels $f_p (p = 0, \ldots, 7)$ in a 3×3 neighborhood with a center value f_c. The output $O(f_p - f_c)$ of the LBP operator is a binary number

$$O(f_p - f_c) = \begin{cases} 1, & f_p \geq f_c \\ 0, & f_p < f_c \end{cases} \tag{6}$$

Then, by assigning a binomial factor 2^p for each $S(f_p - f_c)$, the LBP pattern at the pixel f_c is given as

$$LBP = \sum_{p=0}^{7} O(f_p - f_c) 2^p \tag{7}$$

which characterizes the spatial structure of the local image texture. From the above equation, it is known that the histogram bin of LBP is limited in the region [0, 255] and the bin number is 256. In many applications, the bin number is grouped into 8 or 16 to reduce the dimension of the histogram. In LBIF, we use a four neighborhood of the center pixel, which can reduce the computation complexity and the feature dimension greatly. By this way, the histogram bin is limited from 0 to 15, while the bin number is reduced from 256 to 16. In Fig. 3, we show the four neighborhood operator in LBIF. In Fig. 4, we show the representations with the operator on the images in the third row of Fig. 2.

In the traditional LBP-based method, the neighborhood representations for a given image are generated by dividing the image into several patches and creating histogram in each patch. Usually, all the patches are set to the same size and all the histograms are treated as equivalent. In LBIF, since the sizes of Gabor filters are varied with the band index in the S1 layer, the patch sizes in the S2 layer are also varied with the band index of Gabor filters: for the bands with smaller sizes of Gabor filter, the patch sizes are smaller as well. In LBIF, the overlap between two neighbor patches is set to half of the patch size. The variations of the patch size and overlap influence the histogram number, while at the same time the histogram number L_i is decreased with the increasing of band i. Since the smaller sizes of Gabor filters mean that more details are preserved, we argue that by increasing the weight of the image details, the representation ability of the feature is increased also. In Table 2, we show the parameters in the S2 layer: the first and second rows show the patch sizes and the overlaps for the different bands in LBIF, respectively; the third and fourth rows show the histogram number L_i and the feature dimension for different bands with the specific orientation, respectively.

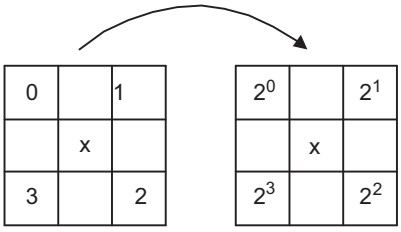

Fig. 3. The neighborhood operator in the S2 layer of LBIF.

For a histogram $\mathbf{h} = (b_0, \ldots, b_{15})$, its sum S_h is computed as follows:

$$S_h = \sum_{i=0}^{15} b_i = w_w \times h_w \tag{8}$$

where w_w and h_w are the width and height of the patch, respectively, which means that S_h is the same as the number of elements in the patch. In LBIF, with the increase of the band index i, w_w and h_w are increased, which means S_h also increases. In some senses, S_h can be seen as the weight of the patch. The larger the patch size, the larger the weight of the patch. From the characteristics of Gabor filters, we can know that the smaller the Gabor filter means more detail of the input image, which is important for pose estimation. So, in LBIF, to emphasize the weight of the smaller patch, the histogram $\mathbf{h}$ is processed as follows:

$$\hat{\mathbf{h}} = \frac{\mathbf{h}}{w_w \times h_w} \tag{9}$$

By this way, for patches with arbitrary sizes, their histograms have the same weight in the final feature. Since the patch number in the smaller bands is larger than that of the larger bands, the number of histogram in the smaller bands is larger than that of the larger bands. So the weight of the smaller bands is increased in the output of the S2 layer and the detail of the input image can be greatly preserved.

Finally, the output $\mathbf{S}$ of the S2 layer can also be computed from Eq. (5):

$$\mathbf{S} = (\mathbf{S}_{11}, \mathbf{S}_{12}, \ldots, \mathbf{S}_{1n}, \mathbf{S}_{21}, \mathbf{S}_{22}, \ldots, \mathbf{S}_{mn}) \tag{10}$$

and

$$\mathbf{S}_{ij} = (\hat{\mathbf{h}}_{ij,1}, \hat{\mathbf{h}}_{ij,2}, \ldots, \hat{\mathbf{h}}_{ij,L_i}) \tag{11}$$

where $\hat{\mathbf{h}}_{ij,l}$ is the histogram of the patch l for the band i and the orientation j. The dimension of $\mathbf{S}$ is 25,216 ($=3152\times8$).

