Robust Kernel Representation with Statistical Local Features for Face Recognition

Meng Yang, *Student Member, IEEE,* Lei Zhang[1], *Member, IEEE*

Simon C. K. Shiu, *Member, IEEE,* and David Zhang, *Fellow, IEEE*

Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong, China

Abstract. Factors such as misalignment, pose variation and occlusion make robust face recognition a difficult problem. It is known that statistical features such as LBP are effective for local feature extraction, while the recently proposed sparse or collaborative representation based classification has shown interesting results in robust face recognition. In this paper, we propose a novel robust kernel representation model with statistical local features (SLF) for robust face recognition. First, multi-partition max pooling is used to enhance the SLF's invariance to image registration error. Then, a kernel based representation model is proposed to fully exploit the discrimination information embedded in the SLF, and robust regression is adopted to effectively handle the occlusion in face images. Extensive experiments are conducted on benchmark face databases, including Extended Yale B, AR, Multi-PIE, FERET, FRGC and LFW, which have various variations of lighting, expression, pose and occlusions, demonstrating the promising performance of the proposed method.

Keywords: robust kernel representation, sparse representation, collaborative representation, face recognition

[1] Corresponding author. Email: cslzhang@comp.polyu.edu.hk. This work is supported by the HK RGC PPR grant (PolyU5019-PPR-11).

1. Introduction

Automatic face recognition (FR) is one of the most active and visible research topics in computer vision, machine learning and biometrics [11] due to its wide range of applications such as access control, video surveillance, and the like. After many years' investigation, FR is still very challenging due to the low quality of face images [1], and the rich variations of facial images from the same or different subjects, e.g., lighting, expression, occlusion, misalignment, etc [11]. In order for different communities to benchmark and verify their FR methods, many large scale face databases, such as FERET [22-23], FRGC [35], LFW [24][32] and PubFig [25], have been established and used as evaluation platforms.

Although facial images have a high dimensionality, their discriminative characteristics usually lie or can be extracted in a lower dimensional subspaces or sub-manifolds. Therefore, subspace and manifold learning methods have been dominantly used in appearance based FR [2-9][42]. Classical methods such as the Eigenface and Fisherface [2-3][42] mainly consider the global scatter of training samples and may fail to reveal the essential data structures nonlinearly embedded in the high dimensional space. The manifold learning methods were proposed to overcome this limitation [5-6], and the representative manifold learning methods include locality preserving projection (LPP) [7], local discriminant embedding (LDE) [8], unsupervised discriminant projection (UDP) [9], etc. In addition, kernel based subspace learning was also proposed for FR. For instance, Yang *et al.* [60] presented a Kernel Fisher discriminant framework for feature extraction and recognition; Zafeiriou *et al*. [4] proposed a robust approach to discriminant kernel-based feature extraction for face recognition and verification.

The subspace or manifold learning methods only consider the holistic feature of face images, which are usually very sensitive to the variations of misalignment, pose, and occlusion. Recent researches have shown that local feature based methods [16-18][43-48][26] are very promising in object recognition, texture classification and uncontrolled FR. Gabor filters, which could effectively extract local directional features on multiple scales, have been successfully used in FR [17-18]. Compared to the holistic feature based approaches such as Eigenface [2] and FisherFace [3], Gabor filtering is less sensitive to image variations (e.g., illumination, expression). Another type of local feature widely used in FR is statistical local feature (SLF), such as histogram of local binary pattern (LBP) [43]. The main idea is that a face image can be seen as a composition of micro-patterns [26]. By partitioning the face image into several blocks, the statistical

feature (e.g., histogram of LBP) of these blocks is extracted, and finally the description of the image is formed by concatenating the extracted features in all blocks. Zhang *et al.* [45-46] proposed to use Gabor magnitude or phase map instead of the intensity map to generate LBP features. New coding technologies on Gabor features have also been proposed. In [47], Zhang *et al.* extracted and encoded the global and local variations of the real and imagery parts in multi-scale Gabor representation. Xie *et al.* [48] proposed local Gabor XOR patterns (LGXP), which utilizes XOR (*exclusive or*) to encode the local variation of Gabor phase, to fuse Gabor magnitude and phase information. These local pattern based statistical features have shown very promising results in large scale face databases, such as FERET [22-23] and FRGC [35].

Apart from the employed features, the employed classifier is also important to the performance of FR. Nearest Neighbor (NN), SVM and Hidden Markov Models are the widely used classifiers in face recognition [43][45-48][59][27]. Moreover, in order to better exploit the prior knowledge that face images from the same subject construct a subspace, nearest subspace (NS) classifiers [19][36-38][51][58] were also developed, which are usually superior to the popular NN classifier. Recently an interesting classifier, namely sparse representation based classification (SRC), was proposed by Wright *et al.* [10] for robust FR. In Wright *et al.*'s work, a testing image is sparsely coded on the whole training set by l_1-norm minimization, and then classified to the class that yields the least coding residual. By assuming that the outlier pixels in the face image are sparse and by using an identity matrix to code the outliers, SRC shows good robustness to face occlusion and corruption. SRC has been attracting much interest and has been widely studied in the computer vision research community [28-31]. Very recently, Zhang *et al.* [33] indicated that the l_1-norm sparsity may not be the key of the success of SRC, and they proposed the collaborative representation based classification (CRC), which uses l_2-norm to regularize the coding coefficients instead of the time consuming l_1-norm, for FR and achieved similar result to SRC but with much less time complexity.

Although the statistical local features (SLF) and SRC/CRC have shown powerful abilities in the field of feature extraction and signal classification, few works have been proposed to integrate them together for better performance. Many works either use NN/NS/SVM as the classifier with SLF as inputs (e.g., NN in [43][45-48]) or use SRC/CRC to do classification with holistic features [10][31][33]. Although the methods [12-13] aim to combine LBP and sparse representation together, no effective representation model was proposed to deal with variations such as occlusion and misalignment, etc.

In this paper, we proposed a novel SLF based robust kernel representation (RKR) model for FR. First, we propose a multi-partition max pooling technology to enhance the invariance of local features to image registration error (e.g., misalignment). Second, we propose a robust kernel representation model, which not only uses kernel representation to fully exploit the discrimination information embedded in the local features, but also adopts a robust regression function as the measure to effectively handle the occlusion in facial images. Compared to the previous classification methods, e.g., NN with SLF features and SRC with holistic features, the proposed SLF based RKR model shows much stronger robustness to various face image variations (e.g., illumination, expression, occlusion and misalignment), as demonstrated in our extensive experiments conducted on benchmark face databases.

The rest of the paper is organized as follows. Section 2 briefly reviews some related work. Section 3 presents the proposed SLF based robust kernel representation algorithm. Section 4 presents the experimental results. Section 5 summarizes the paper.

2. Related Work

2.1. Statistical Local Feature

The extraction of statistical local features (SLF) has three steps: feature map generation, pattern map coding, and histogram computing. The commonly used feature maps include original intensity map [43] and Gabor feature maps (e.g., magnitude [45], phase [46]). LBP [43][45-46], local XOR (exclusive or) operator [48] or others [47][49] could be adopted for pattern map coding. Finally the encoded pattern map is partitioned into non-overlapping blocks, in which the local histogram feature is computed. The descriptor of the input face image is the concatenation of all the histograms computed in each block.

2.2. Sparse Representation or Collaborative Representation based Classifier

Different from Nearest Neighbor (NN) and Nearest Subspace (NS) classifiers [19][36-38][51][58], which forbids representing the query sample across classes, the recently developed l_1-regularized sparse representation [10] or l_2-regularized collaborative representation [33] represents the query image by the training samples from all classes, which could effectively overcome the small-sample-size or overfitting problem of NN and NS. Let $\boldsymbol{X}_i = [\boldsymbol{s}_{i,1}, \boldsymbol{s}_{i,2}, ..., \boldsymbol{s}_{i,n_i}] \in \Re^{m \times n_i}$ denote the set of training samples of the i^{th} object

class, where $\boldsymbol{s}_{i,j}$, j=1,2,…,n_i, is an m-dimensional vector stretched by the j^{th} sample of the i^{th} class[2]. Let $\boldsymbol{y} \in \Re^m$ be a query sample to be classified. The representation model of sparse representation based classifier (SRC) or collaborative representation based classifier (CRC) could be written as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ \left\| \boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{\alpha} \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \right\|_{l_p} \right\} \tag{1}$$

where $\boldsymbol{X}$=[$\boldsymbol{X}_1$, $\boldsymbol{X}_2$, …, $\boldsymbol{X}_c$] and c is the number of classes; $\left\| \cdot \right\|_{l_p}$ is the l_p-norm, and p=1 for SRC in [10], while p=2 for CRC in [33].

The classification of $\boldsymbol{y}$ is done by

$$\text{identity}\left(\boldsymbol{y} \right) = \arg\min_i \left\{ \left\| \boldsymbol{y} - \boldsymbol{X}_i \delta_i \left(\hat{\boldsymbol{\alpha}} \right) \right\|_2 \right\} \tag{2}$$

where $\delta_i \left(\cdot \right) : \Re^n \to \Re^{n_i}$ is the characteristic function that selects from $\hat{\boldsymbol{\alpha}}$ the coefficients associated with the i^{th} class [10]. It is shown in [33] that CRC has very competing accuracy with SRC in FR without occlusion but with much faster speed. In the case of occlusion or corruption, Robust-SRC [10] classifies the occluded face image $\boldsymbol{y}$ by

$$\text{identity}\left(\boldsymbol{y} \right) = \arg\min_i \left\{ \left\| \boldsymbol{y} - \boldsymbol{X}_i \delta_i \left(\hat{\boldsymbol{\alpha}} \right) - \boldsymbol{X}_e \hat{\boldsymbol{\alpha}}_e \right\|_2 \right\} \tag{3}$$

where

$$\left[\hat{\boldsymbol{\alpha}}; \hat{\boldsymbol{\alpha}}_e \right] = \arg\min_{\boldsymbol{\alpha}, \boldsymbol{\alpha}_e} \left\{ \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} - \boldsymbol{X}_e \boldsymbol{\alpha}_e \right\|_2^2 + \lambda \left\| \left[\boldsymbol{\alpha}; \boldsymbol{\alpha}_e \right] \right\|_1 \right\} \tag{4}$$

and $\boldsymbol{X}_e$ is an occlusion dictionary to code the outliers. $\boldsymbol{X}_e$ is simply set as the identity matrix $\boldsymbol{I}$ in [10].

