Contents lists available at ScienceDirect

Pattern Recognition

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

Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person

Meng Yang[a,b,c,*], Xing Wang[b], Guohang Zeng[b], Linlin Shen[b]

[a] *School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China*
[b] *College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China*
[c] *Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China*

ARTICLE INFO

Keywords:
Joint and collaborative representation
Adaptive convolution feature
Sparse representation
Single sample face recognition

ABSTRACT

With the aid of a universal facial variation dictionary, sparse representation based classifier (SRC) has been naturally extended for face recognition (FR) with single sample per person (SSPP) and achieved promising performance. However, extracting discriminative facial features and building powerful representation framework for classifying query face images are still the bottlenecks of improving the performance of FR with SSPP. In this paper, by densely sampling and sparsely detecting facial points, we extract complete and robust local regions and learn convolution features adaptive to the local regions and discriminative to the face identity by using convolutional neural networks (CNN). With this powerful facial description and a generic face dataset with common facial variations, a joint and collaborative representation framework, which performs representation for each local region of the query face image while requires all regions of the query face image to have similar representation coefficients, is presented to exploit the distinctiveness and commonality of different local regions. In the proposed joint and collaborative representation with local adaptive convolution feature (JCR-ACF), both discriminative local facial features that are robust to various facial variations and powerful representation dictionaries of facial variations that can overcome the small-sample-size problem are fully exploited. JCR-ACF has been extensively evaluated on several popular databases including AR, CMU Multi-PIE, LFW and the large-scale CASIA-WebFace databases. Experimental results demonstrate the much higher robustness and effectiveness of JCR-ACF to complex facial variations compared to the state-of-the-art methods.

1. Introduction

With the wide installment of various cameras, such as mobile cameras and CCTV cameras, face images can be captured more conveniently. Face recognition (FR) has always been a hot topic and drawing ever-increasing attention in the community [1] due to its promising applications in information security, access control, surveillance, smart cards, law enforcement, human computer interaction, and entertainment [1,43,44]. For FR in uncontrolled or less controlled scenarios [2,3] many problems remain to be solved. FR with single sample per person (SSPP) is one of the most important topics in the field of ubiquitous biometrics. In the application of face recognition, there is usually very limited training samples (e.g., only a single training sample) in the gallery set. For example, we often meet scenarios (e.g., law enforcement, e-passport, driver license, etc.) where only a single training face image is available for each person. FR with SSPP is very challenging because very limited information can be provided in the single-sample gallery set, leading to the failure of the gallery set to predict various facial variations existing in the query samples. Hence how to overcome the small-sample-size problem and ensure a high performance for FR with SSPP is still open.

The unavailability of sufficient training samples per person can affect the FR performance adversely [4]. Many popular methods requiring multiple training samples per person cannot be directly applied to FR with SSPP, such as discriminative subspace and manifold learning methods (e.g., LDA and its variants [5]), and sparse representation based classification (SRC) method [25]. In addition, most of the classification models based on deep learning [20–22] cannot be directly used to approach the under-sampled FR problem either because they usually require a substantial number of training samples for each class to train the deep classification networks. To handle the under-sampled problem of FR with SSPP many methods are designed specially [4]. And according to the presence of an additional generic training set, these methods are classified into two categories: one needing the generic training set and the other without using the generic training set.

[*] Corresponding author.
E-mail address: yangmengpolyu@gmail.com (M. Yang).

http://dx.doi.org/10.1016/j.patcog.2016.12.028
Received 15 July 2016; Received in revised form 12 November 2016; Accepted 28 December 2016

The methods without using the generic training set usually extract robust local features (e.g., local binary pattern [6] and gradient orientation [7]), perform image partitioning (e.g., multi-manifold learning from local patches [8,34,35], self-organizing maps of local patches [9], local patch based LDA [10] and local structure based multi-phase collaborative representation [36]) or generate additional virtual training samples (e.g., via singular value decomposition [11], geometric transform and photometric changes [12] and lower-upper decomposition [37]). In methods extracting local features, for a given face image local features robust to intra-class variance and discriminative to distinguish among different classes are extracted and concatenated to represent the face image. Then FR is performed by finding the identity of the single gallery sample with the most similar feature to the query face image. In methods performing image partitioning, methods based on multi-manifold learning [8,34,35] view the face image as a manifold and each patch of the face image as a point in the manifold. Then FR can be performed by matching the manifolds. [9] learns the subspace of each subject by using the unsupervised and non-parametric self-organizing maps (SOM). [10] divides the gallery image of each subject into patches and calculates the within-scatter of each subject based on its patches. Then LDA is used to obtain the discriminative feature for each patch. In all previous image partition based methods the patches of the face image are non-overlapping, however, [36] proposed a local structure based multi-phase collaborative representation method to explore the local structure relationship of overlapping patches. For methods generating virtual samples, the virtual training samples of each subject are generated based on the single gallery image of the subject so that the conventional techniques requiring multiple training samples per subject can be applied for FR with SSPP. These methods that don't need the generic training set indeed improve the performance of FR with SSPP to some extent but they neglect to bring additional facial variation information to the gallery set with SSPP. Furthermore, actually very limited new information can be introduced by virtual samples generation because of the high correlation between the virtual samples and their original single gallery sample.

In contrast to the first category of methods, methods requiring an extra generic training set assume that the intra-class facial variations can be shared by different subjects and borrow useful information (e.g. generic intra-class facial variations) from the extra generic set. Due to the easy collection of the generic set, the idea of borrowing useful information from the generic set has been widely used [3,13—16,38–40]. For instance, the expression subspace [15] and pose-invariant subspace [13] were learned from an auxiliary generic training set to acquire the expression-invariant and pose-invariant attributes, respectively. Hu *et al.* [38] proposed to transfer the intra-class variation of a generic training set containing multiple samples per subject to that of the gallery set with single sample per subject. In [39] an equidistant prototypes embedding is proposed for FR with SSPP. This approach intends to learn a linear regression so that the gallery faces and the intra-class face differences in the generic set can be mapped to the equally distant locations and the zero vectors, respectively. Sparse representation based classification (SRC) [25] is proposed for FR and achieves promising performance but it requires sufficient enough samples for each gallery subject. To deal with FR with SSPP, Deng *et al.* [16] extended SRC to FR with SSPP. The Extended SRC (ESRC) extracts the intra-class facial variations from the generic training set and uses the obtained intra-class facial variation matrix together with the gallery dictionary to encode the query samples. Motivated by ESRC, Zhu *et al.* [3] proposed a local generic representation (LGR) based method, which divides the face image into local patches and uses same strategy to build the intra-class variation dictionary for each patch. Based on ESRC, Ding *et al.* [40] presented a variational feature representation-based classification approach where the variational feature is represented by the joint information of the generic set and the gallery set, with a normal feature reserving the identity information obtained more precisely. Different from the above 2D FR scenarios [46] proposed a two-phase weighted collaborative representation classification (TPWCRC) to handle 3D partial FR with SSPP by representing the facial scan with a set of local keypoint-based multiple triangle statistics.

Dictionary learning for patter classification has been extensively studied in the community of image processing and computer vision [41,42]. However, many designed dictionary leaning methods require each class to have multiple training samples. In recent years dictionary learning has been also studied in FR with SSPP. Different from those requiring multiple training samples for each class of interest, the dictionary learning methods [17–19] learn a common facial variation dictionary from an extra generic set to compensate the poor representation ability of single-sample gallery set. Specifically, Yang *et al.* [17] proposed the sparse variation dictionary learning (SVDL) method to learn the intra-class variation dictionary adaptive to the gallery set by jointly learning a sparse variation dictionary and a projection, which is used to project the intra-class variation dictionary learnt from the generic set to the space of the gallery set; Zhuang *et al.* [18] proposed to learn an illumination variation dictionary from the generic set to address image misalignment and pixel corruption; Gao *et al.* [19] introduced a regularized patch-based representation approach (RPR), in which the face image is represented by a collection of overlapping patches and patch based intra-class variation dictionary is learned.

Recently deep learning has been used for face verification and achieved great success [20–22,45]. [20] firstly learns the high-level visual features, termed Deep hidden IDentity features (DeepID), through the deep convolutional neural networks and then uses the joint Bayesian mode [23] to perform the face verification. [21] learns a highly deep face representation and achieves he human-level accuracy. Based on [20], [22] adds the verification supervisory signal in addition to the identification signal to train the convolutional neural networks. [45] uses four million facial images belonging to more than 4000 identities to train a deep network with locally connected layers and closely approaches the human-level performance in face verification. Although exciting performance for face verification has been reported, however, how to effectively apply deep learning based methods to the serious under-sample classification problem, i.e., FR with SSPP, is still an open question.

Indeed, much progress has been made on FR with SSPP, but several issues with previous works are still worth noting. Firstly, Both ESRC [16] and SVDL [17] use the features extracted from the holistic face images, which are susceptible to large facial variations (e.g. caused by corrupted pixels) in local regions of the face image. If some local region is corrupted then the holistic feature will become problematic. Secondly, although LGR [3] uses local features to obtain high robustness, it ignores to use powerful facial feature and exploit the discrimination of particular facial regions (e.g., landmarks such as eyes and nose), which are the most informative part of the face from the viewpoint of human visual perception and prove to have high detection rates in various facial variations [24]. Additionally, LGR [3] encodes each local region independently without exploiting the prior knowledge that the local regions are from the same face image. Thirdly, all these methods use the intensity features which are not discriminative enough. Overall the current sparse representation based methods still perform poorly when dealing with complex facial variations (e.g. variations in expression, occlusion, and pose etc).