2.4. C2 layer

The S2 features are concatenated to form a feature vector for each face image. In the C2 layer, the methods of feature dimensionality reduction are applied to reduce the dimension of the resultant features and enhance the discriminative ability.

Fig. 4. The representations with the neighborhood operator.

Table 2
Parameters in the S2 layer. The first and second rows show the window's sizes and overlaps in the different bands, respectively; The third and fourth rows show the histogram number and the feature dimension for different bands with the specific orientation, respectively.

Band	1	2	3	4	5	6	7	8	Total
Windows size	6×6	8×8	10×10	12×12	14×14	16×16	18×18	20×20	–
Overlap size	3×3	4×4	5×5	6×6	7×7	8×8	9×9	10×10	–
Histogram number	81	49	25	16	9	9	4	4	197
Dimension	1296	784	400	256	144	144	64	64	3152

Principal component analysis (PCA) is a traditional linear transformation technique, which can greatly reduce the dimension of features. In PCA, the projection matrix $\mathbf{W}$ is composed by the orthogonal eigenvectors of the covariance matrix of all the training samples. Since the feature dimension of S2 layer is very high, it is difficult to apply PCA directly to the holistic histogram feature. To address this problem, we turn to ensemble of piecewise PCA classifiers by partitioning the entire feature vector. For example, the output $\mathbf{S}$ of the S2 layer can be rewritten as follows:

$$\mathbf{S}=(\mathbf{H}_1,\mathbf{H}_2,\ldots,\mathbf{H}_K) \tag{12}$$

where $\mathbf{H}_p$ is the p-th feature segment containing a specific number of spatial histograms. Then, for each feature segment $\mathbf{H}_p$, one PCA model is built in order to transform it to a low-dimensional representation $\mathbf{L}_p$ in the subspace

$$\mathbf{L}_p=\mathbf{W}_p^T\mathbf{H}_p \tag{13}$$

Thus, in LBIF, by building K subspaces, the input face image is finally represented as

$$\mathbf{LBIF}=(\mathbf{L}_1,\mathbf{L}_2,\ldots,\mathbf{L}_K) \tag{14}$$

2.5. Supervised local bio-inspired features

The nature of head pose estimation evidently needs discriminating features rather than a pure representation. Therefore, we need a discriminant analysis on the spatial histogram representation of LBIF in order to improve the recognition performance. Generally speaking, the performance of the supervised method is much better than that of the unsupervised method. So, we can use the supervised method, such as linear discriminant analysis (LDA), marginal Fisher analysis (MFA) [31] or locality sensitive discriminant analysis (LSDA) [32], to improve the accuracy of head pose estimation. In this paper, for simplicity, we just use LDA as the method that improves the discriminative ability of LBIF and propose the supervised local bio-inspired features (sLBIF). In fact, LDA can be replaced by other supervised methods in sLBIF, since it has been proven that the performance of other supervised methods is much better than that of LDA.

LDA has been recognized as one of the most successful methods in face recognition [33]. In LDA, the within-class scatter matrices $\tilde{\mathbf{S}}_w$ represents the average scatter of the sample vectors $\mathbf{X}$ of different classes around their respective means $\mathbf{m}_i$. Similarly, the between-class scatter matrices $\tilde{\mathbf{S}}_b$ represent the scatter of the conditional mean vectors $\mathbf{m}_i$ around the overall mean vector $\mathbf{m}$. Through a linear transformation, the original feature representation is projected into a new LDA subspace where $\tilde{\mathbf{S}}_b$ is maximized while $\tilde{\mathbf{S}}_w$ is minimized by maximizing the Fisher separation criterion. The optimal projection matrix $\mathbf{W}^{lda}$ can be obtained by solving a generalized eigenvalue problem.

Replacing $\mathbf{W}_p$ in Eq. 13 by $\mathbf{W}_p^{lda}$, the projection in LDA subspace of $\mathbf{H}_p$ is computed as

$$\mathbf{F}_p=(\mathbf{W}_p^{lda})^T\mathbf{H}_p \tag{15}$$

By this way, LBIF is improved to sLBIF, which can improve the discriminative ability of the S2 features while reduced the dimension of features. The final output of sLBIF is represented as

$$\mathbf{sLBIF}=(\mathbf{F}_1,\mathbf{F}_2,\ldots,\mathbf{F}_K) \tag{16}$$

In LBIF and sLBIF, it must be pointed out that the dimension of $\mathbf{H}_p$ is the constant for p. Since the dimension of $\mathbf{S}_{ij}$ varies with the increase of band index i, the features in $\mathbf{H}_p$ may come from the different bands, while at the same time, the features of $\mathbf{S}_{ij}$ are partitioned into different $\mathbf{H}_p$.