2.3. Robust Sparse Coding

The representation model of Robust-SRC [10] is equivalent to

$$\min_{\boldsymbol{\alpha}} \left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha} \right\|_1 \text{ s.t. } \left\| \boldsymbol{\alpha} \right\|_1 \le \sigma \tag{5}$$

which is actually a Maximum Likelihood Estimation (MLE) of $\boldsymbol{\alpha}$ when the representation residual $\boldsymbol{y}$-$\boldsymbol{X}\boldsymbol{\alpha}$ follows Laplacian distribution. However, for the real occlusion and disguise in practical facial images, the representation residual rarely follows Laplacian model, making robust-SRC less effective to handle occlusions in FR.

Yang *et al.* [31] proposed a robust sparse coding model to achieve robust face recognition with outliers.

[2] More generally, $\boldsymbol{s}_{i,j}$, j=1,2,…,n_i, could be the feature vector extracted from the j^{th} sample of the i^{th} class.

Instead of using l_1-norm to regularize the data fidelity term in the coding model, Yang *et al.* formulated the signal representation as an MLE-like estimator:

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \rho_{\theta}\left(y_i - \boldsymbol{r}_i\boldsymbol{\alpha}\right) \text{ s.t. } \|\boldsymbol{\alpha}\|_1 \le \sigma \tag{6}$$

where $\boldsymbol{r}_i$ is the i^{th} row vector of $\boldsymbol{X}$ and y_i is the i^{th} element of $\boldsymbol{y}$. This robust sparse coding could be efficiently solved by an iterative reweighted sparse coding algorithm. In each iteration, the original robust sparse coding model becomes

$$\min_{\boldsymbol{\alpha}} \left\|\boldsymbol{W}^{1/2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\right)\right\|_2^2 \text{ s.t.} \|\boldsymbol{\alpha}\|_1 \le \sigma \tag{7}$$

where $\boldsymbol{W}$ is a diagonal matrix with $W_{i,i} = \omega(e_i) = 1/\left(1+\exp\left(\mu e_i^2 - \mu\delta\right)\right)$, $e_i = y_i - \boldsymbol{r}_i\boldsymbol{\alpha}$, μ and δ are two automatically updated scalar parameters in the weight function [31]. After the representation coefficient $\hat{\boldsymbol{\alpha}}$ is obtained, the weighted representation residual is used for classification, i.e., identity($\boldsymbol{y}$)=argmin$_i$||$\boldsymbol{W}^{1/2}$($\boldsymbol{y}$-$\boldsymbol{X}_i\delta_i(\hat{\boldsymbol{\alpha}})$)||$_2$.

3. Statistical Local Feature based Robust Kernel Representation

3.1. Multi-partition max pooling (MPMP)

Facial image misalignment caused by factors such as scaling, translation and rotation can make a lot of troubles in less-controlled face recognition system. Even using some advanced face detector (e.g., the Viola and Jones' face detector [53]) to crop and align the query face image, there are still registration errors of several pixels, which will deteriorate much the FR performance [54]. Although there are some pre-processing methods [55][52] to align the query face image to the well cropped training images, it is more interesting that we could improve the robustness of the feature extraction step to face misalignment. In this section, we propose a simple but very effective pooling technique to this end.

Pooling techniques are widely used in object and image classification to extract invariant features. In general, there are two categories of pooling methods, sum pooling [50][57] and max pooling [39][56-57]. Denote by $\boldsymbol{f}_i$ the i^{th} feature vector in a pool, and by $\{\boldsymbol{f}\}_j$ the j^{th} element of the feature vector $\boldsymbol{f}$. In the case of sum pooling, the output feature vector $\boldsymbol{f}_s$ is computed by $\{\boldsymbol{f}_s\}_j = \{\boldsymbol{f}_1\}_j + \{\boldsymbol{f}_2\}_j + \ldots + \{\boldsymbol{f}_n\}_j$, while in the case of max pooling the output feature $\boldsymbol{f}_m$ is $\{\boldsymbol{f}_m\}_j = \max\{|\{\boldsymbol{f}_1\}_j|, |\{\boldsymbol{f}_2\}_j|, \ldots, |\{\boldsymbol{f}_n\}_j|\}$. A simple 1-D example with $f_1 \in [0,1]$ and $f_2 \in [0,1]$ is shown in Fig. 1. It can be seen that the domain of (f_1, f_2) with $f_m=1$ is larger than the

domain of (f_1, f_2) with f_s=1, which indicates that max pooling is more robust to the changes of f_1 or f_2. The experiments in [39][56-57] also show that max pooling is more robust than sum pooling to image spatial variations. In addition, spatial discrimination information can be introduced by using spatial pyramid, which divides the images into multi-scale regions (e.g., 1×1, 2×2, and 4×4 for a total 21 regions in [39][56].)

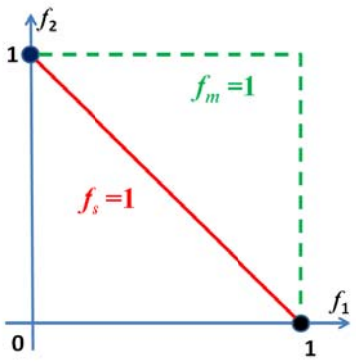

Figure 1: A simple 1-D example for illustrating sum pooling and max pooling.

In the paper we propose a multi-partition max pooling (MPMP) scheme for the statistical feature of local pattern. The main differences between the proposed MPMP and previous max pooling methods lie in the image partition and feature generation. Different from the partition of spatial pyramid, such as 1×1, 2×2, and 4×4, we adopt a more flexible partition. As shown in the first row of Fig. 2, for example, the partition of the pattern map (e.g., LBP) can be made as 2×2, 3×3 and 4×4, respectively, with 29 blocks of three different sizes in total. This kind of partition could flexibly set the number of blocks in each scale and is expected to capture more spatial discrimination information than the spatial pyramid.

In the proposed MPMP based statistical local feature (SLF) extraction, we adopt S+1 level block partition, where s=0, 1, … , S. That is to say, in the s^{th} level, the whole image is divided into $P_s \times Q_s$ blocks, each of which is further partitioned into $p_s \times q_s$ sub-blocks. The pooling technology is operated on a series of local features generated in each partitioned sub-block. Different from the feature generation (e.g., the coding coefficients of local patches' descriptors, such as SIFT or raw intensity) in previous works of pooling [39][50][56], here we extract a sequence of statistical local features (SLF) which are simpler and widely used in face recognition. As shown in the second row of Fig. 2, in each sub-block we first create a sequence of sliding boxes (e.g., the red box shown in Fig. 2), and then compute the histogram of each box's local feature (e.g., LBP). Here the size of the box is smaller than the block, and usually the height and width of the

box are set as $ratio_s$ ($ratio_s < 1$) times of those of the sub-block in the s^{th} scale partition. In this paper, MPMP is defined as the one with the following setting: p_s=2 and q_s=2 for partition scale s=0 and 1; p_s=1 and q_s=1 for s>1; $ratio_s$=1 for s=0; and $ratio_s$ = 0.5 for other values of s.

Take the feature generation in one sub-block as an example. Denote by $\boldsymbol{f}_i$ the feature vector (e.g., the histogram feature) extracted from the i^{th} sliding box, and suppose that there are n feature vectors, $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$, which are extracted from all possible sliding boxes in this sub-block, and then the final output feature vector, denoted by $\boldsymbol{f}$, after max pooling is

$$\{\boldsymbol{f}\}_j = \max\{|\{\boldsymbol{f}_1\}_j|, |\{\boldsymbol{f}_2\}_j|, \ldots, |\{\boldsymbol{f}_n\}_j|\} \tag{8}$$

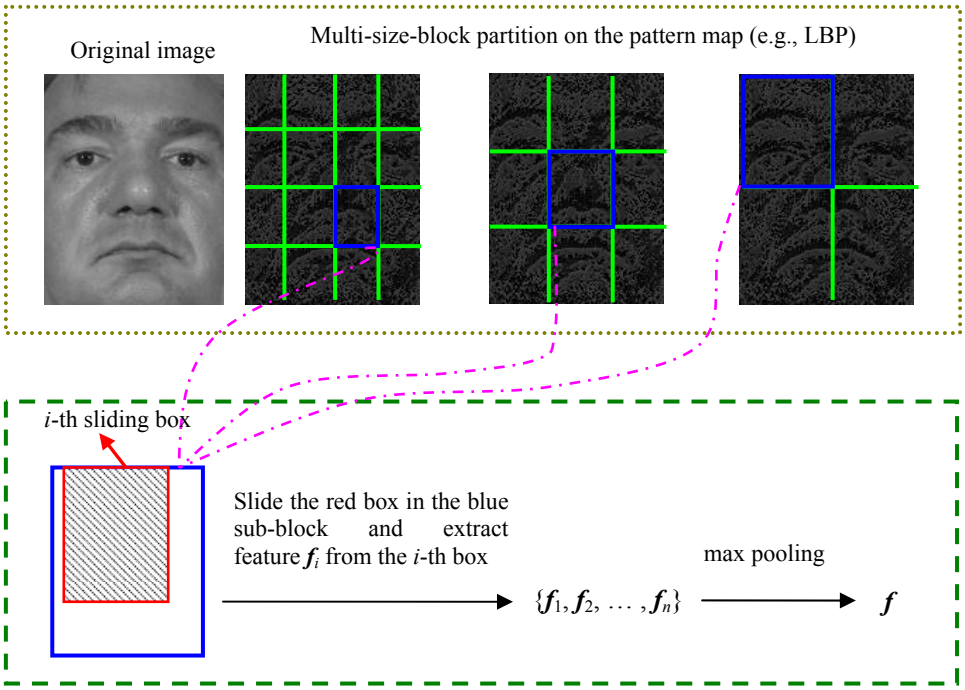

Figure 2: Illustration of the proposed multi-partition max pooling.

