Enhanced Local Gradient Order Features and Discriminant Analysis for Face Recognition

Chuan-Xian Ren, Zhen Lei, *Member, IEEE*, Dao-Qing Dai, *Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*—Robust descriptor-based subspace learning with complex data is an active topic in pattern analysis and machine intelligence. A few researches concentrate the optimal design on feature representation and metric learning. However, traditionally used features of single-type, e.g., image gradient orientations (IGOs), are deficient to characterize the complete variations in robust and discriminant subspace learning. Meanwhile, discontinuity in edge alignment and feature match are not been carefully treated in the literature. In this paper, local order constrained IGOs are exploited to generate robust features. As the difference-based filters explicitly consider the local contrasts within neighboring pixel points, the proposed features enhance the local textures and the order-based coding ability, thus discover intrinsic structure of facial images further. The multimodal features are automatically fused in the most discriminant subspace. The utilization of adaptive interaction function suppresses outliers in each dimension for robust similarity measurement and discriminant analysis. The sparsity-driven regression model is modified to adapt the classification issue of the compact feature representation. Extensive experiments are conducted by using some benchmark face data sets, e.g., of controlled and uncontrolled environments, to evaluate our new algorithm.**

***Index Terms*—Discontinuity, image gradient, order features, sparse representation, subspace learning.**

I. Introduction

ROBUST subspace learning [1]–[4] has been an active area in pattern recognition and machine learning, and it can be roughly categorized by supervised, unsupervised, and semisupervised methods. As one particular case of supervised learning, face recognition has wide applications in public security and commercial developments. It has been shown that a carefully modeled feature representation plays a critical role in robust face recognition [5]. Many algorithms have been proposed to deal with the effectiveness of feature design and extraction [6], [7]; however, the performance of many existing methods is still highly sensitive to variations of imaging conditions, such as outdoor illumination, exaggerated expression, and continuous occlusion. These complex variations are significantly affecting the recognition accuracy in recent years [8]–[10].

Appearance-based subspace learning is one of the simplest approach for feature extraction, and many methods are usually based on linear correlation of pixel intensities. For example, Eigenface [11] uses eigen system of pixel intensities to estimate the lower rank linear subspace of a set of training face images by minimizing the ℓ_2 distance metric. The solution enjoys optimality properties when noise is independent identically distributed Gaussian only. Fisherface [12] will suffer more due to the estimation of inverse within-class covariance matrix [13], thus the performance will degenerate rapidly in the cases of occlusion and small sample size. Laplacianfaces [14] refer to another appearance-based approach which learns a locality preserving subspace and seeks to capture the intrinsic geometry and local structure of the data. Other methods such as those in [5] and [15] also provide valuable approaches to supervised or unsupervised dimension reduction tasks.

A fundamental problem of appearance-based methods for face recognition, however, is that they are sensitive to imaging conditions [10]. As for data corrupted by illumination changes, occlusions, and inaccurate alignment, the estimated subspace will be biased, thus much of the efforts concentrate on removing/shrinking the noise components. In contrast, local feature descriptors [15]–[19] have certain advantages as they are more stable to local changes. In the view of image processing and vision, the basic imaging system can be simply formulated as

$$\mathbf{\Omega}(x, y) = \mathbf{A}(x, y) \times \mathbf{L}(x, y) \tag{1}$$

where $\mathbf{\Omega}(x, y)$ is image pixel value, $\mathbf{A}(x, y)$ is the surface Albedo, and $\mathbf{L}(x, y)$ is the illuminance at each point (x, y). Then the task is to present a robust feature representation from $A(x, y)$ for image $\mathbf{\Omega}$.

