Multi-scale Patch Based Collaborative Representation for Face Recognition with Margin Distribution Optimization

Pengfei Zhu[1], Lei Zhang[1,⋆], Qinghua Hu[2], and Simon C.K. Shiu[1]

[1] Biometric Research Center, Dept. of Computing,
The Hong Kong Polytechnic University
[2] School of Computer Science and Technology, Tianjin University
{cspzhu,cslzhang}@comp.polyu.edu.hk

Abstract. Small sample size is one of the most challenging problems in face recognition due to the difficulty of sample collection in many real-world applications. By representing the query sample as a linear combination of training samples from all classes, the so-called collaborative representation based classification (CRC) shows very effective face recognition performance with low computational cost. However, the recognition rate of CRC will drop dramatically when the available training samples per subject are very limited. One intuitive solution to this problem is operating CRC on patches and combining the recognition outputs of all patches. Nonetheless, the setting of patch size is a non-trivial task. Considering the fact that patches on different scales can have complementary information for classification, we propose a multi-scale patch based CRC method, while the ensemble of multi-scale outputs is achieved by regularized margin distribution optimization. Our extensive experiments validated that the proposed method outperforms many state-of-the-art patch based face recognition algorithms.

1 Introduction

Face recognition (FR) has been an active research topic in computer vision and pattern recognition for many years [1]. In spite of the tremendous achievements, there are still many challenges caused by the large face appearance variations of illumination, expression, pose, noise, occlusion, etc [2]. Particularly, the small sample size (SSS) problem is one of the most fundamental and challenging issues in FR. In many real-world applications such as smart cards, law enforcement, surveillance and access control, the training samples of many subjects are often very limited [3]. Unfortunately, the performance of appearance based FR methods, such as the classical Eigenface [4], Fisherface [5], LPP [6] and the variants of them [7], degrades much with the decrease of training samples.

As a generalization and extension of the nearest neighbor, nearest line, nearest plane and nearest subspace classifiers, the sparse representation based classification (SRC) [8] scheme shows very interesting FR results. SRC represents a query

⋆ Corresponding author.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 822–835, 2012.

face as a sparse linear combination of the training samples from all classes, and classifies it to the class which has the least representation residual. However, in [9] it was indicated that the costly l_1-norm sparse regularization on the representation vector in SRC is not necessary, and l_2-norm regularization can lead to similar FR results but with much lower computational cost. The collaborative representation based classification (CRC) was then proposed in [9] by representing the query sample with non-sparse l_2-regularization. However, both CRC and SRC suffer serious performance degradation when the training sample size is very small and hence the query sample cannot be well represented [10].

To solve the SSS problem, virtual samples and generic training set were used in [11]. On the other hand, the trained classifiers will become unstable and have poor generalization ability when the available samples are insufficient, and hence ensemble learning has been widely applied to FR and has led to significant improvement in recognition rate and robustness [12][13][14]. These methods can be roughly divided into three categories. The first category of methods is patch (or block) based methods, which usually involve steps of local region partition, local feature extraction and classification combination [15][14]. The recognition rate of patch based methods is much affected by patch size, which is often set by experimental experience [16] [12]. Considering that the global and local features can provide complementary information, the second category of methods combines the global and local features for classification [13][17]. Third, a very popular category of methods uses multiple feature extractors to extract different types of facial features, and then uses classifier fusion for classification. For example, in [18][19], local features such as SIFT, LBP, Gabor response and gray values are combined for face verification.

Human faces exhibit distinct structures and characteristics when observed on different scales [13]. Combining the information on different scales could not only lead to much FR improvement but also provide us a simple and effective way for scale-insensitive models. How to combine multi-scale information is essentially an ensemble learning task. AdaBoost [20] is one of the most successful ensemble learning techniques due to its excellent performance and broad applications in face and object detection, visual tracking, etc. The success of AdaBoost actually attributes to margin distribution optimization [21][22][23], and AdaBoost approximately minimizes the loss criterion with l_1-regularization on the coefficient vector [20]. In [24], Shawe-Taylor gave the bound of AdaBoost's generalization error based on margin distribution, which shows that the loss of margin and the norm of coefficient vector could be minimized.

In this paper, to improve the performance of CRC in SSS problem, we propose to conduct CRC on patches, and the so-called patch based CRC (PCRC) classifies the query sample by combining the recognition outputs of all the overlapped patches, each of which is collaboratively represented by the corresponding patches of training samples. Similar to those patch based methods, PCRC is a patch size sensitive method, while the optimal patch size varies with training sample size and databases. In order for a patch size robust scheme, we then propose a multi-scale PCRC (MSPCRC) method by combining the information

on different scales. MSPCRC considers PCRC on each scale as a base classifier and learns scale weights to fuse multi-scale decisions. Scale weights are learned by minimizing the square loss of margin, and sparse l_1-norm regularization is imposed on the weights to get better margin distribution.

The rest of this paper is organized as follows. Section 2 describes PCRC. Section 3 presents the margin distribution optimization for multi-scale ensemble. Section 4 conducts experiments and conclusions are made in Section 5.