Let's suppose that the image is partitioned into B blocks in total. In each block, after extracting the MPMP based SLF of every sub-block, we concatenate the SLFs of all sub-blocks as the output feature vector. Denote by $\boldsymbol{y}_i$ the output feature vector in the i^{th} block. Then the concatenation of all feature vectors extracted from all blocks, i.e., $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_B]$ could be taken as the descriptor of the image. The proposed MPMP based SLF could not only introduce more spatial information to LSF due to its use of multi-partition, but also enhance the robustness of LSF to image misalignment due to its use of max pooling.

3.2. Robust kernel representation

How to measure the similarity of two features is an important issue in pattern classification. The commonly used classifiers, such as the linear SVM, NN and NS classifiers [19][36-38][51][58], as well as the SRC and CRC classifiers [10][33], often adopt the l_2-norm to measure the distance (i.e., Euclidean distance). Apart from l_2-norm based measurement, kernel methods have become increasingly popular for pattern classification, especially face recognition [4][60]. The kernel trick could map the non-linearly separable features into a high dimensional feature space, in which features of different classes can be more easily separated by linear classifiers. From the view of kernel representation, l_2-norm measurement, which could be regarded as a linear kernel, is effective to solve the linearly separable problem. For SLF, more specifically the local histogram feature, it has been shown that histogram intersection and Chi-square distances are more powerful than l_2-norm distance in classification [26][43-48]. Therefore, more discriminant information embedded in SLF could be exploited if the histogram intersection kernel [34] or Chi-square kernel could be adopted in the l_2-norm distance based classifiers such as SRC and CRC. However, directly applying these kernels to SLF based representation may not be robust to facial occlusions. In this section, we propose a new model, namely robust kernel representation, to improve the robustness of SLF based face representation and classification.

Suppose that there exists a kernel function $\kappa(\boldsymbol{v}_j, \boldsymbol{v}_k) = \langle \phi(\boldsymbol{v}_j), \phi(\boldsymbol{v}_k) \rangle$, where $\langle \cdot \rangle$ is the inner product operator, and $\phi : \Re^d \to \Re^h$ is a feature mapping function, which maps the feature vectors $\boldsymbol{v}_j$ and $\boldsymbol{v}_k$ to a higher dimensional feature space. For a matrix $\boldsymbol{Z} = [z_1, z_2, \cdots, z_q] \in \Re^{p \times q}$, we define $\boldsymbol{K}_{\boldsymbol{ZZ}}$ as a $q \times q$ matrix with $\{\boldsymbol{K}_{\boldsymbol{ZZ}}\}_{j,k} = \kappa(z_j, z_k)$ and $\boldsymbol{k}_{\boldsymbol{Zv}}$ as a $q \times 1$ vector with $\{\boldsymbol{k}_{\boldsymbol{Zv}}\}_j = \kappa(z_j, \boldsymbol{v})$. Denote by $\phi(\boldsymbol{Z}) = [\phi(z_1), \cdots, \phi(z_q)]$, we could have

$$\boldsymbol{K}_{\boldsymbol{ZZ}} = \phi(\boldsymbol{Z})^T \phi(\boldsymbol{Z}); \;\; \boldsymbol{k}_{\boldsymbol{Zv}} = \phi(\boldsymbol{Z})^T \phi(\boldsymbol{v}) \tag{9}$$

After the MPMP based SLF extraction on the query image, B blocks of multiple partitions are obtained, and B sub-feature vectors, denoted by $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_B$, are extracted. Similarly, for each of the training samples, we can extract the sub-feature vectors, and let's denote by $\boldsymbol{A}_i$ the matrix formed by all the sub-feature vectors of the i^{th} block from all training samples. Take the i^{th} block as an example, the kernel representation of $\boldsymbol{y}_i$ over the matrix $\boldsymbol{A}_i$ could be formulated as

$$\min_{\boldsymbol{\alpha}} \left\| \phi(\boldsymbol{y}_i) - \phi(\boldsymbol{A}_i)\boldsymbol{\alpha}_i \right\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}_i\|_{l_p} \le \sigma \tag{10}$$

where $\boldsymbol{\alpha}_i$ is the coding coefficient vector in the high dimensional feature space mapped by the kernel function ϕ. If we enforce that $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_j$ for different blocks $i \neq j$, i.e., we assume that the different blocks $\boldsymbol{y}_i$ extracted from the same test sample have the same representation over their associated matrix $\boldsymbol{A}_i$, then kernel representation of the query image by combining all the block features could be written as

$$\min_{\boldsymbol{\alpha}} \left\| \left[\phi(\boldsymbol{y}_1); \phi(\boldsymbol{y}_2); \cdots; \phi(\boldsymbol{y}_B)\right] - \left[\phi(\boldsymbol{A}_1); \phi(\boldsymbol{A}_2); \cdots; \phi(\boldsymbol{A}_B)\right]\boldsymbol{\alpha} \right\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_{l_p} \le \sigma \tag{11}$$

where $\boldsymbol{\alpha}$ is the coding coefficient vector of the query sample. The above model seeks the regularized representation for a mapped feature under the mapped basis in the high dimensional space.

In the kernel representation model Eq. (11), the l_2-norm is used to measure the representation residual. Such a kernel representation is effective when there are no outliers in the query image. However, in FR the facial occlusion and facial disguises (e.g., sunglasses and scarf) can often appear in the query face image. In such case, the block in which outliers appear will have a big representation residual, reducing the role of clean blocks in the final classification. In short, the representation model in Eq. (11) is very sensitive to outliers [31].

In order to make the kernel representation robust to block occlusion and disguises, we propose to adopt some robust fidelity term in the modeling. Denote by $\boldsymbol{e} = [e_1, e_2, \ldots, e_B]$ the representation residual vector, where e_i is the kernel representation residual of the i^{th} block, i.e., $e_i = \sqrt{\left\| \phi(\boldsymbol{y}_i) - \phi(\boldsymbol{A}_i)\boldsymbol{\alpha} \right\|_2^2}$. We assume that e_i is independent from e_j if $i \neq j$ since they represent the representation residuals of different blocks.

The proposed robust kernel representation can then be formulated as

$$\min_{\boldsymbol{\alpha}} \rho(\boldsymbol{e}) \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_{l_p} \le \sigma \tag{12}$$

where $\rho(\boldsymbol{e}) = \sum_{i=1}^{B} \rho(e_i)$ and the cost function $\rho(\cdot)$ is expected to be insensitive to the outliers in the query sample. Usually, we require that $\rho(0)$ is the global minimal of $\rho(x)$ and $\rho(x_1) > \rho(x_2)$ if $|x_1| > |x_2|$. Without loss of generality, we let $\rho(0)=0$.

Obviously, if we define the cost as $\rho(e_i) = (e_i)^2$ (i.e., $\rho(\boldsymbol{e}) = \|\boldsymbol{e}\|_2^2$), the robust kernel representation in Eq. (12) will be reduced to the normal kernel representation in Eq. (11). However, as shown in Fig. 3, this simple setting of $\rho(x)$ will make the representation very sensitive to outliers because the cost (i.e., $\rho(e_i)$) of

those representation residuals corresponding to outliers are often very big. We can also set $\rho(e_i)=|e_i|$ (i.e., $\rho(\boldsymbol{e})=\|\boldsymbol{e}\|_1$). As can be seen in Fig. 3, $\rho(e_i)=|e_i|$ is much less sensitive to outliers than $\rho(e_i)=(e_i)^2$ since the absolute value of an outlier's representation residual is less significant than its square. However, with $\rho(e_i)=|e_i|$ Eq. (12) is difficult to solve because $|e_i|$ is not differentiable, while $|e_i|$ is not bounded with e_i, making $\rho(e_i)$ not robust enough to large outliers. Intuitively, if we can find a function $\rho(e_i)$ such as the blue curve in Fig. 3, which is differentiable and bounded when $|e_i|$ is big, then a good instantiation of the robust kernel representation in Eq. (12) can be implemented.

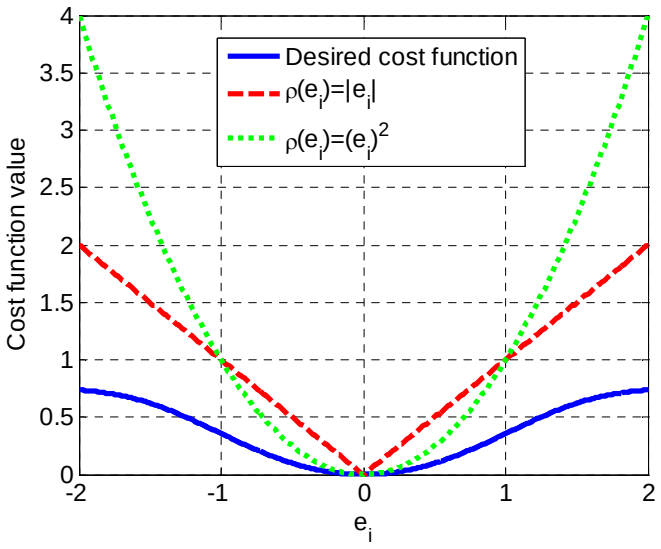

Figure 3: Three typical settings of the cost function $\rho(e_i)$.

