Multimed Tools Appl (2018) 77:10501–10519
DOI 10.1007/s11042-017-4569-1

Face recognition via fast dense correspondence

Quan Zhou[1,2] · Cheng Zhang[3] · Wenbin Yu[2,4] · Yawen Fan[1] · Hu Zhu[1] · Xiaofu Wu[1] · Weihua Ou[5] · Weiping Zhu[1,6] · Longin Jan Latecki[7]

Received: 17 January 2017 / Revised: 27 February 2017 / Accepted: 1 March 2017/
Published online: 11 March 2017

Abstract Face recognition plays a significant role in computer vision. It is well know that facial images are complex stimuli signals that suffer from non-rigid deformations, including misalignment, orientation, pose changes, and variations of facial expression, etc. In order to address these variations, this paper introduces an improved sparse-representation based face recognition method, which constructs dense pixel correspondences between training and

✉ Quan Zhou
quan.zhou@njupt.edu.cn

Cheng Zhang
37500419@qq.com

Wenbin Yu
ywb1518@126.com

Yawen Fan
ywfan@njupt.edu.cn

Hu Zhu
zhuhu@njupt.edu.cn

Xiaofu Wu
xfuwu@njupt.edu.cn

Weihua Ou
ouweihuahust@gmail.com

Weiping Zhu
weiping@ece.concordia.ca

Longin Jan Latecki
latecki@temple.edu

[1] Key Lab of Ministry of Education for Broad Band Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing, China

[2] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China

testing facial samples. Specifically, we first construct a deformable spatial pyramid graph model that simultaneously regularizes matching consistency at multiple spatial extents - ranging from an entire image, though coarse grid cells, to every single pixel. Secondly, a matching energy function is designed to perform face alignment based on dense pixel correspondence, which is very effective to address the issue of non-rigid deformations. Finally, a novel coarse-to-fine matching scheme is designed so that we are able to speed up the optimization of the matching energy function. After the training samples are aligned with respect to testing samples, an improved sparse representation model is employed to perform face recognition. The experimental results demonstrate the superiority of the proposed method over other methods on ORL, AR, and LFWCrop datasets. Especially, the proposed approach improves nearly 4.4 % in terms of recognition accuracy and runs nearly 10 times faster than previous sparse approximation methods.

Keywords Image matching · Face alignment · Face recognition · Deformable spatial pyramid graph · Dense correspondence

1 Introduction

With the rapid progress of big data technique, more and more facial images have been uploaded to Internet. Face recognition, as a most important vision task, is designed to recognize a specific identity from the unknown subjects characterized by the facial images. It has been extensively studied in computer vision [2, 8, 19, 29, 31, 32, 42], and facilitates various real applications, such as robot vision [19], face identification [31], facial emotion recognition [8], video surveillance [2], and biometrics [29], etc. There also exits a large number of public benchmarks [16, 26, 30], providing criteria to evaluate the state-of-the-art face recognition models. However, it is well know that facial images are complex nature stimuli signals that suffer from non-rigid deformations, such as misalignment, orientation, pose changes, and variations of facial expression, etc, which poses a challenging problem to identify an subject in the wild scenario settings.

The traditional face recognition methods mainly include two components: dimensionality reduction and classifier. For the first component, the principal component analysis (PCA), also known as eigenface approach, is widely used for face recognition [33]. This method yields projection directions that maximize the total scatter across all subject classes. An alternative method for dimensionality reduction is linear discriminative analysis (LDA) [18], which is an extension to the conventional PCA by maximizing the discriminative power of projected subspaces. The independent component analysis (ICA), as the generalized version of PCA, is proposed to find the statistically independent basis images, and use

[3] The State Grid HuBei Information and Telecommunication Company, Wuhan, China

[4] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing, China

[5] School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China

[6] Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

[7] Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA

the associated sparse coefficients to deal with the sensitivity to higher order image statistics [3]. For the second component, on the other hand, the mostly common used classifiers include nearest neighbor model [21], linear regression model [27], support vector machines [5, 12–14], minimax probability machine [15] and convolutional neural network [9]. The commonalities of all these classifiers are that they evaluate the similarity between the test image and the training samples to perform classification.

Unlike these representative models, recently, the sparse representation model (SRM) is proposed for the task of face recognition [38], where the query image can be classified by the training samples of all subject classes. The main idea is to reconstruct test samples in an over-complete dictionary whose base elements are *the training facial images themselves*. Once the test image can be approximately linearly represented by the space spanned from all training samples, one can identify the associated category using the sparse reconstruction residue [38]. The SRM achieves impressive results to against the visual variations of illumination, occlusion and corruption of facial images, which attracts great interest in further research on SRM. Therefore, many variations of SRM have been proposed in the last decade, such as Gabor feature based SRM [39], locality-constraint SRM [36], sparse and dense hybrid SRM [17], and dictionary learning SRM [40].

Although the SRMs obtain impressive classification results, they heavily rely on an assumption that the training and testing images are both required to be linearly correlated [38]. A violation of this assumption undoubtedly results in poor performance of the sparse representation-based classification [28, 35]. Especially in the scenario of real application, due to the non-rigid visual deformations (misalignment, orientation, pose changes, and variations of facial expression), the facial images of different subject category tend to be with less correlations than those of the same subject class. In order to enhance the correlation of the training example with the same class, Peng et al. seek a set of optimal image transformations to align facial images [28]. For the practical application, Wagner et al. propose an improved face recognition system [35], without the constraint that the test samples are linearly correlated, but still subject to the correlated training samples. They design an iterative algorithm to address correlation enhancement, where each iteration includes two operation: image alignment and face identification, only considering a parametric affine transformation on the image domain. Overall, previous sparse approximation techniques do not solve the essential problem of SRM when intra-class samples have insufficient correlations. Additionally, the iterative algorithm leads to the potential computational burden. Instead of the conventional SRMs, it seems that we are required to separate the image alignment from the classification procedure.