Motivated by the success of local features [3] and the deep convolution features [20] in face verification we propose to extract the local adaptive convolution features (ACFs) from the local regions of the face image. Specifically, we extract the deep convolution features adaptive to each particular (e.g., eyes and nose) and regular local facial regions (e.g., dense sampling points). The particular regions cover the most informative parts of the face, while the regular regions are used to represent the face completely. To fully exploit the distinctiveness and commonality of different local regions we introduce the joint and collaborative representation (JCR) framework, which jointly and collaboratively represent all adaptive convolution features with requiring them to have similar representation coefficients. We term the proposed method as joint and collaborative representation with adaptive convolution feature (JCR-ACF). The experimental results on four popular databases, including AR, CMU Multi-PIE, LFW and the large-scale CASIA-WebFace databases, validate the higher robustness of JCR-ACF in the various facial variations (e.g. expression, pose and occlusion etc.) than the existing methods.

Our contributions are summarized as follows.

1. We explore how to learn adaptive deep convolutional features in the application of face recognition with SSPP. Although exciting performance for deep learning based face verification has been reported, however, how to effectively apply deep learning based methods, e.g., convolutional neural network, to the serious under-sample classification problem, i.e., FR with SSPP, is still an open question.
2. We require the representation coefficients of ACFs of different local regions to be similar because these local regions come from the same query image and propose a joint collaborative representation model to effectively fuse the local deep feature representations in different locations. Conventional collaborative representation works usually use holistic features or handcraft features without joint representation.
3. We fully exploit the powerful deep feature and collaborative representation based classifier for FR with SSPP for the first time. Higher accuracy and robustness of the proposed JCR-ACF are achieved.

The rest of this paper is organized as follows. Section 2 presents a brief review of the related works. Section 3 gives the proposed JCR-ACF method. Section 4 describes the optimization of JCR-ACF. Section 5 conducts the experiments and Section 6 concludes the paper.

2. Brief review of related works

2.1. Extended sparse representation based classifier (ESRC)

Sparse representation based classification (SRC) [25] has achieved very great success in FR with multiple training samples per person. And following SRC many works have been proposed [16,26,32]. However, SRC cannot be directly applied to FR with SSPP because it requires each person to have sufficient training samples. To deal with FR with SSPP, Deng *et al.* [16] extracted an intra-class variation matrix from an auxiliary generic training set to compensate the poor representation ability of the gallery set with SSPP.

Denote the intra-class variation matrix extracted from the generic training set by $\boldsymbol{V} = [\boldsymbol{V_1}, \boldsymbol{V_2}, ..., \boldsymbol{V_n}]$, where $\boldsymbol{V_i}$ corresponds to the i^{th} generic subject variation matrix, and each column of $\boldsymbol{V_i}$ is obtained by subtracting the i^{th} generic subject's reference image (e.g., the mean face image or the natural face image without facial variations) from other images of the same subject. Let $\boldsymbol{G} = [\boldsymbol{g_1}, \boldsymbol{g_2}, ..., \boldsymbol{g_c}]$ and $\boldsymbol{y}$ denote the gallery set with SSPP and the testing sample, respectively. The procedures of ESRC [16] are described as follows.

1. Sparsely code $\boldsymbol{y}$ on the matrix $[\boldsymbol{G}\ \boldsymbol{V}]$ via l_1-norm minimization:

$$[\hat{\boldsymbol{\rho}};\hat{\boldsymbol{\beta}}] = \arg\min_{\boldsymbol{\rho},\boldsymbol{\beta}} \|\boldsymbol{y} - [\boldsymbol{G}\ \ \boldsymbol{V}][\boldsymbol{\rho};\boldsymbol{\beta}]\|_2^2 + \lambda\|[\boldsymbol{\rho};\boldsymbol{\beta}]\|_1 \tag{1}$$

where $\lambda > 0$ is the regularization parameter, $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ are the coding coefficient vectors associated with $\boldsymbol{G}$ and $\boldsymbol{V}$, respectively.

2. Classify $\boldsymbol{y}$ via

$$\text{identity}(\boldsymbol{y}) = \arg\min_i \|\boldsymbol{y} - \boldsymbol{g}_i\hat{\rho}_i - V\hat{\boldsymbol{\beta}}\|_2 \tag{2}$$

where $\hat{\boldsymbol{\rho}} = [\hat{\rho}_1; \hat{\rho}_2; \cdots; \hat{\rho}_c]$, and $\hat{\rho}_i$ is the coding coefficient associated with class *i*.

Sharing the same spirit as ESRC that the intra-class variation information of the generic set can be used to improve the representation ability of the gallery set, Zhu *et al.* [3] proposed a local generic representation (LGR) based method to take the advantages of both local representation and generic learning. In LGR the face image is partitioned into several overlapping patches regularly and then each patch is coded independently. Finally the classification is performed by checking which class can lead to the minimal construction residuals over all local patches. Additionally, similar to ESRC, LGR constructs the intra-class variation dictionary for each patch by subtracting reference sample from other variation samples of the same class in the generic set.

ESRC uses the holistic feature of the face image. But it is evident that the holistic feature is vulnerable to bad local regions which are impaired by complex facial variations or outliers. LGR uses the local features and can boost the robustness. But there are still several issues worth noting. Firstly, LGR uses the local intensity feature which may not be powerful enough. Secondly, the particular facial regions such as eyes and nose etc., which are the most informative parts of the face and can be robustly detected under various facial variations [24], have not been fully exploited yet. Thirdly, each local facial region is coded independently so the prior knowledge that the local regions come from the same face image isn't utilized.

2.2. Deep hidden IDentity features (DeepID)

Recently, deep learning has been used to learn a deep representation feature for the face image and achieved very great success for face verification [20]. [20] proposes to learn the high-level deep features termed Deep hidden IDentity features (DeepID) through multi-class face identification tasks by using the deep convolutional neural networks. Although the DeepID features are learnt by performing classification, they can be well generalized to other tasks, such as verification, and new identities unseen in the training set. When the deep convolutional neural networks have been trained, images of new identities can be fed to the networks to obtain the DeepID features of the new identities, and then the joint Bayesian mode in [23] is used to perform face verification.

Since deep learning model has a quite huge number of unknown parameters, it requires a large number of subjects and sufficient samples per subject to train the deep convolutional neural networks [20]. So the deep leaning based method cannot be directly applied to FR with SSPP because each gallery class only has a single training sample. Nevertheless, it is worth noting that such learnt high-level features have the maximum discrimination because they are learnt by performing classification.

3. Joint and collaborative representation with local adaptive convolution feature

Inspired by [20] we want to learn the powerful high-level feature for each local region of the face image so that the maximal discrimination for each local facial region can be maintained. Consequently, we propose the joint and collaborative representation with local adaptive convolution feature (JCR-ACF) model, which can adaptively extract discriminative features from local regions and jointly represent each local region of the query image. Specifically, in JCR-ACF, the local adaptive convolution features (ACFs) of both particular (e.g., eyes and nose) and regular facial regions (e.g., dense sampling points) are extracted and a joint and collaborative representation (JCR) framework is proposed for the representation and classification of the query image.

3.1. Local adaptive convolution feature extraction

As illustrated in Fig. 1, for a face image we use the algorithm in [24] to detect 5 particular facial points (i.e., two eye centers, nose tip, and

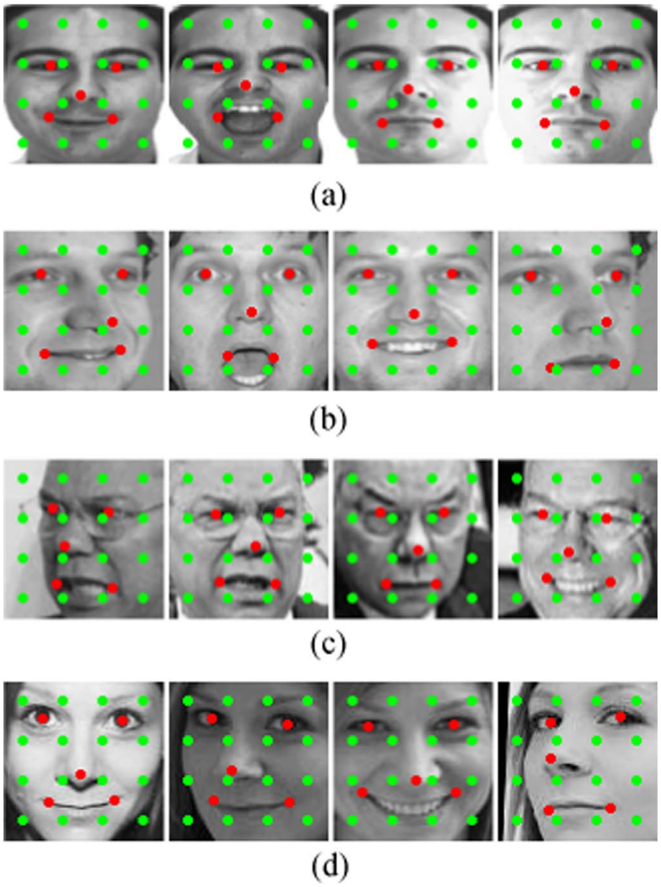

Fig. 1. (a) Facial points on images of the AR database; (b) Facial points on images of the CMU Multi-PIE database; (c) Facial points on images of the LFW database; (d) Facial points on images of the CASIA-WebFace database. The 5 detected particular facial are shown in red and the sampled regular facial points are shown in green. We can see that the detection of the 5 particular points is very robust to various facial variations.

two mouth corners) and sample the other facial points regularly in the face image. In Fig. 1, we can see that in various cases including controlled and uncontrolled environments, the detected particular facial points are very robust to various facial variations. We extract the patch centered around each of the facial points as local regions of the face image. Because of the strong discrimination of the particular facial regions and the availability of robust detection of these particular facial points in various facial variations [24], FR will benefit from the use of these particular local regions. Apart from the robust local particular regions, the dense regular local regions can provide a complete representation for face images to cover the whole face.