3.3. Solution of the robust kernel representation

After doing Taylor expansion of $\rho(\boldsymbol{e})$ in the neighborhood of $\boldsymbol{e}_0$, an approximation of $\rho(\boldsymbol{e})$ could be written as

$$\tilde{\rho}(\boldsymbol{e})=\frac{1}{2}\left\|\boldsymbol{W}^{1/2}\boldsymbol{e}\right\|_2^2+b_{\boldsymbol{e}_0} \tag{13}$$

where $b_{\boldsymbol{e}_0}$ is a scalar constant determined by $\boldsymbol{e}_0$, $\boldsymbol{W}$ is a diagonal matrix and its i^{th} diagonal element is $\boldsymbol{W}_{i,i}=\omega(e_{0,i})=\rho'(e_{0,i})/e_{0,i}$, ρ' is the derivative of ρ, and $e_{0,i}$ is the i^{th} element of $\boldsymbol{e}_0$. According to the property (i.e., $\rho(x_1)>\rho(x_2)$ if $|x_1|>|x_2|$) of ρ, we could see that $\boldsymbol{W}_{i,i}$ is a positive scalar. Clearly, $\omega(\cdot)$ can be viewed as a weight function applied to $\boldsymbol{e}$. A good weight function should be robust to outliers, i.e., $\omega(e_i)$ has big value when $|e_i|$ is small (e.g., blocks without outliers), and small value when $|e_i|$ is big (e.g., blocks with outliers). The widely used logistic function can be chosen as the weight function:

$$\omega(e_i) = 1 \Big/ \left(1 + \exp\left(\mu e_i^2 - \mu\delta\right)\right) \tag{14}$$

The above weight function could effectively assign the outliers with big representation residual low weights, and assign inliers with small representation residual high weights (here the weight value is normalized to the range of [0, 1]). It should be noted that the weight values of each testing sample are estimated online, and there is not a training phase of them. The corresponding cost function ρ to the weight function in Eq. (14) will be differentiable and bounded, as the blue curve shown in Fig. 3.

With the above development, the original robust kernel representation in Eq. (12) could be approximated by

$$\min_{\boldsymbol{\alpha}} \left\| \boldsymbol{W}^{1/2}\boldsymbol{e} \right\|_2^2 \quad \text{s.t.} \quad \left\| \boldsymbol{\alpha} \right\|_{l_p} \le \sigma \tag{15}$$

After some derivation, Eq. (15) could be rewritten as

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{B} \omega_i \left\| \phi(\boldsymbol{y}_i) - \phi(\boldsymbol{A}_i)\boldsymbol{\alpha} \right\|_2^2 \quad \text{s.t.} \quad \left\| \boldsymbol{\alpha} \right\|_{l_p} \le \sigma \tag{16}$$

where ω_i is computed by Eq. (14) with $e_i = \sqrt{\left\| \phi(\boldsymbol{y}_i) - \phi(\boldsymbol{A}_i)\boldsymbol{\alpha}_0 \right\|_2^2}$, and $\boldsymbol{\alpha}_0$ is an known coding coefficient vector. Here μ and δ are scalar parameters, which could be set as a constant value or automatically updated. $\mu\delta$ is usually set as 8 to make the weight close to 1 when e_i=0, δ is set as the $\lfloor \tau B \rfloor$ largest elements of the set $\{ e_i^2 \mid i=1,\ldots, B\}$, where $\lfloor \tau B \rfloor$ outputs the largest integer smaller than τB, and τ is discussed in the section 4.1.

With the defined kernel matrix $\boldsymbol{K}_{\boldsymbol{ZZ}}$ and kernel vector $\boldsymbol{k}_{\boldsymbol{Zv}}$ in Eq. (9), Eq. (16) could be re-written as

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{B} \omega_i k(\boldsymbol{y}_i, \boldsymbol{y}_i) + \boldsymbol{\alpha}^T \sum_{i=1}^{B} \omega_i \boldsymbol{K}_{\boldsymbol{A}_i\boldsymbol{A}_i} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \sum_{i=1}^{B} \omega_i \boldsymbol{k}_{\boldsymbol{A}_i\boldsymbol{y}_i} \quad \text{s.t.} \quad \left\| \boldsymbol{\alpha} \right\|_{l_p} \le \sigma \tag{17}$$

From Eq. (17) we can see that the weighted-sum kernel terms, including $\sum_{i=1}^{B} \omega_i k(\boldsymbol{y}_i, \boldsymbol{y}_i)$, $\sum_{i=1}^{B} \omega_i \boldsymbol{K}_{\boldsymbol{A}_i\boldsymbol{A}_i}$, and $\sum_{i=1}^{B} \omega_i \boldsymbol{k}_{\boldsymbol{A}_i\boldsymbol{y}_i}$, could exploit the discrimination information in the mapped higher dimensional feature space; at the same time, the weight ω_i can effectively remove the outliers' effect on computing the coding vector. The coding vector $\boldsymbol{\alpha}$ is regularized by l_p-norm. In this paper, we discuss two important cases: p=1 for sparse regularization and p=2 for non-sparse regularization. When p=1, l_1-norm minimization methods such as the efficient feature-sign search algorithm [40] could be used to solve the sparse coding problem of Eq. (17). When p=2, a closed-form solution of Eq. (17) could be derived as $\hat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^{B} \omega_i \boldsymbol{K}_{\boldsymbol{A}_i\boldsymbol{A}_i} + \lambda \boldsymbol{I} \right)^{-1} \sum_{i=1}^{B} \omega_i \boldsymbol{k}_{\boldsymbol{A}_i\boldsymbol{y}_i}$, where λ is the Lagrange multiplier.

Because the approximation of $\tilde{\rho}(\boldsymbol{e})$ (i.e., Eq. (13)) is the Taylor expansion of $\rho(\boldsymbol{e})$ in the neighborhood of $\boldsymbol{e}_0$, the solving of robust kernel representation (Eq. (12)) is an iterative and alternative process: the weight value (i.e., ω_i in Eq. (17)) is estimated via Eq. (14) with known coding coefficient, and then the coding coefficient is computed via Eq. (17) with known weight value. After getting the solution $\hat{\boldsymbol{\alpha}}$ after some iterations, the classification of the query sample is done via

$$\text{identity}(\boldsymbol{y}) = \min_j \left\{ \sum_{i=1}^{B} \omega_i \varepsilon_{i,j} \right\} \tag{18}$$

where $\varepsilon_{i,j} = \left\| \phi(\boldsymbol{y}_i) - \phi(\boldsymbol{A}_{i,j}) \hat{\boldsymbol{\alpha}}_j \right\|_2^2$ is the i^{th}-block kernel representation residual associated with the j^{th} class, $\boldsymbol{A}_i = \left[\boldsymbol{A}_{i,1}, \boldsymbol{A}_{i,2}, \cdots, \boldsymbol{A}_{i,c} \right]$ with $\boldsymbol{A}_{i,j}$ being the sub-matrix of $\boldsymbol{A}_i$ associated with the j^{th} class, and $\hat{\boldsymbol{\alpha}} = \left[\hat{\boldsymbol{\alpha}}_1; \hat{\boldsymbol{\alpha}}_2; \cdots; \hat{\boldsymbol{\alpha}}_c \right]$ with $\hat{\boldsymbol{\alpha}}_j$ being the representation coefficient vector associated with the j^{th} class. From Eq. (18) it can be seen that the classification criteria is based on a weight sum of kernel representation residuals, which utilizes both the discrimination power of kernel representation in high dimensional feature space and the insensitiveness of robust representation to outliers. In addition, the kernel representation residual, $\varepsilon_{i,j}$, could be rewritten as $\varepsilon_{i,j} = k(\boldsymbol{y}_i, \boldsymbol{y}_i) + \hat{\boldsymbol{\alpha}}_j^T \boldsymbol{K}_{\boldsymbol{A}_{i,j}\boldsymbol{A}_{i,j}} \hat{\boldsymbol{\alpha}}_j - 2\hat{\boldsymbol{\alpha}}_j^T \boldsymbol{k}_{\boldsymbol{A}_{i,j}\boldsymbol{y}_i}$.

3.4 The algorithm

The whole algorithm of the proposed statistical local feature based robust kernel representation (SLF-RKR) is summarized in Table 1. It includes three steps. The first step extracts the SLF using the proposed MPMP. The second step performs robust kernel representation, and the last step performs classification. Given the feature type (e.g., histogram of LBP) and the partition parameters of MPMP (e.g., S, $ratio_s$, P_s and Q_s), the algorithm of SLF-RKR could be run. The second step is an iterative process. By experiments, we found that this process converges fast. For instance, when there is no occlusion, only 2 or 3 iterations are needed, and when there is occlusion in the query image, about 10 iterations can lead to a good solution. We denote by SLF-RKR_l_1 and SLF-RKR_l_2 the implementations of SLF-RKR model with l_1-norm regularization and l_2-norm regularization, respectively.

The time complexity of SLF-RKR mainly lies in MPMP based SLF extraction and solving the robust kernel representation. According to the characteristics of histogram feature, we can adopt the integral image method [53] to speed up MPMP based SLF extraction. For each pixel in a sub-block, only 2 additions are

needed in computing integral image and 3 additions are needed in computing histogram bin value. So the computing of each histogram bin for this sub-block needs $3hw(1\text{-}ratio)^2$ additions and 1 max operation, where h and w are the height and width of the sub-block, and *ratio* is the parameter of the sliding box. For the robust kernel representation, in the case of FR without occlusion, the weight ω_i in each block could be fixed as 1. Since the matrix inverse in the closed-form solution (i.e., $\hat{\boldsymbol{\alpha}} = \left(\sum_{i=1}^{B} \omega_i \boldsymbol{K}_{\boldsymbol{A}_i \boldsymbol{A}_i} + \lambda \boldsymbol{I}\right)^{-1} \sum_{i=1}^{B} \omega_i \boldsymbol{k}_{\boldsymbol{A}_i \boldsymbol{y}_i}$) of SLF-RKR_l_2 could be computed offline, SLF-RKR_l_2 with ω_i=1 has time complexity of $O(n^2)$, where n is the number of training samples. The solution to SLF-RKR_l_1 with ω_i=1 can be obtained by standard sparse coding. The time complexity of l_1-norm sparse coding with an $m \times n$ dictionary is about $O(m^2 n^{1.5})$ [62], while the l_1-norm minimizers such as the efficient feature-sign search algorithm [40] used in this paper can have a much faster speed in practice. Therefore for FR without occlusion, SLF-RKR_l_2 with ω_i=1 is much faster than SRC [10], while the time complexity of SLF-RKR_l_1 with ω_i=1 is similar to that of SRC [10].