This paper presents a novel approach to face recognition in a more challenging scenario where the training and testing samples might be *both* subject to nonlinear correlations, such as the non-rigid visual variations of poses, expressions and misalignments. A deformable spatial pyramid graph model (DSPGM) is first designed to align training examples with respect to test sample through fast dense matching, then an improved SRM (ISRM) is introduced to perform face recognition. Specifically, the proposed DSPGM regularizes matching consistency of two facial images at multiple spatial extents, ranging from an entire image, to coarse rectangle grid cells, to every single pixel. The basic idea behind our approach is to strike the trade-off between robustness to image variations on one hand, and accurate prediction of pixel correspondences on the other hand. This balance can be achieved through a pyramid graph, whose larger spatial vertices offer greater regularization when appearance matches are ambiguous, while the smaller counterparts help to localize pixel correspondences within fine details. Furthermore, the hierarchical structure of our DSPGM naturally leads to an efficient optimization procedure. After the training samples are roughly aligned

with respect to the query test image, an ISRM and classification algorithm are designed to recognize test images based on the minimized reconstruction error. The proposed approach enables us to align facial images across different visual variations, without requiring the query sample to be a correlated face image. Additionally, there is no iterative process in our classification algorithm, resulting in high computational efficiency. We evaluate our method on three popular face recognition datasets: ORL [30], AR [26], and LFWCrop [16]. Experimental results show that our method is not only robust to non-rigid visual variations, but also achieves better performance in terms of recognition accuracy and implemental efficiency. In summary, the main contributions of this paper are three-folds:

- To address the non-rigid visual variations of facial images, we designed an DSPGM to perform image alignment. One merit of the proposed DSPGM lies in that it regularizes matching consistency using local and global spatial extents, yielding accurate pixel localization and matching correspondence.
- Unlike previous methods that utilize exhaustive searching scheme to densely match pixels, the hierarchical structure of proposed DSPGM allows us to design a coarse to fine matching diagram, which greatly improves the computational efficiency.
- We introduce a novel ISRM and associated classification algorithm based on the proposed DSPGM. On the public evaluation benchmarks, the experimental results demonstrate that our approach outperforms previous top ranked models in terms of recognition accuracy and efficiency.

The remainder of this paper is organized as follows. We first describe the construction of DSPGM in Section 2. Section 3 elaborates on the details of our ISRM and classification algorithm. Experimental results are given in Section 4. Finally, we give concluding remarks and future work in Section 5.

2 Image alignment using DSPGM

In this section, we first introduce our DSPGM, and then elaborate on the matching objective to perform facial image alignment. Finally, we present the coarse to fine optimal algorithm and analyze its complexity.

2.1 DSPGM

The whole construction of our DSPGM is shown in Fig. 1. We first start from the entire image $\mathcal{I}$, which is quartered into four grid cells in a traditional pyramid model [20]. Thereafter, each grid cell is further divided into four smaller rectangular grid cells. This procedure is terminated by some simple criterion such as the predefined number of pyramid levels $\mathcal{L}$ is reached. Therefore, each finest grid cell has $\frac{W \times H}{2^{\mathcal{L}}}$ pixels, where W and H are image width and hight, respectively. Unlike conventional spatial pyramid, however, in addition to those $\mathcal{L}$ partition levels, we further add one more layer, a pixel-level layer, such that the finest cells are one pixel in width.

Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ denote our DSPGM, in which image $\mathcal{I}$ and containing spatial interactions are encoded based on these grid cells and image pixels. As shown in Fig. 1, each grid cell and pixel (denoted as red circle) is defined as a node $v \in \mathcal{V}$. The edge set $\mathcal{E}$ consists of a set of edges $e =< v_i, v_j > \in \mathcal{E}$ (denoted as red line), connecting the neighboring nodes v_i and v_j, where the associated grid cells are within the same level, as well as parent-child nodes across adjacent levels. For the pixel level, however, our DSPGM does not contain links

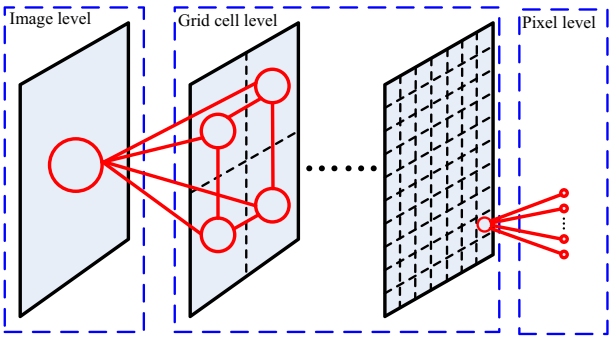


Fig. 1 Illustration of DSPGM (best viewed in color)

between neighbor pixels; each pixel is only linked to its parent cell. This scheme saves us a lot of edge connections in the finest level that would otherwise dominate run-time during optimization.

2.2 Alignment model of facial images

In order to align two facial images, it is required to construct a correspondence among all nodes defined in proposed DSPGM. In this work, we define our matching objective to align all nodes of two images based on SIFT features [24]. This per-node 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 node based on this dense feature representation.

Given two facial images $\mathcal{I}_1$ and $\mathcal{I}_2$ belonging to subject i, the alignment process is similar to matching corresponding nodes in $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. Let $\mathbf{F}_i = (x_i, y_i)$ be the flow vector for node v_i, where x_i and y_i are the flow component in horizontal and vertical directions, respectively. Note we only allow x_i and y_i to be integers. Inspired by [7, 10], the nodes with similar feature appearance are expected to be matched along with the flow vectors $\mathbf{F}_i$, and the flow field is required to be smooth, with discontinuities agreeing with object boundaries. Based on these two criteria, we want to find the optimal flow vector of each node in the first facial image to match it to the second one, through minimizing the following energy function:

$$E(\mathbf{F}) = \sum_{v_i \in \mathcal{V}} D_i(\mathbf{F}_i) + \alpha \sum_{e=<v_i, v_j> \in \mathcal{E}} S_{ij}(\mathbf{F}_i, \mathbf{F}_j) \tag{1}$$

where $D_i(\cdot)$ denotes the *data term* for node v_i, $S_{ij}(\cdot, \cdot)$ represents the *smoothness term* for the connected nodes v_i and v_j, and α is a turned parameter. Note that the edge connections span across the consecutive pyramid levels, as well as within the same pyramid levels.

