Neurocomputing 147 (2015) 403–411

Contents lists available at ScienceDirect

Neurocomputing

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

Sparsely encoded local descriptor for face verification

Zhen Cui [a,b], Shiguang Shan [a,*], Ruiping Wang [a], Lei Zhang [c], Xilin Chen [a]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] School of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[c] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Article history:
Received 14 November 2013
Received in revised form
7 April 2014
Accepted 22 June 2014
Communicated by Qingshan Liu
Available online 10 July 2014

Keywords:
Local descriptor
Sparse coding
Non-negativity
Face verification
Labeled faces in the wild

ABSTRACT

A novel Sparsely Encoded Local Descriptor (SELD) is proposed for face verification. Different from traditional hard or soft quantization methods, we exploit linear regression (LR) model with sparsity and non-negativity constraints to extract more discriminative features (i.e. sparse codes) from local image patches sampled pixel-wisely. Sum-pooling is then imposed to integrate all the sparse codes within each block partitioned from the whole face image. Whitened Principal Component Analysis (WPCA) is finally used to suppress noises and reduce the dimensionality of the pooled features, which thus results in the so-called SELD. To validate the proposed method, comprehensive experiments are conducted on face verification task to compare SELD with the existing related methods in terms of three variable component modules: K-means or K-SVD for dictionary learning, hard/soft assignment or regression model for encoding, as well as sum-pooling or max-pooling for pooling. Experimental results show that our method achieves a competitive accuracy compared with the state-of-the-art methods on the challenging Labeled Faces in the Wild (LFW) database.

1. Introduction

Face recognition has attracted significant attention due to its wide potential applications in public security, law enforcement, etc. Numerous methods or techniques have been developed as surveyed in [1], and considerable progresses have been achieved in the past decades. Currently, state-of-the-art face recognition systems have been able to work well under well-controlled conditions with cooperative users. However, as discovered by LFW evaluation [2], face recognition under uncontrolled environment still remains a great challenge due to complex variations in pose, illumination, expression, aging, etc. To well address this problem, how to discriminatively represent face images plays a key role in the task of unconstrained face recognition.

In the past decade, local descriptors, modeling micro-patterns in images, have formed a blowout in face recognition area [3–9], due to their robustness to identity-irrelevant extrinsic variations. These methods usually fall into two categories: hand-crafted and auto-learned descriptors, which are briefly introduced in what follows.

Many manually designed local patterns have been developed for face recognition. For example, by combining the signs of the differences of central pixel intensity from those of its neighboring pixels, Local Binary Patterns (LBP) [6] implicitly encodes the micro-patterns of the input image such as flat areas, spots, edges, and corners. Because of its invariance to monotonic photometric changes, LBP is robust to lighting variation to some extent. After that, many variants of LBP were proposed. For instance, Zhao and Pietikainen extended LBP to the spatial-temporal domain [10]. In order to make LBP more robust to random and quantization noise in near-uniform face regions, Local Ternary Patterns (LTP) [11] were proposed. By combining Gabor filtering [12] with LBP, Local Gabor Binary Pattern (LGBP) [8] was proposed to endow LBP with capacity of encoding micro-patterns of multi-scale and multi-orientation. Later on, histogram of Gabor phase patterns [7] was further proposed to exploit the Gabor phase information. In addition, some local descriptors widely used in general object classification, such as Histogram of Oriented Gradients (HOG) [13] or Scale Invariant Feature Transform (SIFT) [9], were introduced into face recognition. In spite of its popularity, manually designing local patterns are non-trivial because it has to balance skillfully discriminative power and robustness against data variance.

In contrast to the above hand-crafted approaches, auto-learning based methods typically pursue some codewords (representative local visual primitives) from a large amount of low-level features (e.g. SIFT). Then, given an input image, its low-level features are encoded with these codewords by utilizing hard/soft quantization, followed by pooling operation to form mid-level

* Corresponding author. Tel.: +86 10 62600558.
E-mail address: sgshan@ict.ac.cn (S. Shan).

http://dx.doi.org/10.1016/j.neucom.2014.06.044
0925-2312/

features. By learning the codewords directly on image patches with K-means clustering algorithm, Meng et al. [14] proposed Local Visual Primitives (LVP), which finally represented one face image by concatenating block-based histograms of the learned patterns for face recognition. Ahonen and Pietikainen [15] also tried K-means clustering to build local filter response codebook. Cao et al. [5] argued that quantized codes with K-means usually had uneven distribution and the encoded histogram would be less informative and less compact. To address the problem, they substituted random-projection tree for K-means clustering. In addition, hard quantization may lead to losing a lot of useful information especially subtle textural features in face images, since only one nearest atom is chosen as the agent for one input raw feature. In contrast, soft quantization based methods [16,17] encode the input features with multiple codewords so as to make the representation more accurate. For instance, van Gemert et al. [17] proposed to use Gaussian kernel to deal with visual word ambiguity for object classification.

Another recent progress in face recognition is sparse representation based methods [18–25]. In [18], Wright et al. sparsely encoded one face image by using the training set as the codebook and then sought for the subject whose samples result in the smallest reconstruction error by using their corresponding sparse coefficients. In the case of multiple well-aligned samples per person, they reported impressive results, especially for partially occluded faces. Further, some researchers tried to learn a robust codebook, such as the discriminative codebook [19] and the compact Gabor codebook [20]. Besides, Cui et al [25] apply sparse representation into video-based face recognition. However, these methods mostly focus on holistic representation, and thus are fragile to local appearance variations. Another limitation of these methods is that they only work for the scenario where each subject has multiple enrolled face images, i.e., they cannot be applied to face verification and face identification with single sample per person. To address these problems, more recently, face region descriptor (FRD) [4] is proposed to address still and video images with a similar framework.

Inspired by the above works, in this paper we propose a local descriptor via texton-learning with sparsity constraints. Specifically, our method first learns visual codewords locally on image patches with sparsity constraints. Then, non-negative sparse regression against the visual codewords is exploited to project each pixel-wise raw image patches into more discriminative sparse codes, which is quite different from the existing hard assignment methods [5,14,26] and soft assignment methods [16,17]. In the next step, sum-pooling is exploited to integrate the sparse codes within each image block, and at the same time endow the generated mid-level features more robustness to misalignment. Finally, Whitened Principal Component Analysis (WPCA) [27] is used to further reduce the dimensionality and suppress the noise of the pooled features, eventually resulting in our Sparsely Encoded Local Descriptor (SELD).