For FR with occlusion or disguise, the weight ω_i in each block needs to be updated online. In this case, the time complexity of SLF-RKR_l_2 will increase to about T times of that of SLF-RKR_l_2 with ω_i=1, where T is the total number of iterations to update ω_i. For SLF-RKR_l_1 with updated weight, the step a) (i.e., weighted kernel representation with p=1) is an iterative process itself, and the steps b), c) and d) could be operated in each iteration of step a). Overall, the time complexity of SLF-RKR_l_1 with updated weight is almost the same as that of SLF-RKR_l_1 with ω_i=1, since the former has almost the same solving procedure as the latter with only an additional step to update weight in each iteration. In FR with occlusion/disguise, SRC needs an additional occlusion matrix to code the occlusion, and thus its time complexity is very high.

The running speed of SLF-RKR is very fast. Under the programming environment of Matlab version R2011a in a desktop of 1.86HHz CPU with 2.99G RAM, the running time of SRC (executed by the fast l_1-norm minimizer such as feature-sign search algorithm [40] or Dual ALM [61]) and SLF-RKR is compared in Table 2. In the experiment of AR database with 7 training samples per class (refer to Section 4.2 for the detailed experimental setting), the average running time of SLF-RKR_l_2 and SLF-RKR_l_1 is 0.0418 second and 0.1806 second, respectively; while the average running time of SRC is 0.1239 second. In the experiment of Extended Yale B with 50% occlusion (refer to Section 4.4 for the detailed experimental setting), the average running time of SLF-RKR_l_2 and SLF-RKR_l_1 is 0.8073 second and 0.8439 second, respectively, which are much less than that of SRC (1.8800 seconds).

Table 1: Algorithm of statistical local feature based robust kernel representation (SLF-RKR).

Statistical Local Feature based Robust Kernel Representation (SLF-RKR)

1. Extract statistical local features via multi-partition max pooling.
2. Robust kernel representation:

   Initialize the weight in each block as 1: $\omega_i = 1$.

   While not converge, *Do*

   a). Weighted kernel representation:

   $$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{B} \omega_i k(\boldsymbol{y}_i, \boldsymbol{y}_i) + \boldsymbol{\alpha}^T \sum_{i=1}^{B} \omega_i \boldsymbol{K}_{\boldsymbol{A}_i \boldsymbol{A}_i} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \sum_{i=1}^{B} \omega_i \boldsymbol{k}_{\boldsymbol{A}_i \boldsymbol{y}_i} \quad \text{s.t.} \|\boldsymbol{\alpha}\|_{l_p} \le \sigma$$

   b). Compute the reconstruction residual in each block:

   $$e_i^2 = \left\|\phi(\boldsymbol{y}_i) - \phi(\boldsymbol{A}_i)\hat{\boldsymbol{\alpha}}\right\|_2^2 = k(\boldsymbol{y}_i, \boldsymbol{y}_i) + \hat{\boldsymbol{\alpha}}^T \boldsymbol{K}_{\boldsymbol{A}_i \boldsymbol{A}_i} \hat{\boldsymbol{\alpha}} - 2\hat{\boldsymbol{\alpha}}^T \boldsymbol{k}_{\boldsymbol{A}_i \boldsymbol{y}_i}$$

   c). Estimate the weight value as

   $$\omega_i = 1 \Big/ \left(1 + \exp\left(\mu e_i^2 - \mu\delta\right)\right),$$

   where μ=8/δ, $\delta = \psi_1(\boldsymbol{e})_{\lfloor \tau B \rfloor}$, $\lfloor \tau B \rfloor$ outputs the largest integer smaller than τB, and $\psi_1(\boldsymbol{e})_k$ is the k^{th} largest element of the set $\{e_j^2, j=1,\ldots,B\}$ [31].

   d). Checking convergence condition:

   $$\sum_i \left(\omega_i^{(t)} - \omega_i^{(t-1)}\right)^2 \Big/ \sum_i \left(\omega_i^{(t-1)}\right)^2 < \gamma,$$

   where γ is a small positive scalar and $\omega_i^{(t)}$ is the weight value of block i in iteration t.

   End *While*

3. Do classification:

   $$\text{identity} = \min_j \left\{ \sum_{i=1}^{B} \omega_i k(\boldsymbol{y}_i, \boldsymbol{y}_i) + \hat{\boldsymbol{\alpha}}_j^T \sum_{i=1}^{B} \omega_i \boldsymbol{K}_{\boldsymbol{A}_{i,j} \boldsymbol{A}_{i,j}} \hat{\boldsymbol{\alpha}}_j - 2\hat{\boldsymbol{\alpha}}_j^T \sum_{i=1}^{B} \omega_i \boldsymbol{k}_{\boldsymbol{A}_{i,j} \boldsymbol{y}_i} \right\}$$

   where $\boldsymbol{A}_{i,j}$ is the sub-matrix of $\boldsymbol{A}_i$ associated with the j^{th} class and $\hat{\boldsymbol{\alpha}}_j$ is the representation coefficient vector associated with the j^{th} class.

Table 2: Average running time (second) of SLF-RKR and SRC.

Method	AR database	Extended Yale B with 50% occlusion
SLF+SRC	0.1239	1.8800
SLF-RKR_l_1	0.1806	0.8439
SLF-RKR_l_2	0.0418	0.8073

4. Experimental Results

In this section, we present experimental results on benchmark face databases to illustrate the effectiveness of our method. In section 4.1, we discuss the parameter setting. In section 4.2 we present the experimental results on Extended Yale B [58][20] and AR [21] databases captured in controlled environments. In section 4.3 we demonstrate the robustness of SLF-RKR to pose variation and misalignment. Then in section 4.4, we

test FR against block occlusion and real disguise. Finally, the comprehensive evaluations on large-scale face databases, including FERET [22-23], FRGC [35] and LFW [32], are presented in section 4.5.

4.1 Parameter setting

The proposed SLF-RKR consists of two main procedures: feature extraction and robust kernel representation. If no specific instruction, the parameters of SLF-RKR are set as what shown in Table 3. In feature extraction, the histogram of LBP encoded on the raw image is used as the SLF, and the number of histogram bins for each sub-block is set to 16. In the proposed MPMP based SLF extraction, we set S=0, P_0=5, and Q_0=4 for FR with well aligned images. For FR with registration error (e.g., misalignment and pose), we set S=3, and (P_s, Q_s)={(5,4), (3,2), (4,2), (2,1)} for s={0, 1, 2, 3}. In the procedure of robust kernel representation, the histogram intersection kernel [34] (i.e., $\kappa(\boldsymbol{v}_j, \boldsymbol{v}_k) = \sum_l \min\{v_{j,l}, v_{k,l}\}$ with $v_{j,l}$ and $v_{k,l}$ the l^{th} entry of $\boldsymbol{v}_j$ and $\boldsymbol{v}_k$, respectively) is used as the kernel function. In the online updating of weights, we set τ=0.6 for FR with occlusion and τ=0.8 for FR without occlusion. The Lagrange multiplier λ of SLF-RKR_l_1 (refer to Eq. (17)) is set as 0.005, while the Lagrange multiplier λ of SLF-RKR_l_2 is usually set as a larger value (e.g., 0.1) for l_2-norm regularization is weaker than l_1-norm regularization.

Table 3: Parameter setting of SLF-RKR.

Procedure		parameter setting
Feature extraction	MPMP	P_0=5, Q_0=4 when S=0; (P_0,Q_0)=(5,4), (P_1,Q_1)=(3,2), (P_2,Q_2)=(4,2), (P_3,Q_3)=(2,1) when S=3.
	histogram bin number	16
Robust kernel representation	kernel function	histogram intersection kernel
	weight update	τ=0.6 for occlusion; τ=0.8 for non-occlusion
	Lagrange multiplier	λ=0.005 (SLF-RKR_l_1); λ=0.1(SLF-RKR_l_2)

4.2 Face recognition on Extended Yale B and AR

We first evaluate the performance of the proposed algorithm on two representative face image databases captured in controlled environment: Extended Yale B [58][20] and AR [21]. The original SRC with holistic Eigenface feature [10] is used as the baseline method, and we then apply the proposed MPMP based SLF feature to SRC [10], CRC [33], Linear Regression for Classification (LRC) [38], histogram intersection kernel based Support Vector Machine (HISVM) and Nearest Neighbor (NN) with histogram intersection as its similarity measurement, and compare them with SLF-RKR.

1) *Extended Yale B Database:* The Extended Yale B database consists of 2,432 frontal-face images of 38 individuals (each subject has 64 samples), captured under various laboratory-controlled lighting conditions [58][20]. For each subject, N_{tr} samples are randomly chosen as training samples and 32 of the remaining images are randomly chosen as the testing data. Here the images are normalized to 96×84 and the experiment for each N_{tr} runs 10 times.

The FR results, including mean recognition accuracy and standard variance, of all the competing methods are listed in Table 4. The proposed SLF-RKR achieves the best performance, with more than 2% improvement over all the others when N_{tr} is small (e.g., 5, and 10). When 20 training samples are selected, an accuracy of 99.5% is achieved by SLF-RKR. It could also be seen that those methods based on collaborative representation (e.g., SLF-RKR, SLF+CRC, SLF+SRC and original SRC) are more powerful than other kinds of linear representation methods (e.g., SLF+LRC, SLF+NN).

Table 4: Face recognition results (%) on Extended Yale B database.

