Journal of Electronic Imaging

JElectronicImaging.org

Multiple scales combined principle component analysis deep learning network for face recognition

Lei Tian
Chunxiao Fan
Yue Ming

SPIE. IS&T imaging.org

Lei Tian, Chunxiao Fan, Yue Ming, "Multiple scales combined principle component analysis deep learning network for face recognition," *J. Electron. Imaging* **25**(2), 023025 (2016), doi: 10.1117/1.JEI.25.2.023025.

Multiple scales combined principle component analysis deep learning network for face recognition

Lei Tian, Chunxiao Fan, and Yue Ming*
Beijing University of Posts and Telecommunications, School of Electronic Engineering, Beijing Key Laboratory of Work Safety Intelligent Monitoring, No. 10 Xitucheng Road, Beijing 100876, China

Abstract. It is well known that higher level features can represent the abstract semantics of original data. We propose a multiple scales combined deep learning network to learn a set of high-level feature representations through each stage of convolutional neural network for face recognition, which is named as multiscaled principle component analysis (PCA) Network (MS-PCANet). There are two main differences between our model and the traditional deep learning network. On the one hand, we get the prefixed filter kernels by learning the principal component of images' patches using PCA, nonlinearly process the convolutional results by using simple binary hashing, and pool them using spatial pyramid pooling method. On the other hand, in our model, the output features of several stages are fed to the classifier. The purpose of combining feature representations from multiple stages is to provide multiscaled features to the classifier, since the features in the latter stage are more global and invariant than those in the early stage. Therefore, our MS-PCANet feature compactly encodes both holistic abstract information and local specific information. Extensive experimental results show our MS-PCANet model can efficiently extract high-level feature presentations and outperform state-of-the-art face/expression recognition methods on multiple modalities benchmark face-related datasets. [DOI: 10.1117/1.JEI.25.2.023025]

Keywords: multiple scales feature representation; convolution neural network; binary hashing; spatial pyramid pooling; two-dimensional face recognition; two-dimensional face verification; three-dimensional face recognition.

Paper 15565 received Jul. 15, 2015; accepted for publication Mar. 22, 2016; published online Apr. 21, 2016.

1 Introduction

During the past decades, face recognition has been applied in many challenging scenarios, such as access control, ID authentication, medical nursing, and watch-list surveillance. However, the intraclass variability caused by illumination, poses, nonrigid deformations, and occlusion make it a long way from a face recognition system being applied in those uncontrolled scenarios. In recent years, numerous efforts are being made to manually design low-level features for face recognition, such as scale-invariant feature transform (SIFT)[1] and local binary pattern (LBP).[2] Though the handcrafted features achieve great success for some controlled scenarios through designing low-level features elaborately, they cannot achieve excellent performance when faced with extreme intraclass variability and uncontrolled scenarios.

Learning features from data itself instead of manually designing features is considered a plausible way to overcome the limitation of low-level features, termed as deep learning (DL). The purpose of DL is to extract higher level features with the hope that they have more robustness when faced with more complicated intraclass variabilities. In other words, the high-level feature can describe more complex intraclass variability than the low-level feature if we train the DL model using a large amount of training samples. The presentative examples of DL are convolutional neural network (CNN) and autoencoder (AE). In order to take advantage of the convolution operation, many deep learning frameworks are proposed based on the convolutional architectures.[3–6] The architecture of CNN generally consists of multiple stages, and each stage comprises of a convolutional filter layer, nonlinear processing layer, and pooling layer. What is more, many variations of CNN model whose filter kernels are prefixed have been proposed, such as ScatNet[7] and principle component analysis (PCA) Network.[8,9] The AEs[10] and its variant, such as sparse AE,[11] contractive AE,[12] and denoising autoencoders (DAE),[13] generally comprises "encoder" and "decoder." The architecture of AEs usually is used for nonlinear transformation.

However, the above algorithms do not generalize as well to the face recognition task, which the intraclass variability is often greater than the interclass variability. So, we cast the subspace learning methods into deep learning model with the hope that a set of discriminative and generalized high-level feature representations can be learned. Figure 1 shows the workflow of our two-stage MS-PCANet model. Similar to traditional CNN model, in our model, each stage contains convolutional filter layer, nonlinear processing layer, and feature pooling layer. In the convolutional layer of each stage, the input images are convoluted with PCA-learned based convolutional kernels. Different from traditional CNN model, each feature map in the latter stage is just locally connected with "only one" feature map in the early stage, instead of with several early feature maps. In the nonlinear processing layer, the binary hashing method

*Address all correspondence to: Yue Ming, E-mail: myname35875235@126.com

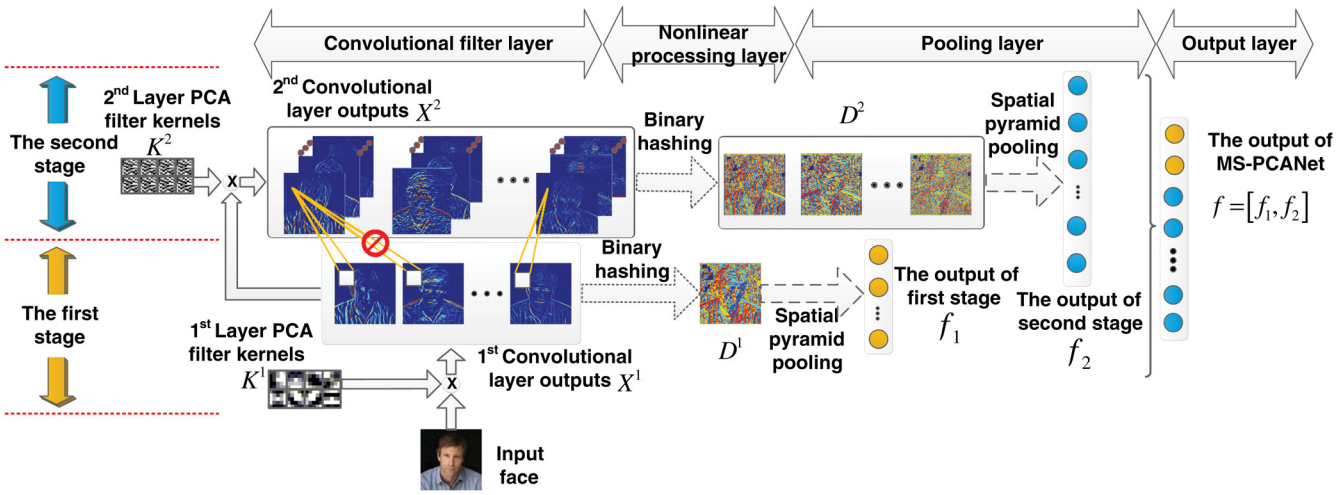


Fig. 1 An illustration of the architecture of our MS-PCANet model.

is used as activation function. Similar to the idea of ReLu,[3] the process of binarization makes some neuron's output equal to 0 and introduces sparse property into our DL model. In the feature pooling layer, the multiscale block-wise histogram features are extracted by the spatial pyramid pooling (SPP) method. In the output layer, it is well known that high-level features which are learned from latter stage tend to be more global and low-level features which are learned from early stage tend to be more local. Therefore, the multiple scales combined feature representations contain both holistic information and local information. We cumulate the output features of several stages as the final output of our model.

The proposed MS-PCANet improves the discriminative and generalization ability of DL model on face recognition tasks. The contributions of this work are summarized as follows:

1. An unsupervised subspace learning method is applied to initialize the filter kernels of our model, the binary hashing are used to nonlinearly process the feature maps. We also introduce the idea of multiscale combined into pixel-block level by using SPP method, the coarser scale can extract macrostructures of the face and the finer scale can extract microstructures of the face. With the multiscaled features on the pixel-block level, more comprehensive face information is explored.
2. We cumulate the output features of several stages as one final feature output of our model. The multiple scales combined representation is critical to feature learning because it can reduce the possible information loss due to successive convolution and down-sampling operations. By incorporating the multistage output features, the multiscaled and discriminative representations of the face image are learned.
3. Extensive experiments are conducted on widely used face datasets to demonstrate the efficiency of our proposed MS-PCANet model when faced with multiple modalities face image [e.g., two-dimensional (2-D) face image and three-dimensional (3-D) face image]. There are various intraclass variabilities in 2-D face databases (such as pose, expression, illumination, occlusion, and aging) and 3-D face database (such as expression, occlusion, face region incomplete, and pixel corruption). In fair experimental comparison, our method is not only better than the state-of-the-art methods on 2-D face recognition/verification scenarios, but also better than most state-of-the-art 3-D methods on 3-D face recognition scenarios.

The rest of the work is organized as follows. Section 2 briefly reviews some recent related work. Then, we detail the architectures of our proposed MS-PCANet model in Sec. 3. We investigate the performance of our model on face benchmark databases in Sec. 4 and conclude this work in Sec. 5.

2 Related Work

Since the topics covered in face recognition literature are numerous, we focus on two most-related aspects: low-level features and high-level features.

2.1 *Low-Level Features*

The low-level features also are termed as handcrafted features, they can achieve excellent performance through elaborately designed methods. Low-level features are generally comprised of the domain knowledge of specific face recognition tasks and cannot be simply adopted to new conditions. The representative examples are LBP-based methods[14] and its variations[15,16] and Gabor.[17–19]

The LBP method has been widely used due to its robustness and efficient computation. Its many variants produce more discriminative LBP's code than the original method for face recognition tasks. In discriminant face descriptor (DFD)[15] and compact binary face descriptor (CBFD),[16] the pixel difference vectors are computed and the compact features are obtained by projecting these pixel difference vectors into low-dimensional vectors. In local quantized patterns (LQP),[14] researchers adopt vector quantization to encode the local pattern values. The above LBP-based features achieve excellent performance on constricted face database. Especially, CBFD and DFD achieve 93.50% and 91.80% accuracy on *dup1* subsets of FERET database,

their excellent performances benefit from their LBP-based descriptors which are obtained by learning discriminative codes from raw pixels. In recent years, several Gabor phase-based features show competitive results and complementary attributes to Gabor magnitude-based features. The binary relationship between neighboring Gabor phases are used for better face representation, the local gabor XOR patterns feature[17] combine with block-Fisher's linear discriminant achieves 92% and 91% accuracy on *dup1* and *dup2* subsets, respectively. What is more, the bag-of-words (BoW) framework[18] and the ordinal measures methods[19] are also adopted to Gabor-based algorithms.

However, with increasing of the complexity of intraclass variability and the advent of uncontrolled face recognition scenarios, some problems of low-level features were revealed, such as poor performance on uncontrolled face recognition scenarios and the sensitivity with extreme intraclass variability.

