Chinese Journal of Electronics
Vol.25, No.6, Nov. 2016

Face Recognition Using Dense SIFT Feature Alignment*

ZHOU Quan[1], Shafiq ur Rehman[2], ZHOU Yu[3], WEI Xin[1], WANG Lei[1] and ZHENG Baoyu[1]

(1. *Key Lab of Ministry of Education for Broad Band Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*)

(2. *Department of Applied Physics and Electronics, Umeå University, Umeå, Sweden*)

(3. *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China*)

Abstract — This paper addresses face recognition problem in a more challenging scenario where the training and test samples are both subject to the visual variations of poses, expressions and misalignments. We employ dense Scale-invariant feature transform (SIFT) feature matching as a generic transformation to roughly align training samples; and then identify input facial images via an improved sparse representation model based on the aligned training samples. Compared with previous methods, the extensive experimental results demonstrate the effectiveness of our method for the task of face recognition on three benchmark datasets.

Key words — Face recognition, Dense SIFT feature alignment, Sparse representation.

I. Introduction

Automatic face recognition, as a high level vision task, is designed to distinguish a specific identity from the unknown objects characterized by facial images. It has been extensively studied in computer vision[1−7], and suffices various practical applications, such as face identification[8], biometrics[9], and video surveillance[10].

From the perspective of establishing face recognition model, the existing methods could be mainly categorized into two classes: supervised and unsupervised. The supervised methods identify human face in a discriminative manner, which utilize projection technique[11] or the trained classifiers[2]. In Refs.[12,13], the authors also utilize the metric learning technique to train discriminative classifiers for the task of face recognition. They are hardly generalized, however, as the human facial images are complex natural stimuli that may suffer from severe variations, such as pose changes, illuminations, occlusions, misalignment and expressions. On the other hand, the unsupervised methods employ reconstructive criterion to estimate unknown objects based on Principal component analysis (PCA)[14,15], Independent component analysis (ICA)[16], Compressed sensing (CS)[17] and Sparse representation models (SRMs)[18,19]. The PCA approach, also known as eigenface method for dimensionality reduction, yields projection directions that maximize the total scatter across all classes. An alternative approach for dimensionality reduction is tensor projection[17], which is an extension to the conventional PCA for image compression and construction. As the generalization of PCA, the ICA methods have been used to find statistically independent basis images or coefficients for the face images to deal with the sensitivity to higher order image statistics[16].

Recently, SRM achieves impressive and robust results to against the visual variations of illumination, occlusion and corruption of facial images. The training and testing images, however, are both required to be well aligned[18]. In order to address this problem, Peng *et al.* seek a set of optimal image transformations via SRM for linearly correlated facial images[20]. For the practical application, Wagner *et al.* propose an improved face recognition system[19], without the constraint of well-aligned test samples, but still subject to the aligned training samples. Overall, previous sparse approximation techniques for face recognition either are not robust to misalignment, or address the misalignment by optimizing a parametric affine transfor-

*Manuscript Received Sept. 14, 2015; Accepted Dec. 10, 2015. This work is supported by the National Natural Science Foundation of China (No.61201164, No.61201165, No.61201326, No.61271240, No.61401228, No.61571240, No.61501253), the State Key Development Program of Basic Research of China (No.2013CB329005), Jiangsu Postdoctoral Program (No.1201014C, No.1501019A), China Postdoctoral Science Foundation (No.2013M531392, No.2015M581841), Specialized Research Fund for the Doctoral Program of Higher Education (No.20113223120002), and University Natural Science Research Project of Jiangsu Province (No.11KJB510016).

DOI:10.1049/cje.2016.10.001

mation on the image domain.

This paper presents a novel approach for face recognition where the training and test samples are both subject to the visual variations of poses, expressions and misalignments. Specifically, our method begins with the image representation that abstracts the SIFT feature[21] for each pixel, then a Dense SIFT feature alignment scheme (DSFA) is designed to automatically perform image alignment via establishing correspondence to match the per-pixel SIFT features[21]. Employing such scheme enabled us to align facial images across different visual variations, without requiring that the query sample is a frontal face image. Finally, after the training samples are roughly aligned with respect to the query test image, an improved SRM is proposed to recognize test images based on the minimized reconstruction error. Experimental results on three face recognition datasets show that our method is not only robust to these visual variations, but also achieves better performance in terms of recognition accuracy.