2 Patch Based CRC

In [9], Zhang et al. proposed to use the regularized least square model for collaborative representation based classification (CRC) of face images. Given a set of training samples, denote by $\boldsymbol{X}_k \in \Re^{m \times n_k}$ the dataset of the k^{th} class, and each column of $\boldsymbol{X}_k$ is a sample of class k. Suppose that we have c classes of subjects, and let $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_c]$. Given a query sample $\boldsymbol{y}$, the collaborative representation of it is

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}\|_2^2 + \lambda \|\boldsymbol{a}\|_2^2\} \tag{1}$$

The solution of CRC is $\hat{\boldsymbol{a}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \cdot \boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The classification of CRC is performed by checking which class yields the minimal regularized reconstruction error. The recognition output of the query sample $\boldsymbol{y}$ is Indentity($\boldsymbol{y}$) = $\text{argmin}_k\{r_k\}$, where $r_k = \|\boldsymbol{y} - \boldsymbol{X}_k \cdot \hat{\boldsymbol{a}}_k\|_2 / \|\hat{\boldsymbol{a}}_k\|_2$ and $\hat{\boldsymbol{a}} = [\hat{\boldsymbol{a}}_1; \hat{\boldsymbol{a}}_2; ...; \hat{\boldsymbol{a}}_c]$.

When the linear system determined by dictionary $\boldsymbol{X}$ is under-determined, the linear representation of the query sample over $\boldsymbol{X}$ can be very accurate while regularization on $\boldsymbol{a}$ is necessary for a unique and stable solution [10]. Once the available samples per subject are very limited, CRC may fail because the linear representation of the query sample $\boldsymbol{y}$ may not be accurate. To alleviate this problem, patch based CRC (PCRC) can be introduced. As shown in Fig. 1, the query image $\boldsymbol{y}$ is firstly divided into a set of overlapped block patches $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_q\}$. Then each patch $\boldsymbol{y}_j$ is represented over local dictionary $\boldsymbol{M}_j$, which is extracted from $\boldsymbol{X}$ at the corresponding location to patch $\boldsymbol{y}_j$. Since the linear system determined by local dictionary $\boldsymbol{M}_j$ tends to be under-determined, the patch based representation is more accurate than the whole image based representation. Finally, plurality or linear weighted combination can be applied to the many patch based recognition outputs for a final classification.

For each local patch, the local features such as LBP and Gabor features can be used in PCRC. Considering that the focus of this paper is to validate the effectiveness of PCRC strategy instead of local features, for simplicity and clarity the raw gray value features in each patch are used. For patch $\boldsymbol{y}_j$, its representation over $\boldsymbol{M}_j$ is obtained by

$$\hat{\boldsymbol{\rho}}_j = \arg\min_{\boldsymbol{\rho}_j} \{\|\boldsymbol{y}_j - \boldsymbol{M}_j\boldsymbol{\rho}_j\|_2^2 + \lambda \|\boldsymbol{\rho}_j\|_2^2\} \tag{2}$$

$\boldsymbol{M}_j$ is a local dictionary. Denote by $\boldsymbol{M}_{jk}$ the sub-dictionary of the k^{th} class, and each column of $\boldsymbol{M}_{jk}$ is a patch of class k. Then $\boldsymbol{M}_j = [\boldsymbol{M}_{j1}, \boldsymbol{M}_{j2}, ..., \boldsymbol{M}_{jc}]$.

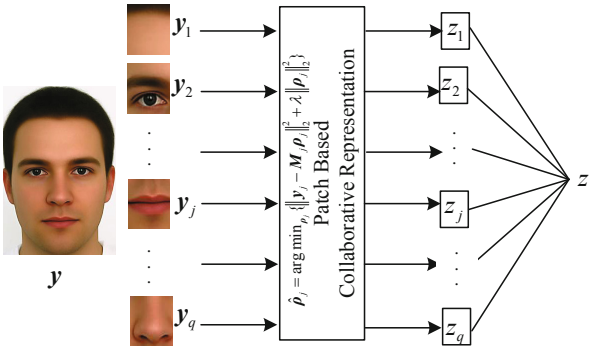

Fig. 1. Diagram of patch based collaborative representation for face classification

The recognition output z_j of patch $\boldsymbol{y}_j$ is Identity($\boldsymbol{y}_j$)= argmin$_k\{r_{jk}\}$, where $r_{jk} = \|\boldsymbol{y}_j - \boldsymbol{M}_{jk} \cdot \hat{\boldsymbol{\rho}}_{jk}\|_2 / \|\hat{\boldsymbol{\rho}}_{jk}\|_2$ and $\hat{\boldsymbol{\rho}}_j = [\hat{\boldsymbol{\rho}}_{j1}; \hat{\boldsymbol{\rho}}_{j2}; ...; \hat{\boldsymbol{\rho}}_{jc}]$.

The classification outputs of all patches can then be combined. Majority voting [15], linear weighted combination [12], kernel plurality [14] and probabilistic model [13] can be employed for the combination. As shown in [15] and [17], the weighted combination leads to little improvement compared to the simple majority voting. Hence, we use the majority voting for the final decision making.