2.2 *High-Level Features*

Learning features from the data itself instead of elaborately designing methods is considered a way to avoid the problems of handcrafted features. Researchers hope that higher level features can represent the abstract semantics of the data and the influence of intraclass variability can be weakened through deep network architecture. In other words, the features learned from deep network are expected to provide more invariance to intraclass variability. Two examples of such methods are CNNs[3–6,20] and AEs.[10,11]

The CNN model usually consists of convolutional filter layer, nonlinear processing layer, and feature pooling layer. In the convolutional filter layer, the parameter of filter kernels and additive bias are randomly initialized, and are iteratively updated by stochastic gradient descent (SGD). In the nonlinear processing layer, the Sigmoid and ReLu[3] function are used. In the feature pooling layer, the max pooling, or mean pooling are used to reduce the resolution of the feature map. The architecture of CNN has been widely applied in face verification task, such as DeepFace,[4] DeepID,[5] DeepID2,[6] and hybrid ConvNet-RBM.[20] The DeepID (DeepID2) and DeepFace feature both are learned from deep network (at least more than three stages). Both works achieve 97.45% (99.15%) and 97.35% performance on LFW by using a large outside database to train their model. Sun et al.[20] proposes a hybrid ConvNet-RBM model to directly learn identity similarities from raw pixels of face pairs.

Similarly, the architecture of deep AEs also widely applied in various visual tasks, which consist of the encoder and the decoder. The encoder straightly maps the input vector into high dimensional feature space and decoder maps from the feature space back into input space and produces a reconstruction signal whose reconstruction errors are both small on train and test examples. The work[11] simplify and speed up the training process of AEs by limited memory Broyden Fletcher Goldfarb Shanno and conjugate gradient with line search. The stacked progressive AE (SPAE)[10] is used to model the complex nonlinear transform from the nonfrontal face images to frontal ones. And it achieves excellent performance when faced with larger pose variations.

However, the architecture of CNN leads to having a large number of parameters, therefore, the performance of CNN framework critically depends on expertise of parameter tuning. What is more, the last hidden layer of CNN model is fully connected to only the last convolutional layer, it may become the bottleneck for information propagation. Similar to the fully connected layer in the CNN model, the AE is also fundamentally incapable of learning discriminative features. It just nonlinearly transforms input feature into feature space, and maps from feature space back into input space. Therefore, in this work, we propose a multiple scales combined DL model to solve the above problems. In order to solve the existing problems of traditional CNN model, we precompute the filter kernels by PCA and cumulate several stages' output feature as the final output feature. And in order to solve the learning disabled problem in AE model, we learn multiscaled discriminative feature representations hierarchically through the architecture of CNN.

3 Architecture of Our Multiscaled Principle Component Analysis Network Model

Our two-staged MS-PCANet contains two convolutional layers to extract features hierarchically, followed by the nonlinear processing layer and feature pooling layer. Let $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_N}]$ of size $m_1 \times m_2$ $(m_1 \times m_2 = m)$ be the training set containing N samples.

3.1 *Cascaded Convolutional Filter Layer*

Inspired by the observations that prefixed filter kernels can extract as effective as features which are extracted from traditional deep CNN, we learn the filter kernels by cascaded-PCA, instead of randomly initializing them and updating them by the SGD algorithms in the convolutional filter layer.

The size of convolutional filter kernels, i.e., the size of extracted eigenvectors, is set to $k_1 \times k_2$ at all stages. In order to learn the convolutional filter kernels, we collect all overlapping $k_1 \times k_2$ patches $\boldsymbol{A_i} = [\boldsymbol{a_1}, \boldsymbol{a_2}, \cdots, \boldsymbol{a_j}, \cdots, \boldsymbol{a_m}]$ from the i'th image, where $\boldsymbol{a_j}$ denotes the j'th mean-removed patch in the i'th original image. By repeating the same process for all training samples, we get

$$\boldsymbol{A} = [\boldsymbol{A_1}, \boldsymbol{A_2}, \cdots, \boldsymbol{A_N}] \in R^{k_1 k_2 \times Nm}. \tag{1}$$

Then, we extract the leading principal component from data matrix $\boldsymbol{A}$, which are used as the filter kernels of convolutional layer. Since the shape and structure of different subject's face are similar, the interclass variability is often smaller than the intraclass variability. Therefore, it is unnecessary that we learn the supervised-based filter kernels by incorporating the information of class labels and we just need the principal component of training samples to "filter" input images. The PCA aims to find a set of orthogonal basis functions that has maximum variance from the data. Its objective function is that

$$\boldsymbol{v}_{\text{opt}} = \underset{\boldsymbol{v}}{\text{argmax}} \sum_{i=1}^{N} (y_i - \bar{y})^2 = \underset{\boldsymbol{v}}{\text{argmax}}\, \boldsymbol{v}^T C \boldsymbol{v}, \tag{2}$$

where $\boldsymbol{y_i} = \boldsymbol{v}^T \boldsymbol{x_i}, \bar{\boldsymbol{y}} = (1/N) \sum \boldsymbol{y_i}$ and C is the data covariance matrix. Assuming that the number of filter kernels in stage i is L_i, we select the L_i leading eigenvectors from v_{opt} which corresponds to the L_i maximum eigenvalues. Therefore, the filter kernels of stage i are expressed as

$$\boldsymbol{K}^i = \{vec2mat_{k_1\times k_2}[f_{l_i}(\boldsymbol{A}\boldsymbol{A}^T)]\}_{l_i=1}^{L_i}, \tag{3}$$

where $f_l(\boldsymbol{A}\boldsymbol{A}^T)$ denotes the l'th principal component of data matrix $\boldsymbol{A}$, and $vec2mat_{k_1\times k_2}(\boldsymbol{v})$ is a function that maps vector $\boldsymbol{v}$ whose dimension is k_1k_2 to a matrix of size $k_1 \times k_2$. It is a remarkable fact that the input of the second stage is "the convolutional output of the first stage." In other words, the filter kernels of the second stage are learned from the patches of the first stage's feature maps. As it goes, the filter kernels of the third stage are learned from the feature maps that come from the convolutional output of the second stage.

Then, our model follows the basic operation flow of a traditional CNN, the input image is convoluted with PCA-learned filter kernels, and the filter outputs of the first stage are expressed as

$$\boldsymbol{X}_i^1 = \{\boldsymbol{X}_i * \boldsymbol{K}_{l_1}^1\}_{l_1=1}^{L_1}, \qquad i = 1,2,\cdots,N. \tag{4}$$

The boundary of $\boldsymbol{X}_i$ is zero-padded before convolution, so that $\boldsymbol{X}_i^1$ has the same size of $\boldsymbol{X}_i$. Almost repeating the same process as the first stage, we can obtain the convolutional filter output $\boldsymbol{X}_i^2(l_1)$

$$\boldsymbol{X}_i^2(l_1) = \{\boldsymbol{X}_i^1(l_1) * \boldsymbol{K}_{l_2}^2\}_{l_2=1}^{L_2}, \qquad i = 1,2,\cdots,N, \tag{5}$$

where $\boldsymbol{X}_i^1(l_1)$ denotes the l_1'th feature maps of the i'th image in the first stage and $\boldsymbol{K}_{l_2}^2$ denotes the PCA-learned kernels of second stage. It is notable that each feature map in the latter stage [i.e., $\boldsymbol{X}_i^2(l_1)$] is just locally connected with only one feature map in the early stage [i.e., $\boldsymbol{X}_i^1(l_1)$], instead of to be locally connected with several early feature maps. One can simply repeat the above process to build deeper PCA-learned convolutional filter architecture.

Similar to traditional CNN model, we also remove patch-mean from each local patch, so that the all local patches are moved to be centered around the original data space. So the learned kernels can better represent variations of the data. However, the filter kernels of our model are learned from the patch of original data directly, we need not iteratively update them by SGD. The training efficiency of our model is improved dramatically. And extensive experiments demonstrate the PCA-learned kernel is even better than BP-learned kernel, though the architecture of our model is not as deep as CNN.

3.2 *Nonlinear Processing Layer*

In the nonlinear processing layer, we use the binary hashing method as activation function to introduce nonlinearity and sparsity into our model. In mathematical terms, to make our model approach any linear/nonlinear function, we use as few layers as possible. So that the output of cascaded-PCA convolutional layer is more suitable for binary hashing method to process, rather than "sigmoid" or "ReLu" function.

In order to extract multiscaled features and feed them to a supervised classifier, we choose all convolutional outputs of each stage (i.e., $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$) as the input of next nonlinear processing layer. Different from the traditional CNN model, the feature maps of the previous stage [i.e., $\boldsymbol{X}_i^1(l_1)$] is convoluted with each filter kernels of the next stage (i.e., $\boldsymbol{K}_{l_2}^2$). We consider that each output feature map contains different local discriminative information of image, and "dropout"[3] is not encouraged to be used in our model, where dropout sets random feature maps to zero during training. Therefore, the number of the second stage's output is L_1L_2.

Then, the question that emerges is how we can represent the feature map of each stage [i.e., $\boldsymbol{X}_i^1(l_1)$ and $\boldsymbol{X}_i^2(l_1)$]. Inspired by "binary hash code" for image retrieval and existing "LBP-based" representation for face analysis, we nonlinearly process the feature maps by binary hash method. Taking the feature maps of the first stage as an example, we first convect the feature maps into binary image by "Heaviside" function

$$\text{Heaviside}(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x \le 0) \end{cases}.$$

Then, we sum L_1 binary code with their bits weighted, and obtain one decimal-valued image

$$\boldsymbol{D}_i^1 = \sum_{l_1=1}^{L_1} 2^{l_1-1} \bullet \max[0, \boldsymbol{X}_i^1(l_1)], \tag{6}$$

whose pixel value range is $[0, 2^{L_1}-1]$. Almost repeating the same process as the first stage, we binarize and sum L_2 feature maps $\boldsymbol{X}_i^2(l_1)$, where $\boldsymbol{X}_i^2(l_1)$ denotes the filter output corresponding to the l_1'th feature map. Therefore, we can get

$$\boldsymbol{D}_{i,l_1}^2 = \sum_{l_2=1}^{L_2} 2^{l_2-1} \bullet \max[0, \boldsymbol{X}_i^2(l_1)], \qquad l_1 = 1,2,\cdots,L_1, \tag{7}$$

where $\boldsymbol{D}_{i,l_1}^2$ denotes the hashing image corresponding to the l_1'th feature map of the i'th image.

In the nonlinear processing layer, we do not follow the common wisdom in building a traditional CNN model. There is no nonlinear processing layer until the end of "all" convolutional filter layer. In other words, it is completely linear with the overall process prior to the first nonlinear processing layer in our MS-PCANet. We had once tried to insert nonlinear function and subsampling layer after each convolutional layer; however, the performance of our model seemed to some degree to decrease, mainly because feature maps are processed by a subsampled operation, which leads to some implicit discriminative information loss.