Note that, in sLBIF, to avoid the singularity of the within-class scatter matrix, PCA is conducted to reduce the dimensionality of the histogram vector to be less than $N-C$, where N is the number of training examples and C is the number of classes. The PCA features are then transformed by LDA for the final classification.

2.6. Classification or regression

Since the extraction of LBIF (sLBIF) can be regarded as the pre-processing step for yaw estimation, it should be combined with the classifier or the regression method to get the yaw of the input image. In this paper, we take the head pose estimation as a yaw classification problem, which is the possible way when the pose angles in the database are not continuous.

In this paper, NC and SVM are selected as the classifier to evaluate the performance of the proposed features. In head pose estimation, since the image difference of the same people with the near angles might be less than the image difference of the different people with the same angle, there is an error which is caused by the subjects inevitably when using the traditional nearest neighbor (NN) classifier. To eliminate this error, we select NC classifier as the classifier to estimate the head pose. In NC classifier, for each angle, the k-means method is applied to find k centroids from the training samples with the same angle. Then we compare the distance of the input feature to each class centroid and take the class with the smallest Euclidean distance as the output label. Besides NC classifier, we also use SVM classifier to evaluate the performance of LBIF (sLBIF), since SVM has been applied in head pose estimation triumphantly.

In fact, head pose estimation can also be considered as a regression problem when the angles are seen as the regression values. For example, the NC classifier or SVM can be replaced by other methods, such as support vector regression or relevance vector machine [34], to get the continuous poses.

3. Experiments

In this section, the proposed LBIF and sLBIF methods are evaluated on two different head pose databases. To demonstrate their effectiveness, we also compare the performance of the proposed methods with other related methods. To show the influence of the parameters in sLBIF, we also repeat the experiments using the different block and orientations number.

Two databases are selected as the experimental data to demonstrate the accuracy of the different methods. The first database is the CAS-PEAL [35] database, which contains 21 poses combining seven yaw angles (-45°, -30°, -15°, 0°, 15°, 30° and 45°) and three pitch angles (30°, 0° and -30°). We use a subset containing totally 4200 images of 200 subjects whose IDs range from 401 through 600. In Fig. 5, we show some face images in the CAS-PEAL database.

The second database is the CMU Multi-PIE face database [36]. In the experiment, we only use the images of the first session. This session contains 3735 images from different subjects. The head angles varies from -90° to 90° with an interval of 15°. In Fig. 6 we show some face images in the MultiPIE database.

In the experiments, we take the yaw poses as the class labels. In this sense, the images with the same yaw pose but different pitch pose belong to the same class. So, the images in CAS-PEAL database belong to seven different classes, and the number of classes for the Multi-PIE database is 13.

For all the input images, the face detector [37] is applied to locate the face region, and then all the face regions are normalized to the same size of 32×32. Finally, histogram equalization is used to reduce the influence of lighting variations. From Figs. 5 and 6, the backgrounds are existed in some images and the head is in the misalignment.

In all the experiments, three-fold cross-validation is used to avoid over-training. Specifically, the images are ranked by their subjects and then divided into three subsets. Two subsets are taken as the training set and the other subset is taken as the testing set. In this way, the persons for training and testing are totally different. Consequently the over-fitting in identity is avoided. Testing is repeated three times, by taking each subset as the testing set. The results are the average of all the tests.

We compare the performance of LBIF with the following unsupervised methods: PCA, Gabor Fourier feature (GF), histogram of oriented gradients (HOG), Local Gabor Binary Patterns (LGBP) and BIF. As one of the baseline methods in face recognition, PCA [38] is also the baseline method in appearance-based pose estimation. GF uses the asymmetry of the head and achieves better results in head pose estimation [16]. In [39], HOG descriptors have been employed as the head image representation and perform much better than the raw Sobel filtered image. LGBP is selected since it can greatly improve pose estimation with a set of multi-class SVMs [40]. Since one of our contributions is the novelty of using BIF features in head pose estimation to improve the accuracy of head pose estimation, we implement the method in [25] by ourselves and compare its accuracy with those of other methods.