As an extension of our previous work [3], we further improve the conference work mainly on three aspects: (1) multiple block-partitioning modes on face images are used to retain more facial configuration information; (2) Distance Metric Learning (DML) is combined with SELD to utilize supervised information; and (3) extensive cross-validation experiments on the three component modules: dictionary learning, encoding and spatial pooling. As a whole, our contributions mainly lie in three folds: (1) propose an auto-learning face descriptor for face verification; (2) conduct extensive cross-validation experiments to validate the role of each module; and (3) achieve a competitive performance on the LFW dataset under its restrict protocol.

As our experiments are mainly conducted on the LFW dataset, here we briefly review the related state-of-the-art methods on it.[1] To achieve competitive, latest methods usually fuse multiple hand-crafted features, such as Gabor, LBP, TPLBP as in [28,29], or learn more efficient features by using Bag-of-Word (BoW) framework [5,4], or turn to deep learning [30]. To measure the similarity of features, distance metric learning methods are popular to enhance discriminability, as in [28,29,31]. Please note that, this paper only focuses on the restrict protocol of LFW, so we do not introduce methods depending on additional external dataset.

The remaining part of this paper is organized as follows. Section 2 presents the details of the proposed SELD, including the detailed description on the whole pipeline and three component modules. Section 3 discusses the fusion of multiple different partition modes, and the combination of Distance Metric Learning and SELD. Results and analysis of comprehensive experiments on LFW are presented in Section 4, followed by discussion and conclusion in the last section.

2. Sparsely encoded local descriptor

In this section, we first give an overview of the proposed SELD. Then we describe its three key components in detail: learning dictionary, encoding image patches and pooling codes. Finally, a discussion of WPCA is given.

2.1. Overview

SELD is essentially an enhanced texton-based method. It aims to learn robust local descriptors from face images. The overall schema of the proposed method is illustrated in Fig. 1. As shown in the figure, before extracting the SELD features, we first roughly align face images by fixing the eyes at the same position for all the face images, and then filter them with a Difference of Gaussian (DoG) so as to remove both high-frequency noises and low-frequency illumination variations. To preserve more texton information, we pixel-wisely sample raw image patches from the images by a pre-defined template. Each raw patch is vectorized into an intensity vector to form the original feature, which is then sparsely encoded into a higher level feature vector using an offline-learned over-complete dictionary (detailed in Section 2.2).

With the above sparse codes computed, the face image is spatially partitioned into a number of cells (or blocks), and the code vectors of all pixels within each cell are sum-pooled together to form a single descriptor for this cell. Finally, in order to suppress the noises of the pooled descriptors, we exploit whitened PCA to project them into a low-dimensional space, which finally results in our SELD.

In the above schema, if different cell-partitioning manners are applied, multiple SELDs can be generated for each face image. Given two face images, we may compute the similarity of their corresponding SELDs in the same cell-partition. The similarity scores from multiple partitioning manners can be either accumulated together followed by the simple Nearest Neighbor (NN) classifier for face identification, or fed into an SVM classifier for face verification.

2.2. Dictionary learning with K-SVD

In theory, sparse representation assumes a signal can be recovered from a very limited number of atoms contained in an over-complete dictionary. Thus, how to construct a good dictionary that can well support the sparse recovery is very crucial for subsequent representation and classification. To produce the

[1] http://vis-www.cs.umass.edu/lfw/results.html.

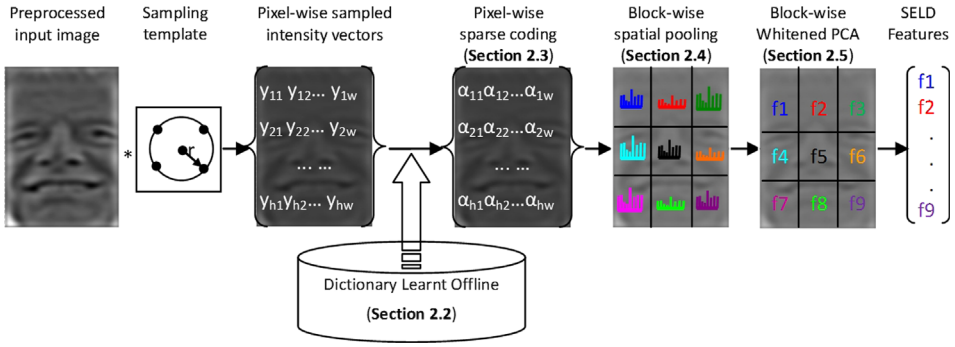


Fig. 1. The proposed framework of sparsely encoded local descriptor.

dictionary, there are two types of methods [32]: mathematical model based methods (e.g. Curvelets, Contourlets, Bandelets, and complex wavelet transforms) and machine learning based methods (e.g. K-means clustering and K-SVD). Recent studies [33] have indicated that the learned dictionary is generally more effective than the manually designed dictionary (e.g. mathematical models) for maintaining sparsity. Therefore, this work only considers machine learning based methods for training dictionary. Among them, one classic method is K-means clustering, which divides all samples into K clusters and assigns each sample to its nearest cluster.

However, K-means clustering assign each sample into only one cluster, which does not match our subsequent Soft Quantization (SQ) or regression based encoding methods. To tackle this problem, we instead use the K-SVD [33] algorithm to learn the dictionary, which naturally represents a sample by several atoms rather than only one, and thus can reduce the representation uncertainty. Below we introduce the K-SVD algorithm.

K-SVD is an iterative method that alternates between sparsely encoding the training samples based on the current dictionary and updating the atoms of the dictionary for better fitting the training data. Formally, given a training set (e.g., sampled patches from training face images) with N samples, $\mathbf{Y}=(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N),\mathbf{y}_i\in\mathbb{R}^n$ to learn an over-complete dictionary matrix $\mathbf{D}=(\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_K),\in\mathbb{R}^{n\times K}(K\gg n)$ contains K prototype signal-atoms. K-SVD's objective function is

$$\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2, \quad s.t. \quad \forall i,\|\mathbf{x}_i\|_0\le T_0, \tag{1}$$

where $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N)$, $\mathbf{x}_i\in\mathbb{R}^K$ is the sparse coefficient vector for the training sample $\mathbf{y}_i$, and $\|\bullet\|_0$ is the l_0 norm.