3 Multi-scale Ensemble

In the proposed PCRC, the patch size, or we call it the patch scale in this paper, will have a great impact on the recognition performance and it is not a trivial work to pre-define an optimal scale for a database. Fig. 2 shows the FR accuracy under different patch sizes and training sample sizes on the Extended Yale B and LFW databases. One can have the following observations. First, the optimal scale varies with the number of training samples per subject. Second,

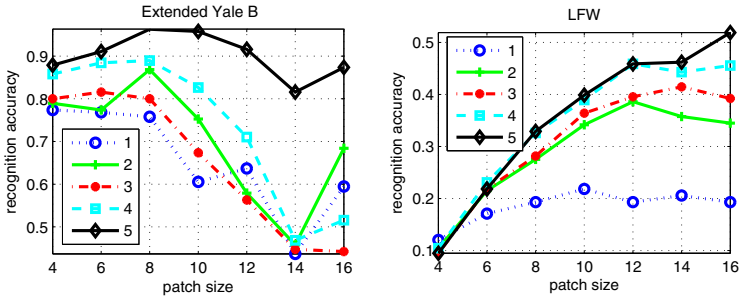

Fig. 2. Impact of patch size on PCRC (1-5 represent the training sample size per subject)

for different databases, the optimal scale also varies a lot. This difficulty can be solved by fusing the multi-scale PCRC results adaptively, via which we can not only be free of the scale selection problem but also exploit the complementary information across scales to improve the FR accuracy and robustness. To this end, we propose an ensemble learning method to combine multi-scale information optimally.

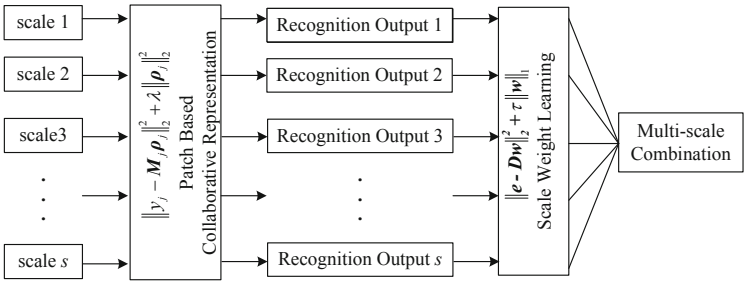

Fig. 3. Flow chart of multi-scale learning for PCRC

The flowchart of the proposed method is given in Fig. 3. On different scales with various patch sizes, we can get the recognition outputs by PCRC. We then find a set of optimal weight $\boldsymbol{w}$ to fuse the outputs. In this paper, we propose to learn $\boldsymbol{w}$ from the training samples by optimizing margin distribution.

3.1 The Objective Function for Ensemble Optimization

The multi-scale ensemble of PCRC outputs can be considered as a special classification task. Suppose there are two scales and two classes labeled as +1 and -1. For a given sample, on each scale we can have a classification output, +1 or -1, and thus the classification output on the two scales of each sample has four possible situations, as shown as the four vertexes in Fig. 4(a). Given a set of training samples, we aim to find a classification line $f = sgn(w_1 z_1 + w_2 z_2)$ that crosses the origin to make all the given samples correctly classified, where z_1 and z_2 represent the classification outputs on the two scales and w_1 and w_2 represent the weights. As to the task in Fig. 4(a), if samples on vertexes $\{A_2, A_4\}$ belong to the first class (+1) and samples on vertexes $\{A_1, A_3\}$ belong to the second class (-1), there are several classification lines that can correctly classify all the samples. Similar to feature selection [25], the importance of one scale is proportional to the weight value assigned to it.

For binary classification problems, given a set of samples $\boldsymbol{S} = \{(\boldsymbol{x}_i, z_i)\}, i = 1, 2, ...n$, $z_i \in \{+1, -1\}$ and s scales, the recognition results on s different scales form a space $\boldsymbol{H} \in \Re^{n \times s}$. Let $\boldsymbol{w} = [w_1, w_2, ..., w_s]$ be the scale weight vector and $\sum_{j=1}^{s} w_j = 1$.

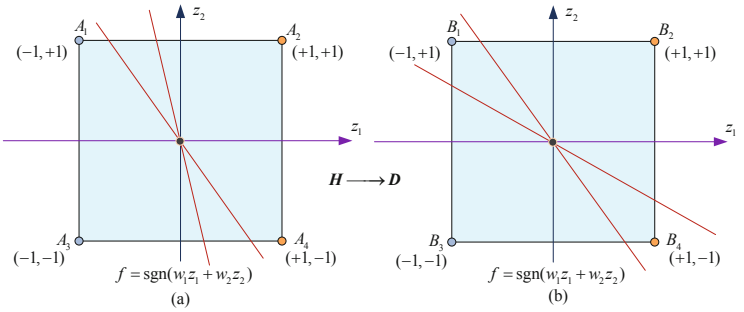

Fig. 4. Illustration of the multi-scale ensemble learning problem

Definition 1. *Given a sample* $\boldsymbol{x}_i \in \boldsymbol{S}$*, the recognition outputs on* s *different scales are* $\{h_{ij}\}, j = 1, 2, ..., s$*. The discriminant function is* $f = \mathrm{sgn}(\sum_{j=1}^{s} w_j h_{ij})$*. The margin of sample* $\boldsymbol{x}_i$ *can be defined as [26]:*