II. Image Alignment Using DSFA

This section first presents image representation using dense SIFT features, and then describes our DSFA scheme.

1. Dense SIFT feature representation

SIFT is a robust descriptor to characterize local gradient information of image pixels[21]. In Ref.[21], it is a sparse feature representation that consists of both feature detection and extraction process. In this paper, however, we only utilize the feature extraction component. For every pixel $\boldsymbol{x} = (x, y) \in \mathcal{I}$, its neighborhood is divided (*e.g.* 16×16) into a 4×4 cell array, and the gradient orientation is quantized into 8 bins in each cell, resulting in a $4 \times 4 \times 8 = 128$-dimensional vector for pixel $\boldsymbol{x}$. We perform this operation for every pixel, then each pixel can be also represented by a 128-dimensional vector. This per-pixel SIFT description is called the dense SIFT feature representation for input facial image $\mathcal{I}$. The goal of our work is to perform alignment for every pixel based on this dense feature representation.

2. Image alignment using DSFA

Given two facial images $\mathcal{I}_1$ and $\mathcal{I}_2$ belonging to subject i, the alignment process is similar to match the corresponding pixels in $\mathcal{I}_1$ and $\mathcal{I}_2$. Let $\boldsymbol{F}(\boldsymbol{x}) = (h(\boldsymbol{x}), v(\boldsymbol{x}))$ be the flow vector at pixel $\boldsymbol{x}$, where $h(\boldsymbol{x})$ and $v(\boldsymbol{x})$ are the flow component in horizon and vertical directions, respectively. Note we only allow $h(\boldsymbol{x})$ and $v(\boldsymbol{x})$ to be integers. Inspired by Refs.[23,24], the SIFT features are expected to be matched along the flow vectors $\boldsymbol{F}(\boldsymbol{x})$, and the flow field is required to be smooth, with discontinuities agreeing with object boundaries. Based on these two criteria, the objective function to align two facial images is defined as:

$$\begin{aligned} E(\boldsymbol{F}) = &\sum_{\boldsymbol{x}} \min(||\mathcal{I}_1(\boldsymbol{x}) - \mathcal{I}_2(\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{x}))||_1, t) \\ &+ \sum_{\boldsymbol{x}} \beta(|h(\boldsymbol{x})| + |v(\boldsymbol{x})|) + \sum_{(\boldsymbol{x},\boldsymbol{y})\in\varepsilon} \min(\alpha||h(\boldsymbol{x}) \\ &- h(\boldsymbol{y})||_1, d) + \min(\alpha||v(\boldsymbol{x}) - v(\boldsymbol{y})||_1, d) \end{aligned} \tag{1}$$

where ε denotes the four-neighborhood system, and α and β are the turned parameters. In Eq.(1), the first term is the data term, which constrains the SIFT features to be matched along with the flow vector $\boldsymbol{F}(\boldsymbol{x})$ within $\mathcal{I}_1$ and $\mathcal{I}_2$. The second term is the small drift term, which constrains the flow vector $\boldsymbol{F}(\boldsymbol{x})$ to be as small as possible when $\mathcal{I}_1(\boldsymbol{x})$ and $\mathcal{I}_2(\boldsymbol{x})$ look similar. The third term is the smoothness term, which constrains the flow vectors to be consistent with the adjacent pixels. The ℓ^1-norm is both used in the data term and the smoothness term to account for flow discontinuities and matching outliers, with t and d as the threshold, respectively.

3. Optimization and complexity analysis

Suppose that an image is with the resolution $M \times N$. In our DSFA diagram, a pixel in one image $\mathcal{I}_1$ can literally match to any pixels in the other image $\mathcal{I}_2$. Therefore, the computation complexity of directly optimizing Eq.(1) is $O((M \times N)^2)$. This optimization scheme scales poorly with respect to image dimension, resulting in very time consuming when there are large number of subjects and number of image samples for each subject.