As mentioned above, the K-SVD optimization is solved by an alternation algorithm, including two stages: in the first stage, $\mathbf{D}$ is fixed, and then the above optimization problem can be solved by some pursuit algorithms. The second stage aims to update the dictionary together with the non-zero coefficients. In this stage, the algorithm updates each column of the dictionary one by one, $\mathbf{d}_k$, and the corresponding coefficients, $\mathbf{x}_R^i$, i.e., the i-th row of $\mathbf{X}$. The objective function (1) can be rewritten as

$$\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2=\|\left(\mathbf{Y}-\sum_{i\neq k}\mathbf{d}_i\mathbf{x}_R^i\right)-\mathbf{d}_k\mathbf{x}_R^k\|_F^2=\|\mathbf{E}_k-\mathbf{d}_k\mathbf{x}_R^k\|. \tag{2}$$

We enforce SVD on $\mathbf{E}_k$, i.e. $\mathbf{E}_k=\mathbf{U}^T\triangle\mathbf{V}$, and then choose the first column of $\mathbf{U}$ and the first column of $\mathbf{V}$ multiplied by $\triangle_{11}$ as the updated $\mathbf{d}_k$, $\mathbf{x}_R^k$ respectively.

In the dictionary learning, an important problem is how many patches should be sampled for training. At first thought, it seems that we should collect as many patches as possible by densely sampling from a large-scale database. In practice, since similar image patches often recur many times inside an image or even across different images, thousands of patches, sparsely sampled from hundreds of training images, are sufficient to learn a robust dictionary, which is validated in our experiments.

2.3. Encoding via non-negative Sparse Coding (nnSC)

After the dictionary is learned by the above K-SVD method, given any input face image, we need to encode its all local patches sampled densely (e.g., pixel by pixel). In previous literature, there are two encoding methods: hard quantization (HQ) and soft quantization (SQ). The former usually chooses one atom as the agent for each input sample (an image patch in our case), which can be formulated as follows:

$$\mathbf{x}\in\{0,1\}^K, x_j=1 \quad \text{if} \quad j=\arg\min_{1\le k\le K}\|\mathbf{y}-\mathbf{d}_k\|^2. \tag{3}$$

Obviously, HQ might lead to visual word ambiguity because two samples with large difference might be assigned to the same atom, which leads to information loss. To tackle the problem, SQ [17,34] seems to be a better choice by assigning a patch fuzzily to several atoms. For this purpose, Gaussian kernel is often used to produce the soft codes. Formally, given an input sample $\mathbf{y}$, its soft codes are computed by

$$x_j=\frac{\exp(-\sigma\|\mathbf{y}-\mathbf{d}_j\|^2)}{\sum_{k=1}^{K}\exp(-\sigma\|\mathbf{y}-\mathbf{d}_k\|^2)}, \quad j=1,\ldots,K, \tag{4}$$

where the parameter σ controls the softness. When $\sigma\to\infty$, the above representation is equivalent to HQ.

Unlike both HQ and SQ, we exploit linear regression to encode the raw patches, which computes the weight coefficients of atoms by using reconstruction. In view of the intrinsic sparse coding mechanism in human visual system [35], we further add sparsity constraints into the regression-based encoding process. In addition, in order to search the atoms with positive correlations rather than negative correlations, non-negativity constraints are imposed on the regression coefficients, which also guarantees pure accumulation without subtraction in the sequent sum-pooling step. Therefore, we finally formulate our encoding method as the following sparse regression model:

$$\min_{\mathbf{x}}\|\mathbf{y}-\mathbf{D}\mathbf{x}\|^2+\lambda\|\mathbf{x}\|_1 \tag{5}$$

$$\text{s.t.}\mathbf{x}\ge 0 \tag{6}$$

The l_1 norm enforces the reconstruction coefficient $\mathbf{x}$ more sparse with the increment of λ. Many optimization algorithms can solve the above model. In this paper, we choose the least angle regression (LAR) [36] for its high efficiency. LAR can be viewed as a kind of democratic version of forward step-wise regression. We name this encoding method as Non-Negative Sparse Coding (nnSC).

Although the above regression procedure is very efficient, it can be still time-consuming when encoding an input image, because we need to solve above regression for a great number of patches (sampled at each pixel). To further accelerate the encoding, an alternative is locality-constrained linear regression, inspired by the locality-constrains linear coding [37]. Specifically, given an image patch, we first find its m nearest neighbors among all the atoms in the dictionary, which are then used for the regression. Formally, the non-zero terms of the code are obtained by solving the following optimization:

$$\min_{\tilde{\mathbf{x}}} \quad \|\mathbf{y}-\mathbf{D}_y\tilde{\mathbf{x}}\|^2, \tag{7}$$

where the subdictionary $\mathbf{D}_y$ only consists of the first m nearest neighbors of $\mathbf{y}$. $\mathbf{D}_y$ is usually full column rank due to $m \ll n$, so the solution of (8) can be computed analytically

$$\tilde{\mathbf{x}} = (\mathbf{D}_y^T\mathbf{D}_y)^{-1}\mathbf{D}_y^T\mathbf{y}. \tag{8}$$

We call this encoding method as Local Least Squares (LLS). Additionally, by adding non-negative constraints to LLS, we can further reach non-negative Local Least Squares (nnLLS), which can be solved by gradient descent algorithms. As has been proved [38], locality naturally leads to sparsity but not necessarily vice versa.

2.4. Sparse codes accumulation via sum-pooling

After the above encoding, each pixel within the input image is associated with a sparse code of K-dimension non-negative vector. However, these codes are position-sensitive, thus not robust to misalignment. To mitigate misalignment and extract more compact features, we partition each face image into several blocks and then integrate the codes within each block to obtain some more robust features. Formally, we sum all the codes within each block $X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_p\}$ to reach a more compact representation of the block, i.e., $\mathbf{z}=\mathbf{x}_1+\mathbf{x}_2+\cdots+\mathbf{x}_p$.

2.5. Dimension reduction via whitened PCA

Even after sum-pooling, the feature is still of high dimensionality due to the use of an over-complete dictionary. To further reduce feature redundancy and find most intrinsic features, we aim at seeking for a compacter representation. A popular method is to use Principle Component Analysis (PCA). However, when a few high-frequent visual words with less discriminability overly dominate the variance, dimension reduction by using PCA might wrongly emphasize those indiscriminating dimensions. For example, smooth facial areas (e.g. the forehead and the cheek) usually contain the same visual words with less discriminability, which however account for too much energy and thus result in features of weak discriminability. To avoid this problem, we resort to WPCA instead. By whitening the variance (via dividing the eignenvalues), WPCA can better suppress the influence of highly frequent but less-discriminant codewords, and thus can extract features of more discriminability.

3. Application to face verification

As a generic descriptor, the above SELD may be used in different tasks. In the case of face verification, as shown in Fig. 2, we may further improve it from two folds: one is to utilize supervised information to SELD, and the other is to adopt multiple block-partitioning modes to generate multiple SELDs. In the following sections, we will introduce them in detail.