We also compare the performance of sLBIF with the following supervised methods: LDA, Gabor Fourier Fisher feature (GFFF), sHOG, sLGBP and sBIF. The LDA-based baseline algorithm, similar to the Fisherfaces method [33], applies first PCA for dimensionality reduction and then LDA for discriminant analysis. GFFF, sHOG, sLGBP and sBIF are the combination of LDA with GF, HOG, LGBP and BIF, respectively. For all these methods, PCA is used after feature extraction to reduce the dimension of features and 95% of the total energy of the eigenvalues is kept.

3.1. Experimental results with NC classifier

First, we evaluate the performances of different method under the NC classifier. In Figs. 7 and 8, we show the accuracies of the head pose estimation with the centroid number k ranging from 1 to 10 on the CAS-PEAL database and the Multi-PIE database, respectively. To show the results more clearly, we take the results of the unsupervised method and the supervised method in different subfigure. In Figs. 7 and 8, the top sub-figures and the bottom subfigures show the accuracies of the unsupervised methods and the supervised methods, respectively. In such figures, the x-axis is the centroid number k of each pose and the y-axis is the accuracy.

From Figs. 7 and 8, we can see the following four points: Firstly, for the unsupervised methods, the accuracies increase along with the centroid number k when k is very small and they become robust when k is large. However, for the supervised methods, the accuracies are much more robust for different k. This actually implies that the excellent compactness of each class in the feature space obtained by LDA and the supervised method can improve the discriminant ability of the features.

Secondly, the accuracies of LBIF are the best in all the unsupervised methods for nearly all k values on both the databases. These results prove that LBIF can be applied in head pose estimation and improve the representation ability of the BIF features. Especially, the accuracies of LBIF are much better than those of BIF, which proves that LBP features reinforce the representation of the BIF features, and the combination of LBP features and BIF features in LBIF is successful.

Thirdly, the accuracies of sLBIF are the best in the supervised methods for all k values on the two databases, which proves again the advantage of sLBIF. And the robustness of sLBIF is much better than other methods. For sBIF, its accuracies are much better than those of other methods except sLBIF on the two databases, which shows that BIF can be applied in head pose estimation. Considering that there are many improvement methods for LDA, the accuracies of sLBIF can be improved further by replacing the LDA method by its improvement, such as MFA and LSDA.

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

Fig. 6. The face images in the MultiPIE database.

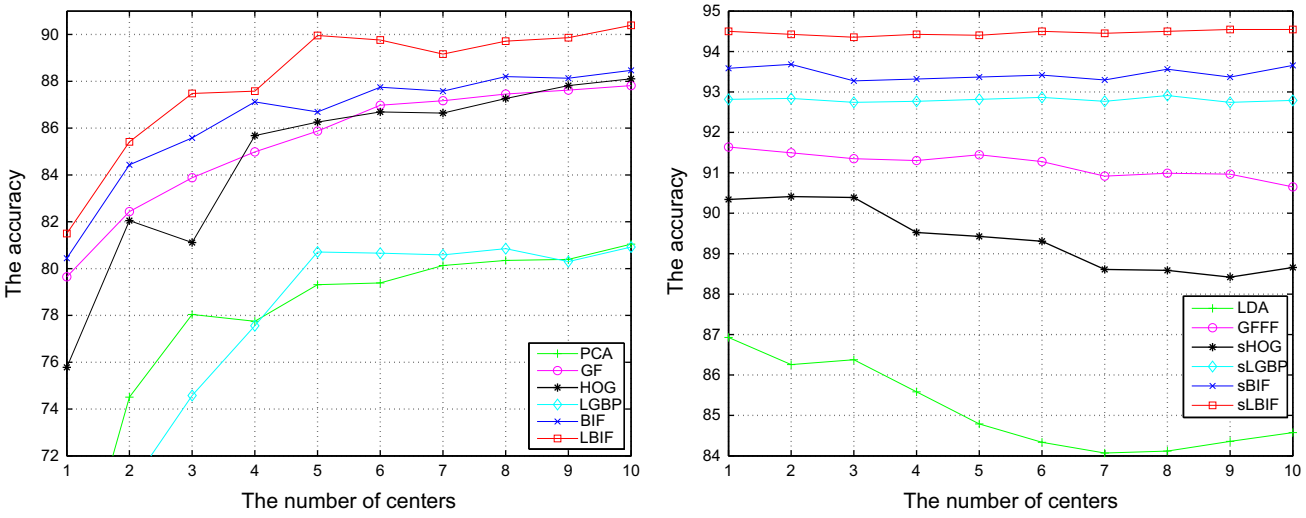

Fig. 7. The accuracy (%) of pose estimation on the CAS-PEAL database. The *x*-axis is the centroid number of each pose and the *y*-axis is the accuracy.