Although deep convolution feature [20] has shown very good performance in face verification, it is still unknown that how to effectively learn deep feature for the challenging FR with SSPP. Since different local regions have contained distinctive discrimination for face recognition, we propose to learn a set of convolution filters for each of local regions independently based on a large scale extra face dataset so that the learned filters are specific to every local region. We use the convolutional neural network in [20] but change its input size (the sizes of following layers are changed accordingly).

The flowchart of learning the ACF network from the extra face dataset and obtaining the ACF of a local region of a face image is shown in Fig. 2. Based on the extra face dataset, we learn a ACF deep network for every local region of the face image. Thanks to the powerful discrimination and representation of deep learning, the feature learning is performed through multi-class face identification for each local region independently so the learned deep convolution feature can be adaptive to each local region and contain the most discrimination embedded in the region. When the ACF deep network for each local region is learned, the input (i.e., the 160-dimensional high-level facial feature) to the final soft-max classification layer is defined as the adaptive convolution feature (ACF) of each local region.

Denote by Φ_k the learned deep convolution neural network on the k^{th} local region. Given a face image $\boldsymbol{y}$, the learned adaptive convolution feature (ACF) in k^{th} local region can be represented as $\boldsymbol{y}_k$=$\Phi_k(\boldsymbol{y})$.

3.2. Joint and collaborative representation

Although the proposed adaptive convolution feature (ACF) is discriminative in face identity due to the powerful deep learning model, the small-sample-size problem (e.g., there is only a single gallery sample) is still not solved. In order to design a discriminative classifier to well handle the small-sample-size problem, in this section we introduce a joint and collaborative representation (JCR) model to effectively exploit the local adaptive convolution features (ACFs). We denote $\boldsymbol{y}$ =[$\boldsymbol{y}_1$, $\boldsymbol{y}_2$,...,$\boldsymbol{y}_K$] the testing sample, where $\boldsymbol{y}_k$ is the local adaptive convolution feature extracted from the k^{th} local region of $\boldsymbol{y}$. Similarly, the variation matrix of a generic training set could also have $\boldsymbol{K}$ local regions, and each local region could have an intra-class variation dictionary $\boldsymbol{D}_k$. And for each local region we can also construct the gallery dictionary $\boldsymbol{G}_k$. For constructing $\boldsymbol{D}_k$, we first obtain the ACFs of all face images of the generic training set, then use the difference-from-reference way to compute the difference of k^{th} ACFs for each generic subject, i.e. by subtracting the ACF of the reference sample from the ACFs of the variation samples of the same subject; and finally concatenate the k^{th} difference matrices of all generic subjects as $\boldsymbol{D}_k$.

In the joint and collaborative representation model, we ask the representation coefficients of ACFs of different local regions to be similar because these local regions come from the same query image and their joint work benefits the final classification. The proposed joint and collaborative representation (JCR) model can be expressed as

$$\min_{\boldsymbol{\alpha}_k} \sum_{k=1}^{K} (\|\boldsymbol{y}_k - [\boldsymbol{G}_k \ \ \boldsymbol{D}_k]\boldsymbol{\alpha}_k\|_2^2 + \lambda\|\boldsymbol{\alpha}_k\|_2^2 + \mu\|\boldsymbol{\alpha}_k - \overline{\boldsymbol{\alpha}}\|_2^2) \tag{3}$$

where $\boldsymbol{\alpha}_k = [\boldsymbol{\rho}_k; \boldsymbol{\beta}_k]$ is the coding coefficient vector of the k^{th} local region of the query sample $\boldsymbol{y}$, $\boldsymbol{\rho}_k$ is the coding sub-coefficient vector associated with $\boldsymbol{G}_k$, and $\boldsymbol{\beta}_k$ is the coding sub-coefficient vector associated with $\boldsymbol{D}_k$. $\overline{\boldsymbol{\alpha}}$ is the mean vector of all $\boldsymbol{\alpha}_k$ (k=1,2,...,K). We use l_2-norm to regularize the representation coefficients, which is motivated by [31]. Because [31] demonstrates that the l_2-norm is comparable to l_1-norm in FR accuracy but is more efficient to solve.

Fig. 3 shows the flowchart of JCR-ACF. The left part of Fig. 3 illustrates the extraction of local adaptive convolution features. The right part shows the classifier based on joint and collaborative representation. From Eq. (3) and Fig. 3, it can be observed that the proposed JCR-ACF model has several advantages to make it powerful for FR with SSPP. First, the powerful deep convolutional feature robust to various face variations and discriminative in face identity is utilized to describe face images; Second, the strategy of divide and conquer, which is effective to FR, is introduced to JCR-ACF (e.g., particular and regular local regions of face images are firstly divided, and then they cooperate with each other based on the last within-class scatter term of Eq. (3)); Last but not the least, an intra-class variation dictionary generated from the generic training set in the model of JCR-ACF can well aid the single-sample gallery to represent the query face image with various variations.

After solving JCR-ACF, the classification is conducted via

$$\text{identity} = \arg\min_i \left\{ \sum_{k=1}^{K} \|\boldsymbol{y}_k - \boldsymbol{g}_k^i \rho_k^i - \boldsymbol{D}_k \boldsymbol{\beta}_k\|_2^2 / \|\rho_k^i; \boldsymbol{\beta}_k\|_2^2 \right\} \tag{4}$$

$\boldsymbol{g}_k^i$ is the ACF of the k^{th} local region of the single gallery sample of class i. $\boldsymbol{\rho}_k = [\rho_k^1; \rho_k^2; \ldots ;\rho_k^c]$ and ρ_k^i is the coding coefficient associated with class i.

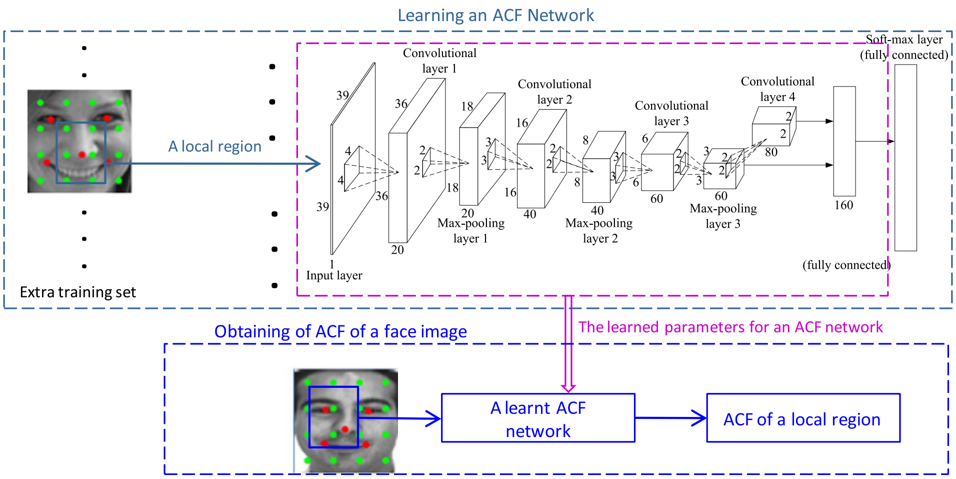

Fig. 2. Learning the ACF network (the first row) and generating the ACF (the second row) for a local region (the convolutional neural network is from [20], we just change its input size). The cuboids denote the input, convolutional, and max pooling layers. The length, width, and height of each cuboid (except the input layer) is the number of maps and the size of each map. The small square in each cuboid denotes the local receptive field. The second last fully connected layer contains the ACF, whose dimensionality is 160.

4. Optimization of JCR-ACF

In this section we tell how to solve the optimization problem of JCR-ACF in Eq. (3). We firstly present the solving algorithm of computing the coding coefficients in Eq. (3) and then analyze the time complexity of the proposed algorithm.