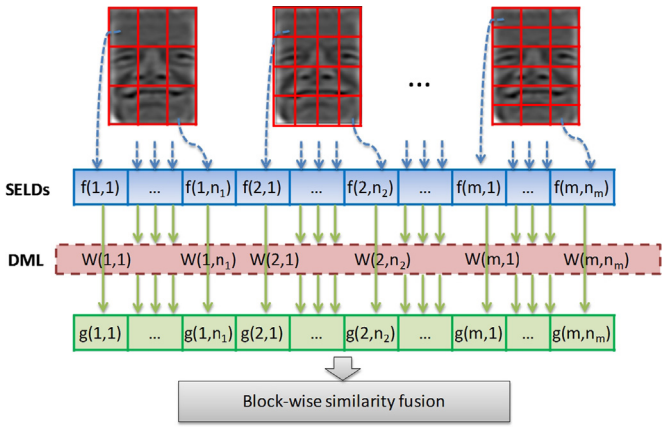


Fig. 2. Illustration of the combination of multiple block-partitioning SELDs with DML.

3.1. Combination with distance metric learning

For face verification task, the goal is to validate whether a pair of face images is from the same subject or not. To utilize the label information, we learn a Mahalanobis distance on SELDs. Formally, given a set of n points $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, we seek a positive definite matrix $\mathbf{A}$ which parameterizes the Mahalanobis distance as follows:

$$\mathrm{dis}_A(\mathbf{x}_i,\mathbf{x}_j)=(\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{A}(\mathbf{x}_i-\mathbf{x}_j) \tag{9}$$

Generally, the degree of freedom of $\mathbf{A}$ is very high when $\mathbf{x}_i$ is of high dimensionality. Moreover, in some cases, only a small number of samples might be available to train the model, e.g. in the LFW evaluation protocol. Obviously, with the small sample size and high model complexity, over-fitting often occurs in many algorithms. One solution is to reduce model complexity by incorporating some prior information. Thus we use Information-Theoretic Metric Learning (ITML) [39], which converts the problem of distance metric learning to learn an entropic objective with constraints on the Mahalanobis matrix. The objective function is

$$\min_{\mathbf{A}} \quad \mathrm{KL}(p(\mathbf{x};\mathbf{A}_0)\,\|\,p(\mathbf{x};\mathbf{A})) \tag{10}$$

$$\text{s.t.} \quad \mathrm{dis}_A(\mathbf{x}_i-\mathbf{x}_j)\le u, \quad (i,j)\in S, \tag{11}$$

$$\mathrm{dis}_A(\mathbf{x}_i-\mathbf{x}_j)\ge l, \quad (i,j)\in D, \tag{12}$$

where $p(\mathbf{x};\mathbf{A})=1/\delta\exp(-\frac{1}{2}\mathrm{dis}_A(\mathbf{x},\boldsymbol{\mu}))$, δ is a normalizing constant, $\boldsymbol{\mu}$ is the mean, KL is the K–L divergence. The set S includes the same-person pairs, D consists of the different-person pairs. u and l are the upper and lower bound respectively, $\mathbf{A}_0$ is an initial matrix, e.g. the identity matrix.

3.2. Combination of multiple block-partitioning SELDs

The face region and component technique [5] have demonstrated promising performance for face verification due to their preservation of spatial layout and robustness to local variations. However, one partitioning mode might segment a whole component into several parts. For instance, the nose component may be divided into two blocks when the 1×2 (horizontal $\times$ vertical) partition is imposed on a face image, but the 1×3 partition may crop the nose into the second block. To handle this problem and make full use of the structural information of different face regions, we simply utilize multiple different partitioning modes for each face image. Therefore, after spatial pooling on different partitioning blocks, we can always collect multiple SELDs although they might come from different partitioning modes. In addition,

face images of different scales usually contain different texton information, so a multi-scale face model may be further used to capture more identity information.

4. Experiments

In this section, we conduct extensive experiments on the LFW dataset. First, we introduce the LFW dataset and the experimental settings. Next, we cross-validate SELD on three modules respectively: dictionary learning, encoding and pooling. Finally, the effectiveness of combination of multiple block-partitioning SELDs with DML is evaluated.

4.1. Database and experimental settings

LFW [2] database is designed for unconstrained face verification with face images containing complex variations in pose, age, expression, race, illumination, etc. Under the restricted mode, the whole standard testing set consists of ten subsets and each subset contains 300 same-person pairs and 300 different-person pairs. The performance of an algorithm is measured by a 10-fold cross validation procedure. The ROC curve or average recognition rate serves as the evaluation criterion. The original size of each image in LFW is 250×250 pixels. In our experiment, all face images are resized to 112×60 pixels after simply cutting out the center from the roughly aligned images provided by Wolf et al. [40]. In all experiments, the parameters of the DoG filter are set to $\sigma_1=0$ and $\sigma_2=2$, where $\sigma_1=0$ means no filtering. The size of the sampling template is set to 9×9. The sampling step is set to 1, i.e. pixel-wise sampling. The default dictionary size is set to 256. The whole face image is partitioned into 8×4 blocks/cells as the default setting, about 13×13 pixels per block. The WPCA preserves the first 20% dimensions.

4.2. Comparisons on three main modules

As described in Section 2, SELD contains three main modules: dictionary learning, encoding and pooling. To evaluate the proposed SELD, below we will discuss the three modules alternately, and try to find the optimal combination strategy by comparing other classic methods.

4.2.1. Comparisons of different dictionary learning methods

We first compare two classic dictionary learning methods, K-means clustering and K-SVD, under different encoding schemes, as shown in Fig. 3(a). Theoretically, K-means clustering minimizes the within-class scatter matrix or maximizes the between-class scatter matrix. Therefore, when the data is a normal distributed and well separated, the centers of clusters can describe the representatives well. K-SVD directly formulates sparse constrains to its object function, and thus can match the sparse coding method well. From Fig. 3(a), we can observe that K-SVD is more matched to nnSC while the K-means algorithm is more adaptive to HQ and SQ. However, the performances of K-SVD and K-means seem comparable under the same encoding strategy, which might be attribute to their nearly representation ability on dictionary learning.

Second, we also attempt to train a dictionary by using FERET database [41] and then apply this dictionary to LFW. The comparison results (in Fig. 3(b)) indicate that the performance is almost independent of the choice of training set because a large amount of repetitive image patches often occur in different face databases.