N_{tr}	5	10	20
Original SRC [10]	80.0±0.82	91.4±0.70	97.3±0.49
SLF+NN	59.7±1.70	76.8±1.30	89.7±0.87
SLF+LRC	59.0±1.70	78.9±1.80	93.3±0.83
SLF+HISVM	72.0±2.20	91.6±0.18	99.0±0.32
SLF+CRC	83.0±1.90	95.5±0.88	99.2±0.32
SLF+SRC	82.8±1.80	95.5±0.91	99.3±0.30
SLF-RKR_l_1	85.6±1.80	97.4±0.76	**99.5±0.18**
SLF-RKR_l_2	**85.8±1.80**	**97.5±0.73**	**99.5±0.18**

2) *AR database:* The AR database consists of over 4,000 frontal images from 126 individuals [21]. For each individual, 26 pictures were taken in two separate sessions. As in [10], in the experiment we chose a subset of the dataset consisting of 50 male subjects and 50 female subjects. For each subject, the seven images with illumination change and expressions from Session 1 were used for training, and the other seven images with only illumination change and expression from Session 2 were used for testing. The size of original face image is 83×60. The recognition rates of all the competing methods versus different number training samples are listed in Table 5. In each test we selected the first N_{tr} training samples as the training data set. We could see that SLF-RKR achieves the highest recognition rates, followed by SLF+SRC and SLF+CRC. In all cases with less than 6 training samples, at least 2% improvement of SLF-RKR over other methods could be achieved. In this experiment, original SRC gets the worst results for that the holistic

features (e.g., Eigenfaces) has much less discrimination information than the statistical local feature (e.g., histogram of LBP) in dealing with variations of expression and time.

Table 5: Face recognition results (%) on the AR database.

N_{tr}	2	3	4	5	6	7
Original SRC [10]	67.0	70.1	77.9	87.4	93.7	93.1
SLF+NN	88.1	88.7	92.3	97.0	98.0	98.3
SLF+LRC	83.3	82.7	85.0	90.0	93.7	94.3
SLF+HISVM	86.7	87.0	90.6	94.1	96.6	96.6
SLF+CRC	87.9	87.4	88.0	93.9	98.3	98.3
SLF+SRC	87.6	88.0	89.9	95.7	98.7	98.8
SLF-RKR_l_1	90.1	91.0	**92.4**	97.0	**99.4**	**99.4**
SLF-RKR_l_2	**90.6**	**91.1**	92.0	**97.4**	**99.4**	**99.4**

Apart from LBP, recently Tzimiropoulos *et al.* [63] utilized image gradient orientation as local feature to perform subspace learning for face recognition. The results in [63] show that image gradient orientation could lead to better performance than LBP. For example, with image gradient orientation feature, the recognition rate could be 95.65% on Extended Yale B by using 5 training samples per subject, while the recognition rate could be 98.66% on AR by using the first 4 training samples of session 1 per subject. From Table 4 and Table 5, we can see that the recognition rate of the proposed SLF-RKR could achieve 99.5% on Extended Yale B and 99.4% on AR by using LBP as its SLF. This clearly shows that the use of robust kernel representation significantly increases the recognition rates. Further improvement could be achieved for SLF-RKR if image gradient orientation is used to design the statistical local feature. In addition, in this experiment the l_1-norm regularization and l_2-norm regularization in SLF-RKR lead to little difference in the recognition rates, but the later has much less time complexity.

4.3 Robustness to misalignment and pose

In this section, we test the robustness of the proposed method to local deformation, including image misalignment introduced by face detector and pose variation. Here the number of histogram bin in each sub-block is set to 30.

Table 6: Face recognition rates (%) on the MPIE databases with misalignment. The numbers in round brackets show the recognition rates of SLF-RKR without MPMP.

Session	Session 2	Session 3	Session 4
Original SRC	33.6	36.3	35.3
SLF+NN	66.0	67.0	64.7
SLF+LRC	60.8	60.4	58.1
SLF+HISVM	67.6	66.6	64.3
SLF+CRC	77.1	74.7	74.0
SLF+SRC	80.2	77.3	77.4
SLF-RKR_l_1	**84.5***(81.4)*	**83.3***(80.2)*	**82.6***(79.0)*
SLF-RKR_l_2	84.3*(79.5)*	82.1*(78.4)*	80.8*(75.0)*

1) *Large-scale Multi-PIE database:* The CMU Multi-PIE database [41] contains images of 337 subjects captured in four sessions with simultaneous variations in pose, expression, and illumination. In the experiments, all the 249 subjects in Session 1 were used. For the training set, we used the 7 frontal images with illuminations {0,1,7,13,14,16,18} and neutral expression. For the testing sets, 10 typical frontal images of even-number illuminations taken with neutral expressions from Session 2 to Session 4 were used. Here the training samples are cropped and normalized to 90×72 based on the coordinates of manually located eye centers; while the testing samples are automatically detected using Viola and Jone's face detector [53] without manual intervention, and thus there are often some misalignments in the testing samples.

Table 6 lists the results of all the competing methods. It can be seen that the proposed SLF-RKR achieves the highest recognition rates, with at least 4%, 5% and 3% improvements than all the other methods in Session 2, Session 3 and Session 4, respectively. The original SRC with Eigenfaces gets the worst recognition rates, much lower than SLF+SRC. This validates that SLF is robust to misalignment to some extent. Collaborative representations (e.g., CRC and SRC) combined with SLF could have about 10% improvements over other kinds of classifiers (e.g., HISVM, LRC and NN). In addition, SLF-RKR_l_1 slightly outperforms SLF-RKR_l_2 in this experiment. In order to show the effectiveness of MPMP, we also give the recognition rate of SLF-RKR without the step of MPMP in Table 6. One can see that even without MPMP, SLF-RKR_l_1 still outperforms SLF+SRC by 1.9% in average, while SLF-RKR_l_2 outperforms SLF+CRC by 2.3%. It can also be observed that the improvement introduced by MPMP is over 3% in each session, which clearly show the effectiveness of the proposed MPMP in dealing with misalignment.

2) *FERET pose database:* In this experiment we use the FERET pose dataset [22-23], which includes 1,400 images from 198 subjects (about 7 each). This subset is composed of the images marked with '*ba*', '*bd*', '*be*', '*bf*', '*bg*', '*bj*', and '*bk*'. Some sample images of one person are shown in Fig. 4. Four tests with

different pose angles were performed, with images marked with '*ba*', '*bj*' and '*bk*' as gallery, and the images with '*bg*', '*bf*', '*be*' and '*bd*' as probes, respectively.

Here the image is cropped and normalized to 80×80. Because the width and height of the face image are equal, we reset the parameters in the MPMP based SLF extraction as follows: $(P_s, Q_s) = \{(4,4), (2,2), (2,2), (1,1)\}$ for $s=\{0, 1, 2, 3\}$. The experimental results on the tests with different pose variations are listed in Table 7. The proposed SLF-RKR significantly outperforms all the methods. In particular, it has at least 6.5%, 7%, and 17.5% improvement over all the other methods in the testing data with -25 degree, 15 degree and 25 degree pose variations, respectively. The original SRC and SLF+HISVM have the worst performance since Eigenface feature is sensitive to pose variation and HISVM cannot learn pose variation from frontal training set. We also give in Table 7 the results of SLF-RKR without MPMP on all poses. Similar conclusion to that in Multi-PIE could be made, i.e., significant improvements (e.g., over 13% improvement when pose degrees are $\pm 25^{o}$) could be achieved by using MPMP.

ba: gallery bj: expression bk: illumination be: +15 bf: -15 bg: -25 bd: +25

Figure 4: Some samples of a subject in the pose subset of the FERET database.

Table 7: Face recognition rates (%) on the FERET pose subset. The numbers in round brackets show the recognition rates of SLF-RKR without MPMP.

Pose (degree)	-25	-15	15	25
Original SRC	32.5	70.5	57.5	28.0
SLF+NN	48.5	94.5	88.0	36.5
SLF+LRC	45.0	94.0	79.0	32.5
SLF+HISVM	12.5	89.0	52.5	13.5
SLF+CRC	34.5	99.0	82.0	35.5
SLF+SRC	31.5	99.5	87.5	37.0
SLF-RKR_l_1	55.0*(42.0)*	**100***(99.5)*	**96.0***(88.0)*	**57.0***(39.0)*
SLF-RKR_l_2	**56.5***(38.5)*	99.5*(99.5)*	95.0*(88.0)*	54.5*(36.0)*

From the above experiments of FR with misalignment and pose variation, we could conclude that the proposed SLF-RKR could not only increase the discrimination and invariance of local features, but also has powerful classification ability due to the use of robust kernel representation. In addition, we find that SLF-RKR_l_1 is slightly better than SLF-RKR_l_2 but with more computational costs.

4.4 Robustness to occlusion and disguise

Facial occlusion and disguise are very challenging issues in FR. One interesting property of SRC [10] is its robustness to face occlusions. In this section, we test the performance of SLF-RKR to various occlusions, including block occlusion and real disguise. In SLF-RKR, the robustness to occlusion mainly comes from its iterative reweighed kernel robust representation. In this section the weight $\boldsymbol{W}$ in each block is automatically updated. The state-of-the-art methods to deal with face occlusion, including the robust version of SRC [10] (i.e., using l_1-norm to characterize the representation residuals), kernel version of SRC (KSRC) [28] (i.e., using RBF kernel to map the original feature to a higher dimensional feature space), kernel version of CRC (KCRC), and Robust Sparse Coding (RSC) [31], are employed to compare with SLF_RKR.

1) *FR with random block occlusion:* In the database of Extended Yale B [58][20], we chose Subsets 1 and 2 (717 images, normal-to-moderate lighting conditions) for training, and Subset 3 (453 images, more extreme lighting conditions) for testing. Similar to the settings in [10], we simulate various levels of contiguous occlusion, from 0% to 60%, by replacing a randomly located square block of each testing image with an unrelated image, as illustrated in Fig. 5, where (a) shows a face image with 30% block occlusion and (b) shows a face image with 40% block occlusion. Here the location of occlusion is randomly chosen for each image and is unknown to each algorithm, and the image size is normalized to 96×84.