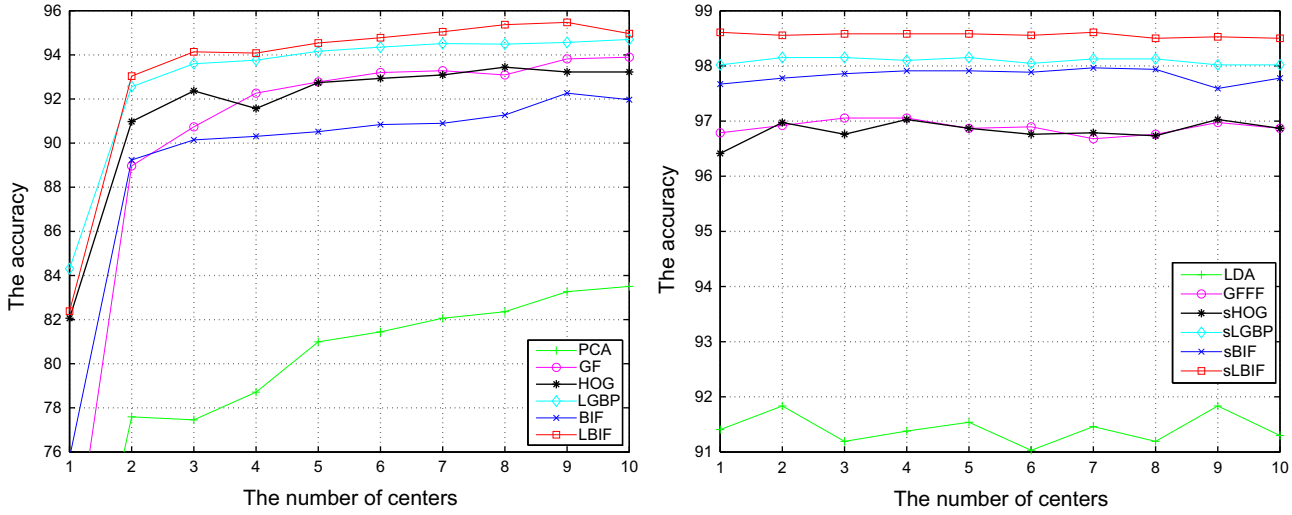

Fig. 8. The accuracy (%) of pose estimation on MultiPIE database. The *x*-axis is the centroid number of each pose and the *y*-axis is the accuracy.

Finally, the accuracies of the supervised methods are much better than those of their corresponding unsupervised methods on both two databases. For example, the accuracies of the supervised sLBIF are much better than those of the unsupervised LBIF. This proves that a supervised method, such as LDA, can improve the discriminative ability of the feature.

3.2. Experimental results with SVM classifier

Besides the NC classifier, we also show the accuracies of the different methods using SVM classifier. SVM can be seen as the baseline classifier in head pose estimation [10]. In the experiments, the kernel parameters of the radial basis function kernel

are obtained by using three-fold cross-validation on the training dataset [41]. Table 4 shows the accuracies of head pose estimation with SVM classifier on the both databases. To compare with the accuracies of the NC classifier, we also show the accuracies of NC classifier when the center number k of each pose is 9 in Table 3.

From Table 4, we can see that for the supervised methods, the performances of the SVM classifier are near to the performances of the NC classifier. But for unsupervised methods, the performances of the SVM classifier are much better than the performances of the NC classifier. This scene means that the SVM classifier is much more robust than the NC classifier.

For the SVM classifier, the accuracies of the unsupervised methods are much better than those of the supervised methods. For example, the accuracies of sLBIF are less than those of LBIF on all the three databases, while the accuracies of PCA, GF, HOG, LGBP, BIF and LBIF are less than those of LDA, GFFF, sHOG, sLGBP, sBIF and sLBIF, respectively. This differs from the general knowledge of supervised methods performing better than those of unsupervised methods. Since there is the limitation that the dimension of LDA is less than $C-1$, the dimension of these supervised methods is only 6 and 12 on the CAS-PEAL database and the Multi-PIE database, respectively. We attribute the decreasing performance of these supervised methods to the fact that the low dimension decreases the classification ability of the SVM classifier.

Finally, LBIF is the best of all methods on both the databases while the accuracies of sLBIF are the best of all the supervised methods on the Multi-PIE database. Compared with the performance of BIF (sBIF), the accuracy improvement of LBIF (sLBIF) shows that the local information of LBP can be combined with BIF features and improve the accuracy of head pose estimation.