In order to address this drawback, we use a dual-layer loopy belief propagation[25] to optimize the objective function defined in Eq.(1). Differently from the traditional formulation of optical flow[23,24], the smoothness term in Eq.(1) is decoupled. This property allows us to optimize the horizontal flow $h(\boldsymbol{x})$ and vertical flow $v(\boldsymbol{x})$ separately in our optimization diagram. As a result, the computation complexity can be reduced from $O((M \times N)^2)$ to $O(M \times N)$ at one iteration of message passing. Specifically speaking, our dual-layer belief propagation on the objective function of Eq.(1) is shown in Fig.1. Firstly, a horizontal layer u and vertical layer v are constructed with exactly the same resolution of facial images, where the data term connecting pixels at the same location. In each iteration of message passing, the intra-layer messages are first updated in $h(\boldsymbol{x})$ and $v(\boldsymbol{x})$ respectively, and then the inter-layer messages are updated between $h(\boldsymbol{x})$ and $v(\boldsymbol{x})$. Since the functional form of the objective function has truncated ℓ^1-norm, the complexity can be further reduced using the distance transform function[25].

Fig.2(a) illustrates three examples of aligned facial images on ORL dataset[26] according to randomly selected query image. Notice how the pose changes, expression variations and misalignments of other images are rectified to the query image. It also shows that using DSFA is not sensitive to whether the query image is a frontal facial image or not (see third example), which makes our method more flexible for face recognition. We also perform DSFA between the samples with different object categories, and the results are exhibited in Fig.2(*b*). It shows that DSFA approach achieves good intra-class alignment, while poor inter-class alignment.

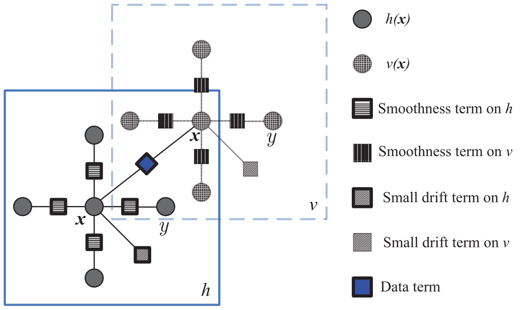
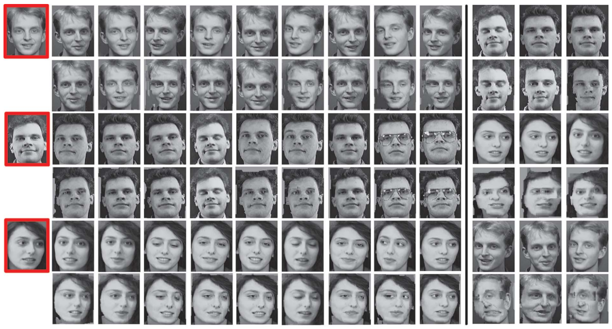


Fig. 1. Illustration of our dual-layer belief propagation on objective function defined in Eq.(1), where $\boldsymbol{x}$ and $\boldsymbol{y}$ denote adjacent pixels



(*a*) (*b*)

Fig. 2. Three examples of aligned facial images with respect to the query image (as shown in first column) on ORL dataset[26], which is able to account for the different visual variations

III. The Theory

In this section, we first introduce the improved SRM, and then elaborate on our classification algorithm for face recognition.

1. The proposed SRM

Let $\boldsymbol{A}_i = [\boldsymbol{X}_{i,1}, \boldsymbol{X}_{i,2}, \cdots, \boldsymbol{X}_{i,n_i}] \in \mathbb{R}^{m\times n_i}$ be a series of training samples for the ith object class, where $\boldsymbol{X}_{i,j} \in \mathbb{R}^m$ denotes a vector stacked from all m pixels of facial image $\mathcal{I}$. The task of face recognition is to identify the object class i of any test samples $\boldsymbol{Y} \in \mathbb{R}^m$. In the beginning, however, the membership i of $\boldsymbol{Y}$ is not known, we thus define a new matrix $\boldsymbol{A}$ for the whole training facial images as the concatenation of the N training samples of all K object categories:

$$\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_K] = [\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \cdots, \boldsymbol{X}_{K,N}] \tag{2}$$