In (1), the data term constrains the appearance matching cost for node v_i. It is defined as the average dissimilarity between local pixel SIFT features within node v_i in the first image $\mathcal{I}_1$ and those located within a grid cell of the same scale in the second image $\mathcal{I}_2$, after shifting by $\mathbf{F}_i$:

$$D_i(\mathbf{F}_i) = \frac{1}{|n_i|} \sum_{\boldsymbol{p}} \min(||d_1(\boldsymbol{p}) - d_2(\boldsymbol{p} + \mathbf{F}_i)||_1, \lambda) \tag{2}$$

where $|n_i|$ is the total number of pixels contained in grid cell associated with v_i, $\boldsymbol{p}$ denotes pixel coordinates within the node v_i from which local SIFT descriptors were extracted, and

$d_1(\cdot)$ and $d_2(\cdot)$ are the descriptors extracted at the locations $\boldsymbol{p}$ and $\boldsymbol{p}+\mathbf{F}_i$ in the first and second image, respectively. Note in the pixel level of DSPGM, $|n_i| = 1$, thus (2) can be simplified to:

$$D_i(\mathbf{F}_i) = \min(||d_1(\boldsymbol{p}) - d_2(\boldsymbol{p}+\mathbf{F}_i)||_1, \lambda) \tag{3}$$

On the other hand, the smoothness term in (1) regularizes the energy function by penalizing large discrepancies in the matching locations of neighboring nodes, which constrains the flow vectors to be consistent within the adjacent nodes:

$$S_{ij}(\mathbf{F}_i, \mathbf{F}_j) = \min(||\mathbf{F}_i - \mathbf{F}_j||_1, \gamma) \tag{4}$$

The ℓ^1-norm is both used in the *data term* and the *smoothness term* to account for flow discontinuities and matching outliers, with λ and γ as the thresholds, respectively.

Our matching objective has three main advantages. First of all, the hierarchical structure of our DSPGM is defined by grid cells of varying spatial extents, allowing us to overcome matching ambiguities without committing to a single spatial scale. Secondly, our data term aggregates ensemble local SIFT matches within each node, as opposed to using a single match from each individual pixel. This greatly enhances robustness with respect to non-rigid image variations. Thirdly, the nodes are explicitly linked within different spatial extents to impose matching smoothness, striking a balance between strong regularization by the larger nodes and accurate localization by the finer nodes.

2.3 Model optimization and complexity analysis

We minimize the matching objective function defined in (1) using loopy belief propagation [11] to find the optimal correspondence $\mathbf{F}$. For dense matching, it is required to consider computation complexity for scalability.

Generally speaking, there are two major factors that take most of the calculation time: (1) computing SIFT feature distances at every possible flow position and (2) optimization via belief propagation (BP) algorithm. For the first factor, the computational complexity is $O(mlk)$, where m is the number of SIFT features abstracted in the first image, and l is the number of all potential matching position, and k is the feature dimension. For the second factor, the generalized distance transform technique proposed in [11] is utilized, which is able to reduce the computing cost of message passing between nodes from $O(l^2)$ to $O(l)$. Even so, the overall run-time complexity of BP algorithm is $O(nl)$, where n is the number of nodes in our DSPGM. Thus, the total cost of this optimized scheme is $O(mlk+nl)$. Note that n, m, and l are all based on the order of the number of pixels. Obviously, it is far from efficient if solving (1) at once. Therefore, we propose a two stage hierarchical diagram to further improve computational efficiency, as shown in Fig. 2. In the first stage, the coarse solution of (1) is initialized using BP for all nodes in our DSPGM except the pixel-level ones. In the second stage, the initialized matching results are refined at the pixel-level nodes to get the finest alignment correspondence.

Observing Fig. 1, the hierarchical structure of DSPGM allows us to solve BP on larger grid cells, it essentially narrows down the possible solution space as it performs to the small cells, reducing the number of potential matching positions l. Moreover, we also observe that sparse descriptor sampling is enough for the image-level and grid cell level BP: as long as a grid cell includes more than 100 of local SIFT descriptors, its average descriptor distance for the data term defined in (2) provides a reliable and robust matching cost. As a result, it is not required to compute dense descriptors in the image-level and grid cell level BP, substantially

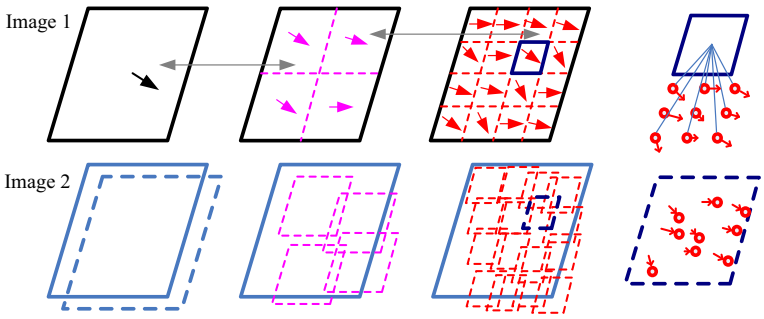


Fig. 2 Sketch of our optimization procedure. The flow vectors of nodes in image level and grid cell level are represented using *single arrows* with different color. The *double arrow* denotes he parent-child connections across two adjacent levels (best viewed in color)

reducing m. Finally, there is no loopy graph in the pixel-level of our DSPGM, which means the solution of finest level can be calculated very efficiently in a non-iterative manner. Once we have obtained the initialized matching results at the coarse level, the optimal alignment position $\mathbf{F}_i$ for the ith pixel-level node is simply determined as:

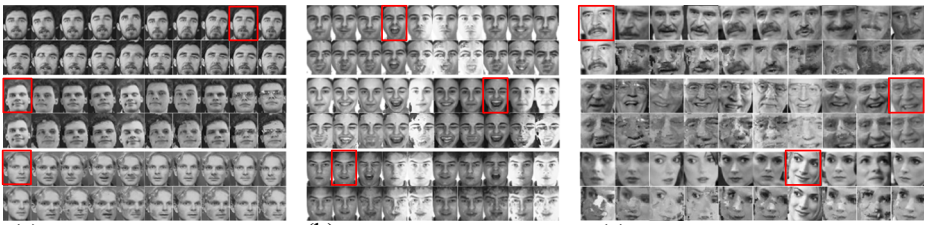
$$\mathbf{F}_i = \arg\min_{\mathbf{F}}[D_i(\mathbf{F}) + \alpha S_{ij}(\mathbf{F}, \mathbf{F}_j)] \tag{5}$$

where $\mathbf{F}_j$ is the flow vector of parent node v_j, which is connected to node v_i.