Table 8 lists the FR results versus various levels of occlusions. Here λ of SLF-RKR_l_1 is set as 0.1. From Table 8, we can see that almost all methods could correctly classify all the testing samples when occlusion level is from 0% to 20%. However, when occlusion percentage is larger than 20%, the advantage of SLF-RKR over other methods becomes significant. For instance, when occlusion is 50%, SLF-RKR could achieve at least 94% recognition accuracy, compared to at most 87.4% for other methods. For SLF-RKR_l_1, when there is 60% block occlusion, it can still achieve a recognition rate of over 84%. This clearly demonstrates the effectiveness of the proposed SLF-RKR method to deal with face occlusion. In addition, both KCRC and KSRC could get better performance than CRC, but worse performance than SRC and SLF-RKR. This is because the l_1-norm fidelity term of the robust version of SRC can deal with outliers to some extent; however, the RBF kernel is sensitive to signal's outliers.

(a) (b)

Figure 5: The examples of face images with occlusion: (a) 30% block occlusion; (b) 40% block occlusion.

Table 8: Face recognition rates (%) of different methods under different levels of block occlusion.

Occlussion	0%	10%	20%	30%	40%	50%	60%
Robust SRC[10]	100	100	99.8	98.5	90.3	65.3	27.8
RSC [31]	100	100	**100**	**99.8**	96.9	87.4	60.8
SLF+NN	100	100	**100**	98.9	96.0	85.9	57.4
SLF+LRC	100	100	98.9	96.0	89.8	69.5	35.3
SLF+HISVM	100	99.6	98.7	97.8	90.3	77.5	52.3
SLF+CRC	100	99.8	98.7	93.6	85.4	61.8	33.6
SLF+KCRC	100	100	99.1	96.9	88.5	66.4	34.7
SLF+SRC	100	100	99.8	98.0	95.1	82.3	50.6
SLF+KSRC	100	99.8	98.0	94.9	88.5	67.3	36.0
SLF-RKR_l_1	**100**	**100**	**99.6**	**99.6**	**99.6**	**96.7**	**84.8**
SLF-RKR_l_2	**100**	**100**	**99.6**	**99.6**	98.5	94.0	77.9

2) *FR with disguise:* A subset of 50 males and 50 females are selected from the AR database [21]. For each subject, 7 samples without occlusion from session 1 are used for training, with all the remaining samples with disguises for testing. These testing samples (including 3 samples with sunglass in Session1, 3 samples with sunglass in Session 2, 3 samples with scarf in Session 1 and 3 samples with scarf in Session 2 per subject) not only have disguises, but also have variations of time and illumination. Here the image size is normalized to 83×60. Table 9 lists the FR results on the four test sets with disguise. It can be seen that in the two tests of Session 1, the proposed methods achieve 100% recognition accuracy, much higher than the state-of-the-art results reported in literature, for examples, 83.3% (Sunglass-S1) and 48.7% (Scarf-S1) for original SRC, and 94.7% (Sunglass-S1) and 91.0% (Scarf-S1) for RSC. In the two tests of Session 2, the improvement of SLF-RKR over all the other methods is at least 6%, which clearly shows the superior classification ability of SLF-RKR. The SLF-RKR_l_1 is slightly better than SLF-RKR_l_2 in the tests of Session 2, which again shows that l_1-norm regularization could introduce more discrimination into the coding coefficients but at the price of speed. For large-scale database, SLF-RKR_l_2 can be a good candidate to balance the recognition accuracy and running speed. In addition, like in the experiments of FR with block occlusion, one can see that KCRC and KSRC have lower recognition rates than SLF-RKR and SRC since the standard RBF kernel is not robust to outliers.

Table 9: Face recognition rates (%) on the challenging datasets with real disguise.

	Sunglass-S1	Scarf-S1	Sunglass-S2	Scarf-S2
Robust SRC [10]	83.3	48.7	49.0	29.0
RSC [31]	94.7	91.0	80.3	72.7
SLF+NN	98.7	98.0	82.3	88.7
SLF+LRC	96.7	92.0	68.7	68.7
SLF+HISVM	97.0	95.7	70.3	78.7
SLF+CRC	99.7	98.7	80.3	86.7
SLF+KCRC	**100**	98.3	82.7	88.0
SLF+SRC	**100**	99.0	85.0	90.7
SLF+KSRC	**100**	98.3	84.0	86.7
SLF-RKR_l_1	**100**	**100**	**93.0**	**97.6**
SLF-RKR_l_2	**100**	**100**	91.3	96.0

4.5 Face recognition on large-scale face database

Finally, we verify the performance of SLF-RKR on three large scale face databases: FERET [22-23], FRGC [35] and LFW [32]. To demonstrate the effectiveness of SLF-RKR, we also report the results of KSRC and KCRC with RBF kernel. Considering that SLF-RKR_l_2 has similar recognition accuracy to SLF-RKR_l_1 but has much lower time complexity, in this section we only report the results of SLF-RKR_l_2. We update the weight of SLF-RKR_l_2 and set the number of histogram bin in each sub-block as 30.

1) *FERET database:* The FERET database [22-23] is often used to validate an algorithm's effectiveness because it contains many kinds of image variations. By taking Fa subset as a gallery, the probe subsets Fb and Fc were captured with expression and illumination variations (the images in Fc were captured by a different camera). Especially, Dup1 and Dup2 consist of images that were taken at different times. For some people, more than two years elapsed between the gallery set and Dup1 or Dup2 set.

The image size is normalized to 150×130. Table 10 lists the FR results of competing methods. Because each subject in the gallery set has only one training sample, the LRC is equivalent to NN so that we only report the result of NN classifier. The proposed SLF-RKR_l_2 achieves the best performance in all tests. Especially, it achieves much higher performance than the competitors on Dup 1 and Dup 2. The proposed RKR has 7.5% and 5.1% average improvement over CRC and SRC, respectively. Standard kernel (e.g., RBF) could improve the performance of CRC and SRC, but still 6.1% and 3.5% lower than SLF-RKR_l_2 in average. It is also interesting that the collaborative representation based classifiers (e.g., SRC, CRC, KSRC, KCRC, and RKR) still have much higher recognition rates than NN and HISVM in the case that each subject has only one training sample.

Table 10: Face recognition rates (%) on FERET database.

Method	Fb	Fc	Dup1	Dup2
SLF+NN	94.6	50.5	59.7	45.7
SLF+HISVM	95.3	54.6	61.8	41.9
SLF+CRC	98.1	85.1	74.4	60.3
SLF+KCRC	98.2	87.1	75.1	63.2
SLF+SRC	98.2	85.6	77.7	66.2
SLF+KSRC	98.3	88.6	78.3	68.8
SLF-RKR_l_2	**99.2**	**89.2**	**81.9**	**77.8**

Another widely used statistical local feature is the histogram of LBP encoded on the Gabor magnitude [45]. In addition, the block based Fisher's linear discriminant (BFLD) proposed in [48] has shown powerful ability to extract the discriminative low-dimensional feature in each block. Therefore, here we compare the proposed SLF-RKR (using Gabor magnitude based SLF) with the state-of-the art methods on FERET database. The feature dimensionality extracted by BFLD in each block is set to 400 and Gaussian kernel $\kappa(\boldsymbol{\nu}_j, \boldsymbol{\nu}_k) = \exp\left\{-\left\|\boldsymbol{\nu}_j - \boldsymbol{\nu}_k\right\|_2^2 \Big/ 2\xi^2\right\}$ is used in SLF-RKR. The results of SLF-RKR_l_2, SLF+NN, SLF+SVM and other state-of-the-art methods are listed in Table 11. It shows that the proposed SLF-RKCR_l_2 not only outperforms SLF+NN and SLF+SVM in all cases, but also has better performance than the best methods reported in literature. Especially, SLF-RKCR_l_2 has recognition accuracies of 96.3% and 94.4% in Dup1 and Dup2, respectively, which may be the best results so far.

Table 11: Face recognition rates (%) of SLF-RKR and other state-of-the-art methods on the FERET database.

Method	Fb	Fc	Dup1	Dup2
SLF+NN	**99.7**	98.5	92.8	91.0
SLF+SVM	99.7	98.5	92.8	91.0
Tan's [14]	98.0	98.0	90.0	85.0
Zou's [15]	99.5	**99.5**	85.0	79.5
Xie's [48]	99.0	99.0	94.0	93.0
SLF-RKR_l_2	**99.7**	**99.5**	**96.3**	**94.4**

2) *FRGC 2.0:* FRGC version 2.0 [35] is a large-scale face databases designed with uncontrolled indoor and outdoor settings. We use the subset (352 subjects having no less than 15 samples in the original target set) of Experiment 4, which is the most challenging dataset in FRGC 2.0 with large lighting variations, ageing and image blur. Some examples are shown in Fig. 6. The selected target set contains 5,280 samples, and the query set has 7,606 samples. The image is normalized to 168×128. The feature dimensionality extracted by BFLD in each block is set to 220 and Gaussian kernel is used in SLF-RKR.

Three tests with 5, 10 and 15 target samples for each subject are made in the experiments. The recognition rates of SLF+NN, SLF+LRC, SLF+HKSVM, SLF+CRC, SLF+SRC, SLF+KCRC, SLF+KSRC, and the proposed SLF-RKR are listed in Table 12. Again, SLF-RKR performs the best, though the improvement is not significant since there are no occlusion, misalignment and pose variations in the query set.

(a) (b)

Figure 6: Samples of FRGC 2.0. (a) and (b) are samples in target and query sets, respectively.

Table 12: Face recognition result (%) with Gabor magnitude based SLF on the FRGC database.