Gradient-based methods have been used for texture description and image classification due to its robustness to local variations and efficiency to computation. In the gradientface method [8], after the orientation-based feature generation, a ℓ_1-type metric is exploited for feature matching. By integrating both time and storage requirement, Vu and Caplier [20]

Manuscript received May 20, 2015; revised August 21, 2015; accepted September 20, 2015. Date of publication October 26, 2015; date of current version October 13, 2016. This work was supported in part by the National Science Foundation of China under Grant 11171354, Grant 61203248, Grant 61375033, and Grant 61572536, in part by the Ministry of Education of China under Grant SRFDP-20120171120007 and Grant 20120171110016, in part by the Natural Science Foundation of Guangdong Province under Grant S2013020012796, in part by the Fundamental Research Funds for the Central Universities under Grant 13lgpy26, and in part by the Open Project Program of the National Laboratory of Pattern Recognition. This paper was recommended by Associate Editor S. Zafeiriou. *(Corresponding author: Dao-Qing Dai.)*

C.-X. Ren and D.-Q. Dai are with the Intelligent Data Center, School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou 510275, China (e-mail: stsddq@mail.sysu.edu.cn).

Z. Lei and S. Z. Li are with the Center for Biometrics and Security Research and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2015.2484356

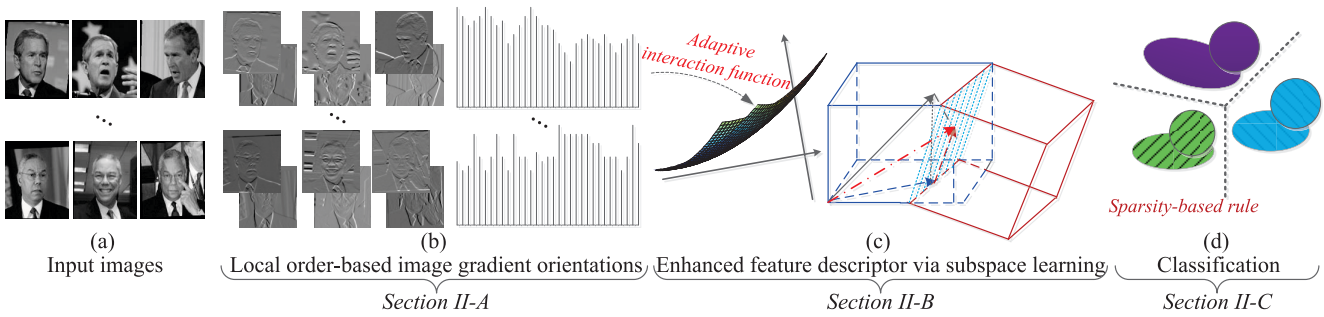


Fig. 1. Block diagram of our algorithm. (a) Input images. (b) Two directional gradient responses are generated, and then gradient orientations and block-wise histograms are used to formulate the basic feature representation. (c) AIFs are exploited to enhance the subspace learning/fusion and address the possible discontinuous problem. (d) Sparsity-based classifier is proposed to complete the recognition.

proposed to enhance the face recognition performance by optimizing the patterns of oriented edge magnitudes descriptor. Huang *et al.* [21] extracted the histograms of second-order gradients to capture the curvature-related geometric properties of the neural landscape. Tzimiropoulos *et al.* [3] first generated the image gradient orientation (IGO) features, and then studied the properties of IGO in the cosine-based subspace learning method. The robustness of IGO to occlusions is experimentally displayed to validate the effectiveness for feature extraction. However, simple orientation features are not sufficient to characterize the complete variations in robust and discriminant subspace learning. As shown in later, the local order-based gradients can also be validated effective in feature description. Meanwhile, the discontinuous scenarios in edge alignment and feature match are not been deeply discussed. Motivated by that gradient information can enhance the edge response, Lei *et al.* [22] proposed a novel local gradient order pattern (LGOP), taking into account the ordinal relationship of gradient responses in local region. However, properties of the feature subspace have not been further addressed, and the ability to deal with discontinuity, e.g., brought by occlusion and expression variations, is not considered.