In many practical face recognition scenarios, the training samples $\boldsymbol{A}$ and test sample $\boldsymbol{Y}$ might be both subject to some visual variations (*e.g.*, pose changes, expressions and misalignments). Let $\boldsymbol{\tau}$ be a generic transformation set acting on the image domain, and "$\circ$" denote a non-linear operator, therefore, we can apply $\boldsymbol{\tau}$ to $\boldsymbol{A}$ to produce a series of warped training samples $\boldsymbol{A}' = \boldsymbol{A}\circ\boldsymbol{\tau}$. If the training samples with the same class of $\boldsymbol{Y}$ is able to be roughly aligned to $\boldsymbol{Y}$ using transformation set $\boldsymbol{\tau}$ while the others are not, then $\boldsymbol{Y}$ lies in the linear space spanned by the entire training set $\boldsymbol{A}'$, plus a sparse error $\boldsymbol{e} \in \mathbb{R}^m$ due to the corrupted pixels:

$$\boldsymbol{Y} = \boldsymbol{A}'\boldsymbol{x} + \boldsymbol{e} = (\boldsymbol{A}\circ\boldsymbol{\tau})\boldsymbol{x} + \boldsymbol{e} \tag{3}$$

where $\boldsymbol{x} \in \mathbb{R}^N$ is a sparse coefficient vector that the most entries are zero except those associated with the same category of $\boldsymbol{Y}$. The sparse characteristic of $\boldsymbol{x}$ provides a strong cue to find the appropriate deformation set $\boldsymbol{\tau}$: one would like to seek $\boldsymbol{\tau}$ that allows the sparsest representation, solving the following ℓ^1-norm optimization problem:

$$\begin{aligned}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{e}}) &= \arg\min ||\boldsymbol{x}||_1 + ||\boldsymbol{e}||_1 \\ \text{s.t.}\ \ \boldsymbol{Y} &= (\boldsymbol{A}\circ\boldsymbol{\tau})\boldsymbol{x} + \boldsymbol{e}\end{aligned} \tag{4}$$

However, Eq.(4) has many local minima due to two facts: there are multiple faces in the matrix $\boldsymbol{A}$, and each training sample might need different transformation to perform alignment with respect to $\boldsymbol{Y}$. In our implementation, we employ the vector flow $\boldsymbol{F}(\boldsymbol{x})$ introduced in DSFA as the generic transformation $\boldsymbol{F}(\boldsymbol{x}) = \tau_{k,n} \in \boldsymbol{\tau}$ to align the training sample $\boldsymbol{X}_{k,n} \in \boldsymbol{A}$ with respect to $\boldsymbol{Y}$. Once the best transformation has been applied for each training sample, a global sparse representation problem can again be solved to obtain a discriminative representation in terms of the entire training facial images. Thus Eq.(4) can be rewritten as:

$$\begin{aligned}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{e}}) &= \arg\min ||\boldsymbol{x}||_1 + ||\boldsymbol{e}||_1 \\ \text{s.t.}\ \ \boldsymbol{Y} &= \boldsymbol{B}\boldsymbol{x} + \boldsymbol{e}\end{aligned} \tag{5}$$

where $\boldsymbol{B} = [\boldsymbol{X}_{1,1}\circ\tau_{1,1}, \boldsymbol{X}_{1,2}\circ\tau_{1,2}, \cdots, \boldsymbol{X}_{K,N}\circ\tau_{K,N}]$.

2. Classification based on the proposed SRM

In order to better harness the subspace structure associated with aligned images in face recognition, we classify test sample $\boldsymbol{Y}$ based on how well the coefficients associated with all aligned training samples of each object recover $\boldsymbol{Y}$. For each class i, let $\delta_i : \mathbb{R}^N \rightarrow \mathbb{R}^N$ be the characteristic function that selects the coefficients associated with class i. For $\hat{\boldsymbol{x}} \in \mathbb{R}^N$, $\delta_i(\hat{\boldsymbol{x}})$ is a new vector whose only nonzero entries are the entries in $\hat{\boldsymbol{x}}$ that are associated with class i. Using only the coefficients associated with the ith class, one can approximate $\boldsymbol{Y}$ as the sparse construction: $\hat{\boldsymbol{y}} = \boldsymbol{B}\delta_i(\hat{\boldsymbol{x}}) + \hat{\boldsymbol{e}}$. Then $\boldsymbol{Y}$ can be classified based on these approximations via assigning the object class that minimizes the residual between $\boldsymbol{Y}$ and $\hat{\boldsymbol{y}}$:

$$\min_i r_i(\boldsymbol{Y}) = ||\boldsymbol{Y} - \hat{\boldsymbol{e}} - \boldsymbol{B}\delta_i(\hat{\boldsymbol{x}})||_2 \tag{6}$$

where $||\cdot||_2$ is ℓ^2-norm. However, before classifying a test sample $\boldsymbol{Y}$ in the real-world scenarios, we have to first decide whether it is an outlier from none of the classes in the dataset. To this end, we also adopt the Sparsity concentration index measurement (SCIM)[17] to evaluate how concentrated the coefficients are on a single class in the dataset, after the training samples are roughly aligned:

$$SCIM(\hat{\boldsymbol{x}}) \doteq \frac{K \cdot \max_i ||\delta_i(\hat{\boldsymbol{x}})||_1/||\hat{\boldsymbol{x}}||_1 - 1}{K-1} \qquad (7)$$

The complete recognition procedure is summarized in Algorithm 1. Our implementation minimizes the ℓ^1-norm via a primal-dual algorithm for linear programming based on Ref.[27].

Algorithm 1 Classification algorithm based on proposed SRM

```
Input: a matrix of training samples: $\boldsymbol{A}$ for $K$ classes,
a test sample $\boldsymbol{Y} \in \mathbb{R}^m$, threshold $\lambda \in [0,1]$
Output: identity($\boldsymbol{Y}$)
Align training samples with respect to $\boldsymbol{Y}$ using DSFC
to get matrix $\boldsymbol{B}$ based on Section II;
Normalize the columns of $\boldsymbol{B}$ to have unit $\ell^2$-norm;
Solve the $\ell^1$-minimization problem:
    $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{e}}) = \arg\min ||\boldsymbol{x}||_1 + ||\boldsymbol{e}||_1$, s.t. $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{e}$;
Compute the residuals $r_i(\boldsymbol{Y}) = ||\boldsymbol{Y} - \hat{\boldsymbol{e}} - \boldsymbol{B}\delta_i(\hat{\boldsymbol{x}})||_2$;
if SCIM($\hat{\boldsymbol{x}}$) $\geq \lambda$ then
    Set identity($\boldsymbol{Y}$) $= \arg\min r_i(\boldsymbol{Y})$ for all $K$ classes
end
```

IV. Experimental Evaluation

To demonstrate the effectiveness of our method, we have conducted several experiments on man-made and real-life face recognition datasets.

1. Dataset

The ORL face database[26] contains 400 gray images of 40 subjects. All the images were taken against a homogeneous background, and some were taken at different times. This database includes frontal views of upright faces with facial expression (open or closed eyes/mounthes, smiling or nonsmiling), misalignment, facial occlusions (glasses or no glasses) and pose variations. The main reason behind employing ORL dataset is that this dataset has facial images with different subject gender and input stimulus variety.

The AR face dataset[28] contains over 4000 color images of 126 subjects (70 males and 56 females), including frontal views of faces with different facial expressions (neutral, smile, anger, and scream), luminance alterations (left light on, right light on, and all side lights on) and occlusion modes (sunglass and scarf). In this research, we address two fundamental challenges of face recognition, *i.e.*, the natural variations of misalignment in the head orientation and the changes in facial expressions.

The LFW face dataset[30] is collected from the web for the unconstrained face recognition. There are 13,233 color facial images from 5,749 different persons, with large pose, occlusion, expression variations. In this experiment, we crop the images with the size of 150 × 130 from the original images, and evaluate our method on this dataset to address the non-rigid deformations of poses and facial expressions.