Table 3
The accuracy (%) of pose estimation with NC classifier on different databases ($k=9$).

Method	CAS-PEAL	Multi-PIE
PCA	80.39	83.27
GF	87.63	93.82
HOG	87.82	93.23
LGBP	80.30	94.56
BIF	88.13	92.26
LBIF	**89.86**	**95.48**
LDA	84.36	91.83
GFFF	90.97	96.97
sHOG	88.42	97.03
sLGBP	92.74	98.02
sBIF	93.37	97.59
sLBIF	**94.55**	**98.53**

Table 4
The accuracy (%) of pose estimation with SVM classifier on different databases.

Method	CAS-PEAL	Multi-PIE
PCA	90.97	95.42
GF	92.12	97.75
HOG	92.62	97.27
LGBP	93.06	97.94
BIF	93.73	97.70
LBIF	**94.57**	**98.74**
LDA	87.36	92.42
GFFF	90.73	96.68
sHOG	90.65	97.16
sLGBP	92.53	97.94
sBIF	**93.25**	97.67
sLBIF	92.72	**98.45**

3.3. Experimental results with different number of blocks

In LBIF and sBIF, one important parameter is the number of blocks. To show the effectiveness of the block number, we also repeat the experiments with the different block number. In Fig. 9, we show the accuracies of LBIF and sLBIF when the number of blocks is 4, 8 and 16 on the CAS-PEAL database.

From the above figure, it is easily seen that for LBIF, the performance is robust to the block number. But for sLBIF, the performance decreases with the increase of the block number. We attribute this to the dimension of the blocks. The smaller the block number, the higher the feature dimension of the block. In other words, the smaller number of blocks means the information in the block are better adapted to pose estimation. In the extreme, we can take the full feature of LBIF as one block, and the performance is maximized since all the information is retained within the feature. In fact, considering the memory of the computer, it is difficult for only one block. Based on this scenerio, we set the block number to 4 in LBIF and sLBIF.

3.4. Experimental results with different number of orientations

Besides the block number, the other parameter in LBIF and sLBIF is the number of orientations of the Gabor filters. To show the effectiveness of the number of orientations, we also repeat the experiments by using the different number of orientations on the CAS-PEAL database. In Fig. 10, we show the accuracies of head pose estimation with the orientations number are 4, 8 and 16 on the CAS-PEAL database.

From the above figure, we can notice that the accuracies of LBIF are nearly the same for the different orientation numbers. But for sLBIF, the accuracies are increased with the increasing of the orientation number. The accuracies of four orientations are much less than those of 8 and 16 orientations. Though the accuracy with 16 orientations is better than that of eight orientations, the difference between them is very little. Considering the advantage of eight orientations in the computation complexity and the restoration requirement, we still select eight orientations in Gabor filters in LBIF and sLBIF.

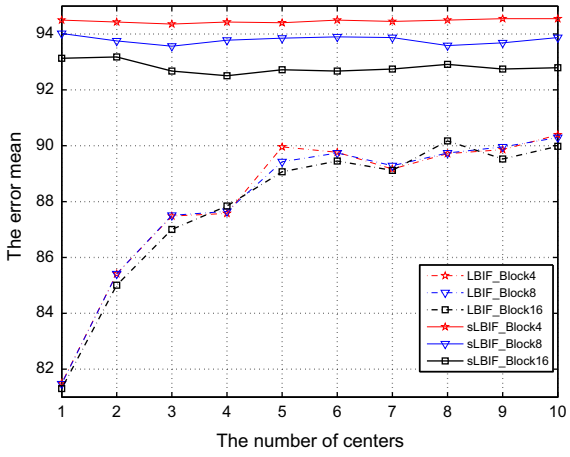

Fig. 9. The accuracy (%) of LBIF and sLBIF on the CAS-PEAL database when the block number is 4, 8 and 16. The x-axis is the centroid number of each pose and the it y-axis is the accuracy.

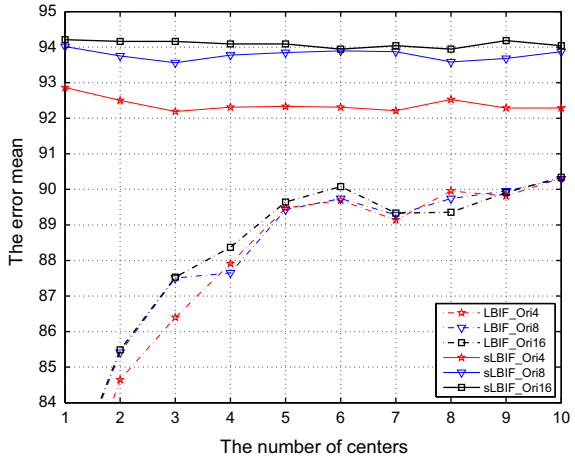

Fig. 10. The accuracy (%) of LBIF and sLBIF on the CAS-PEAL database when the orientation number is 4, 8 and 16. The *x*-axis is the centroid number of each pose and the *y*-axis is the accuracy.