4.1. Solving algorithm of JCR-ACF

The coding coefficients of different local regions are correlated to each other due to the last within-class scatter term of Eq. (3). Thus the optimization problem in Eq. (3) can be solved as follows: the representation coefficient vector $\boldsymbol{\alpha}_k$ can be derived as

$$\boldsymbol{\alpha}_k = \boldsymbol{\alpha}_{k,0} + \mu \boldsymbol{P}_k \bar{\boldsymbol{\alpha}} \tag{5}$$

where

$$\boldsymbol{\alpha}_{k,0} = \boldsymbol{P}_k [\boldsymbol{G}_k \ \ \boldsymbol{D}_k]^T \boldsymbol{y}_k \tag{6}$$

and

$$\boldsymbol{P}_k = ([\boldsymbol{G}_k \ \ \boldsymbol{D}_k]^T [\boldsymbol{G}_k \ \ \boldsymbol{D}_k] + (\lambda + \mu)\boldsymbol{I})^{-1} \tag{7}$$

Due to $K\bar{\boldsymbol{\alpha}} = \sum_{k=1}^{K} \boldsymbol{\alpha}_k$ and by summing $\boldsymbol{\alpha}_k$ we can get

$$K\bar{\boldsymbol{\alpha}} = \sum_{k=1}^{K} \boldsymbol{\alpha}_k = \sum_{k=1}^{K} \boldsymbol{\alpha}_{k,0} + \mu \sum_{k=1}^{K} \boldsymbol{P}_k \bar{\boldsymbol{\alpha}} \tag{8}$$

and we can derive

$$\bar{\boldsymbol{\alpha}} = \left(K\boldsymbol{I} - \mu \sum_{k=1}^{K} \boldsymbol{P}_k\right)^{-1} \sum_{k=1}^{K} \boldsymbol{\alpha}_{k,0} \tag{9}$$

Based on Eq. (6) and Eq. (9), the closed-form solution for $\boldsymbol{\alpha}_k$ can be got by Eq. (5). Note that $\boldsymbol{P}_k$ in Eq. (7) and $(K\boldsymbol{I} - \mu \sum_{k=1}^{K} \boldsymbol{P}_k)^{-1}$ in Eq. (9) can be pre-computed before the testing stage because they do not involve the testing sample. So in the testing stage only some matrix multiplication operations are involved, which makes JCR-ACF very efficient. The algorithm of JCR-ACF is summarized in Table 1.

4.2. Computational complexity

We denote $[\boldsymbol{G}_k \ \ \boldsymbol{D}_k] \in \mathbb{R}^{d \times n}$ where d is the dimension of ACF and n is number of atoms in $[\boldsymbol{G}_k \ \ \boldsymbol{D}_k]$. The computational cost of JCR-ACF is mainly spent on the least squares problem in Eq. (3). According to the solving algorithm of JCR-ACF the computation cost of solving Eq. (3) lies on Eq. (5). $\boldsymbol{P}_k$ doesn't involve the testing sample thus can be pre-computed offline. After computing $\boldsymbol{P}_k$, the term $(K\boldsymbol{I} - \mu \sum_{k=1}^{K} \boldsymbol{P}_k)^{-1}$ in Eq. (9) can also be pre-computed offline. Thus $\bar{\boldsymbol{\alpha}}$ can be calculated very efficiently online. Then the calculation of $\boldsymbol{\alpha}_k$ for k=1, 2, … K only involve the matrix multiplication operations with the time complexity of $KO(nd+3n^2)$).

5. Experiments

In this section, we firstly train the deep convolutional networks using the CASIA-WebFace dataset [30] (The details of this dataset will be given in Section 5.6). We then perform FR with SSPP on four benchmark face databases, including the AR database [27], the CMU Multi-PIE database [28], the Labeled Faces in the Wild (LFW) database [29] and the large-scale CASIA-WebFace dataset [30] to evaluate the performance of JCR-ACF. For FR experiments on the AR, CMU Multi-PIE, and LFW databases we use all the subjects in the CASIA-WebFace database to train the deep convolutional networks. And for the FR experiment on the CASIA-WebFace database we choose a subset of this

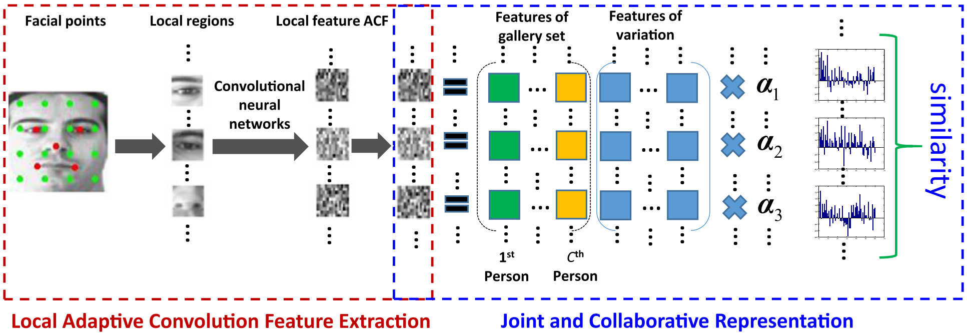

Fig. 3. The flowchart of the proposed JCR-ACF, which includes local adaptive convolution feature (ACF) extraction and joint and collaborative representation (JCR).

Table 1
The algorithm of JCR-ACF.

```
Input: The query sample $\boldsymbol{y}$, the gallery set and the generic set
Output: The class label of the query sample $\boldsymbol{y}$
1. Extract the $K$ ACFs for each sample of the gallery set and the generic set
2. Construct the gallery dictionary $\boldsymbol{G}_k$ and the intra-class variation dictionary $\boldsymbol{D}_k$ for
   each local region $k$
3. Extract the $K$ ACFs for the query sample $\boldsymbol{y}$, which are denoted as $\boldsymbol{y}_k$ ($k$=1, 2,..., $K$).
4. Calculate the representation coefficient vector $\boldsymbol{\alpha}_k$ for each ACF $\boldsymbol{y}_k$ via Eq. (5)
5. Calculate the class label of the query sample by via.(4)
```

dataset to train the deep convolutional networks and use another subset to evaluate the FR performance (Details about the partition of the dataset can be found in Section 5.6).

We first discuss the experimental setting in Section 5.1, in Section 5.2 we investigate the discrimination of the particular facial regions. In Section 5.3 we test the proposed JCR-ACF on the AR database, in Section 5.4 we test JCR-ACF on the CMU Multi-PIE database. In Section 5.5 and Section 5.6 we test JCR-ACF on the LFW database and the CASIA-WebFace database, respectively. We compare JCR-ACF with several state-of-the-art methods on FR with SSPP, including Regularized Patch-based Representation (RPR) [19], Local Generic Representation (LGR) [3], Extended SRC (ESRC) [16], Adaptive Generic Learning (AGL) for Fisher faces [33], Discriminative Multi-Manifold Analysis (DMMA) [8], Sparse Variation Dictionary Learning (SVDL) [17], and two baseline methods such as SRC [25] and Support Vector Machine (SVM). Among all these methods, SVM, SRC and DMMA do not use the generic training set, while AGL, ESRC, SVDL, RPR, LGR and the proposed JCR-ACF need the extra generic training set. In addition, following [19] we only conduct experiments for SVDL on the AR and CMU Multi-PIE databases because SVDL requires all subjects in the generic set to have images for a given type of variation. This requirement cannot be fulfilled in the LFW and CASIA-WebFace databases.

5.1. Experimental setting

For all face images, we align the face image to a face template in which two eyes are horizontal. The images of AR database published in [27] have already been cropped and normalized based on the locations of facial landmarks. We only resize them from 165×120 to 80×80. For the CMU Multi-PIE databases we manually detect the center of eyes; then align all images to make them have the same locations of eyes; and finally crop the images and resized the sizes of them to 80×80. LFW-a [29] is its aligned version by using a commercial face alignment software. However, in the dataset of LFW-a the facial variations are totally unconstrained, so we align the face images by mapping 5 facial points (e.g., two eye centers, nose tip, and two mouth corners detected by the algorithm [24]) to the facial points in the face template (two eyes in the template are horizontal) via similarity transformation. Finally the face image in LFW-a is cropped and resized to 80×80. For the CASIA-WebFace database [30], we also align the face image by using the 5 facial points provided by the database itself to map the face image to the face template via similarity transformation.

Besides the 5 detected particular points we sample 16 additional facial points regularly on the face image. Then there are total 21 facial points in the face image, as illustrated in Fig. 1. The used deep convolutional neural network is based on [20] and we train the deep ACF networks using different local regions of face images. The specific structure of the network is shown in Fig. 2. We train 21 independent convolutional networks of which one corresponds to one local region.

The size of facial patches depends on the training of deep ACF networks and local representation. For training the deep ACF network for each patch, the size of facial patches should be big enough (i.e., enough discrimination should be contained in the patch) to make the training of deep ACF network converge. For local representation, facial patches should be small to avoid global information and preserve local information. To make a balance between the training of deep ACF networks and local representation, the size of facial patches is set as 39×39. It is obvious that there are overlaps between different facial patches. The procedure of selecting these facial patches is described as follows. We first detect the regular and particular facial points; and then with these facial points as the centers, facial patches with the size of 39×39 are extracted. In the training phase, the size of face images is set as 120×120 to make sure that all facial patches can be extracted. The images used for training the convolutional networks are from the CASIA-WebFace dataset. In total we can get 21 local ACFs for a face image. In order to make the representation of testing sample as local as possible, if no specific instruction, we only keep the pixel values in the central part with the size of 19×19 in each facial patch, with the pixel values in the other part as zero.