Figure 3 illustrates some examples of aligned facial images on ORL [30], AR [26] and LFWCrop [16] datasets according to randomly selected query image (denoted as red rectangle). Notice how the pose changes, expression variations and misalignments of other images are rectified to the query image. It also shows that using our matching method to perform alignment is not sensitive to whether the query image is a frontal facial image or not (see aligned results on LFWCrop dataset), which makes our method more flexible for face recognition.

3 Improved sparse representation model (ISRM) for face recognition

In this section, we first introduce the ISRM, and then elaborate on the associated classification algorithm for face recognition.



(a) aligned results on ORL dataset (b) aligned results on AR dataset (c) aligned results on LFWCrop dataset

Fig. 3 Visual examples of the alignment results of ORL (**a**), AR (**b**), and LFWCrop (**c**) using our two stage hierarchical matching diagram (best viewed in color)

3.1 The ISRM

Let $\mathbf{A}_i = [\mathbf{X}_{i,1}, \mathbf{X}_{i,2}, \cdots, \mathbf{X}_{i,n_i}] \in \mathbb{R}^{m \times n_i}$ be a series of training samples for the ith object class, where $\mathbf{X}_{i,j} \in \mathbb{R}^m$ 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 $\mathbf{Y} \in \mathbb{R}^m$. In the beginning, however, the membership i of $\mathbf{Y}$ is unknown, we thus define a new matrix $\mathbf{A}$ for the whole training facial images as the concatenation of the N training samples of all K object categories:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_K] = [\mathbf{X}_{1,1}, \mathbf{X}_{1,2}, \cdots, \mathbf{X}_{K,N}] \tag{6}$$

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

$$\mathbf{Y} = (\mathbf{A} \circ \boldsymbol{\tau})\boldsymbol{x} + \boldsymbol{e} \tag{7}$$

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 $\mathbf{Y}$. According to [38], 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:

$$\begin{aligned} &(\hat{\boldsymbol{x}}, \hat{\boldsymbol{e}}) = \arg\min ||\boldsymbol{x}||_1 + ||\boldsymbol{e}||_1 \\ &\text{subject to} \quad \mathbf{Y} = (\mathbf{A} \circ \boldsymbol{\tau})\boldsymbol{x} + \boldsymbol{e} \end{aligned} \tag{8}$$

However, directly optimizing (8) is very hard since it has many local minima. On one hand, there are multiple faces in the matrix $\mathbf{A}$, and on the other hand, each training sample might need different transformation to perform alignment with respect to $\mathbf{Y}$. In our implementation, we align the training sample $\mathbf{X}_{k,n} \in \mathbf{A}$ using the vector flow $\mathbf{F} = \tau_{k,n} \in \boldsymbol{\tau}$ introduced in Section 2.2. Once the best transformation has been applied to each training sample, a global sparse representation problem can be solved to obtain a discriminative representation in terms of the entire training facial images. Thus (8) can be rewritten as:

$$\begin{aligned} &(\hat{\boldsymbol{x}}, \hat{\boldsymbol{e}}) = \arg\min ||\boldsymbol{x}||_1 + ||\boldsymbol{e}||_1 \\ &\text{subject to} \quad \mathbf{Y} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e} \end{aligned} \tag{9}$$

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

3.2 Classification algorithm

In order to better harness the subspace structure associated with aligned images in face recognition, we classify test sample $\mathbf{Y}$ based on how well the coefficients associated with all aligned training samples of each object recover $\mathbf{Y}$. For each class i, let $\delta_i : \mathbb{R}^N \to \mathbb{R}^N$ be the binary 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 $\mathbf{Y}$ as the

sparse construction: $\hat{\boldsymbol{y}} = \mathbf{B}\delta_i(\hat{\boldsymbol{x}}) + \hat{\boldsymbol{e}}$. Then $\mathbf{Y}$ can be classified via assigning the object class that minimizes the residual between $\mathbf{Y}$ and $\hat{\boldsymbol{y}}$:

$$\min_i r_i(\mathbf{Y}) = ||\mathbf{Y} - \hat{\boldsymbol{e}} - \mathbf{B}\delta_i(\hat{\boldsymbol{x}})||_2 \tag{10}$$

where $||\cdot||_2$ is ℓ^2-norm. 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 [6].

Algorithm 1 Classification algorithm based on proposed ISRM

Input: Matrix of training samples $\mathbf{A} = [\mathbf{X}_{1,1}, \mathbf{X}_{1,2}, \cdots, \mathbf{X}_{K,N}]$; a test sample $\mathbf{Y} \in \mathbb{R}^m$
Output: identity($\mathbf{Y}$)

1 **for** $k = 1$ *to* K **do**
2 Align training samples $\mathbf{X}_{k,n}$ with respect to $\mathbf{Y}$ according to generic transformation $\mathbf{F} = \tau_{k,n} \in \tau$ to get matrix $\mathbf{B}$;
3 **end**
4 Normalize the columns of $\mathbf{B}$ to have unit ℓ^2-norm;
5 Solve the ℓ^1-minimization problem: $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{e}}) = \arg\min ||\boldsymbol{x}||_1 + ||\boldsymbol{e}||_1$, subject to $\mathbf{Y} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e}$;
6 Compute the residuals $r_i(\mathbf{Y}) = ||\mathbf{Y} - \hat{\boldsymbol{e}} - \mathbf{B}\delta_i(\hat{\boldsymbol{x}})||_2$;
7 Set identity($\mathbf{Y}$) $= \arg\min r_i(\mathbf{Y})$ for all K classes.