2. Experimental setup

To show the advantages of our approach, we selected 8 state-of-the-art models as baselines for comparison, namely, TPFRS[19], GIST + nearest neighbor (GNN)[22], FDP[6], SML[7], PCA[15], Kernel PCA (KPCA)[29], ICA[16], and Fisher face (FF)[14], in terms of recognition accuracy. Each original image on three datasets is first downsampled into a low-resolution image (with 16 × 16 pixels). In order to reduce the effect with special choice of the training data, we report performance on LFW dataset using the same split settings which are randomly generated and provided by the organizers[30]. For the rest two datasets, we conducted our experiment over 10-fold cross validation with random splits that a γ ($\gamma \in [0,1]$) portion of the samples for each subject for training, and the rest $1-\gamma$ portion for testing. The settings of parameters were $\gamma = 0.5$, $\lambda = 0.7$, $\alpha = 500$, $\beta = 1.275$, $t = 765$, $d = 10^4$ to achieve the best results on three datasets. For the baseline TPFRS[19], we first use RASL[20] to align training samples, then employ Ref.[18] to perform recognition.

3. Overall results

Table 1 reports the average and standard deviation of the recognition accuracies, and demonstrates our method outperforms other models on three datasets. This probably dues to the reconstructive approaches (*e.g.*, PCA, ICA and SRM) are sensitive to the visual variations of poses, expressions and misalignments, while our approach employs DSFA, which is robust to these variations. In Ref.[21], the authors also use DSFA and describe the aligned images using GIST feature[31], but employ nearest neighbor to perform recognition. Results show that our model works better than nearest neighbor over three datasets. Among compared models, TPFRS[19] performed higher than the rest, GNN[22], SML[7], KPCA[29] and FDP[6] achieve comparable results, while PCA[15] and ICA[16] are ranked at the bottom.

4. Parameter analysis

We also analyze how the parameter γ affects recognition accuracy of our approach. Fig.3 plots the performance along with the increasing number of training samples per person on ORL[26] and AR[28] datasets. We use TPFRS[19], SML[7], and FDP[7], as baselines. Clearly, our method outperforms other models since it benefits from the advantages of our proposed SRM and DSFA. We also

Table 1. Performance comparison on ORL, AR, and LFW datasets in terms of recognition accuracy

Methods		Ours	TPFRS[19]	GNN[22]	SML[7]	FDP[6]	KPCA[29]	FF[14]	PCA[15]	ICA[16]
ORL[26]	Mean	**100%**	97.2%	95.5%	94.7%	92.4%	93.4%	92.3%	92.8%	88.1%
	Std.	**0%**	0.35%	0.58%	0.24%	0.45%	0.43%	0.82%	0.67%	0.59%
AR[28]	Mean	**91.5%**	89.5%	86.4%	84.3%	85.7%	82.7%	82.6%	80.6%	81.5%
	Std.	**0.13%**	0.51%	0.33%	0.36%	0.28%	0.69%	0.27%	0.42%	0.48%
LFW[30]	Mean	**88.3%**	85.3%	84.4%	83.9%	80.8%	80.5%	75.1%	71.2%	64.6%
	Std.	1.7%	**1.3%**	4.8%	3.3%	2.6%	2.7%	5.4%	5.5%	7.9%

observe that our model achieves 100% recognition rate on ORL dataset[26] when $\gamma = 0.5$, while others are not.

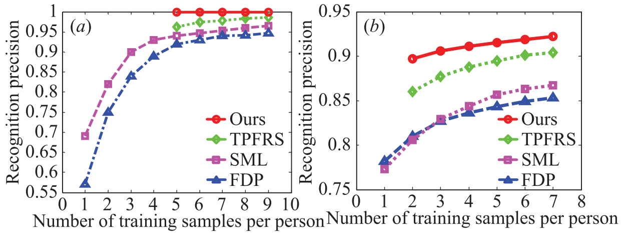


Fig. 3. The recognition accuracy changes with the number of training samples per person on (*a*) ORL dataset and (*b*) AR dataset

V. Conclusions and Future Work

In this paper, we improved SRM for robust face recognition by overcoming its sensitivity to the visual variations (*e.g.*, pose changes, expressions, and misalignments) of facial images. Our method first utilizes the DSFA to roughly align the training samples with respect to the test sample, then the improved SRM is employed to distinguish a specific face from the unknown objects. The experimental results show that our method outperforms the competing models on ORL, AR and LFW face recognition dataset in terms of the recognition accuracy.

In the future, we plan to employ robust matching scheme via vector field consensus[32] to enhance performance. We are also interested in extending the current work to address other visual variations, such as illuminations, occlusions and corruptions.