When performing FR, for the generic set and the testing set one patch centered around each facial point is extracted, and for the gallery set one patch centered around each facial point as well as its 8 neighbor patches are extracted to boost the robustness to local deformation and misalignment. The neighbor size is set as 1 for all the experiments. In JCR-ACF there are two regularization parameters, λ and μ in Eq. (3). λ regularizes the l_2-norm of the representation coefficients and μ controls the similarity of the representation coefficients in the joint and collaborative representation. In all experiments we fix λ=0.005. In experiments on AR database and the CMU Multi-PIE database we set μ =0.005. In experiments on the LFW database and the CASIA-WebFace database where the facial variations are more uncontrolled we set μ =0.05.

For methods involving learning the intra-class variation dictionary the size of the dictionary is set as follows: For SVDL [17], we follow [17] and set the size of the dictionary as 400 in the initialization in the experiments on the AR database and the CMU Multi-PIE database. For RPR [19], as [19] different sizes for different databases are used. The size of the dictionary is set as 120 for the AR database and 500 for the other three databases.

Also note that we follow the experimental setting on the AR database and LFW database with 158 subjects from the work of LGR [3] and the setting on CMU Multi-PIE database from the work of SVDL [17], respectively. Thus the results of SVM, SRC [25], DMMA [8], AGL [33], ESRC [16] and SVDL [17] are kept the same to those in previous work. For all methods in other databases, including LFW with 901 subjects and the large-scale CASIA-WebFace database, we get the results of all methods by running the corresponding codes provided by the authors because none of the existing methods have been evaluated in these cases. In addition, for the recent RPR [19] method we get the results on all databases by running the code provided by the author. Since the work of DeepID [20] doesn't publish their code, we carefully realize it and report its results by tuning its parameters.

5.2. Investigating the discrimination of particular facial regions

According to the visual perception of the human being, the particular facial regions (e.g., eyes, nose and mouth) of the face are more discriminative than the other parts of the face. It should be beneficial for boosting the FR performance to exploit these particular facial regions. We demonstrate the strong discrimination of the particular facial regions by conducting experiments on AR database with illumination variations (Please refer to Section 5.3 for details about the database and the experimental protocol). The recognition rates of using the 5 particular facial regions and 5 regular facial regions selected randomly out of the 16 regular facial regions are listed in Table 2. We perform the random selection 6 times. From Table 2 we can see that using the 5 particular regions performs better in almost all trails. Hence the benefits of utilizing the 5 particular facial regions can be convinced.

Table 2
Recognition accuracy (%) on AR database with illumination variations. 5P means 5 particular facial regions are used. 5R-i means 5 regular regions in the i^{th} random selection are used.

Session	5P	5R-1	5R-2	5R-3	5R-4	5R-5	5R-6
1	92.9	90.4	86.7	87.1	81.3	87.9	93.3
2	88.3	78.3	75.8	67.9	79.6	77.9	80.4

5.3. Evaluation on variations of AR database

In this section, we evaluate the proposed JCR-ACF on the AR database [27]. The images in this database are taken in two sessions. As [3], a subset of AR database including 50 male and 50 female subjects with 26 images per subject is used for the experiments. Four types of facial variations including illumination, expression, disguise, and illumination + disguise are involved for each subject. The first 80 subjects and the remaining 20 subjects in Session 1 are used for the gallery set and the generic set, respectively. For the gallery set the neutral face image without disguise and illumination of each subject is used. For the generic set the neutral face image without disguise and illumination of each subject is used as a reference image, with the other images of each subject for the variation images. The face images with four types of facial variations from both two sessions are used for testing. Fig. 4 shows some samples in Session 1 of the AR database.

The recognition rates of the competing methods for testing images from Session 1 and Session 2 are listed in Tables 3, 4, respectively. We can see that ESRC and SVDL perform fairly well for illumination variations, however, their performance degrades much for the tough conditions i.e. for variations in expression and disguise. This is because they use the holistic information of the face image, which is vulnerable to complex variations (e.g. expression) and outliers (e.g. disguise). On the contrary, the performance of RPR, LGR and the proposed JCR-ACF remain fairly stable. Because these methods use local features thus can tolerate gross facial variations to some extent. Furthermore, we can notice that JCR-ACF achieves much better performance (e.g., 11.3% improvement over RPR on Exp of Session 2 and 4% improvement over LGR on Illu+Dis of Session 2) than RPR and LGR for the variations except the illumination variations, which validates the high robustness of JCR-ACF to non-linear challenging facial variations. And for the illumination variations JCR-ACF performs slightly worse than RPR and LGR. The better performance of RPR and LGR for illumination variations may owe to the use of intensity features, which perfectly fit the linear additivity embedded in the representation based model.

Table 3
Recognition accuracy (%) on AR database with illumination (Illu), expression (Exp), disguise (Dis) and illumination+disguise (Illu+Dis) (Session1).

Method	Illu	Exp	Dis	Illu+Dis
SVM	55.8	90.4	43.1	29.4
SRC [25]	80.8	85.4	55.6	25.3
DMMA [8]	92.1	81.4	46.9	30.9
AGL [33]	93.3	77.9	70.0	53.8
ESRC [16]	99.6	85.0	83.1	68.6
SVDL [17]	98.3	86.3	86.3	79.4
RPR [19]	99.6	96.3	98.8	93.8
LGR [3]	**100**	97.9	98.8	96.3
JCR-ACF	99.2	**100**	**100**	**99.4**

Table 4
Recognition accuracy (%) on AR database (Session2).

Method	Illu	Exp	Dis	Illu+Dis
SVM	40.0	58.8	26.9	14.4
SRC [25]	55.8	68.8	29.4	12.8
DMMA [8]	77.9	61.7	28.1	21.9
AGL [33]	70.8	55.8	40.6	30.7
ESRC [16]	87.9	70.4	59.4	45.0
SVDL [17]	87.1	74.2	61.3	54.1
RPR [19]	**97.5**	82.9	91.9	82.2
LGR [3]	**97.5**	85.0	93.8	88.8
JCR-ACF	95.0	**94.2**	**96.3**	**92.8**

5.4. Evaluation on variations of CMU-MPIE database

The face images in the CMU Multi-PIE database [28] are taken in four sessions with simultaneous variations in pose, expression, and illumination. For a particular pose and a particular expression of each

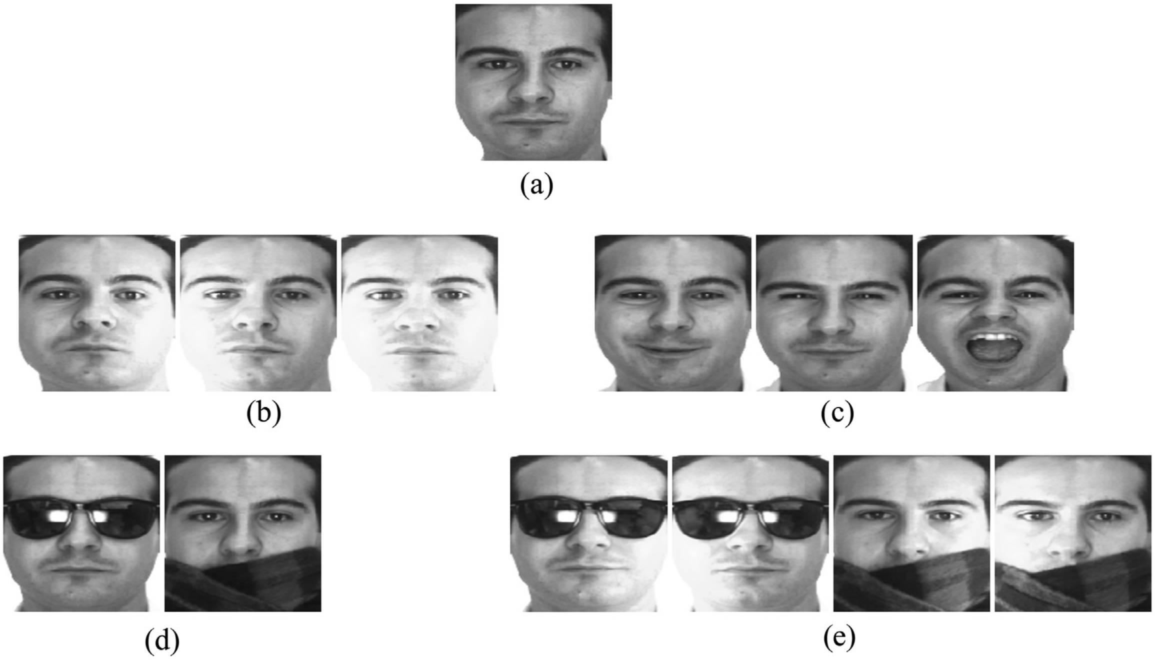

Fig. 4. Samples in Session 1 of the AR database. (a) the single gallery sample; (b) testing samples with illumination variations; (c) testing samples with expression variations; (d) testing samples with disguise variations; (e) testing samples with illumination and disguise variations.

subject in each session there are 20 illuminations with index from 0 to 19. Session 1 contains 249 subjects in total, of which the first 100 subjects are used as the gallery and query subjects. And the other 149 subjects in Session 1 are used for the generic subjects. The images in Session 1 are used to form the gallery set. Especially the frontal image with illumination 7 and neutral expression of each gallery subject is used as a gallery image. For the reference subset of the generic set, the frontal image with illumination 7 and neutral expression of each subject is chosen as the reference face image of that subject. In the following three testing conditions, the same gallery set and reference set of the generic training set are used. And the variation subsets of the generic set are different for each testing condition. The experimental protocol is consistent with [17,3].