3.3 *Feature Pooling Layer*

The multiscale concept of our proposed MS-PCANet is not only applied to the layer level, but is also applied to the pixel-block level. In the feature pooling layer, inspired by SPP network (SPP-net)[21] for visual object recognition, we pool each spatial block whose scale is fine-to-coarse (throughout this work the max-pooling is used) for each hashed image $\boldsymbol{D}_i^1$ (or $\boldsymbol{D}_i^2$). The finest pyramid level is similar to the size of filter kernel, this is in fact a "local" pooling operation which can maintain local spatial information. And the coarsest pyramid level has a single bin that covers the entire image; it is in fact a "global" pooling operation which is more like the traditional BOW method. In fact, incorporating the global pooling feature extracted from

the entire cropped face image into the final output feature might degenerate the performance of MS-PCANet model. Since the relevance of local patches of a cropped face image is stronger than a cropped object, the "words" in the BOW model that are applied to visual object recognition are unsuited to describe these face local patches.

Taking one hashed image of first stage $\boldsymbol{D}_i^l$ as an example. Assuming that $\boldsymbol{D}_i^l$ is partition into $b_1 \times b_2$ nonoverlapping blocks, i.e., $b_1 = \text{floor}(m_1/w)$ and $b_2 = \text{floor}(m_2/h)$, where w and h is the size of the finest-scale block. What is more, our experimental results suggest that nonoverlapping between neighboring blocks are appropriate for face image. We concatenate the histogram of all fine-to-coarse blocks into one vector and define it as the final output feature of the first stage f_1. In each spatial bin, we maximally pool the responses of each PCA-based filter. We further define the final output of the second stage f_2. Figure 2 shows the spatial pyramid pooling method.

Different from the CNN model, we extract multiscale blockwise histograms by using SPP method. It already introduces sufficient invariance in our final feature that the use of "binary hashing" plus "multiscale blockwise histogram pooling" after several cascade convolutional layers. This is another reason that we do not insert nonlinear function and subsampling layer after the convolutional layer.

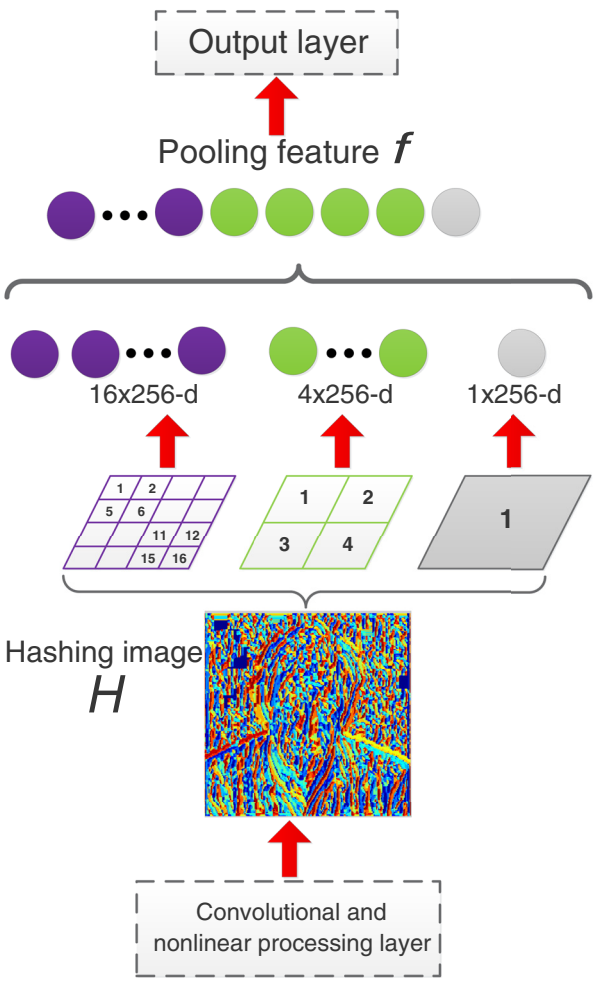


Fig. 2 A structure of an SPP layer, where SPP size is $[4 \times 4, 2 \times 2, 1 \times 1]$.

3.4 *Output Layer*

In the last output layer, we cumulate the histogram feature of multiple stages. Therefore, for each input image X_i $(i = 1,2, \cdots, N)$, the final feature of our proposed MS-PCANet model is expressed as

$$f_i = [f_{i,1}; f_{i,2}] \in \mathbb{R}^{(2^{L_1}+L_1 2^{L_2}) \times (b_1 \bullet b_2 + \cdots + 1)}. \tag{8}$$

In the traditional CNN model, there may be the bottleneck problem that arises from successive subsampling operations along a cascade. The last convolutional layer of CNN model contains too few neurons to describe image's information. Therefore, the multiscale feature representations are learned from several stages, and they are cumulated as the final output feature in our MS-PCANet model. It allows the supervised classifier (or soft-max layer) to use the holistic features and the detailed features simultaneously.

4 Experiments

Now, we evaluate the performance of the proposed MS-PCANet in various face-related tasks, including 2-D face recognition, 2-D face verification and 3-D face/expression recognition.

4.1 *Two-Dimensional Face Recognition on Several Datasets*

In this section, we first focus on the problem of 2-D face recognition. We investigate the performance of our proposed MS-PCANet on widely used 2-D face benchmark databases, such as MultiPIE,[22] FERET,[23] AR,[24] and Extended Yale-B.[25] Specifically, the MultiPIE dataset is employed to show the excellent performance of our model when it is faced with the pose variability, the FERET dataset is used to show the effectiveness of our model when it is faced with the expression, illumination, and aging variabilities, the AR and the Extended Yale-B datasets are used to show the robustness of our model when it is faced with various kinds of occlusion.

4.1.1 *Evaluation on MultiPIE dataset*

The MultiPIE database contains 337 subjects who are diverse across poses, expressions, and illumination. In this work, the images of a subject at −45 deg to +45 deg with step size 15 deg with neural expression and frontal illumination are used. Of these 337 subjects, we select 200 subjects (subject ID 001 to 200, in total 5600 images) are used for training our MS-PCANet model. And the rest of the 137 subjects are used for testing. All images are aligned to 80×64 pixels by using the supervised descent method[26] to locate five facial landmarks automatically. Some cropped example images are shown in Fig. 3. Similar to SPAE,[10] the frontal face images from the first session for the 137 images are used as gallery set, and the rest of the testing images (in total 2706 images) are used as the probe set.

An issue worth mentioning is that the feature learns from our proposed MS-PCANet model and does not contain any supervised information. Also, it is expected that the performance of our model is improved significantly by encoding discriminating label information. Therefore, we apply the supervised LDA to these unsupervised output features for dimensionality reduction. We also test our method with

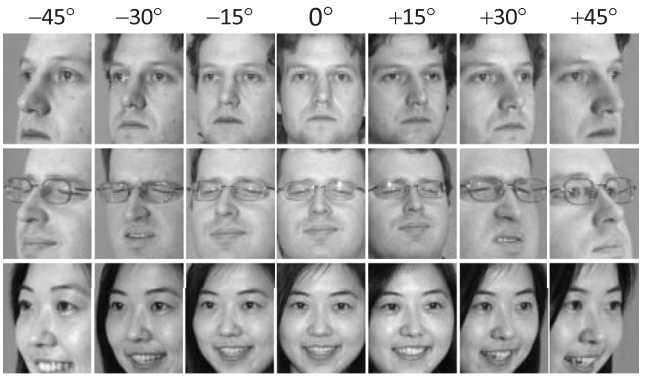


Fig. 3 Several cropped cross-pose face examples from the MultiPIE dataset.

various classifiers, and there is no significant difference in performance among various classifiers. Therefore, we adopt nearest neighbor (NN) classifier with cosine distance measure for all experiments below due to its simplicity.

Impact of the number of filter kernels. We first investigate the impact of filter kernel number on cross poses test sets. The size of filter kernel is $[k_1, k_2] = [5,5]$, the block size $[w, h] = [8,6]$, and the finest pyramid level is extracted (i.e., the size of SPP is $[10^2]$). We vary the number of filter kernels in both the first L_1 and the second stage L_2 from four to eight for our MS-PCANet model. From the view of mathematics, more kernels imply a more flexible network structure, which can achieve a better mapping from input images to output features. And the size of output feature of our model is $2^{L_1} + L_1 2^{L_2}$, larger $[L_1, L_2]$ can contain more discriminative information. The results are shown in Fig. 4. As expected, the more kernels, the better performance. One can see that the accuracy rate of MS-PCANet improves with the increase of $[L_1, L_2]$. However, $[L_1, L_2] = [9,9]$ just slightly improves $[L_1, L_2] = [8,8]$ but uses double the memory. In order to balance the high accuracy and computational complexity, we set $[L_1, L_2] = [8,8]$ for all experiments later.

Impact of the block size. Next, we investigate the impact of the block size (throughout this section we use the finest pyramid level for feature pooling). The parameters of our model are set to $[k_1, k_2] = [5,5]$ and $[L_1, L_2] = [8,8]$. The block size 6×4, 8×6, 12×9, and 16×12 are investigated. Figure 5 shows the experiment results. It is observed that small block size provides more robustness against pose variability. The reasons mainly come from two aspects. On the one hand, the image is partitioned into more local blocks with small block size, therefore, this output feature is longer than those with large block size. The longer feature means it has more discriminative information. On the other hand, the large block covers more area of the feature map, it easily lead to the fact that some uncorrelated structures of the face are forcibly pooled together. And the pose variability are generated by a continuously and nonlinearly movement, it usually cannot characterize well by a simple pooling operation. So the more nonlinear intraclass variability will appears in the large block, and the performance is more easily affected by pose, especially a larger pose variability. Therefore, we set the block size $[w, h]$ to be between 10th and 8th of the image's size on all experiments later.

Impact of the spatial pyramid pooling method. Then, we investigate the impact of the SPP and its scales on our MS-PCANet model. The parameters of our model are set to $[k_1, k_2] = [5,5]$, $[L_1, L_2] = [8,8]$ and the block size is 8×6. We pool the hashing image D_i^1 (or D_i^2) with different pyramid scales. The four scales $[10^2]$, $[10^2, 5^2]$, $[10^2, 5^2, 2^2]$, and $[10^2, 5^2, 2^2, 1^2]$ are considered. $[10^2]$ means the hashing image is partitioned into 10×10 blocks and only the finest-scale blocks are pooled. $[10^2, 5^2]$ means that we concatenate the pooled features that extracted from 10×10 and 5×5 blocks, respectively. The results are shown in Fig. 6. One can see that the size of SPP $[10^2, 5^2]$ achieves the best performance instead of $[10^2, 5^2, 2^2, 1^2]$. It does not improve the performance of our model with the incorporation of a larger-scale pyramid pooling feature, and even deteriorates our performance. Compared with cropped object image, the relevance of local blocks of a cropped face image is stronger, therefore, they are unfit for being understood as

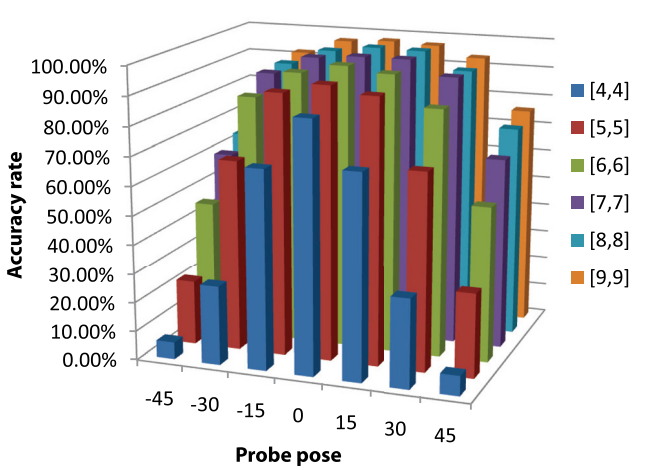
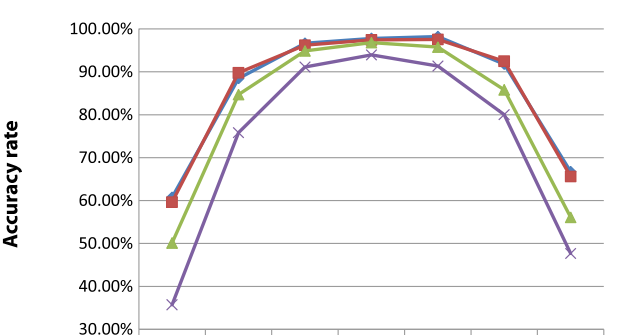


Fig. 4 Rank-one recognition rate of MS-PCANet for varying number of filter kernels $[L_1, L_2]$.