$$\varepsilon(\boldsymbol{x}_i) = z_i \sum_{j=1}^{s} w_j h_{ij} \tag{3}$$

Obviously, if $\varepsilon(\boldsymbol{x}_i) > 0$, then sample $\boldsymbol{x}_i \in \boldsymbol{S}$ is correctly classified; if $\varepsilon(\boldsymbol{x}_i) < 0$, then sample $\boldsymbol{x}_i \in \boldsymbol{S}$ is misclassified; if $\varepsilon(\boldsymbol{x}_i) = 0$, we cannot decide the label of sample $\boldsymbol{x}_i$. It is similar to linear classifiers (e.g., LSVM). Since Definition 1 is only suitable for binary classification, we define the following decision matrix in order for multi-class classification tasks.

Definition 2. *As to multi-class classification, given a sample* $\boldsymbol{x}_i \in \boldsymbol{S}$*, the recognition outputs on* s *different scales are* $\{h_{ij}\}, j = 1, 2, ..., s$*. The decision matrix* $\boldsymbol{D} = \{d_{ij}\}, i = 1, 2, ..., n, j = 1, 2, ..., s$*, is defined as:*

$$d_{ij} = g(z_i, h_{ij}) = \begin{cases} +1, \text{if } z_i = h_{ij} \\ -1, \text{if } z_i \neq h_{ij} \end{cases} \tag{4}$$

where z_i *is the label of sample* $\boldsymbol{x}_i$*.*

Clearly, $d_{ij} = +1$ means that $\boldsymbol{x}_i$ is correctly classified on the j^{th} scale. Otherwise, it is misclassified.

Definition 3. *Given a sample* $\boldsymbol{x}_i \in \boldsymbol{S}$*, the classification outputs on* s *different scales are* $\{h_{ij}\}, j = 1, 2, ..., s$*. The ensemble margin of* $\boldsymbol{x}_i \in \boldsymbol{S}$ *can be defined as:*

$$\varepsilon(\boldsymbol{x}_i) = \sum_{j=1}^{s} w_j d_{ij} \tag{5}$$

Ensemble margin reflects the misclassification degree in classifier fusion. Samples with positive margin are correctly classified. As shown in Fig. 4(b), +1 and -1 represent the elements in the decision matrix $\boldsymbol{D}$, and then the margin of samples on vertex B_2 is 1 (i.e., correctly classified on all scales) , while the margin of samples on vertex B_3 is -1 (i.e., misclassified on all scales). The margin of samples on vertexs B_1 and B_4 is between -1 and +1. In this case, how should we choose the scale weights to get better combination result? We should make the ensemble

margin as larger as possible by scale weight learning. Margin maximization is usually converted into a loss minimization problem [27][20][22].

If the ensemble margin of a sample $\boldsymbol{x}_i$ is $\varepsilon(\boldsymbol{x}_i)$, then the ensemble loss of sample $\boldsymbol{x}_i$ is

$$l_{x_i} = l(\varepsilon(x_i)) = l(\textstyle\sum_{j=1}^{s} w_j d_{ij}) \tag{6}$$

We adopt the square loss used in CRC [9], SRC [8], LS-SVM [27] and least square regression [28]. For a sample set $\boldsymbol{S}$, the ensemble square loss is

$$\begin{aligned} l(\boldsymbol{S}) &= \textstyle\sum_{i=1}^{n} l_{x_i} = \sum_{i=1}^{n} \left[1 - \varepsilon(x_i)\right]^2 \\ &= \textstyle\sum_{i=1}^{n} \left[1 - \sum_{j=1}^{s} w_j d_{ij}\right]^2 = \|\boldsymbol{e} - \boldsymbol{D}\boldsymbol{w}\|_2^2 \end{aligned} \tag{7}$$

where $\boldsymbol{e}$ is a vector whose elements are 1 and length is s.

3.2 Constrained l_1-Regularized Optimization

To learn the optimal scale weights, we should minimize the ensemble loss in Eq. (7). However, there may be many solutions that can minimize the loss for the given task, as illustrated in Fig. 4. Clearly, we should regularize the objective function in Eq. (7) in order for a unique and robust solution. In [20], Saharon et al. showed that AdaBoost approximately minimizes its loss criterion with l_1-regularization imposed on the coefficient vector. In [23], it was shown that AdaBoost optimizes margin distribution rather than minimum margin. Shawe-Taylor gave the bound on generalization error based on margin distribution for linear classifiers ($f = \boldsymbol{w}\boldsymbol{x} + b$) and showed that both the square loss (when $\sum_{j=1}^{s} w_j = 1$ and $x \in \{+1, -1\}$) and the norm of $\boldsymbol{w}$ should be minimized to improve the generalization ability [24].