(21) **Illumination Variations:** Here the frontal face images with neutral expression and 20 various illumination changes in Session 2, 3, and 4 are separately used to evaluate the proposed JCR-ACF. The variation subset of the generic set consists of all the frontal face images with neutral expression of the generic subjects in Session 1 except images with illumination 7 (as mentioned before images with illumination 7 have been used for the reference subset). Fig. 5 shows the single gallery simple and the some testing samples with illumination variations of a subject. The recognition rates for the testing sets from the three sessions are list in Table 5.

It can be seen that JCR-ACF achieves the best performance in all three sessions. Specifically, the improvements of JCR-ACF over the runner-up method LGR are 1.3%, 4.2% and 1.6%, respectively. Recall that in the previous experiments with illumination variations on the AR database JCR-ACF performs a little worse than LGR and RPR, but here JCR-ACF can outperform them.

(2) **Expression and Illumination Variations:** The frontal face images with smile in Session 1, smile in Session 3, and surprise in Session 2 and with various lighting changes are used for testing. For each testing case, all frontal face images with the corresponding expression of the generic subjects in the same session as the testing set are used to construct the variation subset. Fig. 6 shows some testing samples with expression and illumination variations. Table 6 presents the recognition rates of all the competing methods.

It can be seen that JCR-ACF outperforms the other methods by visible margins, which indicates that compared to other methods JCR-ACF has the higher robustness to complex facial variations. In addition, it can be easily noticed that all methods perform much better on Smile-S1 than on Smile-S3 and Surprise-S2. This is because the testing set of Smile-S1 and the gallery set are from the same data session.

(3) **Pose, Illumination and Expression Variations:** In this section, we evaluate JCR-ACF on more difficult conditions where the testing samples are composed of face images with Pose ‘05_0′ in Session 2 (P05_0-S2) and Pose ‘04_1′ in Session 3 (P04_1-S3), and face images with Pose ‘04_1′ and smile expression in Session 3 (Smi-P04_1-S3). For each testing case, all face images with the corresponding expression and pose of the generic subjects are used to form the variation subset. Fig. 7 shows some testing samples with pose, expression and illumination variations. We list the recognition results of all competing methods in Table 7.

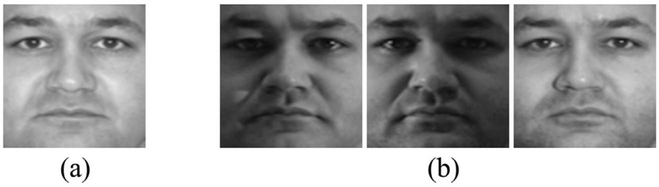

Fig. 5. (a) The single gallery sample; (b) testing samples with illumination variations in Sessions 2, 3 and 4, respectively.

Table 5
Recognition accuracy (%) on Multi-PIE database with illumination variations.

Session	Session 2	Session 3	Session 4
SVM	45.3	40.2	43.7
SRC [25]	52.4	46.7	49.5
DMMA [8]	63.2	55.4	60.4
AGL [33]	84.9	79.4	78.3
ESRC [16]	92.6	84.9	86.7
SVDL [17]	94.8	87.7	91.0
RPR [19]	93.2	84.1	89.5
LGR [3]	96.9	90.5	94.4
JCR-ACF	**98.2**	**94.7**	**96.0**

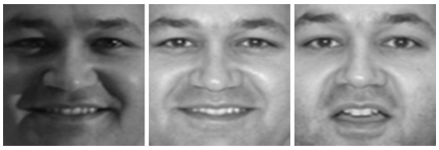

Fig. 6. Testing samples with expression and illumination variations from Smile-S1, Smile-S3 and Surprise-S2, respectively.

Table 6
Recognition accuracy (%) on Multi-PIE database with expression and illumination variations.

Expression	Smile-S1	Smile-S3	Surprise-S2
SVM	46.9	28.8	18.0
SRC [25]	49.6	28.1	20.4
DMMA [8]	58.2	31.5	22.0
AGL [33]	84.9	39.3	31.3
ESRC [16]	81.6	50.5	49.6
SVDL [17]	88.8	58.6	54.7
RPR [19]	88.5	61.9	67.7
LGR [3]	90.7	65.2	74.5
JCR-ACF	**95.5**	**77.7**	**76.6**

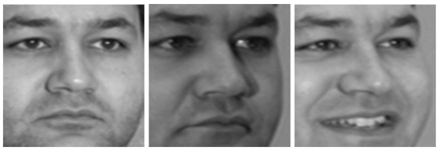

Fig. 7. Testing samples with pose, expression and illumination variations from P05_0-S2, P04_1-S3 and Smi-P04_1-S3, respectively.

Table 7
Recognition accuracy (%) on Multi-PIE database with pose, expression and illumination variations.

Pose	P05_0-S2	P04_1-S3	Smi-P04_1-S3
SVM	26.0	8.7	12.0
SRC [25]	25.0	7.3	10.3
DMMA [8]	27.1	5.3	11.0
AGL [33]	66.7	24.9	23.9
ESRC [16]	63.9	31.8	26.9
SVDL [17]	77.8	38.3	34.4
RPR [19]	62.1	28.0	23.0
LGR [3]	79.1	39.5	36.3
JCR-ACF	**90.3**	**66.7**	**54.2**

It can be observed that JCR-ACF performs much better than the other methods. Specifically, the improvements of JCR-ACF over the second best method LRG are 11.2%, 27.2% and 17.9% for

Fig. 8. Samples on LFW database. (a) the single gallery sample; (b) the testing samples.

Table 8
Recognition accuracy (%) on LFW database with 158 subjects.

Method	SVM	SRC [25]	DMMA [8]	AGL [33]	ESRC [16]	RPR [19]	LGR [3]	DeepID-19200 [20]	JCR-ACF
Accuracy	11.6	20.4	17.8	19.2	27.3	33.1	30.4	70.7	**86.0**

Table 9
Recognition accuracy (%) on LFW database with 901 subjects.

Method	ESRC [16]	RPR [19]	LGR [3]	DeepID-150 [20]	DeepID-19200 [20]	JCR-ACF
Accuracy	20.2	15.6	9.1	44.4	48.9	**65.3**

the three tests, respectively. The impressive improvements of JCR-ACF over the other competing methods demonstrate the superiority of the JCR-ACF in handling the tough and challenging facial variations.

5.5. Evaluation on variations of LFW database

The LFW database [29] contains images of 5749 subjects. This database is built by using images taken in the uncontrolled environment. So it is very challenging. As [3] we use the LFW-a dataset, which is the aligned version of LFW. We choose a subset of 158 subjects with more than 10 images (inclusive) per subject for this experiment. Although face alignment has been conducted severe misalignment still exists due to the wild poses. The first 50 subjects are selected to form the gallery set and the query set, and for each subject the first image is used for the gallery image with the remaining images for testing. The other 108 subjects in the database are used as the generic training set. Fig. 8 shows some samples in the LFW database. Since there is no frontal neutral face for each subject, we use the mean of ACFs of the face images of each subject as the reference in the generic set, which is consistent with [3] where the mean face is used as the reference image.

Recently deep learning has been used for face verification and achieved great success. As a representative deep learning model for face recognition, DeepID net [20] learns deep convolutional features and combines 60 overlapping regions to conduct the final verification. In this challenging dataset, we also compare our proposed JCR-ACF with DeepID in the application of FR with SSPP. As [20] we extract features from 60 face patches with ten regions, three scales, and RGB or gray channels. For each face patch, two 160-dimensional DeepID features are extracted from a particular patch and its horizontally flipped counterpart. Then all DeepID features are concatenated to form the 19200-dimensional (160×60×2) feature, which is denoted by DeepID-19200. The Nearest Neighbor with cosine distance is used to perform the classification. Also as [20], we use PCA to reduce the feature dimensionality to 150 for reference (The data mean and the projection matrix used by PCA are computed on the gallery set and applied to both the gallery set and test set) and the obtained 150 dimensional feature is denoted by DeepID-150. Since the local regions of DeepID are much overlapped, for fair comparison we also use overlapped patches in JCR-ACF by keeping the pixel values of the whole patch with the size of 39×39 (Note that making the patches overlap with each other is not always beneficial because a big-size patch will contain more global information).

We list the FR rates of different methods in Table 8. Since there are only 50 subjects in the gallery set, we report the performance of DeepID with 19200-dimensional (160×60×2) feature. Because the facial variations in this database are uncontrolled, all methods except JCR-ACF and DeepID fail to achieve satisfactory performance (e.g., RPR only gets 33.1% accuracy, LGR only gets 30.4% accuracy). In the test, our proposed JCR-ACF has achieved 86.0% accuracy, about 15% improvement over the second best method, DeepID. It can also be noticed that JCR-ACF surpasses the other methods by very large margins (e.g., around 50% improvement over the method, RPR). When JCR-ACF uses non-overlapped regions, its accuracy is 52.6%, much lower than JCR-ACF with overlapped regions, which shows that the information contained in the patches with a suitable size (e.g., 39×39) is beneficial to the final classification. Nevertheless, as shown in tables in the previous sections, the proposed JCR-ACF has achieved much better performance than the methods without deep features even we don't require different facial patches to overlap.