	-45	-30	-15	0	15	30	45
BlockSize=[6,4]	60.75%	88.47%	96.67%	97.77%	98.23%	91.80%	66.74%
BlockSize=[8,6]	59.65%	89.80%	96.23%	97.45%	97.56%	92.46%	65.63%
BlockSize=[12,9]	50.11%	84.70%	94.90%	96.82%	95.79%	85.81%	56.10%
BlockSize=[16,12]	35.70%	75.83%	91.13%	93.95%	91.35%	80.04%	47.67%

Fig. 5 Rank-one recognition rate of MS-PCANet for different block size $[w, h]$.

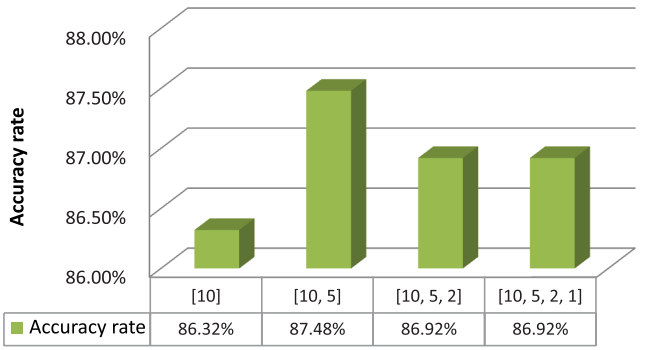


Fig. 6 The performance of our MS-PCANet model with different scale SPP. Here, the square of the number in the brackets represents the block number of hashing image. [10] means the hashing image is partitioned into 10×10 blocks.

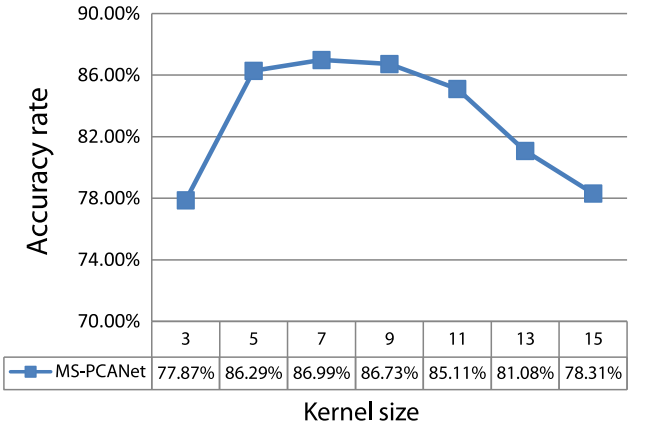


Fig. 8 The performance of different kernel sizes. Here, 3 denotes $k_1 = k_2 = 3$.

independent-semantics "words" of the BOW model. In other words, larger-scale pooling would lead to the fact that some uncorrelated blocks are forcibly pooled together.

Impact of multiscale features. In order to verify the effect of multiscale representation, we examine the impact of the output features that learn from multiple stages of our MS-PCANet model, i.e., from the last stage to the earlier stage. The parameters are the same as the above experiments, and the SPP size is $[10^2, 5^2]$. The results are shown in Fig. 7. The leftmost bar shows the performance of the feature extracted from the original intensity and the rightmost bar shows the performance of the cumulated features learned from the first and the second stage. We can see that the performance improves when the features from one more stage are cumulated. Compared with other single-stage output features, the cumulated structure $[f_1, f_2]$ encodes both holistic abstract and local specific information, simultaneously. Therefore, the cumulated features achieved better performance. This demonstrates that the cumulated features learn from our MS-PCANet model and can achieve a better performance when combined with supervised LDA.

Impact of filter kernel size. At last, we investigate the impact of the kernel size. The parameters of MS-PCANet

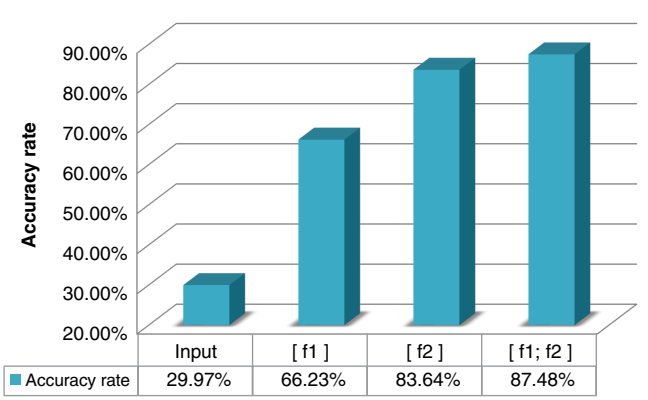


Fig. 7 The performance of cumulated features from multiple stages. Here, the character string in the brackets is the indexes of stages for cascade, and the "input" means the original intensity image.

are set to $[L_1, L_2] = [8,8]$, $[w, h] = [8,6]$ and the SPP size is $[10^2, 5^2]$. The cumulated features $[f_1, f_2]$ are used for feature representation. We vary the kernels size $[k_1, k_2]$ (where $k_1 = k_2$) from 3 to 15 with step size 2. The results are shown in Fig. 8. Similar to CNN model, the local receptive fields whose size is too small or too large would weaken the model's described ability for basic geometric structure of the face image, such as edge and corner. Therefore, the MS-PCANet model can achieve better performance when kernel size is set to between $[5,5]$ and $[9,9]$.

Comparison with state-of-the-art methods. In this section, we investigate the performance of our MS-PCANet model when it is faced with intraclass pose variation. The parameters are set to $[k_1, k_2] = [5,5]$, $[L_1, L_2] = [8,8]$, and $[w, h] = [8,6]$. We compare our MS-PCANet models with some state-of-the-art methods. These existing methods are briefly described as follows.

- Canonical correlation analysis (CCA)[27] aims at projecting samples which have different intraclass variabilities to a common subspace where the cross correlation between them are maximized.
- Partial least squares (PLS)[28] is similar to CCA, which is a pairwise classification method. It tries to linearly project samples of different variabilities into a latent subspace. Here, one sample is used as a regressor and another is used as a response.
- Multiview discriminant analysis (MvDA)[29] jointly learning multiple view-specific linear transforms by optimizing a generalized Rayleigh quotient to seek for a discriminant common space. In this space, the interclass variation is maximizing and the intraclass variation is minimizing. It is a nonpairwise method.
- DAE[13] has the same structure as AE deep networks. But it is trained locally to denoise corrupted versions of input images. Similar to our MS-PCANet, the output feature of DAE is also purely unsupervised. The supervised LDA is applied to dimensionality reduction.

Another issue worth mentioning is that pose estimation is necessary for the methods above. The CCA, PLS, and

Table 1 Comparison with the state-of-the-art methods on the cross-pose MultiPIE dataset.

Methods	Probe pose							
	−45 deg	−30 deg	−15 deg	+15 deg	+30 deg	+45 deg	Average	Pose estimation
CCA[27]	53.30%	74.20%	90.00%	90.00%	85.50%	48.20%	73.50%	Yes
PLS[28]	51.10%	76.90%	88.30%	88.30%	78.50%	56.50%	73.30%	Yes
MvDA[29]	**75.00%**	74.50%	82.70%	92.60%	87.50%	65.20%	79.60%	Yes
DAE[13]	69.90%	81.20%	91.00%	91.90%	86.50%	**74.30%**	82.50%	No
MS-PCANet (Spp = $[10^2]$)	63.64%	90.47%	96.01%	97.78%	91.80%	72.73%	85.40%	No
MS-PCANet (Spp = $[10^2, 5^2]$)	63.19%	**92.46%**	**97.34%**	**98.00%**	**94.24%**	72.51%	**86.29%**	No
MS-PCANet (Spp = $[10^2, 5^2, 2^2]$)	61.42%	92.24%	96.90%	97.78%	94.68%	70.73%	85.63%	No

Note: Bold values represent the best performance on one subset or one dataset.

MvDA methods all assume that the poses of the test images are already known, which means that they may degenerate when the pose is unavailable. And the DAE and our MS-PCANet can perform well without any pose estimation.

The results are shown in Table 1. As seen, our MS-PCANet model which has two pyramid levels achieves best performance when pose variations are smaller than −45 deg. As mentioned, larger scale SPP is not able to fully describe the intrinsic relevance of local blocks of a cropped image. Therefore, the MS-PCANet model which has three pyramid levels does not achieve better than those that have two pyramid levels. The CCA and PLS perform the worst since they do not contain any label information and it is unfavorable for the recognition task. The MvDA method is designed to specially solve the problem of intraclass pose variability, and it achieves better performance than our method when faced with larger pose variations. But it is necessary for the MvDA method to have the pose estimation of the test image, and our model can achieve excellent performance without any pose estimation. Similar to our model, the DAE has deep network structure, but it performs worse when faced with smaller pose variations. In order to model the complicated pose transforms, the DAE model uses the nonfrontal images as input and the frontal images as output. But the transforms from the nonfrontal pose to frontal pose is a highly nonlinear process and the DAE is easier to trap into local minima. Compared with other existing methods, our MS-PCANet learns discriminative high-level features by PCA-based filter kernels. PCA can be viewed as the simplest class of AEs, which can minimize the reconstruction error between input signal and output signal. So these learned filter kernels can preserve discriminative identity information as much as possible by using a relatively small dimensional feature. Our experimental conclusion coincides with what have been observed in SPAE.[10] Therefore, our MS-PCANet model can effectively extract discriminative high-level features which is robust against pose variations.