Inspired by the principle of AdaBoost, we propose the following constrained l_1-regularized least square optimization to minimize the ensemble loss and solve the weights:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \arg\min_{\boldsymbol{w}} \{\|\boldsymbol{e} - \boldsymbol{D}\boldsymbol{w}\|_2^2 + \tau \|\boldsymbol{w}\|_1\} \\ &\quad s.t. \textstyle\sum_{j=1}^{s} w_j = 1, w_j > 0, j = 1, 2, ..., s \end{aligned} \tag{8}$$

where τ is the regularization parameter.

For the constraint $\sum_{j=1}^{s} w_j = 1$, it equals to $\boldsymbol{e}\boldsymbol{w} = 1$, where $\boldsymbol{e} = [1; 1; ...; 1]$ is a column vector, and then

$$\|\boldsymbol{e} - \boldsymbol{D}\boldsymbol{w}\|_2^2 = \|\boldsymbol{e} - \boldsymbol{D}\boldsymbol{w} + 1 - \boldsymbol{e}\boldsymbol{w}\|_2^2 = \|[\boldsymbol{e}; 1] - [\boldsymbol{D}, \boldsymbol{e}]\boldsymbol{w}\|_2^2 \tag{9}$$

Let $\hat{\boldsymbol{e}} = [\boldsymbol{e}; 1]$, $\hat{\boldsymbol{D}} = [\boldsymbol{D}, \boldsymbol{e}]$, then we can get

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \{\left\|\hat{\boldsymbol{e}} - \hat{\boldsymbol{D}}\boldsymbol{w}\right\|_2^2 + \tau\|\boldsymbol{w}\|_1\} \; s.t. \; w_j > 0, \; j = 1, 2, ..., s \tag{10}$$

Since the size of the decision matrix is very small (e.g., the size of decision matrix for the LFW database is 632×7 when the training sample size per subject is 5

and 7 scales are selected), $\boldsymbol{w}$ can be easily solved by some representative l_1-minimization approaches [29]. In this paper l_1_ls is used for its accuracy and stable solution [30]. The proposed ensemble learning algorithm for multi-scale PCRC (MSPCRC) is summarized in Table 1. After scale weight learning, for a query sample $\boldsymbol{x}_i$, the recognition output is $z_i = \arg\max_k \{\sum w_j | h_{ij} = k\}$.

It should be noted that though the form of multi-scale ensemble in Eq. (10) is similar to the step of coding in CRC (Eq. (1)) and SRC, their physical meanings are different. The square loss in CRC and SRC is the reconstruction error while in multi-scale ensemble learning the square loss is the function of classification margin. The l_1-norm regularization used in SRC is to sparsify the coding coefficient to enhance classification accuracy, while the l_1-norm regularization used in multi-scale ensemble learning is to suppress the effect of less-useful scales.

Table 1. The algorithm of multi-scale ensemble learning for PCRC

1: Choose s patch sizes $\delta = \{\delta_1, \delta_2, ..., \delta_s\}$
2: Get recognition outputs $\{h_{ij}\}$ by PCRC
3: Get the decision matrix
$$d_{ij} = g(z_i, h_{ij}) = \begin{cases} +1, if\ z_i = h_{ij} \\ -1, if\ z_i \neq h_{ij} \end{cases}$$
4: Learn scale weights
$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \left\| \hat{\boldsymbol{e}} - \hat{\boldsymbol{D}}\boldsymbol{w} \right\|_2^2 + \tau \left\| \boldsymbol{w} \right\|_1 \ s.t.\ w_j > 0, j = 1, 2, ..., s$$

4 Experimental Analysis

We use the Extended Yale B [31], Multi-PIE [32] and AR [33] databases in controlled environments together with the LFW database [34] in uncontrolled environments to test the FR performance of the proposed method.

The baseline CRC, SRC and NN methods, and the state-of-the-art patch based methods including BlockFLD [16], Volterrafaces [15] and patch based nearest neighbor (PNN) classifier [14] are used for comparison. As the average accuracy improvement of kernel plurality [14] compared to vote is only about 1%, we report the result of PNN and Volterrafaces with majority voting. For Volterrafaces, the best recognition performance is reported with different kernel sizes and patch sizes. As linear kernel outperforms quadratic kernel on all the four databases, we only report the performance of liner kernel for Volterrafaces. For BlockFLD [16], the performance of CS2 (combine outputs of different blocks), which is better than CS1 (combine projected blocks as a feature), is reported.

In all the following experiments, the program is run for 20 times on each database and the average results are reported. Seven scales are used in our MSPCRC method and the patch sizes are 4×4, 6×6, 8×8, 10×10, 12×12, 14×14, 16×16. In single scale based PCRC and PNN, the patches are overlapped and the patch size is set as 10×10 (overlap is 5 pixels). The parameter λ used in SRC, CRC, PCRC and MSPCRC are set as 0.001, 0.005, 0.001 and 0.001, respectively.

Parameter τ (Eq. (10)) is set as 0.1 for MSPCRC. For BlockFLD, we tried three different sizes (4×4, 8×8, 10×10 for 32×32 image and 10×10,15×15, 20×20 for 80×80 image) and report the result of the best size 8×8 (32×32 image) and 10×10 (80×80 image) for all the databases.