In addition, we conduct another experiment on the LFW database with more gallery and testing subjects. Based on the above experiments we add another 743 new subjects with 3–9 images per subject to the gallery set and testing set. As before the first image of each subject is used as the gallery image and the remaining images are used for the testing set. The same intra-class variation dictionary as the above experiment is used. Now there are 901 total subjects in this experiment. Some competitive competitors (e.g., ESRC, RPR, LGR and DeepID) verified in previous experiments are compared against. Table 9 shows the corresponding results. From Table 9 we can see that the proposed JCR-ACF outperforms the other competitors very much. Compared to the second best method, DeepID, at least 16% improvement are achieved. Other methods, such as ESRC, RPR and

Fig. 9. Samples on CASIA-WebFace database. (a) the single gallery sample; (b) the testing samples.

Table 10
Recognition accuracy (%) on CASIA-WebFace database.

Method	SVM	SRC [25]	DMMA [8]	AGL [33]	ESRC [16]	RPR [19]	LGR [3]	JCR-ACF
Accuracy	3.5	8.3	6.4	5.6	9.8	10.6	5.9	**15.0**

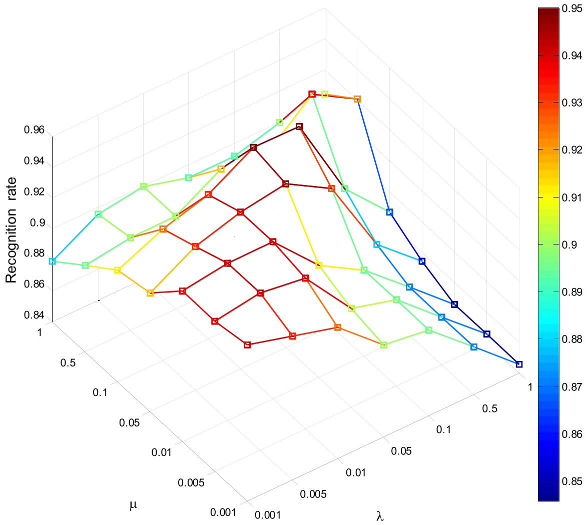

Fig. 10. The recognition rates of JCR-ACF versus different parameter values in the experiment on session 2 of AR database with expression variations.

Table 11
Comparison among JCR-ACF, LGR and RPR in the patch number, intra-class variation dictionary size and feature dimensionality.

Method	Patch number	Intra-class variation dictionary size	Feature dimensionality
RPR [19]	49	120	64
LGR [3]	49	240	400
JCR-ACF	21	240	160

Table 12
Average computational time (seconds) for each testing image on AR database (Session 1).

Method	Illu	Exp	Dis	Illu+Dis
RPR [19]	0.8467	0.7996	0.8030	0.8562
LGR [3]	2.9983	3.0441	2.9437	2.9674
JCR-ACF	**0.1510**	**0.1517**	**0.1533**	**0.1547**

LGR, have quite low recognition rates in this challenging experiment, which verifies the effectiveness of the learned adaptive convolutional feature and joint collaborative representation model in the proposed JCR-ACF.

Table 13
Recognition accuracy of JCR-ACF with only keeping the pixel values in the 19×19 central part on the LFW database with 158 subjects when assigning different weights to different local regions (%).

Case	1	2	3	4	5	6
Accuracy	52.6	53.9	53.0	52.7	51.6	52.4

5.6. Evaluation on variations of the CASIA-webface database

The CASIA-WebFace database [30] contains 10,575 subjects with 494,414 images. We use 8575 subjects containing 40,1993 images for training the deep convolutional neural networks. We use another 500 subjects with 5 images per subject as the gallery set and the query set where the first image of each subject is used for the gallery set and the other 4 images for the query set. In addition, another 100 subjects with 1930 images are used to form the generic set. Fig. 9 shows some samples in the CASIA-WebFace database. As the previous experiment on the LFW database, we use the mean of ACFs of the face images of each subject as the reference in the generic set, since there is no frontal neutral face for each subject. The recognition rates of all methods are listed in Table 10. Although all methods perform quite poorly in this challenging case (e.g., RPR gets 10.6% recognition rate compared to 0.2% guess accuracy), the proposed JCR-ACF is still much better that all the others (e.g., 4.4% improvement over the second best one, RPR).

5.7. Parameter evaluation

In this section we conduct the experiment on Session 2 of AR database with expression variations to evaluate the robustness of JCR-ACF to different parameter values. The recognition rates of JCR-ACF versus different parameter values are shown in Fig. 10. It can be seen that the performance of JCR-ACF remains fairly stable when $\lambda\in[0.001, 0.1]$ and $\mu\in[0.001, 0.5]$, which indicate the high robustness of JCR-ACF to different parameter values. Accordingly, based on our experience we fix λ=0.005 for all the experiments, set μ=0.005 for the AR and CMU Multi-PIE database and set μ=0.05 for the LFW and CASIA-WebFace databases.

5.8. Running time comparison

Here we compare the computational time of JCR-ACF with that of the other two most competitive methods including LGR and RPR. The used desktop is of 4.0 GHz CPU with a 16 G RAM. To make a clear demonstration the patch numbers, the intra-class variation dictionary sizes and the feature dimensionality of the three methods are reported in Table 11. The average consuming time (seconds) for each testing

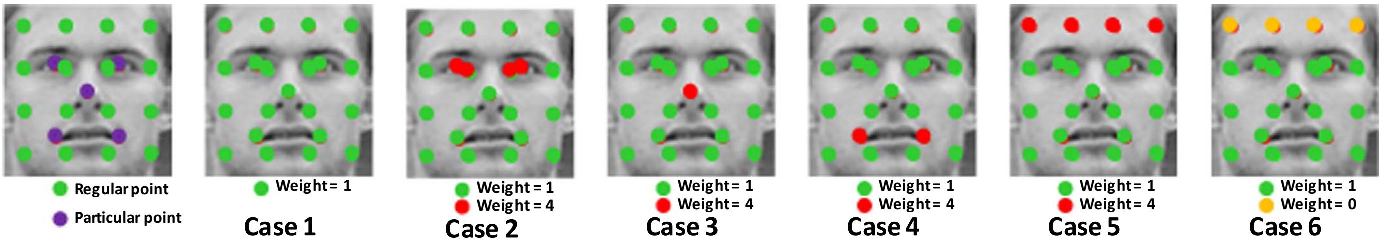

Fig. 11. Weights of local regions centered around the red or yellow facial points are varied.

image on AR database (Session1) is list in Table 12. We can see that JCR-ACF is the most efficient. This benefits from the advantage that only matrix multiplication operations are involved in the testing stage for JCR-ACF. RPR uses about 2.3 times as many patches as JCR-ACF and less dictionary atoms and feature dimensionality but is more than 5 times slower than JCR-ACF. Because the optimization problem in RPR has to be iteratively solved by using the augmented Lagrange multiplier method, which usually needs several steps to obtain the optimum [19]. LGR is very time-consuming because it has to calculate the matrix inversion for each patch and the patch number and the feature dimensionality are both large [3].

5.9. Discussion on weighting different local regions

Here we evaluate the FR performance by assigning different weights to different local regions. Specifically, we vary the weights of local regions centered around the 11 facial points, i.e. 4 points on the forehead, 2 points around eyes, 2 eye centers, 2 mouth corners and 1 nose tip in Fig. 11. We compare 6 Cases as shown in Fig. 11 (Note that the default weight for any local region whose weight is not specified explicitly is 1): Case 1: weights of all local regions are 1; Case 2: weights of local regions centered around the eyes are 4; Case 3: weights of local regions centered around the nose are 4; Case 4: weights of two local regions centered around the 2 mouth corners are 4; Case 5: weights of local regions centered around the 4 points on the forehead are 4; Case 6: weights of local regions centered around the 4 points on the forehead are 0. The FR results are shown in Table 13. From Table 13, we can notice that increasing weights of local regions around the eyes and the nose improves the FR accuracy when comparing Case 2 and 3 with Case 1 (e.g., 1.3% improvement from Case1 to Case 2 and 0.4% improvement from Case 1 to Case 3). However, increasing weights of regions on the forehead in Case 5 degrades the FR performance in Case 1 from 52.6 to 51.6%. And compared to Case 5 if we remove the 4 local regions on the forehead in Case 6 the FR accuracy can be improved by 0.8% in reverse. These observations demonstrate that different local regions have different discrimination and some regions don't contribute to or even mislead the recognition. It is expected that the performance can be further boosted by a weighting strategy in which the roles of the discriminative regions can be automatically emphasized whereas the roles of regions which are not discriminative or even misleading are inhibited. This weighting strategy serves as an important direction for our future research.