4.1.2 *Evaluation on FERET dataset*

Then, we apply the MS-PCANet model to evaluate the performance on the FERET dataset. The FERET dataset consists of 1196 different individuals who are diverse across age, ethnicity, and gender. The complete dataset is partitioned into six disjoint sets: *training, fa, fb, fc, dup1, dup2*. Here, all face images of these six sets are aligned and cropped into 128×128. Figure 9 shows some cropped example images from the FERET dataset. We perform model learning on the "training" set and MultiPIE generic training set which comprises around 100,000 images of size 80×64. We take *fa* as the gallery set and the rest of the sets as probe sets. The parameters of MS-PCANet are set to $[L_1, L_2] = [8,8]$, $[w, h] = [15,15]$ and the SPP size is $[8^2, 4^2]$. In order to compare fairly with existing methods, we apply whitened principal component analysis (WPCA) to reduce the dimensions of our MS-PCANet features to 1000. Here the projection matrix is learned from the *fa* set and the NN classifier with the cosine distance is used.

The results are listed in Table 2. We can see that both MultiPIE-learned model whose kernel size is $[k_1, k_2] = [5,5]$ (denoted as MultiPIE in the parentheses) and FERET-learned (KS denotes kernel size) model achieved excellent performance on different probe sets. On the one hand, compared with other low-level methods, DL-based methods (PCA Network-2, DCT-Net, and ours) achieved better performance, this is because DL-based methods can learn high-level feature representations from raw data by elaborately learned kernels and hierarchical nonlinear mappings. These abstract high-level feature representations can provide more invariance and robustness to various intraclass variations. And our model learns AE-like filter kernels so that discriminative information can be retained as much as

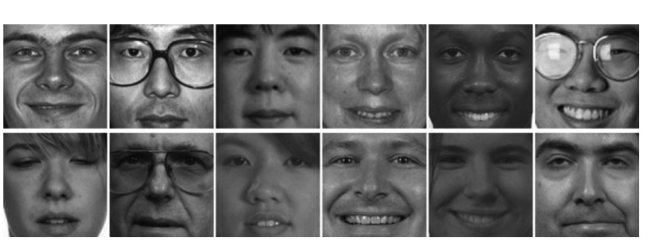


Fig. 9 Several aligned and cropped face examples from the FERET dataset.

Table 2 Comparison of face recognition rate (%) of various methods on FERET dataset.

Method	fb	Fc	dup1	dup2	Year
LBP[2]	93.0	51.0	61.0	50.0	2006
GV-LBP[30]	98.1	98.5	80.9	81.2	2011
POEM[31]	99.6	99.5	88.8	85.0	2012
sPOEM[32]	99.4	**100.0**	91.7	90.2	2013
DFD[15]	99.4	**100.0**	91.8	92.3	2014
PRF[33]	99.7	98.5	89.1	85.5	2014
SBGPM[34]	—	—	—	89.7	2015
PCA Network-2[8]	99.2	**100.0**	92.9	91.5	2015
CBFD+WPCA[16]	**99.8**	**100.0**	93.5	93.2	2015
DCT-Net[35]	99.7	**100.0**	93.3	91.5	2015
MS-PCANet (MultiPIE)	99.7	**100.0**	**94.0**	**93.6**	—
MS-PCANet (KS = [5,5])	99.2	**100.0**	92.9	91.5	—
MS-PCANet (KS = [7,7])	**99.8**	**100.0**	**94.0**	92.3	—
MS-PCANet (KS = [9,9])	**99.8**	**100.0**	92.9	91.5	—

Note: Bold values represent the best performance on one subset or one dataset.

possible. On the other hand, since the cumulation of multiscale features can effectively and comprehensively encodes the detailed and abstract semantics of the data, compared with other deep learning model, such as PCANetwork-2 and DCT-Net, our proposed method obtained better accuracies. Therefore, our model achieves better results than other state-of-the-art methods. Another issue worth mentioning is that training our MS-PCANet from a face dataset can be very effective to capture the abstract representation of another dataset. The generalization ability of our method would further be validated on unconstrained scenarios (Sec. 4.2.2). What is more, MultiPIE-learned model is slightly better than FERET-learned model, and it benefits from the intraclass variabilities in MultiPIE training set that are richer than the FERET training set.

4.1.3 *Evaluation on AR dataset*

Now, we further evaluate the robustness of our MS-PCANet model when faced with real malicious occlusions using AR dataset. AR dataset consists of around 4000 frontal images of 126 individuals. For each individual, pictures are taken under four different expressions, three illumination conditions, two disguise cases, four illumination and disguise cases during two separate sessions. We choose a subset of the database consisting of 50 male and 50 female subjects. The facial images are cropped into 165×120 and converted to gray-scale. The images of frontal illumination and neural are used as the gallery images and the rest all form the testing set. The parameters of MS-PCANet are set to $[k_1, k_2] = [5,5]$, $[L_1, L_2] = [8,8]$ and $[w, h] = [8,6]$. The scale of SPP is $[20^2, 10^2, 5^2, 2^2, 1^2]$. To compare fairly with some existing handcrafted feature, the dimension of our MS-PCANet feature is reduced to 200 by WPCA. We use the nearest neighbor classifier with the cosine distance.

Table 3 Comparison of face recognition rate (%) of various methods on AR dataset.

Method	Illum	Exps	Disguise	Dis+Illum	Year
LBP[2]	93.83	81.33	91.25	79.63	2006
P-LBP[36]	97.50	80.33	93.00	88.58	2010
PRF[33]	99.30	—	99.25	—	2014
SBGPM[34]	99.50	—	97.15	—	2015
SRC[37]	99.20	88.30	71.90	55.30	2009
LLC+SRC[38]	—	—	80.03	80.03	2011
Extended-SRC[39]	100.00	94.20	94.40	92.20	2012
S-SRC[40]	—	—	90.90	90.90	2013
PCA Network-2	99.50	85.00	97.00	95.00	2014
MS-PCANet	**100.00**	**96.17**	**99.75**	**98.50**	—

Note: Bold values represent the best performance on one subset or one dataset.

The results are given in Table 3. We can see that the recognition rate of our MS-PCANet is almost perfect when it is faced with various condition test sets, and achieve 100% for illumination variations test set. What is more, the performance of our model is much better than other methods on cross-expression test set. The experimental results are consistent with that on FERET dataset, it is that our MS-PCANet model is insensitive to illumination and expression. And the results also demonstrate that our model is robust against malicious occlusion, and even outperform sparse-based methods. It is well known that sparse-based methods work very well for an occlusion-related problem. It is assumed that the test image can be approximated by a sparse linear combination of the training images, however, this prior knowledge would be invalid when faced with large intraclass variations, such as pose. Different from the solution of sparse-based methods for an occlusion problem, our MS-PCANet learn feature representation by training the filter kernels of the deep network, and the learned filters contains basic geometric structure of occluded face image. Therefore, the contribution from valid region would be enhanced after multiple stages learning. What is more, our model can also solve an occlusion-related problem when faced with some extreme intraclass variations (see Secs. 4.2.1 and 4.2.2).

4.1.4 *Evaluation on extended Yale B dataset*

Finally, we test the ability of our MS-PCANet model to cope with random partial face recognition on the Extended Yale B

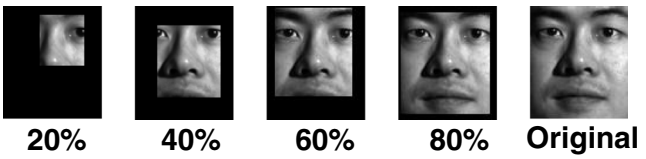


Fig. 10 The illustration of varying percentages of partial face image.

dataset. The extended Yale B dataset consists of 2414 frontal images for 38 individuals. The facial images are captured under various illumination conditions. We normalize and crop the origin images into 96×84, and choose subset 1 and 2 (from normal to moderate lighting conditions) for training and subset 3 (more extreme lighting conditions) for testing. It is a relatively easy face recognition task if integral face images are used. To challenge ourselves, the testing face images were represented by the face patches with different sizes at random locations. To compare easily with some existing partial face recognition methods, we retain different percents (0% to 80%) pixel values and set other pixel values to be 0 in the testing set, so that the size of partial face images is also equal to 96×84. Figure 10 shows some partial examples from the Extended Yale B dataset. The parameters of MS-PCANet are set to $[k_1, k_2] = [9,9]$, $[L_1, L_2] = [8,8]$, $[w, h] = [8,7]$ and the finest pyramid level is used (i.e., SPP $= [12^2]$). At last, the nearest neighbor classifier with the cosine distance is used for classification.

Table 4 shows the recognition rate of all methods. Our MS-PCANet significantly outperforms the other sparse-based methods for all partial levels, the corresponding reason is that PCA-based learned filters can minimize the reconstruction error. Our features contain more identity information compared with those sparse-based methods when the dimension of output feature is fixed. In addition, in this experiment, the partial region randomly appear on the holistic face image, therefore, the performance of sparse-based methods are worse compare with the work of fixed partial face region.[37] Compared with another DL-based method (i.e., PCA Network-2), since the multiple scales features are cumulated in our proposed MS-PCANet model, it can not only learn high-level features from holistic structure, but also can encode discriminative representations from local structure (such as partial edge, corner, and texture) information of the face. Therefore, the contribution from occluded region (i.e., pixel values is 0) would somehow be weakened after multiscaled feature cumulating. It is the reason why our model can achieve better performance when only 20% valid face region is provided.

Table 4 Recognition rate (%) of various methods faced with partial face recognition problem on Extended Yale-B dataset.

Method	20%	40%	60%	80%	100%	Year
SRC[37]	13.19	33.63	67.25	90.99	100.00	2009
G-SRC[41]	26.37	64.18	95.16	99.34	100.00	2010
CRC[42]	14.29	47.91	74.73	89.89	100.00	2011
I-RSC[43]	3.52	23.30	66.16	95.60	100.00	2011
GISA ($q = 0.5$, $p = 0.5$)[44]	12.53	35.60	67.03	93.63	100.00	2013
PCA Network-2[8]	50.33	**97.36**	99.78	**100.00**	**100.00**	2014
MS-PCANet	**52.31**	**97.36**	**100.00**	**100.00**	**100.00**	—

Note: Bold values represent the best performance on one subset or one dataset.

Therefore, a prominent message is concluded from extensive experiments in the above section that our MS-PCANet model can be very effective to learn the discriminative feature representations of various datasets. The MS-PCANet model achieves excellent performance when faced with pose, expression, illumination, and aging, and also perform well when faced with various levels of a partial face recognition problem.