For scale weight learning, we divide the training set into subset1 (one image per individual is selected) and subset2 (the rest of the training set). Then samples from subset1 are classified by PCRC using subset2 as the training set on seven scales so that the weights can be learned. Obviously, as least two samples per subject are needed to learn the scale weights. Hence, we first test the performance of PCRC and MSPCRC with 2 to 5 training samples per subject. Then when there is only one sample per person, only the result of PCRC is reported.

The Matlab code of the proposed method can be downloaded at: *http : //www4.comp.polyu.edu.hk/ cslzhang/code.htm.*

4.1 Extended Yale B Database

The Extended Yale B face database [31] contains 38 human subjects under 9 poses and 64 illumination conditions. All frontal-face images marked with P00 were used in our experiment. The face images are resized to 32×32. We randomly choose 2∼5 samples from the first 32 images for training and choose 5 samples from the other 32 images for test. The experimental results are shown in Table 2. It can be clearly seen that MSPCRC achieves the highest recognition rate on all experiments with the training sample size increasing from 2 to 5. Compared to PCRC, MSPCRC leads to much better results, validating the effectiveness of multi-scale ensemble learning.

Table 2. Recognition accuracy (%) on the extended Yale B database

Method	2	3	4	5
CRC[9]	61.3±16.6	74.0±15.5	81.4±17.6	87.8±13.7
SRC[8]	64.2±17.2	74.2±15.2	82.6±16.8	89.0±12.5
NN	49.8±17.3	55.8±16.6	63.7±17.2	68.4±16.8
PNN[14]	60.8±14.4	65.6±15.1	73.8±15.8	79.7±14.6
BlockFLD[16]	79.5±8.4	83.8±7.8	88.3±5.4	90.7±5.5
Volterra[15]	69.8±12.9	79.5±12.3	84.0±9.6	86.4±9.6
PCRC	75.7±12.6	82.8±12.4	88.7±8.4	92.0±8.2
MSPCRC	**83.0±9.2**	**88.4±10.1**	**92.5±6.8**	**95.0±6.6**

4.2 Multi-PIE Database

The Multi-PIE database [32] contains a total of more than 750,000 images from 337 individuals, captured under 15 viewpoints and 19 illumination conditions in four recording sessions. A subset that contains images of 164 subjects from session 3 is selected, and there are 10 images with neutral expression and 10 images with smile expression per person. To make the FR problem more challenging, we randomly choose 2∼5 samples per subject from images with neutral expression for training

and randomly choose 3 samples from images with smile expression for test. The face images are resized to 32×32. The FR results are listed in Table 3. Similar to the results on the Extended Yale B database, PCRC and MSPCRC lead to much improvement in FR rate compared with the other methods. MSPCRC is always better than PCRC since it combines the multi-scale decisions.

Table 3. Recognition accuracy (%) on the Multi-PIE database

Method	2	3	4	5
CRC[9]	62.6±13.8	74.3±6.3	78.5±5.2	80.4±3.7
SRC[8]	61.9±14.0	73.2±8.9	78.6±6.5	80.8±4.2
NN	54.9±14.5	64.7±12.1	71.9±9.9	74.5±8.8
PNN[14]	54.4±14.9	63.2±14.0	72.3±10.7	76.7±8.8
BlockFLD[16]	66.1±6.9	71.1±5.7	76.4±4.6	79.2±3.2
Volterra[15]	52.2±11.3	57.6±7.6	62.4±6.0	65.4±4.8
PCRC	68.8±10.9	76.0±6.2	79.4±4.8	81.3±3.7
MSPCRC	**72.4±10.5**	**79.6±5.9**	**83.6±4.0**	**84.6±2.6**

4.3 AR Database

The AR face database [33] contains over 4,000 color face images of 126 people, including frontal views of faces with different facial expressions, lighting conditions and occlusions. As in [9], a subset with only illumination and expression changes that contains 50 male subjects and 50 female subjects was chosen from the AR dataset in our experiments. For each subject, we randomly choose 2∼5 samples from session 1 for training and choose 3 samples from session 2 for test. The face images are resized to 32×32.

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

Method	2	3	4	5
CRC[9]	69.9±12.6	80.6±10.4	83.8±9.6	89.1±6.2
SRC[8]	69.7±14.8	79.0±10.6	83.5±8.9	88.2±5.7
NN	48.5±9.5	54.7±9.0	58.5±9.1	63.2±7.0
PNN[14]	72.7±14.2	82.4±9.3	87.6±8.0	92.2±6.0
BlockFLD[16]	71.5±11.5	78.6±9.8	84.2±8.7	87.6±4.2
Volterra[15]	65.4±12.0	74.9±11.1	79.8±10.5	85.2±6.8
PCRC	82.2±11.3	87.7±9.4	89.9±8.5	92.9±6.7
MSPCRC	**82.3±11.5**	**87.8±10.5**	**90.2±9.1**	**93.6±7.6**

The recognition accuracy on the AR database is shown in Table 4. The proposed methods show superior performance to all the other methods. Different from the results on the Extended Yale B and Multi-PIE databases, multi-scale ensemble learning in MSPCRC only leads to a little improvement over PCRC.