4 Experimental evaluation

In order to demonstrate the effectiveness of the proposed method, we have conducted several experiments on ORL [30], AR [26], and LFWCrop [16] face recognition datasets.

4.1 Datasets

The ORL face dataset [30] contains 400 gray images of 40 subjects, and some visual examples are shown in Fig. 4. 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/mouthes, smiling or non-smiling), 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.

Fig. 4 Some visual examples from ORL facial database

Fig. 5 Some visual examples from AR facial database

The AR face dataset [26] 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. Some facial images are shown in Fig. 5.

The LFWCrop dataset is the crop version of LFW facial dataset [16], which is collected from the web for the unconstrained face recognition. The face detection is first performed using [34], when the detected faces are normalized with the uniform resolution of 64 $\times$ 64. There are 13,233 color facial images from 5,749 different persons, with large pose, occlusion, expression variations. Some cropped facial images are illustrated in Fig. 6. We evaluate our method on this dataset to address the non-rigid deformations of poses and facial expressions.

4.2 Experimental setup

To show the advantages of our approach, we selected 8 state-of-the-art models as baselines for comparison, namely, TPFRS [35], LBP [1], GIST + nearest neighbor (GNN) [43], PCA [33], ICA [3], SF [23], Fisher face (FF) [4], and Deep Learning (DL) [37], in terms of recognition accuracy and efficiency. In our experience, the facial images of ORL is downsampled into a low-resolution image with 16 $\times$ 16 pixels, while the images of AR dataset is downsampled with resolution of 60×80 pixels. In order to reduce the effect with special choice of the training data, we report performance over 10-fold cross validation on LFWCrop dataset, and utilize the same split settings provided by [16]. For the rest two datasets, we conducted our experiment over 30-fold cross validation with random splits that a η ($\eta \in [0, 1]$) portion of the samples for each subject for training, and the rest $1 - \eta$ portion for testing. The settings of parameters were $\eta = 0.5$, $\lambda = 128$, $\gamma = 64$, $\alpha = 600$, and number of level in DSPGM $\mathcal{L} = 4$ to achieve the best results on three datasets. For the method of TPFRS [35], we first use RASL [28] to align training samples, then employ [38] to perform recognition.

Fig. 6 Some visual examples from LFWCrop facial database

4.3 Overall results

Table 1 reports the average and standard deviation of the recognition accuracies, compared with the baseline approaches. It demonstrates our method outperforms other models on three datasets, especially achieving 100 % recognition accuracy in ORL and AR dataset. It is surprising that our method performs better than DL approach, probably due to the fact that the deep learning models are sensitive to the non-rigid visual deformations. We also observe that, compared with the results on LFWCrop dataset, higher performance is obtained among all recognition models on ORL and AR dataset, probably because the facial images on LFWCrop dataset have great visual variations than other two datasets.

Compared with other state-of-the-art models, our method improves the recognition accuracy by 6.8 %, 5.9 %, and 9.1 % on three datasets, respectively, and the average performance is enhanced by 7.3 %. Especially when compared with conventional SRM model [35], the recognition accuracy is improved by 4.4 % on all datasets. Among all the baseline methods, DL, TPFRS and SF model obtain the best results, but they are hard to solve the non-rigid visual variations, leading to the drastic drop of performance (especially 10 % on LFWCrop dataset). On the other hand, GNN [43], SF [23], FF[4] and LBP [1] achieve comparable results, while PCA [33] and ICA [3] are ranked at the bottom. This is probably due to the fact that they are sensitive to the visual variations of poses, expressions and misalignments. Additionally, we also discover that our method achieves highest improvement on LFWCrop dataset. Compared with other two datasets, this dataset has more non-rigid deformations, which also demonstrates that using DSPGM in our approach is robust to these variations.

4.4 Efficiency

In Table 2, we also compare the implemental efficiency between our method and other baseline methods in terms of average running time per image. All the methods are executed on a dual-core I5 personal computer with 2.6 GHz CPU and 16 GB memory. The average running time of our method on three datasets are 0.22 s, 0.27 s, and 0.21 s, respectively. As shown in Table 2, TPFRS [35] achieves the highest recognition accuracy among all baseline methods, yet it is executed very slower, yielding 2.31 s, 2.74 s, and 2.57 s running time on three dataset, respectively. This is probably because it requires very long time to execute its iterative algorithm. One the other hand, due to the hard training of deep learning model and

Table 1 Performance comparison on ORL [30], AR [26] and LFWCrop [16] datasets in terms of recognition accuracy

Method	Recognition accuracy (%)		
	ORL [30]	AR [26]	LFWCrop [16]
Ours	**100** $\pm$ **0**	**100** $\pm$ **0**	**95.2** $\pm$ 2.4
DL [37]	99.4 $\pm$ 0.52	98.1 $\pm$ 0.15	89.6 $\pm$ 1.57
TPFRS [35]	98.6 $\pm$ 0.35	97.2 $\pm$ 0.51	86.3 $\pm$ **1.30**
SF [23]	97.3 $\pm$ 1.36	94.2 $\pm$ 2.18	83.5 $\pm$ 2.32
GNN [43]	96.5 $\pm$ 0.58	96.4 $\pm$ 0.33	88.4 $\pm$ 4.80
FF [4]	94.7 $\pm$ 0.67	96.6 $\pm$ 0.42	85.2 $\pm$ 5.50
LBP [1]	93.7 $\pm$ 0.24	92.3 $\pm$ 0.36	83.9 $\pm$ 3.34
PCA [33]	86.6 $\pm$ 0.82	90.6 $\pm$ 0.27	72.1 $\pm$ 5.30
ICA [3]	85.1 $\pm$ 0.59	91.5 $\pm$ 0.48	64.6 $\pm$ 7.90

Table 2 Performance comparison on ORL [30], AR [26] and LFWCrop [16] datasets in terms of implemental efficiency