4.2 *Two-Dimensional Face Verification on Several Datasets*

In this section, we focus on the problem of 2-D face verification. We first investigate the performance of our proposed MS-PCANet on widely used LFW dataset,[45] then test the efficiency of our model when it is faced with the extreme intraclass variability on point-and-shoot cameras (PaSC) dataset.[46]

4.2.1 *Evaluation on LFW dataset*

In this section, we evaluate the performance of our MS-PCANet model on the LFW dataset for uncontrolled 2-D face verification scenarios. The LFW dataset contains 13,233 face images of 5749 subjects collected from the web with large pose, occlusion, expression, and illumination variations. In this experiment, we use the aligned images (LFW-a)[47] and crop the images with the size of 150×80 from the original images. The cropped examples are shown in Fig. 11. We test on the View 2 set of LFW, which consists of 10 folds of 300 intraclass pairs and 300 interclass pairs. In our model, the kernels size, the number of kernels and the nonoverlapping block size are set to $[k_1, k_2] = [7,7]$, $[L_1, L_2] = [8,8]$, and $[w, h] = [15,13]$. The scale of SPP is set to $[10 \times 6, 5 \times 3, 1 \times 1]$. The MS-PCANet features are projected onto 3200 dimensions by WPCA and the NN classifier with cosine distance metric is used.

We use "unsupervised" setting (i.e., the class label information is not involved in classifier or metric learning), since "unsupervised" setting does not depend on any metric

Fig. 11 Several cropped face examples from LFW database.

Table 5 Comparison of mean recognition accuracy (±std) (%) and AUC (%) on LFW under unsupervised setting.

Methods	Accuracy (±std)	AUC	Year
SIFT[1]	64.10 ± 0.6	54.07	2002
LBP[2]	69.45 ± 0.5	75.47	2006
POEM[31]	82.70 ± 0.6	—	2012
G-LQP[14]	82.1 ± 0.26	—	2012
OCLBP[48]	82.78 ± 0.4	—	2013
MRF-MLBP[49]	80.08	89.94	2013
High-dim HOG[50]	84.98	—	2013
High-dim LBP[50]	84.08	—	2013
DFD[15]	84.02 ± 0.4	—	2014
PCA Network-2[8]	85.20 ± 1.5	—	2014
Spartans[51]	—	**94.28**	2015
SLBFLE[52]	85.62	92.00	2015
MS-PCANet	86.81 ± 1.5	93.16	—

Note: Bold values represent the best performance on one subset or one dataset.

learning and supervised classifier training. Therefore, the performance of different methods under the "unsupervised" setting just represents the discrimination ability of the extracted feature from different methods. It is the best choose to evaluate the performance of our model on uncontrolled scenarios that using "unsupervised" setting. The mean accuracy (±std) and area under curve (AUC) are adopted to compare our MS-PCANet with other methods previously reported on LFW dataset. Table 5 lists the mean accuracy and AUC of various methods and Fig. 12 shows the ROC curves of our model and some methods which are available on the LFW website. One can see that MS-PCANet whose accuracy and AUC are 86.81% and 93.16%, respectively is quite competitive to the state-of-the-art methods. The experiment results show that the proposed model not only effectively learn high-level feature representations in controlled scenarios, but also have excellent discrimination ability for face images that collected in uncontrolled conditions. Several low-level methods only achieved good performance in controlled conditions. This is mainly due to low-level methods that import strong prior knowledge about face-related field, therefore, it lacks sufficient generalization ability when faced with complicated intraclass variabilities which are different from the intravariabilities of imported prior knowledge. On the one hand, similar to AE network, our proposed model can obtain identity-bearing representation by minimizing the reconstruction error between input face images and output features. On the other hand, the learned high-level and low-level features are cumulated together as the final representation. Our MS-PCANet has sufficient discriminative and generalized ability when faced with extreme intraclass variability or uncontrolled scenarios. Another issue worth mentioning is our MS-PCANet model achieves 86.81% accuracy without any outside data. In preparation of this work, we have read some impressive works[4–6] which also employ DL model for face verification. These works achieve excellent results on LFW, but they employed the identity information and large amounts of outside data to train their model. By contrast, our MS-PCANet model has less fine-tuned parameters and can learn discriminative feature representations as much as possible when less training samples are provided.

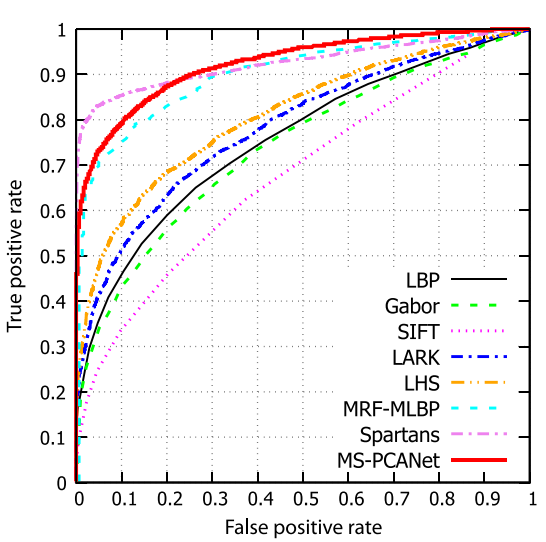


Fig. 12 ROC curves over View 2 on LFW database. The performance of other methods is directly cited from the website.[53]

4.2.2 *Evaluation on point-and-shoot cameras dataset*

The experiment in Sec. 4.2.1 can be considered as an evaluation of our model's discriminative ability in uncontrolled scenarios. In this section, we next evaluate the effectiveness of our model's discriminative and generalization ability (i.e., learned from constrained images and test on unconstrained images) in real applications. The PaSC dataset[46] consists of 9376 still images from 293 subjects balanced with respect to distance, alternative sensors, poses, and varying location. There are 4688 images in both "query set" and "target set." We align and crop each image into 128×128 pixels according to the eye coordinates provided by the PaSC database website (eye coordinates of each face image can be found at Ref. 54). Figure 13 shows some aligned and cropped face images from PaSC dataset. There are some complicated intraclass variabilities in PaSC dataset, such as pose, lighting, expression, motion blur, and poor focus.

In order to evaluate the discriminative and generalization ability of our MS-PCANet when faced with real unconstrained scenarios, we learn the MS-PCANet models on the "constrained" FERET, MultiPIE dataset, and "unconstrained" PaSC itself, and test them on the PaSC dataset. In our model, the kernels size, the number of kernels, and the nonoverlapping block size are set to $[k_1, k_2] = [7,7]$, $[L_1, L_2] = [8,8]$, and $[w, h] = [8,8]$. The scale of SPP is set to $[16^2, 8^2]$. We compare our MS-PCANet model with other state-of-the-art learning methods. At last, the learned features are reduced into 500-dimensional feature by

Fig. 13 Aligned and cropped face examples from the PaSC dataset. (a) Frontal or near-frontal face images and (b) nonfrontal face images.

Table 6 Verification rate (%) at FAR = 0.01 on PaSC dataset for all images and frontal images.

Method	All	Frontal	Year
LBP[2]	17.6	29.6	2006
LRPCA[46]	10.0	19.0	2011
CohortLDA[46]	8.0	22.0	2012
BSIF[55]	14.3	24.9	2014
CBFD[16]	19.4	36.0	2015
JFL[56]	32.6	—	2015
SLBFLE[52]	29.2	—	2015
PCA Network-2[8]	28.6	53.0	2015
MS-PCANet (FERET)	33.4	53.8	—
MS-PCANet (MultiPIE)	33.2	**54.0**	—
MS-PCANet (Itself)	**33.6**	**54.0**	—

Note: Bold values represent the best performance on one subset or one dataset.

WPCA as the final representation. Table 6 tabulates the verification rate at false accept rate(FAR) = 0.01 for all images and only frontal images. Figure 14 shows the ROC curves of some descriptors for all images and frontal images, respectively. As can be seen, our proposed model significantly outperforms the other state-of-the-art methods and baseline methods. Though many low-level methods can work well in controlled conditions, such as LBP and CBFD, their accuracies are even less than a random guess when faced with some extreme intraclass variations, such as large pose, poor lighting, motion blur, and poor focus. Our model learns information-preserving filter kernels by deep network structure, these learned PCA-based filter kernels can better capture essential characteristics of face images in a dataset. Therefore, although the face appearances of FERET and MultiPIE are very different from that on PaSC, the MS-PCANet model learned from FERET and MultiPIE datasets can still work very well in the unconstrained case. This experiment once again demonstrated that our model can not only learn discriminative feature representations on constrained scenarios, but also show excellent generalization ability and robustness when faced with real unconstrained scenarios.

4.3 *Three-Dimensional Face/Expression Recognition on Several Datasets*

In this section, we focus on the problem of 3-D face or expression recognition. The FRGC v2 dataset[57] is employed

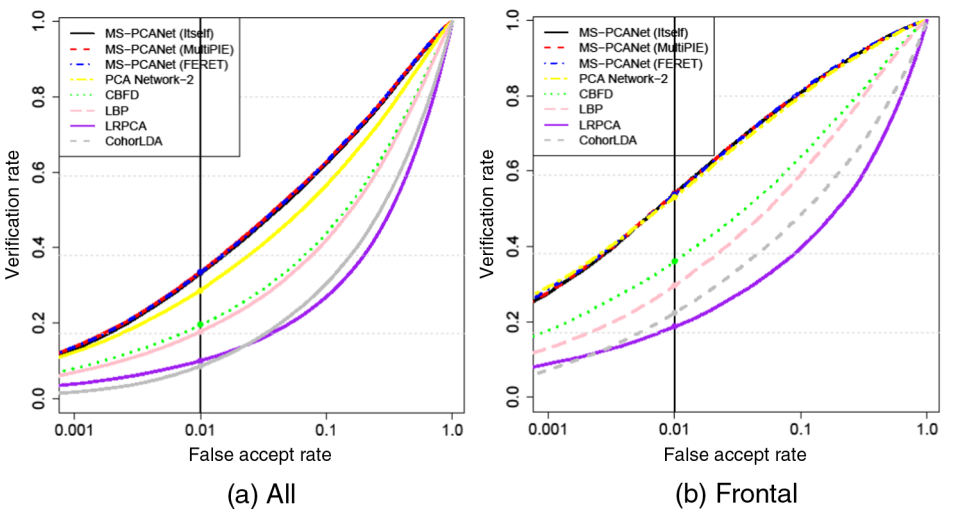


Fig. 14 The ROC curves of different methods on the PaSC dataset for all and frontal images scenarios, respectively.

to show the excellent performance of MS-PCANet model on 3-D face dataset, and the BU-3DFE dataset[58] is used to show that our MS-PCANet also can achieve state-of-the-art performance on 3-D expression dataset.

4.3.1 *Face recognition evaluation on FRGC v2 dataset*

In this section, the FRGC v2 dataset is used to evaluate the performance of our MS-PCANet model when faced with various intraclass variabilities in 3-D face images. The FRGC v2 dataset consists of 4007 3-D face images from 466 subjects who are diverse across age, gender, and ethnicity. There are also large expression variations, face region incompleteness and hair occlusion in the FRGC v2 dataset. The examples are shown in Fig. 15. Some pixel corruptions are present during the image acquisition, such as the absence of a nose and holes around the eyes and mouth. In our experimental setting, we employ FRGC v1 dataset (in total 942 images) to train our MS-PCANet. The parameters of our model are set to $[k_1, k_2] = [5,5]$, $[L_1, L_2] = [8,8]$, and $[w, h] = [15,15]$. The two pyramid scale $[8^2]$ and $[8^2, 4^2]$ are used. The nearest neighbor classifier with the cosine distance is used for our MS-PCANet model and with the Euclidean distance for other methods. The NN classifier with different distance measure is to secure the best performances of respective methods.