Method	5	10	15
SLF+NN	0.948	0.966	0.974
SLF+SVM	0.676	0.742	0.797
SLF+LRC	0.951	0.969	0.977
SLF+CRC	0.946	0.967	0.976
SLF+KCRC	0.942	0.965	0.973
SLF+SRC	0.948	0.970	0.977
SLF+KSRC	0.952	0.974	0.980
SLF-RKR_l_2	**0.955**	**0.976**	**0.981**

3) *LFW:* Labeled Faces in the Wild (LFW) [32] is a large-scale database of face photographs designed for unconstrained FR with variations of pose, illumination, expression, misalignment and occlusion, etc. Some examples are shown in Fig. 7. Two subsets of aligned LFW [24] are used in the experiments. In subset 1 (LFW6) which consists of 311 subjects with no less than 6 samples per subject, we use the first 5 samples as training data and the remaining samples as testing data. In subset 2 (LFW11) which consists of 143 subjects with no less than 11 samples per subject, we use the first 10 samples as training data and the remaining samples as testing data.

(a) (b)

Figure 7: Samples of LFW. (a) and (b) are samples in training and testing sets, respectively.

Table 13 lists the FR results of competing methods with the MPMP based SLF. The image is normalized to 127×116. We can see that SLF-RKR still achieves the best performance. Compared with the second best method, SLF+KCRC/SLF+KSRC, the improvements of SLF-RKR are about 6% in LFW6 and 3% in LFW11, respectively, which clearly demonstrates the powerful classification ability of the proposed robust kernel representation.

Table 13: Face recognition rates (%) on the LFW database.

Method	SLF+ NN	SLF+ LRC	SLF+ HISVM	SLF+ SRC	SLF+ KSRC	SLF+ CRC	SLF+ KCRC	**SLF-RKR_l_2**
LFW6	0.302	0.338	0.443	0.535	0.555	0.540	0.549	**0.619**
LFW11	0.459	0.529	0.630	0.755	0.779	0.768	0.788	**0.819**

5. Conclusion

In this paper, we proposed a statistical local feature based robust kernel representation (SLF-RKR) model for face recognition. A robust representation model to image outliers (e.g., occlusion and real disguise) was built in the kernel space, and a multi-partition max pooling technology was proposed to enhance the invariance of local pattern feature to image misalignment and pose variation. We evaluated the proposed method on different conditions, including variations of illumination, expression, misalignment and pose, as well as block occlusion and disguise occlusion. One big advantage of SLF-RKR is its high face recognition rates and robustness to various occlusions. The extensive experimental results demonstrated that SLF-RKR is superior to state-of-the-arts and has great potential to be applied in practical face recognition systems.

References

[1] H. Huang and H.T. He, "Super-Resolution Method for Face Recognition Using Nonlinear Mappings on Coherent Features," *IEEE Trans. Neural Networks*, vol. 22, no. 1, pp. 121-130, 2011.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-86, 1991.

[3] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriengman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 19, no. 7, pp. 711-720, 1997.

[4] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 526-534, 2012.

[5] J.B. Tenenbaum, V. deSilva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, 2000.

[6] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2325, 2000.

[7] X. He, S. Yan, Y. Hu, P. Niyogi, and H.J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328-340, 2005.

[8] H.-T.Chen, H.-W. Chang, and T.-L. Liu, "Local discriminant embedding and its variants," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.
[9] J. Yang, D. Zhang, J.-Y. Yang, and B. Niu, "Globally maximizing, locally minimizing: Unsupervised discriminant projection with applications to face and palm biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 4, pp. 650-664, 2007.
[10] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210 – 227, 2009.
[11] W.Y. Zhao, R. Chellppa, P.J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Survey*, vol. 35, no. 4, pp. 399-458, 2003.
[12] R. Min, and J. L. Dugelay, "Improved combination of LBP and sparse representation based classification (SRC) for face recognition," *Proc. IEEE Int'l Conf. Multimedia and Expo*, 2011.
[13] C.C. Kang, S.C. Liao, S.M. Xiang, and C.H. Pan, "Kernel sparse representation with local patterns for face recognition," *Proc. IEEE Int'l Conf. Image Processing*, 2011.
[14] X. Tan and B. Triggs, "Fusing gabor and LBP Feature sets for kernel-based face recognition," *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2007.
[15] J. Zou, Q. Ji, and G. nagy, "A comparative study of local matching approach for face recognition," *IEEE Trans. Image Processing*, vol. 16, no. 10, pp. 2617-2628, 2007.
[16] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C.v.d. Malsburg, R.P. Würtz and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, no. 3, pp.300-311, 1993.
[17] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Processing*, vol. 11, no. 4, pp.467-476, 2002.
[18] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition", *Pattern Analysis and Application*, vol. 9, no. 10, pp.273-292, 2006.
[19] Y.G. Liu, S.Z. Sam Ge, C.G. Li, and Z.S. You, "K-NS: A classifier by the distance to the nearest subspace," *IEEE Trans. Neural Networks*, vol. 22, no. 8, pp. 1256-1268, 2011.
[20] A.Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 6, pp. 643-660, 2001.
[21] A. Martinez and R. benavente, "*The AR face database*," CVC Tech. Report No. 24, 1998.
[22] P.J. Phillips, H. Wechsler, J. Huang, P. Rauss, "The FERET database and evaluation procedure for face recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295-306, 1998.
[23] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, "The FERET evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090-1104, 2000.
[24] L. Wolf, T. Hassner, and Y. Taigman, "Similarity scores based on background samples," *Proc. Asian Conf. Computer Vision*, 2009.
[25] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar, "Attribute and simile classifiers for face verification," *Proc. Int'l Conf. Computer Vision*, 2009.
[26] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037-2041, 2006.
[27] D. Bouchaffra, "Conformation-based hidden markov models: Application to human face identification," *IEEE Trans. Neural Networks*, vol. 21, no. 4, pp. 595-608, 2010.
[28] S.H. Gao, I.W-H. Tsang, and L-T. Chia, "Kernel sparse representation for image classification and face recognition," *Proc. European Conf. Computer Vision*, 2010.
[29] Z. Zhou, A. Wagner, H. Mobahi. J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," *Proc. Int'l Conf. Computer Vision*, 2009.
[30] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.
[31] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.
[32] G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in uncontrained environments," *Technical Report 07-49, University of Massachusetts*, 2007.
[33] L. Zhang, M. Yang, and X. C. Feng, "Sparse representation or collaborative representation which helps face recognition?" *Proc. Int'l Conf. Computer Vision*, 2011.
[34] A. Barla, F. Odone, and A. Verri, "Histogram intersection kernel for image classification," *Proc. IEEE Int'l Conf. Image Processing*, 2003.
[35] P.J. Phillips, P.J. Flynn, W.T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W.J. Worek, "Overview of the face recognition grand challenge," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005.
[36] S.Z. Li, "Face recognition based on nearest linear combinations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998.
[37] S.Z. Li and J. Lu, "Face recognition using nearest feature line method," *IEEE Trans. Neural Network*, vol. 10, no. 2, pp. 439-443, 1999.

[38] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, 2106-2112, 2010.
[39] J.C. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
[40] H. Lee, A. Battle, R. raina, and A.Y. Ng, "Efficient sparse coding algorithm," *Proc. Neural Information and Processing Systems*, 2006.
[41] R. Gross, I. Matthews. J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, pp. 807–813, 2010.
[42] J.W. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, "Face recognition using LDA-based algorithms," *IEEE Trans. Neural Networks*, vol. 14, no. 1, pp. 195-200, 2003.
[43] A. Timo, H. Abdenour, and P. Matti, "Face recognition with local binary patterns," *Proc. European Conf. Computer Vision*, 2004.
[44] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971 -987, 2002.
[45] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," *Proc. Int'l Conf. Computer Vision*, 2005.
[46] W. Zhang, S. Shan, X. Chen, and W. Gao, "Are gabor phases really useless for face recognition?" *Proc. Int'l Conf. Pattern Recognition*, 2006.
[47] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of gabor phase patterns (HGPP): A novel object representation approach for face recognition," *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 57-68, 2006.
[48] S.F. Xie, S.G. Shan, X.L. Chen, and J. Chen, "Fusing local patterns of gabor magnitude and phase for face recognition," *IEEE Trans. Image Processing*, vol. 19, no. 5, pp. 1349-1361, 2010.
[49] M. Yang, L. Zhang, L. Zhang and D. Zhang, "Monogenic binary pattern (MBP): A novel feature extraction and representation model for face recognition," *Proc. Int'l Conf. Pattern Recognition,* 2010.
[50] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.
[51] J.T. Chien, and C.C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1644-1649, 2002.
[52] A. Wagner, J. Wright, A. Ganesh, Z. H. Zhou, H. Mobahi, and Y. Ma, "Towards a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no, 2, pp. 372-386, 2012.
[53] P. Viola and M.J. Jones, "Robust real-time face detection," *Int'l J. Computer vision*, vol. 57, no. 2, pp. 137-154, 2004.
[54] S.G. Shan, Y. Chang, W. Gao, B. Cao, and P. Yang, "Curse of misalignment in face recognition: Problem and a novel misalignment learning solution," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2004.
[55] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681-685, 2001.
[56] J.C. Yang, K. Yu, T. Huang, "Supervised translation invariant sparse coding," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.
[57] Z.L. Jiang, Z. Lin, L.S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.
[58] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684-698, 2005.
[59] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machine: Global versus component-based approach," *Proc. Int'l Conf. Computer Vision*, 2001.
[60] J. Yang, A.F. Frangi, J.Y. Yang, D. Zhang, and J. Zhong, "KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 230-244, 2005.
[61] A.Y. Yang, A. Ganesh, Z.H. Zhou, S.S. Sastry, and Y. Ma, "*A review of fast l_1-minimization algorithms for robust face recognition*," arXiv:1007.3753v2, 2010.
[62] Y. Nesterov, A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM Philadelphia, PA, 1994.
[63] G. Tzimiropoulos, S. Zafeiriou, M. Pantic, "Subspace Learning from Image Gradient Orientations," *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(12):2454-2466, 2012.