Method	Efficiency(train(s)/test(s))		
	ORL [30]	AR [26]	LFWCrop [16]
Ours	0.22 ± 0.08/**0.21** $\pm$ **0.03**	$0.27 \pm$ **0.03**/0.29 ± 0.05	**0.21** $\pm$ **0.02**/0.23 ± 0.06
DL [37]	$9.77 \pm 3.75/0.37 \pm 0.12$	$9.55 \pm 3.30/0.34 \pm 0.10$	$9.39 \pm 2.84/0.39 \pm 0.11$
TPFRS [35]	$2.31 \pm 0.17/2.48 \pm 0.22$	$2.74 \pm 0.13/2.87 \pm 0.20$	$2.57 \pm 0.22/2.32 \pm 0.19$
SF [23]	$0.41 \pm 0.03/0.45 \pm 0.05$	$0.45 \pm 0.07/0.43 \pm$ **0.04**	$0.43 \pm 0.09/0.40 \pm 0.07$
GNN [43]	$3.51 \pm 0.97/0.49 \pm 0.06$	$3.27 \pm 0.68/0.49 \pm 0.11$	$3.38 \pm 1.05/0.41 \pm 0.04$
FF [4]	$0.35 \pm 0.04/0.38 \pm 0.04$	$0.40 \pm 0.06/0.35 \pm 0.09$	$0.45 \pm 0.10/0.42 \pm 0.11$
LBP [1]	$0.41 \pm 0.11/0.49 \pm 0.07$	$0.38 \pm 0.13/0.48 \pm 0.10$	$0.33 \pm 0.07/0.35 \pm 0.13$
PCA [33]	**0.19** $\pm$ **0.02**/0.22 ± 0.06	**0.21** ± 0.04/**0.25** ± 0.05	$0.34 \pm 0.10/0.32 \pm$ **0.02**
ICA [3]	$0.37 \pm 0.09/0.39 \pm 0.04$	$0.39 \pm 0.11/0.38 \pm 0.07$	0.21 ± 0.07/**0.18** ± 0.04

the fine-turning in postprocess, the DL model [37] performs slowest among all the baseline approaches, requiring nearly 10 s to train the discriminative model. Especially, it can be seen that our method runs nearly ten times faster than this traditional SRM approach. Notice SF model [23] also employs a hierarchical matching scheme, but our approach still runs 5 times faster than [23].

4.5 Parameter analysis

In our experiment, two factors directly affecting the performance are the portion of training data η and the number of levels $\mathcal{L}$ in DSPGM. We evaluate the recognition accuracy of our method by changing the values of these two parameters. Specifically, we use TPFRS [35], GNN [43], SF [23], and FF [4], as baselines. The selection of these two parameters illustrates the trade-off between model complexity and the recognition precision.

We first evaluate the effect of η by sequentially increasing the number of training data. Figure 7a and b show the plot of recognition accuracy vs. the number of training data on ORL and AR datasets, respectively. Clearly, our method outperforms baseline models since it benefits from the advantages of the proposed DSPGM and ISRM. We also observe that our model achieves 100 % recognition rate on ORL dataset [30] when $\eta = 0.5$, while others do not. In Fig. 7c, we also display the recognizable accuracy along with the increasing

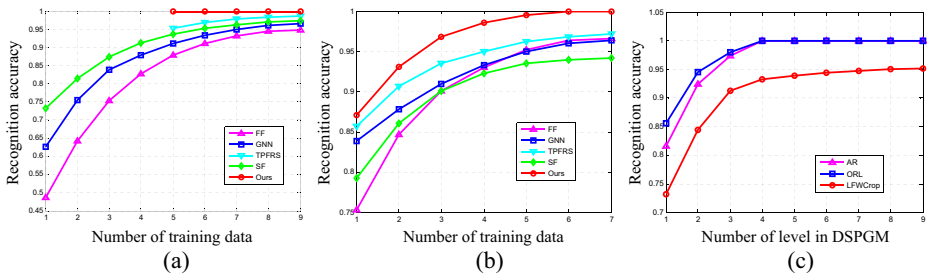


Fig. 7 Effects on performance using different number of training data on ORL (**a**) and AR (**b**) dataset, and different number of level in DSPGM (**c**) (best viewed in color)

number of $\mathcal{L}$, ranging from 1 to 9. The performance of our method peaks when $\mathcal{L} = 4$ for ORL and AR datasets. While in LFWCrop dataset, any refinement to this parameter will result in slightly improvement of performance. To balance the computational efficiency and recognition accuracy, we thus choose $\mathcal{L} = 4$ in our experience.

5 Conclusions and future work

In this paper, we improved SRM for robust face recognition by overcoming its sensitivity to the nonlinear correlation of facial images, such as non-rigid visual variations of pose changes, expressions, and misalignments. Our method first employs the DSPGM to perform image alignment, where the matching consistency is regularized using local and global spatial extents. Thereafter, an ISRM and associated classification algorithm are proposed for face recognition. The experimental results show that our method outperforms the competing models on ORL, AR and LFWCrop datasets in terms of the recognition accuracy and efficiency.

Although our method has achieved promising results, there are two directions that we plan to improve upon in the future. In spite of achieving high efficiency to align facial images, we are still required to align every query sample with all samples in the dataset, which might limit the scalability of our method to large databases. Therefore, one future work includes eliminating the inter-class training samples with different subject category of query, further saving computational time. We are also interested in extending our model to identify facial images in a spatio-temporal domain (e.g., video sequence), under water image classification [22], and medical image processing [25, 41].

Acknowledgments The authors would like to thank the associated editor and all the anonymous reviewers for their valuable comments and suggestions. This work was partly supported by the National Science Foundation (Grant No. IIS-1302164), and the National Natural Science Foundation of China (Grant No. 61401228, 61402122, 61571240, 61501247, 61501259, 61671253), and China Postdoctoral Science Foundation (Grant No. 2015M581841), and Natural Science Foundation of Jiangsu Province (Grant No. BK20160908), and Postdoctoral Science Foundation of Jiangsu Province (Grant No. 1501019A), and the Priority Academic Program Development of Jiangsu Higer Education Institutions(PAPD), and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology(CICAEET), and Nanjing University of Information Science and Technology Research Foundation for Talented Scholars (Grant No. 2015r014).