Some recent published methods divided the original dataset into different test sets, including non-neutral versus neutral (non-N versus N), all versus neutral (A versus N) for the expression variations. The configurations of test sets are listed in Table 7. Similar to the work of Ming[59] we also evaluate the performance of various 3-D face recognition algorithms with the FAR of 0.1%. Table 8 shows the recognition results and Fig. 16 shows the cumulative matching characteristic (CMC) curves of our MS-PCANet with different pyramid scale for two test sets. The results illustrate the following conclusions. On the one hand, our model significantly improve the performance of existing popular methods, and achieve almost perfect performance when faced with *A vs. N* test set. On the other hand, other state-of-the-art methods which import strong prior knowledge cannot work very well when large face region incompleteness and hair occlusion are collected in gallery set. Compared with these methods, our model learns filter kernels from the original data and import little prior knowledge. These learned kernels can extract basic spatial structure information of 3-D image and provide sufficient robustness to these extreme situations, the reason for robustness has been explained in Secs. 4.1.3 and 4.1.4. In addition, the multiscaled feature representations also provide identity information for 3-D face recognition. Therefore, our model achieves excellent performance when these non-neural faces are used as gallery set.

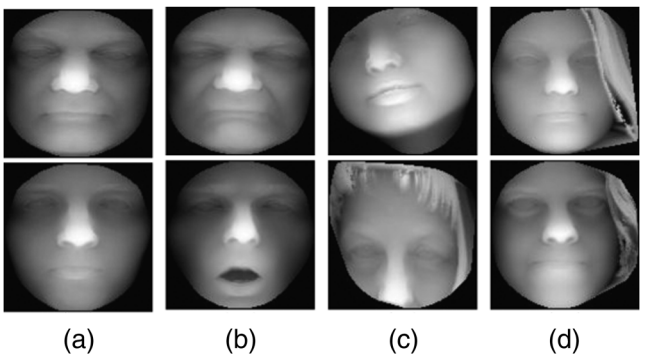


Fig. 15 Some examples of depth images in FRGC v2 dataset. (a) The regular face images. (b) The face images with nonrigid deformations. (c) The face images with some missing face parts. (d) The face images with hairs occlusions.

Table 7 The configurations of different test sets.

Test set	Gallery set	Probe set
Non-N versus N	Non-N (2469)	N (1538)
A versus N	A (4007)	N (1538)

Table 8 The recognition rate (%) of various 3-D face recognition algorithms on different test set.

Method	Test set	
	Non-N versus N	A versus N
Maurer et al.[60]	—	92.00
Passalis et al.[61]	79.40	85.10
Berretti et al.[62]	91.40	95.50
Ming et al.[59]	93.14	95.24
Al-Osaimi et al.[63]	97.80	98.14
Mian et al.[64]	97.00	97.54
Ours (spp = $[8^2]$)	**98.44**	**100.00**
Ours (spp = $[8^2, 4^2]$)	**98.44**	**99.93**

Note: Bold values represent the best performance on one subset or one dataset.

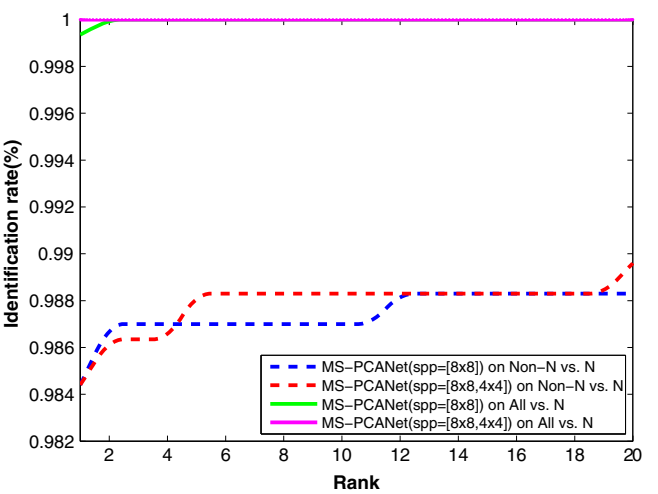


Fig. 16 The CMC curves of our proposed MS-PCANet model on the different FRGC test set.

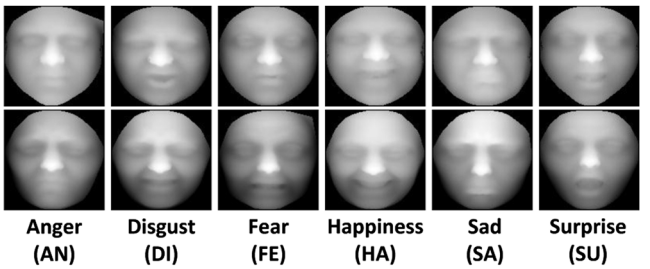


Fig. 17 Some cropped examples of various expressions on BU-3DFE dataset.

4.3.2 *Expression recognition evaluation on BU-3DFE database*

This database contains 100 subjects, 56 female and 44 male with a variety of ethnic/racial ancestries. The subjects display the six universal expressions, i.e., anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU) in four levels of intensity. The face images are cropped into 128×128, Fig. 17 shows some examples of BU-3DFE face images. In this experiment, we use the similar experimental settings as the work.[65] That is, we randomly partition the 100 subjects into a training dataset and a testing dataset with the size of 80 versus 20. Both training and testing datasets cover all six expressions, four intensities, therefore, there are 1920 training samples and 480 testing samples. We average the results over 10 splits as the final recognition accuracy. The parameters of our model are set to $[k_1, k_2] = [7,7]$, $[L_1, L_2] = [8,8]$, and $[w, h] = [15,15]$. The two pyramid scales $[8^2]$ and $[8^2, 4^2]$ are used.

Table 9 shows the average recognition accuracies versus the standard deviations (in bracket) with various expressions of different methods. We also provide Fig. 18, which shows the corresponding confusion matrix of our proposed MS-PCANet. Our MS-PCANet model achieved better performance than other state-of-the-art methods on FE and SA subset. It is well known that these two subsets are more difficult to be recognized than other expressions due to the facial muscle deformations of both expressions being relatively small. And those specialized 3-D expression recognition methods which introduced strong prior knowledge are difficult to classify these small-deformation expressions. However, our DL-based model can learn discriminative filters which contain the basic spatial information of 3-D face images and find the difference between these two expressions due to the combination of low-level and high-level features. Therefore, as long as enough training samples

Table 9 The results of average recognition accuracies versus the standard deviations with different methods on the BU-3DFE Ddatabase.

Method	Expression number	AN	DI	FE	HA	SA	SU	Average
Sparse-SIFT[66]	6 (4 intensities)	80.25	75.00	61.75	87.50	72.12	91.12	77.96
SGPR[67]	6 (2 intensities)	69.10	71.50	58.00	80.10	59.10	80.50	69.72
CSGPR[67]	6 (2 intensities)	68.30	72.70	63.20	80.60	63.40	81.50	71.62
GSRRR-LBPu2[65]	6 (4 intensities)	64.50 (5.2)	65.10 (8.2)	46.50 (6.9)	83.90 (6.2)	63.60 (6.0)	86.30 (7.1)	68.3 (2.6)
GSRRR-Sparse-SIFT[65]	6 (4 intensities)	**81.30 (4.2)**	**77.8 (4.0)**	58.4 (7.7)	87.20 (6.9)	75.60 (6.1)	**93.3 (5.8)**	**78.9 (2.5)**
GSRRR-83-Points[65]	6 (4 intensities)	68.5 (8.8)	62.6 (3.9)	35.9 (7.7)	**91.9 (2.7)**	71.9 (7.3)	89.5 (5.8)	70.0 (1.8)
MS-PCANet (spp $= [8^2]$)	6 (4 intensities)	72.4 (5.1)	74.3 (5.0)	**68.9 (3.6)**	85.3 (5.3)	**79.0 (5.0)**	79.7 (5.7)	76.4 (1.1)
MS-PCANet (spp $= [8^2, 4^2]$)	6 (4 intensities)	71.1 (4.2)	74.0 (3.9)	66.3 (4.5)	83.6 (5.3)	77.4 (5.1)	78.5 (5.3)	75.0 (1.4)

Note: Bold values represent the best performance on one subset or one dataset.

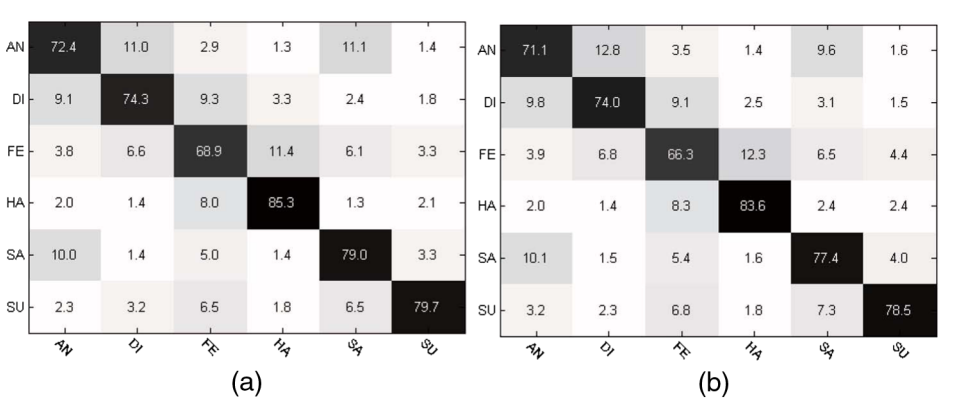


Fig. 18 The confusion matrices of our MS-PCANet model with different pyramid scale on the BU-3DFE dataset. (a) SPP $= [8^2]$ and (b) SPP $= [8^2, 4^2]$.

are provided, our model not only can encode holistic semantic information, but can also encode the micropattern of both 2-D and 3-D face images. And our MS-PCANet model also achieved quite competitive performance compared with state-of-the-art methods on other expression subsets (AN, DI, HA, and SU).

The above-mentioned experiments show that our MS-PCANet model not only achieved excellent performances on face recognition task, but also work very well on other face-related tasks, such as expression recognition.

5 Conclusion

This work proposes a multiple scales combined deep learning model to learn high-level feature representation for face recognition. Different from the traditional CNN, our model convolute input image with the prefixed filter kernels which are learned by PCA, nonlinearly process feature maps by binary hashing and pool them using SPP method. It further enhances the face recognition performance by cumulating the output features of several stages. Multiscale feature, which are fed to the classifier, efficiently encodes both the holistic semantic information and local semantic information of original data. The MS-PCANet is examined on both 2-D face datasets and 3-D face/expression datasets. The results show that MS-PCANet outperforms the state-of-the-art face recognition methods in most cases, and demonstrate the effectiveness of our model in 2-D and 3-D face-related recognition/verification scenarios. We can expected that our MS-PCANet model can be deployed into real scenarios if our model is trained on a large enough database, which consists of large intraclass and interclass variations.