4.2.2. Comparisons of varying encoding methods

Firstly, we compare nnSC with the previous hard quantization and soft quantization methods under the dictionary learning with K-SVD. By varying the dictionary size, the mean accuracies of the three methods are reported in Fig. 4, where the block-wise cosine scores are accumulated into the final similarity (Fig. 4(a)) or directly fed into the SVM classifier to predict the similarity (Fig. 4(b)). As in the two figures, we can find that the sparse coding based method outperforms the traditional soft quantization method, which is superior over the hard quantization method due to the uncertainty for the latter. Meanwhile, we can find that the accuracies are further promoted with the increase of dictionary size, which may be attributed to the stronger representation ability of the larger size dictionary.

Secondly, since nnSC belongs to a linear regression model, we may substitute nnSC for other reconstruction methods in SELD, such as least squares (LS) and sparse coding (SC) without non-negative constraints. The comparison results are reported in Fig. 5 (a). Note that, the direct sum-pooling on the codes generated from LS and SC largely degrades the performance in our experiments due to the trade-off between negative and positive codes, thus we use absolute codes in pooling for LS and SC. Besides, as a substitute of nnSC, nnLLS can not only achieve a comparable performance, but also have a faster encoding procedure.

4.2.3. Comparisons of different pooling methods

In SELD, we employ sum-pooling on the codes to extract more abstract features. Except sum-pooling, another classic pooling method is max-pooling, which is often used to integrate sparsely

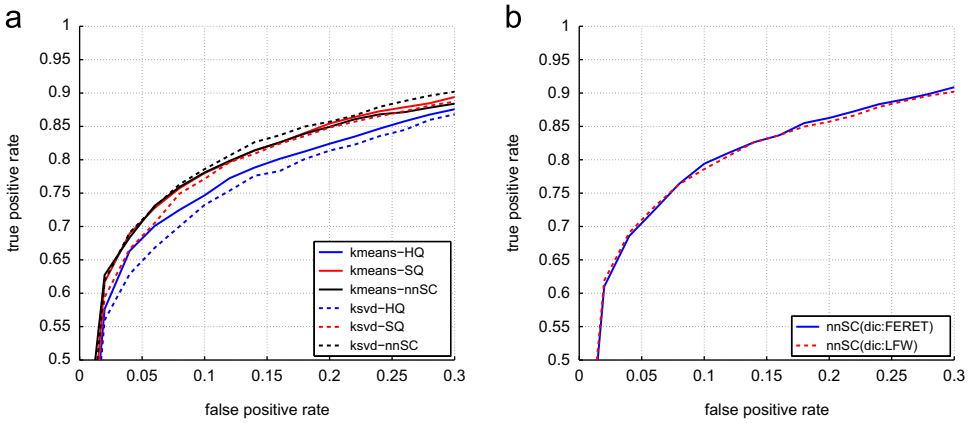


Fig. 3. Performance comparison of dictionary learning methods. (a) Comparisons on two typical dictionary learning methods: K-means clustering and K-SVD, with different encoding schemes. (b) Comparisons of different dictionaries trained on different datasets by K-SVD.

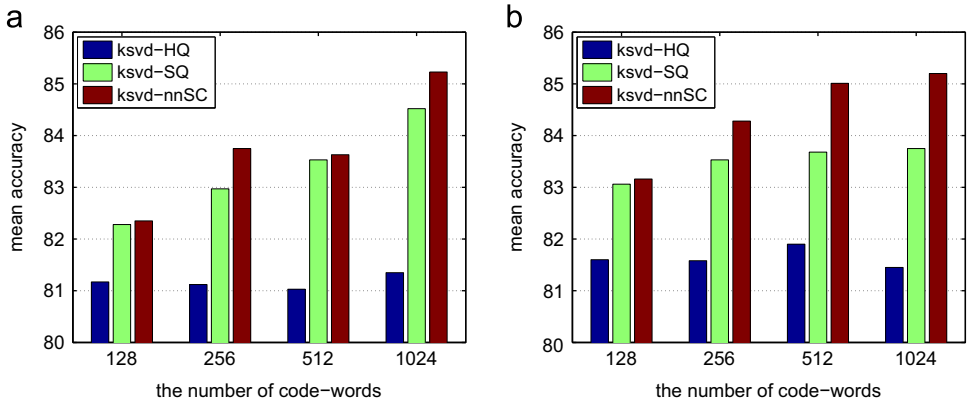


Fig. 4. Performance comparisons of HQ, SQ and nnSC. (a) The accumulation of block-wise cosine scores is used for the distance between two face images. (b)The block-wise cosine scores between two images are fed into the SVM classifier.

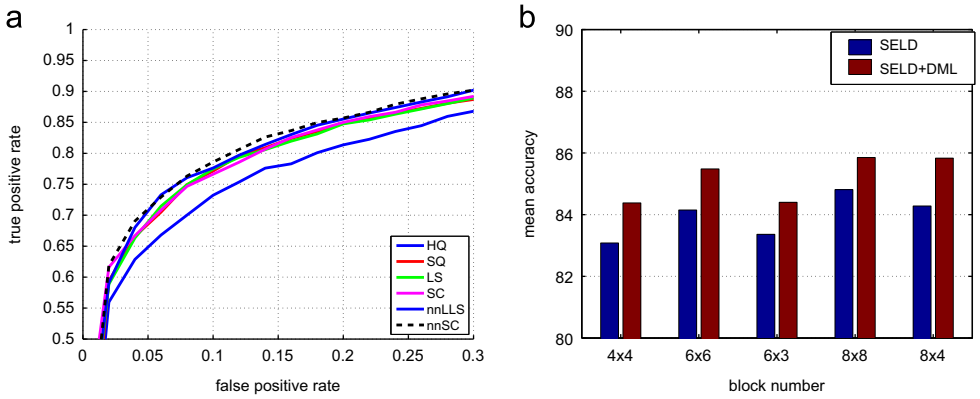


Fig. 5. (a) Performance comparison of different encoding methods. (b) Performance of SELDs with DML. The *X*-axis shows the partitioning modes on the whole face image, e.g. 6×3 means 6 parts in vertical direction and 3 parts in horizontal direction.

sampled codes in object classification [34,37]. To compare the two pooling methods, we conduct six comparison experiments, as shown in Fig. 6. We can find that sum pooling works better than max-pooling for face images in the SELD framework. To illustrate the reason, below we give a theoretical analysis following the recent work [42].

Given two classes C_1 and C_2, we can model the distribution of image patches as conditional distributions $p(f_m|C_1)$ and $p(f_m|C_2)$, and $p(f_s|C_1)$ and $p(f_s|C_2)$, where f_m and f_s are patch features respectively from two pooling methods: max-pooling and sum-pooling. Generally, better separability can be achieved by either making the centers of two class conditional distributions far away or reducing their variances.