References

[1] G. Matheron, *Handbook of Face Recognition*, Spring, pp.1–30, 2011.

[2] L. Jun, *et al.*, "Local Gabor dominant direction pattern for face recognition", *Chinese Journal of Electronics*, Vol.24, No.1, pp.245–250, 2015.

[3] W. Hao and W. Youkui, "Face recognition using spatially smooth and maximum minimum value of manifold preserving", *Chinese Journal of Electronics*, Vol.22, No.1, pp.71–75, 2013.

[4] H. Haifeng, "Face recognition with image sets using locally grassmannian discriminant analysis", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.24, No.9, pp.1461–1474, 2014.

[5] W. Weihong, *et al.*, "Face recognition based on deep learning", *Human Centered Computing*, Spring, pp.812–820, 2015.

[6] L. Wolf, *et al.*, "Face recognition in unconstrained videos with matched background similarity", *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.529–534, 2011.

[7] C. Qiong, *et al.*, "Similarity metric learning for face recognition", *Proc. of IEEE Conference on Computer Vision*, pp.2408–2415, 2013.

[8] W. Shen, *et al.*, "Face identification using reference-based features with message passing model", *Neurocomputing*, Vol.99, No.6, pp.339–346, 2012.

[9] R. Tokola, *et al.*, "3D face analysis for demographic biometrics", *Proc. of IEEE Conference on Biometrics*, pp.201–207, 2015.

[10] K. Assaleh, *et al.*, "Combined features for face recognition in surveillance conditions", *Neural Information Processing*, pp.503–514, 2014.

[11] C. Wang, *et al.*, "Singular value decomposition projection for solving the small sample size problem in face recognition", *Journal of Visual Communication and Image Representation*, Vol.26, No.6, pp.265–274, 2015.

[12] X. Ben, *et al.*, "Kernel coupled distance metric learning for gait recognition and face recognition", *Neurocomputing*, Vol.120, No.7, pp.577–589, 2013.

[13] X. Ben, *et al.*, "An improved biometrics technique based on metric learning approach", *Neurocomputing*, Vol.97, No.7, pp.44–51, 2012.

[14] N. Peter, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.19, No.7, pp.711–720, 1997.

[15] M. Turk and A. Pentland, "Eigenfaces for recognition", *Journal of cognitive neuroscience*, Vol.3, No.1, pp.71–86, 191.

[16] M.S. Bartlett, *et al.*, "Face recognition by independent component analysis", *IEEE Transactions on Neuron Network*, Vol.13, No.6, pp.1450–1464, 2002.

[17] B. Du, *et al.*, "Hyperspectral biological images compression based on multiway tensor projection", *Proc. of IEEE Conference on Multimedia and Expo*, pp.1–6, 2014.

[18] J. Wright, *et al.*, "Robust face recognition via sparse representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.31, No.2, pp.210–227, 2009.

[19] A. Wagner, *et al.*, "Toward a practical face recognition system: Robust alignment and illumination by sparse representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.34, No.2, pp.372–386, 2012.

[20] Y. Peng, *et al.*, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.34, No.11, pp.2232–2246, 2012.

[21] D.G. Lowe, *et al.*, "Distinctive image features from scale invariant keypoints", *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110, 2004.

[22] C. Liu, *et al.*, "Sift flow: Dense correspondence across scenes and its applications", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.33, No.5, pp.978–994, 2011.

[23] A. Bruhn, *et al.*, "Lucas/kanade meets horn/schunck: Combining local and global optic flow methods", *International Journal of Computer Vision*, Vol.61, No.3, pp.211–231, 2005.

[24] T. Brox, *et al.*, "High accuracy optical flow estimation based on a theory for warping", *Proc. of Europe Conference on Computer Vision*, pp.25–36, 2004.

[25] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient belief propagation for early vision", *International Journal of Computer Vision*, Vol.70, No.1, pp.41–54, 2006.

[26] F.S. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification", *Proc. of IEEE Workshop on Applications of Computer Vision*, pp.138–142, 1994.

[27] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, London, England, pp.1–326, 2004.

[28] A.M. Martinez, "The ar face database", *CVC Technical Report*, 1998.