Our proposed MS-PCANet model is a general deep learning model. Ideally, it can learn discriminative feature representations of various visual recognition tasks as long as extensive training samples are provided. Therefore, a possible extension of our work can apply to other computer vision applications, such as expression recognition and visual tracking to further demonstrate its effectiveness.

Acknowledgments

The work presented in this paper was supported by the National Natural Science Foundation of China (Grant Nos. NSFC-61402046 and NSFC-61170176), Fund for Beijing University of Posts and Telecommunications (Grant Nos. 2013XZ10 and 2013XD-04), Fund for the Doctoral Program of Higher Education of China (Grant No. 20120005110002).

References

1. D. G. Lowe, "Object recognition from local scale-invariant features," in *The Proc. of the Seventh IEEE Int. Conf. on Computer Vision 1999*, Vol. 2, pp. 1150–1157, IEEE (1999).
2. T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 2037–2041 (2006).
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *26th Annual Conf. on Neural Information Processing Systems 2012, NIPS 2012, December 3, 2012—December 6, 2012, Advances in Neural Information Processing Systems*, Vol. 2, pp. 1097–1105, Neural Information Processing System Foundation (2012).
4. Y. Taigman et al., "Deepface: closing the gap to human-level performance in face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, pp. 1701–1708 (2014).
5. S. Yi, W. Xiaogang, and T. Xiaoou, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, pp. 1891–1898 (2014).
6. Y. Sun et al., "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, pp. 1988–1996 (2014).
7. J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1872–1886 (2013).
8. T.-H. Chan et al., "Pcanet: a simple deep learning baseline for image classification?" arXiv:1404.3606 (2014).
9. L. Tian et al., "Stacked PCA network (SPCANet): an effective deep learning for face recognition," in *IEEE Int. Conf. on Digital Signal Processing (DSP 2015)*, pp. 1039–1043, IEEE (2015).
10. M. Kan et al., "Stacked progressive auto-encoders (SPAE) for face recognition across poses," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2014)*, pp. 1883–1890 (2014).
11. J. Ngiam et al., "On optimization methods for deep learning," in *Proc. of the 28th Int. Conf. on Machine Learning (ICML-11)*, pp. 265–272 (2011).
12. S. Rifai et al., "Contractive auto-encoders: explicit invariance during feature extraction," in *Proc. of the 28th Int. Conf. on Machine Learning (ICML-11)*, pp. 833–840 (2011).
13. P. Vincent et al., "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
14. S. U. Hussain, T. Napoléon, and F. Jurie, "Face recognition using local quantized patterns," in *British Machive Vision Conf.*, pp. 11 (2012).
15. Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(2), 289–302 (2014).
16. J. Lu et al., "Learning compact binary face descriptor for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 2041–2056 (2015).
17. S. Xie et al., "Fusing local patterns of Gabor magnitude and phase for face recognition," *IEEE Trans. Image Process.* **19**, 1349–1361 (2010).
18. J. Yi and F. Su, "Histogram of log-Gabor magnitude patterns for face recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2014)*, pp. 519–523 (2014).
19. Z. Chai et al., "Gabor ordinal measures for face recognition," *IEEE Trans. Inf. Forensics Secur.* **9**, 14–26 (2014).
20. Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *IEEE Int. Conf. on Computer Vision (ICCV 2013)*, pp. 1489–1496 (2013).
21. K. He et al., "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916 (2015).
22. R. Gross et al., "Multi-pie," *Image Vision Comput.* **28**(5), 807–813 (2010).
23. P. J. Phillips et al., "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(10), 1090–1104 (2000).
24. A. M. Martinez, "The AR face database," *CVC Technical Report*, Vol. **24** (1998).
25. A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(6), 643–660 (2001).
26. X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2013)*, pp. 532–539, IEEE (2013).
27. W. Yang et al., "2D–3D face matching using CCA," in *8th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG'08)*, pp. 1–6, IEEE (2008).
28. A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2011)*, 593–600, IEEE (2011).
29. A. Sharma et al., "Generalized multiview analysis: a discriminative latent space," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2012)*, pp. 2160–2167, IEEE (2012).
30. Z. Lei et al., "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Trans. Image Process.* **20**(1), 247–256 (2011).
31. N.-S. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Trans. Image Process.* **21**(3), 934–945 (2012).
32. N.-S. Vu, "Exploring patterns of gradient orientations and magnitudes for face recognition," *IEEE Trans. Inf. Forensics Secur.* **8**(2), 295–304 (2013).
33. F. Shen et al., "Face image classification by pooling raw features," *Pattern Recogn.* **54**, 94–103 (2016).
34. W. Huang and H. Yin, "Robust face recognition with structural binary gradient patterns," arXiv:1506.00481 (2015).
35. C. J. Ng and A. B. J. Teoh, "DCTNet: a simple learning-free approach for face recognition," arXiv:1507.02049 (2015).
36. X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.* **19**, 374–383 (2010).
37. J. Wright et al., "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 210–227 (2009).

38. Y.-W. Chao et al., "Locality-constrained group sparse representation for robust face recognition," in *18th IEEE Int. Conf. on Image Processing (ICIP 2011)*, pp. 761–764, IEEE (2011).
39. W. Deng, J. Hu, and J. Guo, "Extended SRC: undersampled face recognition via intraclass variant dictionary," *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(9), 1864–1870 (2012).
40. W. Deng, J. Hu, and J. Guo, "In defense of sparsity based face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2013)*, pp. 399–406, IEEE (2013).
41. M. Yang and L. Zhang, "Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary," in *European Conference on Computer Vision (ECCV 2010)*, pp. 448–461, Springer (2010).
42. L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: which helps face recognition?" in *IEEE Int. Conf. on Computer Vision (ICCV 2011)*, pp. 471–478, IEEE (2011).
43. M. Yang et al., "Robust sparse coding for face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2011)*, pp. 625–632, IEEE (2011).
44. W. Zuo et al., "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 217–224 (2013).
45. G. B. Huang et al., "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Tech. Report 07-49, University of Massachusetts, Amherst (2007).
46. J. R. Beveridge et al., "The challenge of face recognition from digital point-and-shoot cameras," in *IEEE Sixth Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 2013)*, pp. 1–8, IEEE (2013).
47. L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," in *Asian Conf. on Computer Vision (ACCV 2009)*, pp. 88–97, Springer (2009).
48. O. Barkan et al., "Fast high dimensional vector multiplication face recognition," in *IEEE Int. Conf. on Computer Vision (ICCV 2013)*, pp. 1960–1967, IEEE (2013).
49. S. R. Arashloo and J. Kittler, "Efficient processing of MRFS for unconstrained-pose face recognition," in *IEEE Sixth Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 2013)*, pp. 1–8, IEEE (2013).
50. D. Chen et al., "Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2013)*, pp. 3025–3032, IEEE (2013).
51. F. Juefei-Xu, K. Luu, and M. Savvides, "Spartans: single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios," *IEEE Trans. Image Process.* **24**(12), 4780–4795 (2015).
52. J. Lu, V. Erin Liong, and J. Zhou, "Simultaneous local binary feature learning and encoding for face recognition," in *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 3721–3729 (2015).
53. G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: updates and new reporting procedures," Amherst Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, Massachusetts.
54. J. R. Beveridge et al., "Point-and-Shoot Face Recognition Challenge Users Guide," http://www.cs.colostate.edu/~vision/pasc/pasc_support.php (2013).
55. J. Ylioinas et al., "An in-depth examination of local binary descriptors in unconstrained face recognition," in *22nd Int. Conf. on Pattern Recognition (ICPR 2014)*, pp. 4471–4476, IEEE (2014).
56. J. Lu et al., "Joint feature learning for face recognition," *IEEE Trans. Inf. Forensics Secur.* **10**, 1371–1383 (2015).
57. P. J. Phillips et al., "Overview of the face recognition grand challenge," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, Vol. 1, pp. 947–954, IEEE (2005).
58. L. Yin et al., "A 3D facial expression database for facial behavior research," in *7th Int. Conf. on Automatic Face and Gesture Recognition (FGR 2006)*, pp. 211–216, IEEE (2006).
59. Y. Ming, "Rigid-area orthogonal spectral regression for efficient 3D face recognition," *Neurocomputing* **129**, 445–457 (2014).
60. T. Maurer et al., "Performance of geometrix activeID TM 3D face recognition engine on the FRGC data," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (CVPR 2005) Workshops*, pp. 154–154, IEEE (2005).
61. G. Passalis et al., "Evaluation of 3D face recognition in the presence of facial expressions: an annotated deformable model approach," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition-Workshops (CVPR 2005) Workshops*, 171–171, IEEE (2005).
62. S. Berretti, A. D. Bimbo, and P. Pala, "3D face recognition using isogeodesic stripes," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(12), 2162–2177 (2010).
63. F. Al-Osaimi, M. Bennamoun, and A. Mian, "An expression deformation approach to non-rigid 3D face recognition," *Int. J. Comput. Vision* **81**(3), 302–316 (2009).
64. A. S. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2D-3D hybrid approach to automatic face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(11), 1927–1943 (2007).
65. W. Zheng, "Multi-view facial expression recognition based on group sparse reduced-rank regression," *IEEE Trans. Affective Comput.* **5**(1), 71–85 (2014).
66. W. Zheng et al., "A novel approach to expression recognition from non-frontal face images," in *IEEE 12th Int. Conf. on Computer Vision 2009*, pp. 1901–1908, IEEE (2009).
67. O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1357–1369 (2013).

Lei Tian received his MS degree in the Department of Electronic Engineering from Beijing University of Posts and Telecommunications, China, in 2015. He is currently working toward the PhD in Beijing Key Laboratory of Work Safety Intelligent Monitoring, the Department of Electronic Engineering, Beijing University of Posts and Telecommunications. His research interests include computer vision, pattern recognition, and deep learning.

Chunxiao Fan is currently a professor and the director of the Center for information electronic and intelligence system. She served as a member of ISO/IEC JTC1/SC6 WG9, ASN.1 (since 2006) and the Chinese Sensor Network working group. Her research interests include heterogeneous media data analysis, Internet of things, data mining, communication software, and so on. She has published more than 30 papers in international journals and conferences, authored and edited three books and has authorized several patents for inventions.

Yue Ming received her BS degree in communication engineering, her MS degree in human-computer interaction engineering, and her PhD in signal and information processing from Beijing Jiaotong University, China, in 2006, 2008, and 2013, respectively. She worked as a visiting scholar at Carnegie Mellon University, USA, between 2010 and 2011. Her research interests are in the areas of biometrics, computer vision, computer graphics, information retrieval, pattern recognition, and so on.