For sum-pooling (or average pooling) over a block area (n locations), i.i.d. Bernoulli variables with the mean value α, the sum follows a binomial distribution $B(n,\alpha)$. So the expectation is $\mu_s=\alpha$ and the variance is $\sigma_s^2=\alpha(1-\alpha)/n$. Thus, when the number n of pooled codes becomes larger, sum-pooling is more robust to class separability because the variance decreases with the increase of n.

For max-pooling, however, the variance increases when the number of pooled codes increases. In max-pooling, the expectation is $\mu_m=1-(1-\alpha)^n$ and the variance is $\sigma_m^2=(1-(1-\alpha)^n)(1-\alpha)^n$. So the mean increases monotonically from 0 to 1 with the increase of n. For better separability between two classes, it should uniformize the codes, which contradicts with sparse codes. Moreover, the variance first increases and then decreases with the maximum 0.5 at $\log(2)/|\log(1-\alpha)|$. Consequently, in the case of face image, max-pooling is not as efficient as sum-pooling due to densely sampling codes (i.e. a larger n).

4.3. Combining multiple SELDs with distance metric learning

As described in Sections 3.1 and 3.2, we can apply SELD for face verification by utilizing label information and multiple block-partitioning modes. Below we first conduct the experiment of SELD combined with DML, then try to fuse multiple SELDs from different block-partitioning modes, and finally give a competing performance on LFW dataset.

As a descriptor, SELDs can be concatenated with those machining learning methods to further promote the performance by using supervised information. In the case of face verification, we employed the DML method [39] as introduced in Section 3.1. For DML, $\mathbf{A}_0$ is initiated to be the identity matrix, which corresponds to Euclidean distance. After sorting the similarities on the training set with ascending trend, the values at 10%, 90% are assigned to u and l respectively. The results are reported in Fig. 5(b) with five different block partitioning modes: 4×4, 6×6, 6×3, 8×4, 8×8. From the figure, we can find DML gets a promotion of more than one percent.

In addition, to make full use of spatial structures, we may fuse the similarities under different block-partitioning modes across different scales. As reported in Fig. 7(a), we crop face images into two scales, 110×60 and 75×40 pixels, and then partition each face image into five modes: 4×4, 6×6, 6×3, 8×4, 8×8. The fusion (by the SVM classification) can greatly improve the performance, about 4 percent at false positive rate 0.1.

Finally, we provide the comparisons with the state-of-the-art methods on LFW database in Table 1. It can be seen that our method is comparable with the current best methods. However, the most state-of-the-art methods combined multiple local

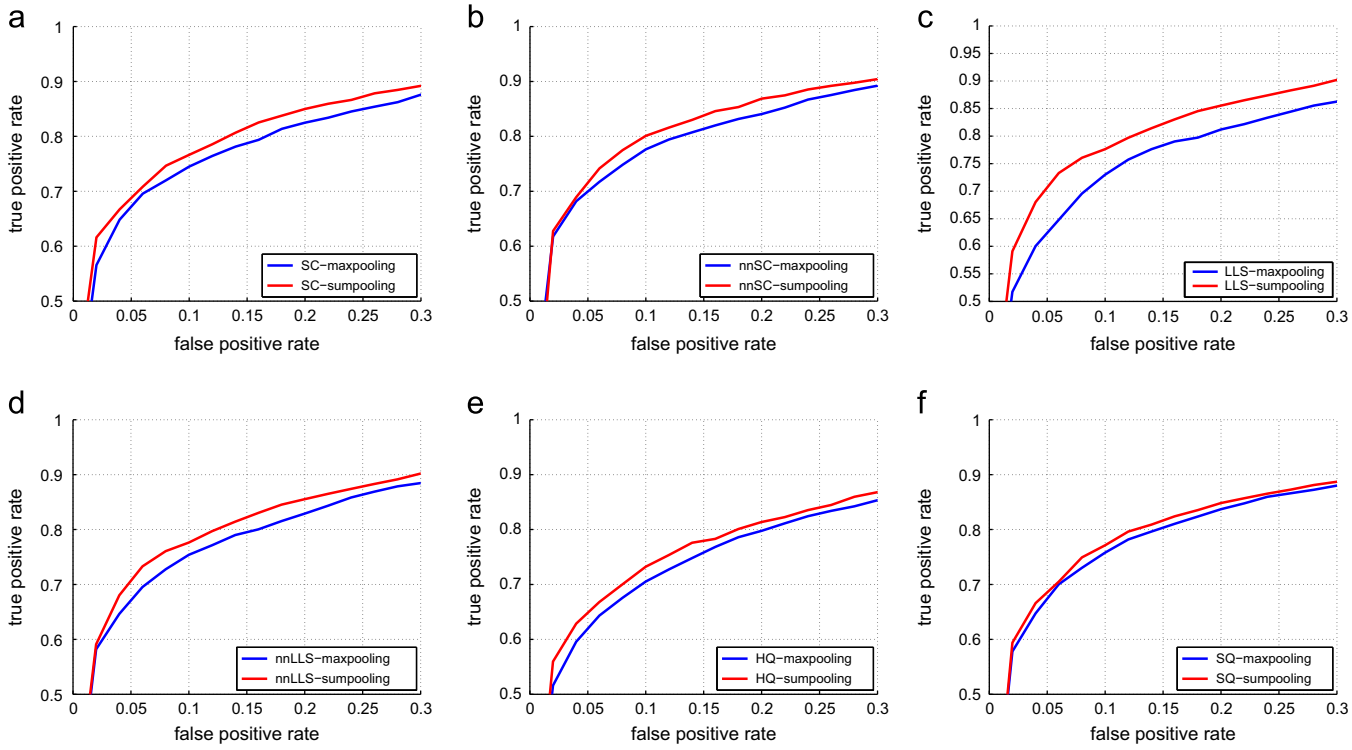


Fig. 6. Performance comparison of different combinations of encoding methods with pooling methods.