4. Conclusion

Motivated by the successful application of the BIF feature in many different scenarios, this paper argues that it can also be applied on the head pose estimation problem for the first time ever. We proposes the novel LBIF representation for head yaw estimation. In LBIF, the combination of the local information of the LBP operator and the visual information of the BIF feature can improve the accuracy of head pose estimation. In addition, the ensemble of the dimension reduction method is finally applied to enhance the discriminant ability and reduce the feature dimension. Extensive experimental results illustrate the advantages of the proposed method.

There are several aspects to be further studied in the future work. First, the dimension of the final feature is still very high even it has been reduced; therefore, more effort should be put to further reduce the dimension. Second, it is also worth studying how to set the optimal parameters of Gabor filters in order to emphasize some regions and further reduce the dimension of features. Finally, this paper only uses a simple supervised method to improve the discriminant ability of LBIF. More work may be done to find which supervised methods are more adapted to the LBIF feature and the problem of head pose estimation.

Acknowledgment

This paper is partially supported by National Natural Science Foundation of China under Contract nos. 61003103, 61001193 and 61173065, Research Fund for the Doctoral Program of Higher Education under Contract no. 20100142120029 and Hubei Provincial Natural Science Foundation under Contract nos. 2010CDB02304 and 2010CDB02305.

References

[1] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Comput. Surv. 35 (2003) 399–458.
[2] E. Murphy-Chutorian, M.M. Trivedi, Head pose estimation in computer vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. April 31 (4) (2009) 607–626.
[3] L. Chen, L. Zhang, Y. Hu, M. Li, H. Zhang, Head pose estimation using Fisher manifold learning, Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (2003) 203–207.
[4] Q. Ji, R. Hu, 3D face pose estimation and tracking from a monocular camera, Image Vision Comput. 20 (7) (2002) 499–511.
[5] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models—their training and application, Comput. Vision Image Understanding 61 (1) (1995) 38–59.
[6] T.F. Cootes, G. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[7] N. Krüger, M. Pötzsch, C.V.D. Malsburg, Determination of face position and pose with a learned representation based on labeled graphs, Image Vision Comput. 15 (8) (1997) 665–673.
[8] A. Nikolaidis, I. Pitas, Facial feature extraction and determination of pose, Proceedings of the NOBLESSE Workshop on Nonlinear Model Based Image Analysis (1998) 257–262.
[9] T. Darrell, B. Moghaddam, A.P. Pentland, Active face tracking and pose estimation in an interactive room, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1996.
[10] Y. Li, S. Gong, H. Liddel, Support vector regression and classification based multi-view face detection and recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 300–305.
[11] J.N.S. Kwong, S. Gong, Learning support vector machines for a multi-view face model, Proceedings of the British Machine Vision Conference (1999) 300–305.
[12] Y. Fu, T. S. Huang, Graph embedded analysis for head pose estimation, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2006, pp. 3–8.
[13] A.H. Gee, Roberto Cipolla, Determining the gaze of faces in images, Image Vision Comput. 12 (10) (1994) 639–647.
[14] Stan.Z. Li, X.G. Lu, X.W. Hou, X.H. Peng, Q.S. Cheng, Learning multiview face subspaces and facial pose estimation using independent component analysis, IEEE Trans. Image Process. 14 (6) (2005) 705–712.
[15] Y. Wei, L. Fradet, T. Tan, Head pose estimation using Gabor eigenspace modeling, in: Proceedings of the IEEE International Conference on Image Processing, 2002, pp. 281–284.
[16] Bingpeng Ma, Shiguang Shan, Xilin Chen, Wen Gao, Head yaw estimation from asymmetry of facial appearance, IEEE Trans. Systems Man Cybern. Part B: Cybern. 38 (6) (2008) 1501–1512.
[17] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[18] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[19] Mikhail Belkin, Partha Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 15 (2001) 585–591.
[20] N. Hu, W. Huang, S. Ranganath, Head pose estimation by non-linear embedding and mapping, in: Proceedings of the IEEE International Conference on Image Processing II-342-5, 2005.
[21] B. Raytchev, I. Yoda, K. Sakaue, Head pose estimation by nonlinear manifold learning, in: Proceedings of the IEEE International Conference on Pattern Recognition, vol. 4, 2004, pp. 462–466.
[22] Maximilian Riesenhuber, Tomaso Poggio, Hierarchical models of object recognition in cortex, Nature Neurosci. 2 (11) (1999) 1019–1025.
[23] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, June 2005, pp. 994–1000.
[24] Ethan Meyers, Lior Wolf, Using biologically inspired features for face processing, Int. J. Comput. Vision 76 (1) (2008) 93–104.
[25] Guodong Guo, Guowang Mu, Yun Fu, T.S. Huang, Human age estimation using bio-inspired features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 112–119.
[26] Dongjin Song, Dacheng Tao, Biologically inspired feature manifold for scene classification, IEEE Trans. Image Process. 19 (2010) 174–184.
[27] J. Mutch, D. Lowe, Object class recognition and localization using sparse features with limited receptive fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 11–18.
[28] C. Cortes, V. Vapnik, Support vector network, Mach. Learn. 20 (1995) 273–297.
[29] L. Wiskott, J.M. Fellous, Norbert Krüger, C.V.D. Malsburg, Face recognition by elastic bunch graph matching, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 775–779.
[30] T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on feature distributions, Pattern Recognition 29 (1) (1996) 51–59.
[31] Shuicheng Yan, Dong Xu, Benyu Zhang, Hongjiang Zhang, Qiang Yang, Stephen Lin, Graph embedding and extension: a general framework for dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 40–51.
[32] Deng Cai, Jiawei Han, Xiaofei He, Kun Zhou, Hujun Bao, Locality sensitive discriminant analysis, in: Proceedings of the International Joint Conference on Artifical Intelligence, 2007, pp. 1713–1726.
[33] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 20 (7) (1997) 711–720.
[34] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244.
[35] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, Debin Zhao, The CAS-PEAL large-scale chinese face database and baseline evaluations, IEEE Trans. Syst. Man Cybern. Part A 38 (1) (2008) 149–161.
[36] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Guide to the cmu Multi-pie Database, Technical Report, Carnegie Mellon University, 2007.