References

1. Ahonen T, Hadid A, Pietikainen M (2006) Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Anal Mach Intell 28(12):2037–2041
2. Assaleh K et al (2014) Combined features for face recognition in surveillance conditions. In: Proceedings of international conference on neural information processing, pp 503–514
3. Bartlett MS, Movellan JR, Sejnowski TJ (2002) Face recognition by independent component analysis. IEEE Trans Neural Netw 13(6):1450–1464
4. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
5. Bin G, Sheng VS, Li S (2015) Bi-parameter space partition for cost-sensitive SVM. In: Proceedings of the 24th international conference on artificial intelligence, pp 3532–3539
6. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press, Cambridge
7. Bruhn A, Joachim W, Christoph S (2005) Lucas/kanade meets horn/schunck: combining local and global optic flow methods. Int J Comput Vis 61(3):211–231
8. Caesar H, Uijlings J, Ferrari V (2016) Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. IEEE Acess 4(2016):8375–8385

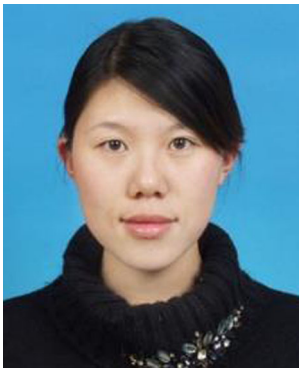
9. Changxing D, Dacheng T (2015) Robust face recognition via multimodal deep face representation. IEEE Trans Multimed 17(11):2049–2058
10. Chen Y, Hao C, Wu W, Wu E (2016) Robust dense reconstruction by range merging based on confidence estimation. SCIENCE CHINA Inf Sci 59(9):1–11
11. Felzenszwalb PF, Huttenlocher DP (2006) Efficient belief propagation for early vision. Int J Comput Vis 70(1):41–54
12. Gu B, Sheng VS (2016) A robust regularization path algorithm for -support vector classification. IEEE Trans Neural Netw Learn Syst. doi:10.1109/TNNLS.2016.2527796
13. Gu B, Sheng VS, Tay KY, Romano W, Li S (2015) Incremental support vector learning for ordinal regression. IEEE Trans Neural Netw Learn Syst 26(7):1403–1416
14. Gu B, Sheng VS, Wang Z, Ho D, Osman S, Li S (2015) Incremental learning for -Support Vector Regression. Neural Netw 67:140–150
15. Gu B, Sun X, Sheng VS (2016) Structural minimax probability machine. IEEE Trans Neural Netw Learn Syst. doi:10.1109/TNNLS.2016.2527796
16. Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. University of Massachusetts, Amherst, Tech. Rep. 7-49
17. Jiang X, Lai J (2015) Sparse and dense hybrid representation via dictionary decomposition for face recognition. IEEE Trans Pattern Anal Mach Intell 37(5):1067–1079
18. Jiang X, Mandal B, Kot A (2008) Eigenfeature regularization and extraction in face recognition. IEEE Trans Pattern Anal Mach Intell 30(3):383–394
19. Kazuhiro F, Osamu Y (2005) Face recognition using multi-viewpoint patterns for robot vision. In: Proceedings of the eleventh international symposium on robotics research, pp 192–201
20. Li SZ, Jain AK (2011) Handbook of face recognition. Springer, Berlin
21. Li SZ, Lu J (1999) Face recognition using the nearest feature line method. IEEE Trans Neural Netw 10(2):439–443
22. Li Y, Lu H, Li J, Li X, Li Y., Serikawa S (2016) Underwater image de-scattering and classification by deep neural network. Comput Electr Eng 2016(54):68–77
23. Liu C, Yuen J, Torralba A (2011) Sift flow: dense correspondence across scenes and its applications. IEEE Trans Pattern Anal Mach Intell 33(5):978–994
24. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
25. Lu HM, Li B, Zhu YJ, Li Y, Xu X, He L, Li X, Li JR, Serikawa S (2016) Wound intensity correction and segmentation with convolutional neural networks. Concurrency and computation: practice and experience
26. Martinez AM (1998) The AR face database. CVC Techique Report
27. Naseem I, Togneri R, Bennamoun M (2010) Linear regression for face recognition. IEEE Trans Pattern Anal Mach Intell 32(11):2106–2112
28. Peng Y, Ganesh A, Wright J, Xu W, Ma Y (2012) Rasl: robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Trans Pattern Anal Mach Intell 34(11):2233–2246
29. Phillips PJ, Moon H, Rizvi S, Rauss PJ (2000) The FERET evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22(10):1090–1104
30. Samaria F, Harter A (1994) Parameterization of a stochastic model for human face identification. In: Proceedings of IEEE workshop on applications of computer vision
31. Shao CB, Song XN, Shu X, Wu XJ (2016) Converted-face identification: using synthesized images to replace original images for recognition. Multimed Tools Appl 75:1–21
32. Shen F, Yang WK, Li H, Zhang H, Shen HT (2016) Robust regression based face recognition with fast outlier removal. Multimed Tools Appl 75:12535–12546
33. Turk M, Pentland A (2010) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
34. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(3):137–154
35. Wagner A, Wright J, Ganesh A, Zhou Z, Mobahi H, Ma Y (2012) Toward a practical face recognition system: Robust alignment and illumination by sparse representation. IEEE Trans Pattern Anal Mach Intell 34(2):372–386
36. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality constrained linear coding for image classification. In: Proceedings of IEEE international conference on computer vision and pattern recognition, pp 3360–3367
37. Wang W, Yang J, Xiao J, Li S, Zhou D (2015) Face recognition based on deep learning. Human Centered Computing 812–820
38. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227