That is because in this experiment the average weight value (over different training sample sizes) for scale 10×10 is about 0.9, which indicates that 10×10 is a very suitable patch size for PCRC in the AR database.

4.4 LFW Database

The LFW database [34] contains images of 5,749 different individuals in unconstrained environment. LFW-a is a version of LFW after alignment using commercial face alignment software [35]. We gathered the subjects including no less than ten samples and then get a dataset with 158 subjects from LFW-a. For each subject, 2∼5 samples are randomly chosen for training and another 2 samples for test. The images are firstly cropped to 121×121 and then resized to 32×32. The FR rates on the LFW dataset are listed in Table 5. One can see that PCRC and MSPCRC work much better than other methods, while the recognition performance is greatly improved by MSPCRC.

Table 5. Recognition accuracy (%) on the LFW database

Method	2	3	4	5
CRC[9]	24.7±2.1	31.9±2.4	37.8±2.6	42.0±3.2
SRC[8]	24.4±2.4	32.7±3.2	38.7±2.4	44.1±2.6
NN	9.3±1.7	11.4±1.8	13.0±1.7	14.3±1.9
PNN[14]	23.1±2.4	28.1±3.1	33.2±3.1	37.4±2.7
BlockFLD[16]	18.0±2.1	22.3±2.1	26.2±2.6	28.4±2.5
Volterra[15]	26.0±3.0	32.0±3.4	36.4±3.3	40.3±2.7
PCRC	32.0±1.9	37.0±2.8	40.2±2.5	42.9±2.6
MSPCRC	**35.0±1.6**	**41.1±2.8**	**46.0±3.0**	**49.0±2.9**

4.5 Single Sample Per Person (SSPP)

As there is only one sample per person, the proposed ensemble learning cannot be conducted. We report the recognition accuracy of PCRC on one scale for all the databases. The images are resized to 32×32 and 80×80, and the corresponding patch size is set as 8×8 and 20×20, respectively, for PCRC. When the image size is 80×80, the neighbor patches are used to construct the local dictionary. Since volterrafaces cannot deal with SSPP problem, its performance is not reported. BlockFLD (CS2) [16], AGL [11] and FLDA_single [36], which are methods specially designed for SSPP problem are compared. The results are listed in Table 6. The performance of PCRC is much better than SRC, CRC, NN, PNN, FLDA_single, and BlockFLD. Compared with AGL (adaptive generic learning) method, which uses an additional generic set to learn the projection matrix, the proposed PCRC shows better performance on the MPIE, AR and LFW databases without using any additional information apart from the training set.

Table 6. Recognition accuracy (%) for SSPP

32×32	Yale B	Multi-PIE	AR	LFW
CRC[9]	39.8±20.5	47.2±19.0	42.9±14.6	15.5±22.0
SRC[8]	38.7±20.5	48.2±18.6	44.9±14.8	14.7±1.9
NN	35.4±19.8	42.9±17.0	35.4±12.0	7.0±1.6
PNN[14]	45.1±18.3	40.1±17.7	54.4±19.5	15.8±2.0
BlockFLD[16]	63.1±15.0	56.9±9.7	52.1±19.8	11.8±1.4
FLDA_single[36]	39.9±21.4	43.5±14.8	37.2±10.4	6.7±1.5
AGL[11]	**75.9±12.2**	58.9±14.8	52.1±15.9	14.3±1.4
PCRC	66.5±16.3	**59.1±13.3**	**65.4±20.9**	**21.1±2.2**
80×80	Yale B	Multi-PIE	AR	LFW
CRC[9]	42.0±20.2	49.2±18.2	46.8±17.2	14.6±2.4
SRC[8]	39.3±19.6	48.3±16.7	42.0±13.3	12.6±1.8
NN	37.2±20.2	44.5±17.3	36.8±12.3	7.0±1.5
PNN[14]	57.9±18.6	49.1±17.3	61.0±19.3	16.0±2.3
BlockFLD[16]	65.7±13.3	51.9±5.6	41.9±17.8	4.9±1.3
FLDA_single[36]	41.2±20.9	39.3±10.5	32.9±12.0	8.7±1.8
AGL[11]	**79.1±12.7**	58.5±24.8	51.7±16.7	12.6±2.1
PCRC	76.7±17.4	**69.5±10.4**	**69.5±22.6**	**25.0±1.8**

5 Conclusion

In order for a more effective face recognition when the number of training samples per class is small, in this paper we proposed a patch based CRC (PCRC) method and consequently the multi-scale version of it, i.e., MCPCRC, by margin distribution optimization. The query image was partitioned into a set of overlapped patches and each patch is collaboratively represented over the corresponding set of patches of all training samples. The classification outputs of all patches were then combined by voting. However, the patch size will have a great impact on the final classification result of PCRC. Therefore, we proposed to use multiple patch sizes and then optimally combine the multi-scale outputs by margin distribution optimization with l_1-norm regularization. Our experimental results on controlled and uncontrolled face databases showed that MSPCRC outperforms not only much the CRC and SRC benchmarks, but also state-of-the-art patch based methods such as BLDA and Volterrafaces, especially when the training samples size is very small.