[29] K.I. Kim, *et al.*, "Face recognition using kernel principal component analysis", *IEEE Signal Processing Letters*, Vol.9, No.2, pp.40–42, 2002.

[30] B. Gary, *et al.*, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments", *Technical Report*, University of Massachusetts, Amherst, 2007.

[31] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope", *International Journal of Computer Vision*, Vol.42, No.3, pp.145–175, 2001.

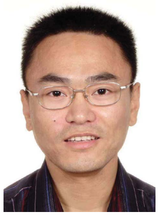
[32] J.Y. Ma, *et al.*, "Robust point matching via vector field consensus", *IEEE Transactions on Image Processing*, Vol.23, No.4, pp.1706–1721, 2014.

ZHOU Quan was born in Ezhou, Hubei, China. He received the B.S. degree in electronics and information engineering in 1998 from China University of Geosciences, Wuhan, China, and M.S. degree and Ph.D. degree in communication and information system in 2006 and 2013, respectively, from Huazhong University of Science and Technology, China. He is now an assistant professor of Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition. He has published more than 10 articles in top journal (*e.g.*, IEEE TIP, EL, and sensors) and conference (ICIP, ICASSP, ACCV, and ICPR) in image processing and computer vision. He now serves as TPC member of many international conferences and reviewer for a series of SCI journals, including IEEE TIP, IEEE TC, IEEE TSP, IEEE CSVT, PR, and Neurocomputing. He is member of IEEE. (Email: quan.zhou@njupt.edu.cn)

Shafiq ur Rehman received the Ph.D. degree in multimodal signal processing at Umeå University, Sweden, in 2010. After several year in industry, currently he is associate professor and director of i2lab at Department of Applied Physics and Electronics, Umeå University, Sweden. He is interested in theoretical and experimental research in areas of mobile computing, and multimodal signal processing based on image analysis and computer vision methods. As author of several book chapters and international conference/journal papers in recent years, he has been emphasizing of new detection and tracking methods and techniques using monocular camera as well as 3D sensors in order to make human-machine interaction richer and more expressive. (Email: shafiq.urrehman@umu.se)

ZHOU Yu was born in Xiangcheng, Henan, China. received the B.S. degree in electrical engineering from Wuhan Polytechnic University (WHPU), Wuhan, China in 2007, and the M.S. and Ph.D. degrees both in electronics and information engineering from Huazhong University of Science and Technology (HUST), Wuhan, China in 2010 and in 2014, respectively. From January 2012 to January 2013, he worked in the Department of Computer Sciences and Information, Temple University. Now he is a post-doctoral research fellow in School of Computer Sciences, Beijing University of Posts and Telecommunications (BUPT). His research interests include computer vision and pattern recognition. (Email: yu.zhou@bjupt.edu.cn)

WEI Xin was born in Nanjing, Jiangsu, China. He received the B.S. degree in electronic science and technology from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in 2005, and the Ph.D. degree in information and communication engineering from the Southeast University, Nanjing, China, in 2009. From 2009 to 2011, he was a postdoctoral researcher with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Now, he was an associate professor with college of Communication and Information Engineering, NUPT, Nanjing, China. His current research interests include machine learning, pattern recognition, and statistical signal processing. (Email: xwei@njupt.edu.cn)



WANG Lei was born in Qitaihe City, Heilongjiang, China. He received the M.S. degree and the Ph.D. degree in telecommunications and information engineering from Nanjing University of Posts and Telecommunications (NUPT), China, in 2007 and 2010, respectively. From 2012 to 2013, he was a postdoctoral research fellow at the Department of Electrical Engineering, Columbia University, USA. He is currently an associate professor at the College of Telecommunications and Information Engineering, NUPT, China. His research interests include multimedia information process. (Email: wanglei@njupt.edu.cn)

ZHENG Baoyu was born in Fuzhou, Fujian, China. He received the B.S. degree in electronic science and technology from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China. He is currently a full professor at the College of Telecommunications and Information Engineering, NUPT, China. His research interests include signal processing in communications and quantum signal processing. He also served as chair for communication theory and signal processing of China communication academy, associate chair of Nanjing Communication chapter of IEEE. He is senior member of IEEE. (Email: zby@njupt.edu.cn)