39. Yang M, Zhang L (2010) Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary. In: Proceedings of Europe conference on computer vision, pp 448–461
40. Yang M, Zhang L, Feng X, Zhang D (2014) Sparse representation based Fisher discrimination dictionary learning for image classification. In: Proceedings of IEEE international conference on computer vision, pp 209–232
41. Zhang DY, Wang S, Phillips P, Yang J, Yuan TF (2016) Three-dimensional eigenbrain for the detection of subjects and brain regions related with alzheimer's disease. J Alzheimers Dis 50(4):1163–1179
42. Zhang L, Zhou WD, Li FZ (2015) Kernel sparse representation-based classifier ensemble for face recognition. Multimed Tools Appl 74:123–137
43. Zhao W, Chellappa R, Phillips PJ, Rosenfeld A (2003) Face recognition: a literature survey. ACM Comput Surv 35(4):399–458

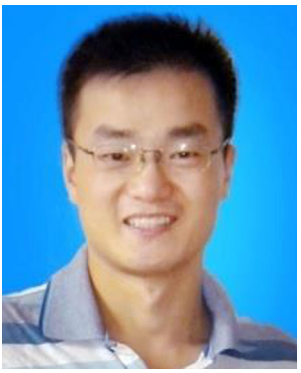
Quan Zhou He received the M.S. degree and Ph.D. degree in communication and information system in 2006 and 2013, respectively, from Huazhong University of Science and Technology, China. He is now an associate professor of Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition. He has published over 20 research papers in SCI journals (e.g., IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, Pattern Recognition) and conference (ICIP, ICASSP, ACCV, and ICPR) in image processing and computer vision. He now serves as TPC member or chair of many international conferences and reviewer for a series of SCI journals, including IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Circuits System for Video Technology, Pattern Recognition, and Neurocomputing. He is member of IEEE.

Cheng Zhang He received BS and M.S. degree in electronic and electricity engineering in 2005 and 2008, respectively, from Huazhong University of Science and Technology, China. He is now the engineering centre director of the state grid HuBei information and telecommunication company, China. His research interests include signal processing and applications in smart grid.

Wenbin Yu He received Ph.D. degree in the field of quantum information processing. He is a lecturer in school of computer science and software, Nanjing University of Information Science and Technology, China. His research interests focus on quantum information processing, signal and information processing. He is a leading researcher of National Science Foundation of China, Youth Program.



Yawen Fan She received BE and MS degrees in electronic engineering from Hohai University, Nanjing, China in 2003 and 2006, respectively. She has received the doctor degree in EE department from Shanghai Jiao Tong University, Shanghai, China. She is now an assistant professor at Nanjing University of Posts and Telecommunications, Nanjing, P. R. China. Her research interests are intelligent video surveillance and video analysis and understanding.

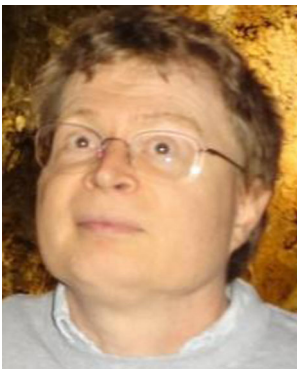


Hu Zhu He received his B.S. degree in mathematics and applied mathematics from Huaibei Coal Industry Teachers College, Huaibei, China, in 2007, and received his M.S. and Ph.D degree in computational mathematics and pattern recognition and intelligent systems respectively from Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2013. In 2013, he joined the Nanjing University of Posts and Telecommunications, Nanjing. His research interests are pattern recognition, image processing and spectral data processing.

Xiaofu Wu He received the B.S. and M.S. degrees in electrical engineering from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1996 and 1999, respectively, and Ph. D. degree in electrical engineering from Peking University, Beijing, China, in 2005. From 2005 to 2007, he has with the Southeast University as a Post-Doctoral reseacher at the National Mobile Communication Research Laboratory. Since 2012, he has been with the Nanjing University of Posts and Telecommunications, where he is currently a full Professor. His research interests are in coding and information theory, information-theoretic security, machine learning and computer vision.

Weihua Ou He received the M.S. degree in Mathematics from the Southeast University, Nanjing, China in 2006 and the Ph.D. degree in Information and Communication Engineering from Huazhong University of Science and Technology (HUST), China in 2014, respectively. Currently, he is an Associate Professor at the School of Big data and Computer Science in Guizhou Normal University, Guiyang, China. His current research interests include sparse representation, multi-view learning, and image processing and computer vision.

Weiping Zhu He received the B.E. and M.E. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1982 and 1985, respectively, and the Ph.D. degree from Southeast University, Nanjing, in 1991, all in electrical engineering. From 1991 to 1992 and from 1996 to 1998, he was a Post-Doctoral Fellow and a Research Associate, respectively, with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada. From 1993 to 1996, he was an Associate Professor with the Department of Information Engineering, Nanjing University of Posts and Telecommunications. From 1998 to 2001, he was with high-tech companies in Ottawa, ON, Canada, including Nortel Networks and SR Telecom Inc. Since 2001, he has been a full-time Faculty Member with the Department of Electrical and Computer Engineering, Concordia University, where he is currently a Full Professor. Since 2008, he has been also an Adjunct Professor with the Nanjing University of Posts and Telecommunications. His research interests include digital signal processing fundamentals, speech and audio processing. He served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–I and–II. He was also a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He was the Secretary of Digital Signal Processing Technical Committee (DSPTC) of the IEEE Circuits and System Society during June 2012-May 2014, and the Chair of the DSPTC during June 2014–May 2016.



Longin Jan Latecki He received the PhD degree in computer science from Hamburg University, Germany, in 1992. He is a professor of computer science at Temple University, Philadelphia. His main research interests include shape representation and similarity, object detection and recognition in images, robot perception, machine learning, and digital geometry. He has published over 230 research papers and books. He is an editorial board member of Pattern Recognition, Computer Vision and Image Understanding and the International Journal of Mathematical Imaging. He received the annual Pattern Recognition Society Award, together with Azriel Rosenfeld, for the best article published in the journal Pattern Recognition in 1998. He is the recipient of the 2000 Olympus Prize, the main annual award from the German Society for Pattern Recognition (DAGM). He is a senior member of the IEEE.