Acknowledgement. This work is partially supported by NSFC under Grant 61222210.

References

1. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: A literature survey. Acm Computing Surveys (CSUR) 35(4), 399–458 (2003)
2. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 947–954. IEEE (2005)

3. Tan, X., Chen, S., Zhou, Z., Zhang, F.: Face recognition from a single image per person: A survey. Pattern Recognition 39(9), 1725–1745 (2006)
4. Zhang, J., Yan, Y., Lades, M.: Face recognition: eigenface, elastic matching, and neural nets. Proceedings of the IEEE 85(9), 1423–1435 (1997)
5. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
6. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 328–340 (2005)
7. Yan, S., Xu, D., Zhang, B., Zhang, H., Yang, Q., Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(1), 40–51 (2007)
8. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 210–227 (2009)
9. Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: Int. Conf. on Comput. Vis. (2011)
10. Wright, J., Ganesh, A., Yang, A., Zhou, Z., Ma, Y.: Sparsity and robustness in face recognition. Arxiv preprint arXiv:1111.1014 (2011)
11. Su, Y., Shan, S., Chen, X., Gao, W.: Adaptive generic learning for face recognition from a single sample per person. In: CVPR 2010, pp. 2699–2706. IEEE (2010)
12. Tan, X., Chen, S., Zhou, Z., Zhang, F.: Recognizing partially occluded, expression variant faces from single training image per person with som and soft k-nn ensemble. IEEE Transactions on Neural Networks 16(4), 875–886 (2005)
13. Lin, D., Tang, X.: Recognize high resolution faces: From macrocosm to microcosm. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1355–1362. IEEE (2006)
14. Kumar, R., Banerjee, A., Vemuri, B.C., Pfister, H.: Maximizing all margins: Pushing face recognition with kernel plurality. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2375–2382 (November 2011)
15. Kumar, R., Banerjee, A., Vemuri, B.: Volterrafaces: Discriminant analysis using volterra kernels. In: CVPR 2009, pp. 150–155. IEEE (2009)
16. Chen, S., Liu, J., Zhou, Z.: Making flda applicable to face recognition with one sample per person. Pattern Recognition 37(7), 1553–1555 (2004)
17. Su, Y., Shan, S., Chen, X., Gao, W.: Hierarchical ensemble of global and local classifiers for face recognition. IEEE Transactions on Image Processing 18(8), 1885–1896 (2009)
18. Wolf, L., Hassner, T., Taigman, Y.: Effective unconstrained face recognition by combining multiple descriptors and learned background statistics. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(10), 1978–1990 (2011)
19. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 498–505. IEEE (2009)
20. Rosset, S., Zhu, J., Hastie, T.: Boosting as a regularized path to a maximum margin classifier. The Journal of Machine Learning Research 5, 941–973 (2004)
21. Reyzin, L., Schapire, R.: How boosting the margin can also boost classifier complexity. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 753–760. ACM (2006)
22. Shen, C., Li, H.: Boosting through optimization of margin distributions. IEEE Transactions on Neural Networks 21(4), 659–666 (2010)

23. Shen, C., Li, H.: On the dual formulation of boosting algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(12), 2216–2231 (2010)
24. Shawe-Taylor, J., Cristianini, N.: Robust bounds on generalization from the margin distribution. In: 4th European Conference on Computational Learning Theory, Citeseer (1999)
25. Gilad Bachrach, R., Navot, A., Tishby, N.: Margin based feature selection-theory and algorithms. In: Proceedings of the Twenty-First International Conference on Machine Learning, p. 43. ACM (2004)
26. Schapire, R., Freund, Y., Bartlett, P., Lee, W.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26(5), 1651–1686 (1998)
27. Van Gestel, T., Suykens, J., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., De Moor, B., Vandewalle, J.: Benchmarking least squares support vector machine classifiers. Machine Learning 54(1), 5–32 (2004)
28. Ramsey, J.: Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society. Series B (Methodological) 31(2), 350–371 (1969)
29. Yang, A., Ganesh, A., Zhou, Z., Sastry, S., Ma, Y.: A review of fast l1-minimization algorithms for robust face recognition. Arxiv preprint arXiv:1007.3753 (2010)
30. Kim, S., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: An interior-point method for large-scale l1-regularized least squares. IEEE Journal of Selected Topics in Signal Processing 1(4), 606–617 (2007)
31. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6), 643–660 (2001)
32. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multipie. Image and Vision Computing 28(5), 807–813 (2010)
33. Martinez, A.: The ar face database. CVC Technical Report 24 (1998)
34. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst (October 2007)
35. Wolf, L., Hassner, T., Taigman, Y.: Similarity Scores Based on Background Samples. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009, Part II. LNCS, vol. 5995, pp. 88–97. Springer, Heidelberg (2010)
36. Gao, Q., Zhang, L., Zhang, D.: Face recognition using flda with single training image per person. Applied Mathematics and Computation 205(2), 726–734 (2008)