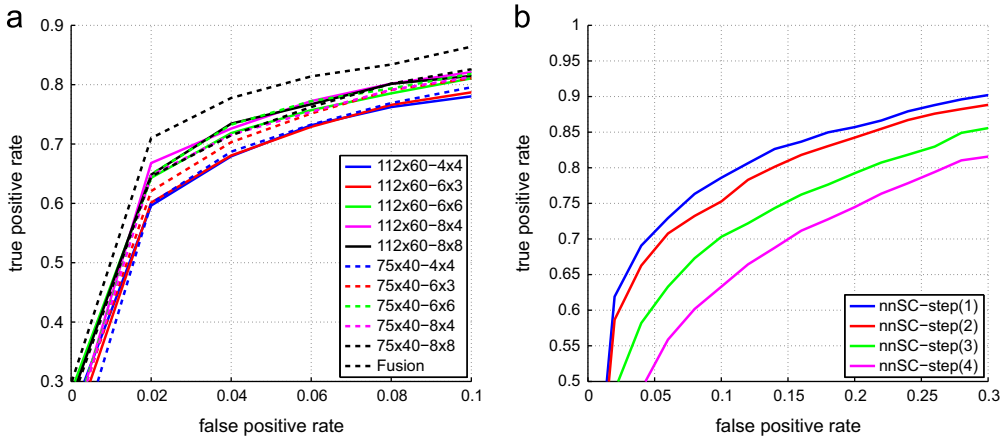


Fig. 7. (a) Performance of fusion on multiple block-wise SELDs. (b) Mean accuracy with different sampling steps. The sampling step is set to 1,2,3,4 pixels.

Table 1
Face verification comparison with the state-of-the-art methods on the LFW benchmark under restrict protocol (mean accuracy).

Method	Mean accuracy (%)
V1-like/MKL funneled [45]	79.35
Hybrid, aligned [46]	83.98
LDML, funneled [31]	79.27
Attibute and Simile classifiers [47]	85.29
Multiple LE + comp [5]	84.45
CSML + SVM, aligned [28]	88.00
DML-eig combined, funneled aligned [29]	85.65
SELDs+DML	**88.40**

features, e.g. "CSML+SVM" [28] employed three types of features: intensity, LBP and Gabor, and used DML with the cosine distance. In addition, it is worth pointing out that we do not compare with Associate-Predict [43] and Tom-vs-pet [44] which use an external data set and thus do not follow the standard protocol.

5. Discussion and conclusion

As shown in the above experiments, the proposed SELD is impressively better than the similar methods based on hard or soft quantization. So, what is the source of gain of the performance? To answer this question, we need to analyze the main differences between our method and previous ones. As mentioned above, the main differences lie in several folds.

First, pix-wise sampling collects more invariant characteristics of one subject. As in Fig. 7(b), the accuracy rapidly decreases with the sampling step increasing. In theory, the dense sampling leads to better separability for pooled features because the variance of

conditional distributions decreases with the increase of samplings (refer the details in Section 4.2.3).

Second, sparse regression chooses multiple visual words to reconstruct each patch and thus avoids the ambiguity of representation. Compared with hard quantization which assigns the single nearest atom as the agent, the traditional soft quantization avoids the uncertainty by weighting the local neighbors. Different from HQ and SQ, SELD computes the contributions of atoms by a regression model, where the atom with more contribution is endowed with a larger weight.

Third, sum-pooling not only provides the statistical information of one subject but also weakens the effect of misalignment. Theoretical analysis (in Section 4.2.3) demonstrates that sum-pooling is more suitable for pooling pixel-wise codes than max-pooling.

Besides, in the task of face verification, we can reach some additional conclusions: (1) the choice of training data seems not crucial to dictionary learning due to patches repetition across different datasets; (2) the dictionary learnt from K-means clustering has a comparable performance with that from K-SVD for those classic encoding methods; and (3) compared with sparse coding, local linear regression not only achieves a comparable performance but also has a faster encoding procedure.

Furthermore, by combining with DML, multiple block-partitioning SELDs achieve a competitive accuracy against the state-of-the-art methods on LFW database. Nevertheless, SELD is not limited to face verification, and also can be used in object classification.

Acknowledgments

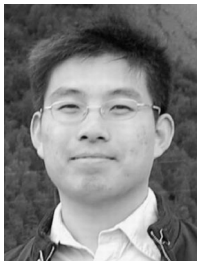
The work is partially supported by Natural Science Foundation of China under contracts Nos. 61025010, 61222211, 61202297, and 61390511.