6. Conclusion

This paper presented a joint and collaborative representation (JCR) with local adaptive convolution feature (ACF) for FR with SSPP. ACFs contains discriminative local high-level features for classification and local regular regions and particular regions provide a complete and robust description of face images. With ACFs, the proposed JCR has fully exploited the distinctiveness and commonality of different local regions of the face image by doing representation for each ACF and requiring different ACFs from different local regions to have similar representation coefficients. Extensive experiments demonstrate that JCR-ACF is more robust to complex and challenging facial variations compared with the state of the art methods. Additionally, only matrix multiplication operations are involved in the testing stage, which makes JCR-ACF efficient.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Grant no. 61402289), National Natural Science Foundation of China (Grant no. 61272050), Natural Science Foundation of Guangdong Province (Grant no. 2014A030313558 and 2014A030313556), and Shenzhen Scientific Research and Development Funding Program (Grant no. JCYJ20140509172609171), Scientific Research Fund for new teachers of Shenzhen University (Grant no. 201536), Scientific Research Fund for Advanced Talents of Shenzhen University (Grant no. 000070).

References

[1] W. Zhao, R. Chellppa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Comput. Surv. 35 (4) (2003) 399–458.

[2] L. Wolf, T. Hassner, Y. Taigman, Effective face recognition by combining multiple descriptors and learned background statistics, IEEE TPAMI 33 (19) (2011) 1978–1990.

[3] P.F. Zhu, M. Yang, L. Zhang, I.Y. Lee, Local generic representation for face recognition with single sample per person, in: Proceedings of the ACCV, 2014.

[4] X. Tan, S. Chen, Z. Zhou, F. Zhang, Face recognition from a single image per person: a survey, Pattern Recognit. 39 (9) (2006) 1725–1745.

[5] P.N. Belhumeur, J. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE TPAMI 19 (7) (1997) 711–720.

[6] T. Ahonen, A. Hadid, M. Pietikainen, Face recognition with local binary patterns, in: ECCV, 2004.

[7] G. Tzimiropoulos, S. Zafeiriou, M. Pantic, Subspace learning from image gradient orientations, IEEE TPAMI 33 (12) (2012) 2454–2466.

[8] J.W. Lu, Y.P. Tan, G. Wang, Discriminative multimanifold analysis for face recognition from a single training sample per person, IEEE TPAMI 35 (1) (2013) 39–51.

[9] X. Tan, S. Chen, Z. Zhou, F. Zhang, Recognizing partially occluded expression variant faces from single training image per person with som and soft k-nn ensemble, IEEE NN 16 (4) (2005) 875–886.

[10] S. Chen, J. Liu, Z. Zhou, Making flda applicable to face recognition with one sample per person, Pattern Recognit. 37 (7) (2004) 1553–1555.

[11] D. Zhang, S. Chen, Z. Zhou, A new face recognition method based on svd perturbation for single example image per person, Appl. Math. Comput. 163 (2) (2005) 895–907.

[12] S. Shan, B. Cao, W. Gao, D. Zhao, Extended fisherface for face recognition from a single example image per person, in: Proceedings of the ISCAS, 2002.

[13] A.N. Li, S.G. Shan, W. Gao, Coupled bias-variance tradeoff for cross-pose face recognition, IEEE TIP 21 (1) (2012) 305–315.

[14] J. Wang, K. Plataniotis, J. Lu, A. Venetsanopoulos, On solving the face recognition problem with one training sample per subject, Pattern Recognit. 39 (9) (2006) 1746–1762.

[15] H. Mohammadzade, D. Hatzinakos, Expression subspace projection for face recognition from single sample per person, IEEE Affect. Comput. 4 (1) (2013) 69–82.

[16] W.H. Deng, J.N. Hu, J. Guo, Extended SRC: undersampled face recognition via intra-class variant dictionary, IEEE TPAMI 34 (9) (2012) 1864–1870.

[17] M. Yang, L.V. Gool, L. Zhang, Sparse variation dictionary learning for face recognition with a single training sample per person, in: Proceedings of the ICCV, 2013.

[18] L. Zhuang, A. Yang, Z. Zhou, S.S. Sastry, Y. Ma, Single-sample face recognition with image corruption and misalignment via sparse illumination transfer, in: Proceedings of the CVPR, 2013.

[19] S. Gao, K. Jia, L. Zhuang, Y. Ma, Neither global nor local: regularized patch-based representation for single sample per person face recognition, Int. J. Comput. Vis. 111 (3) (2015) 365–383.

[20] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10, 000 classes, in: Proceedings of the CVPR, 2014.

[21] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Proceedings of the CVPR, 2014.

[22] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Proceedings of the NIPS, 2014.

[23] D. Chen, X. Cao, L. Wang, F. Wen, J. Sun, Bayesian face revisited: A joint formulation, in: Proceedings of the ECCV, 2012.

[24] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, in: Proceedings of the CVPR, 2013.

[25] J. Wright, A.Y. Yang, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE TPAMI 31 (2) (2009) 210–227.

[26] M. Yang, L. Zhang, X.C. Feng, D. Zhang, Sparse representation based Fisher discrimination dictionary learning for image classification, Int. J. Comput. Vis. 109 (3) (2014) 209–232.

[27] A.M. Martinez, R. Benavente, The AR face database, Technical Report 24, CVC, 1998.

[28] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput 28 (5) (2010) 807–813.

[29] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[30] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch arXiv preprint arXiv:1411.7923, 2014.

[31] L. Zhang, M. Yang, X.C. Feng, Sparse representation or collaborative representation: Which helps face recognition, in: Proceedings of the ICCV, 2011.

[32] M. Yang, L. Zhang, J. Yang, D. Zhang, Regularized robust coding for face recognition, IEEE Trans. Image Process. 22 (5) (2013) 1753–1766.

[33] Y. Su, S. Shan, X. Chen, W. Gao, Adaptive generic learning for face recognition from a single sample per person CVPR, 2010.

[34] H. Yan, J. Lu, X. Zhou, Y. Shang, Multi-feature multi-manifold learning for single-sample face recognition, Neurocomputing 143 (2014) 134–143.

[35] P. Zhang, X. You, W. Ou, C.L. Chen, Y. Cheung, Sparse discriminative multi-manifold embedding for one-sample face identification, Pattern Recognit. 52 (2016) 249–259.

[36] F. Liu, J. Tang, Y. Song, Y. Bi, S. Yang, Local structure based multi-phase collaborative representation for face recognition with single sample per person, Inf. Sci. 346 (2016) 198–215.

[37] C. Hu, M. Ye, S. Ji, W. Zeng, X. Lu, A new face recognition method based on image decomposition for single sample per person problem, Neurocomputing 160 (2015) 287–299.

[38] J. Hu, J. Lu, X. Zhou, Y. Tan, Discriminative transfer learning for single-sample face recognition, in: Proceedings of the International Conference on Biometrics, 2015.

[39] W. Deng, J. Hu, X. Zhou, J. Guo, Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning, Pattern Recognit. 47 (12) (2014) 3738–3749.

[40] R. Ding, D.K. Du, Z. Huang, Z. Li, K. Shang, Variational Feature Representation-based Classification for face recognition with single sample per person, J. Vis. Commun. Image Represent. 30 (2015) 35–45.

[41] R. Rubinstein, A. Bruckstein, M. Elad, Dictionary learning for sparse representation modeling, Proc. IEEE 98 (6) (2010) 1045–1057.

[42] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing over-complete dictionaries for sparse representation, IEEE SP 54 (11) (2006) 4311–4322.

[43] S.Z. Li, A.K. Jain, Handbook of Face Recognition, Second ed., Springer, 2011.

[44] R. Jafri, H.R. Arabnia, A survey of face recognition techniques, J. Inf. Process. Syst. 5 (2) (2009) 41–68.

[45] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: Closing the Gap to Human-Level Performance in Face Verification, in: Proceedings of the CVPR, 2014.

[46] Y. Lei, Y. Guo, M. Hayat, et al., A Two-phase weighted collaborative representation for 3D partial face recognition with single sample, Pattern Recognit. 52 (2016) 218–237.

Meng Yang was an associate professor at School of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China. He received his Ph.D degree from The Hong Kong Polytechnic University in 2012. Before joining Shenzhen University, he has been working as Postdoctoral fellow in the Computer Vision Lab of ETH Zurich. His research interest includes sparse coding, dictionary learning, object recognition and machine learning. He has published 11 ICML/AAAICVPR/ICCV/ECCV papers and several IJCV, IEEE TNNLS and TIP journal papers. Now his Google citation is over 3300, and his homepage is www.yangmeng.org.cn.

Xing Wang is currently a research assistant at School of Computer Science and Software Engineering, Shenzhen University. He received his B.Eng. and M. Eng. degrees from the College Of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, in 2009 and 2012, respectively. His research interest is sparse representation for face recognition.

Guohang Zeng is currently a senior student at School of Computer Science and Software Engineering, Shenzhen University. His research interest is sparse representation and deep learning.

Linlin Shen is currently the Director of Computer Vision Institute and professor at School of Computer Science & Software Engineering, Shenzhen University. He received his Ph.D. degree from University of Nottingham, UK in 2005. Before joining Shenzhen University, he has been working as Research Fellow on MRI brain image processing at Medical school, University of Nottingham. His research interest covers pattern recognition, medical image processing and biometrics. His is the recipient of Most Cited Paper Award by the journal of Image and Vision Computing, the winner of Competition on Cells Classification by Fluorescent Image Analysis organized by ICIP 2013, and Highly Cited Author in China listed by Elsevier in 2015.