[37] S. Yan, S. Shan, X. Chen, W. Gao, J. Chen, Matrix-structural learning (MSL) of cascaded classifier from enormous training set, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[38] M.A. Turk, A.P. Pentland, Eigenfaces for recognition, J. Cognitive Neurosci. 3 (1) (1991) 71–86.
[39] E. Murphy-Chutorian, A. Doshi, M. M. Trivedi, Head pose estimation for driver assistance systems: a robust algorithm and experimental evaluation, in: IEEE Intelligent Transportation Systems Conference, 2007, pp. 709–714.
[40] Bingpeng Ma, Wenchao Zhang, Shiguang Shan, Xilin Chen, Wen Gao, Robust head pose estimation using LGBP, in: Proceedings of the International Conference on Pattern Recognition, vol. 2, 2006, pp. 512–515.
[41] C. Chang, C. Lin, LIBSVM: a library for support vector machines, Software available at 〈http://www.csie.ntu.edu.tw/cjlin/libsvm 〉, 2001.

Bingpeng Ma received the B.S. degree in Mechanics, in 1998 and the M.S. degree in Mathematics, in 2003 from HuaZhong University of Science and Technology, respectively. He received Ph.D. degree at the Institute of Computing Technology, Chinese Academy of Sciences, PR China, in 2009. He was a Post-doctorial Researcher in the University of Caen, France, from 2011 to 2012. His current interests are in the areas of mathematical programming, machine learning and pattern recognition.

Xiujuan Chai received the B.S., M.S., and Ph.D. degrees in Computer Science from the Harbin Institute of Technology, Harbin, China, in 2000, 2002, and 2007, respectively. She was a Post-doctorial Researcher in Nokia Research Center (Beijing), from 2007 to 2009. She joined the Institute of Computing Technology, Chinese Academy Sciences, Beijing, in July 2009 and now she is an Assistant Professor. Her research interests include computer vision, pattern recognition, and multimodal human computer interaction.

Tianjiang Wang received the B.S. degree in Mathematics, in 1982 and the Ph.D. degree in computer software, in 2002 from HuaZhong University of Science and Technology (HUST), and now he is a Professor of HUST. He has been working at media computing and intelligent computing including machine learning and pattern recognition.