References

[1] W. Zhao, R. Chellappa, P. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Comput. Surv. (CSUR) 35 (4) (2003) 399–458.
[2] 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.
[3] Z. Cui, S. Shan, X. Chen, D. Zhang, Sparsely encoded local descriptor for face recognition, in: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, 2011, pp. 149–154.
[4] Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, Fusing robust face region descriptors via multiple metric learning for face recognition in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3554–3561.
[5] Z. Cao, Q. Yin, X. Tang, J. Sun, Face recognition with learning-based descriptor, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2707–2714.
[6] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[7] B. Zhang, S. Shan, X. Chen, W. Gao, Histogram of gabor phase patterns (HGPP): a novel object representation approach for face recognition, IEEE Trans. Image Process. 16 (1) (2007) 57–68.
[8] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang, Local gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition, in: IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 786–791.
[9] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[10] G. Zhao, M. Pietikainen, Local binary pattern descriptors for dynamic texture recognition, in: International Conference on Pattern Recognition, vol. 2, 2006, pp. 211–214.
[11] X. Tan, B. Triggs, Enhanced local texture feature sets for face recognition under difficult lighting conditions, in: Analysis and Modeling of Faces and Gestures, Springer, 2007, pp. 168–182.
[12] J. Daugman, et al., Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, Opt. Soc. Am. A: Opt. Image Sci. 2 (1985) 1160–1169.
[13] A. Albiol, D. Monzo, A. Martin, J. Sastre, Distinctive image features from scale-invariant keypoints, Pattern Recognit. Lett. 29 (2008) 1537–1543.
[14] X. Meng, S. Shan, X. Chen, W. Gao, Local visual primitives (LVP) for face modeling and recognition, in: International Conference on Pattern Recognition, 2006, pp. 1–12.
[15] T. Ahonen, M. Pietikäinen, Image description using joint distribution of filter bank responses, Pattern Recognit. Lett. 30 (4) (2009) 368–376.
[16] H. Jégou, M. Douze, C. Schmid, On the burstiness of visual elements, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1169–1176.
[17] J. van Gemert, C. Veenman, A. Smeulders, J. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.
[18] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[19] Q. Zhang, B. Li, Discriminative K-SVD for dictionary learning in face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2691–2698.
[20] M. Yang, L. Zhang, Gabor feature based sparse representation for face recognition with gabor occlusion dictionary, Eur. Conf. Comput. Vis. (2010) 448–461.
[21] Y. Xu, D. Zhang, J. Yang, J. Yang, A two-phase test sample sparse representation method for use with face recognition, IEEE Trans. Circuits Syst. Video Technol. 21 (9) (2011) 1255–1262.
[22] Y. Xu, W. Zuo, Z. Fan, Supervised sparse representation method with a heuristic strategy and face recognition experiments, Neurocomputing 79 (2012) 125–131.
[23] J. Yin, Z. Liu, Z. Jin, W. Yang, Kernel sparse representation based classification, Neurocomputing 77 (1) (2012) 120–128.
[24] Z. Cui, S. Shan, H. Zhang, S. Lao, X. Chen, Structured sparse linear discriminant analysis, in: IEEE International Conference on Image Processing, 2012, pp. 1161–1164.
[25] Z. Cui, H. Chang, S. Shan, B. Ma, X. Chen, Joint sparse representation for video-based face recognition, Neurocomputing 135 (2014) 306–312.
[26] S. Xie, S. Shan, X. Chen, X. Meng, W. Gao, Learned local gabor patterns for face representation and recognition, Signal Process. 89 (12) (2009) 2333–2344.
[27] C. Liu, H. Wechsler, Evolutionary pursuit and its application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 22 (6) (2000) 570–582.
[28] H. Nguyen, L. Bai, Cosine similarity metric learning for face verification, Asian Conf. Comput. Vis. (2011) 709–720.
[29] Y. Ying, P. Li, Distance metric learning with eigenvalue optimization, J. Mach. Learn. Res. 13 (2012) 1–26.
[30] G.B. Huang, H. Lee, E. Learned-Miller, Learning hierarchical representations for face verification with convolutional deep belief networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2518–2525.
[31] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? Metric learning approaches for face identification, in: IEEE International Conference on Computer Vision, 2009, pp. 498–505.
[32] R. Rubinstein, A. Bruckstein, M. Elad, Dictionaries for sparse representation modeling, Proc. IEEE 98 (6) (2010) 1045–1057.
[33] M. Aharon, M. Elad, A. Bruckstein, K-SVD: design of dictionaries for sparse representation, Proc. SPARS 5 (2005) 9–12.
[34] Y. Boureau, F. Bach, Y. LeCun, J. Ponce, Learning mid-level features for recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2559–2566.
[35] B. Olshausen, D. Field, et al., Sparse coding with an overcomplete basis set: a strategy employed by VI? Vis. Res. 37 (23) (1997) 3311–3326.
[36] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[37] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360–3367.
[38] J. Yang, L. Zhang, Y. Xu, J.-y. Yang, Beyond sparsity: the role of l_1-optimizer in pattern classification, Pattern Recognit. 45 (3) (2012) 1104–1118.
[39] J. Davis, B. Kulis, P. Jain, S. Sra, I. Dhillon, Information-theoretic metric learning, in: International Conference on Machine Learning, 2007, pp. 209–216.
[40] L. Wolf, T. Hassner, Y. Taigman, Similarity scores based on background samples, Asian Conf. Comput. Vis. (2010) 88–97.
[41] P. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[42] Y. Boureau, J. Ponce, Y. LeCun, A theoretical analysis of feature pooling in visual recognition, in: International Conference on Machine Learning, 2010, pp. 111–118.
[43] Q. Yin, X. Tang, J. Sun, An associate-predict model for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 497–504.
[44] T. Berg, P. Belhumeur, Tom-vs-pete classifiers and identity-preserving alignment for face verification, in: British Machine Vision Conference, vol. 1, 2012, p. 5.
[45] N. Pinto, J. DiCarlo, D. Cox, How far can you get with a modern face recognition test set using only simple features?, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 2591–2598.
[46] Y. Taigman, L. Wolf, T. Hassner, Multiple one-shots for utilizing class label information, in: BMVC, 2009, pp. 1–12.
[47] N. Kumar, A. Berg, P. Belhumeur, S. Nayar, Attribute and simile classifiers for face verification, in: IEEE International Conference on Computer Vision, 2009, pp. 365–372.

Zhen Cui received the B.S. degree from Shandong Normal University, Jinan, China, in 2004, and then the M.S. degree from Sun Yatsen University, Guangzhou, China, 2006. Currently, he is a Ph.D candidate in Institute of Computing Technology, Chinese Academy of Science since 2009, and also a Lecturer in Huaqiao University since 2006. From June 2012 to December 2012, he was a Research Associate in Nanyang Technological University, Singapore. His research interests cover pattern recognition and computer vision, especially face recognition and image super resolution based on recently emerged theories, such as Sparse Coding, Manifold Learning, and Deep Learning.

Shiguang Shan received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences, Beijing, in 2004. He has been with ICT, CAS since 2002 and has been a Professor since 2010. He is also the Executive Director of the Key Lab of Intelligent Information Processing of CAS. His research interests cover image analysis, pattern recognition, and computer vision. He is focusing especially on face recognition related research topics. He received the China's State Scientific and Technological Progress Awards in 2005 for his work on face recognition technologies.



Ruiping Wang received the B.S. degree in applied mathematics from Beijing Jiaotong University, Beijing, China, in 2003, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2010. He was a postdoctoral researcher with the Department of Automation, Tsinghua University, Beijing, from July 2010 to June 2012. He also spent one year working as a Research Associate with the Computer Vision Laboratory, Institute for Advanced Computer Studies (UMIACS), at the University of Maryland, College Park, from November 2010 to October 2011. He has been with the faculty of the Institute of Computing Technology, Chinese Academy of Sciences, since July 2012, where he is currently an Associate Professor. His research interests include computer vision, pattern recognition, and machine learning.

Lei Zhang received the B.S. degree in 1995 from Shenyang Institute of Aeronautical Engineering, Shenyang, P.R. China, the M.Sc. and Ph.D degrees in Control Theory and Engineering from Northwestern Polytechnical University, Xi'an, P.R. China, respectively in 1998 and 2001. From 2001 to 2002, he was a research associate in the Dept. of Computing, The Hong Kong Polytechnic University. From January 2003 to January 2006 he worked as a Postdoctoral Fellow in the Dept. of Electrical and Computer Engineering, McMaster University, Canada. In 2006, he joined the Dept. of Computing, The Hong Kong Polytechnic University, as an Assistant Professor. Since Sept. 2010, he has been an Associate Professor in the same department. His research interests include Image and Video Processing, Computer Vision, and Pattern Recognition and Biometrics. He has published about 200 papers in those areas. He is currently an Associate Editor of IEEE Transactions on CSVT and Image and Vision Computing. He was awarded the 2012–13 Faculty Award in Research and Scholarly Activities. More information can be found in his homepage http://www4.comp.polyu.edu.hk/~cslzhang/.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1988, 1991, and 1994 respectively. He was a Professor with the HIT from 1999 to 2005 and was a Visiting Scholar with Carnegie Mellon University from 2001 to 2004. He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, since August 2004. His research interests include image processing, pattern recognition, computer vision, and multimodal interface. He has received several awards, including the China's State Scientific and Technological Progress Award in 2000, 2003, 2005, and 2012 for his research work.