To address these problems, a enhanced IGO descriptor is proposed in this paper. A flowchart of our enhanced feature representation system is displayed in Fig. 1. As the difference-based gradient filters explicitly consider the local contrasts within neighboring pixel points, the proposed feature transformation enhances the local texture description ability, thus further discovers intrinsic structure of facial images. The utilization of adaptive interaction function (AIF) suppresses outliers in each dimension for robust similarity measurement and discriminant analysis. Different weights are automatically determined and assigned to the two-modal features to boost discriminant subspace learning. Accordingly, our learning system consists of three sequential stages, including local order-based image descriptor, enhanced subspace fusion, and sparse identity coding. The main contributions are summarized as follows.

1) A enhanced IGO descriptor is proposed to robust feature extraction. The local order features describe the spacial information, neighbor contrasts, and region histograms, thus they are predominant in dealing with complex image data.
2) AIF is used as a robust similarity measurement between two high-dimensional vectors. As a second-order statistics, it reflects the closeness in a local region, and thus it has the potential ability to deal with the discontinuity problem.
3) It presents a natural regularization in similarity measurement and descriptor enhancement. The structured regularization weights and the discriminant subspace are simultaneously modeled in a supervised criterion, and then alternatively optimized.

It is worth noting that the convolutional neural networks (CNNs)-based deep learning methods [23] are proved effective in the case of large-scale data. However, the main drawbacks of CNNs include their high computational complexity and difficult parameter tuning. Moreover, visualization of the filtered images is also inconvenient. Thus our method will not be widely compared with the large-scale data-based deep learning methods.

The rest of this paper is organized as follows. Section II presents our new algorithm, i.e., the successive stages including local feature generation, discriminant subspace fusion, and the least angle-based identity prediction. The algorithm is evaluated from several technical aspects in Section IV. In Section V, extensive experiments are conducted to compare the new algorithm with several state-of-the-art methods, and the results on the benchmark sets (controlled/uncontrolled lighting, disguise, etc.) show the competitive performance of our approach.

II. Enhanced Image Feature Descriptor

A. Local Order-Based Feature Generation

It is well known that the human visual system is more sensitive to local changes in contrast than the absolute magnitudes of the signal [16]. The difference-based gradient features have biological justifications and they can improve the texture description performance.

The features are generated by two steps. Let us focus the image structure on (1). For simplicity, the position index (x, y) in this section is omitted. First, taking the x- and y-axis gradient responses of image I, we have

$$\mathbf{\Omega}_x = (\mathbf{A} \times \mathbf{L})_x \approx \mathbf{A}_x \times \mathbf{L} + \mathbf{A} \times \mathbf{L}_x$$
$$\mathbf{\Omega}_y = (\mathbf{A} \times \mathbf{L})_y \approx \mathbf{A}_y \times \mathbf{L} + \mathbf{A} \times \mathbf{L}_y.$$

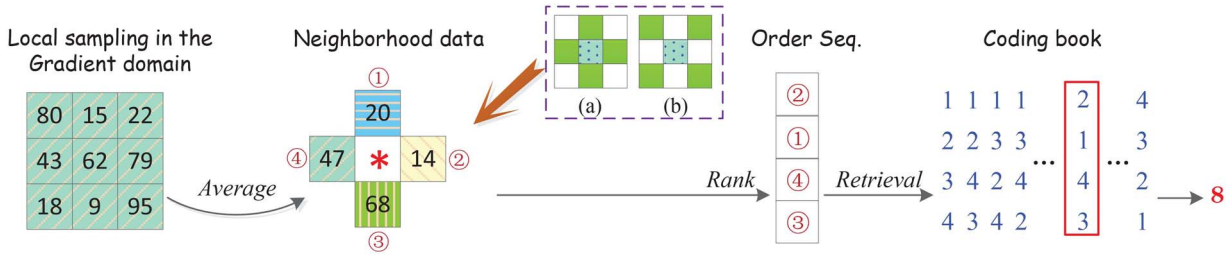


Fig. 2. Basic generation procedure of LGOP. The order-based feature is first generated by average filtering and neighborhood sampling in the gradient domain. Note that two ways of neighbor sampling are shown in the dashed square. The serial vector is retrieved by using a permutation-based coding book, and then the order number is defined as the final feature code.

According to the hypothesis of Albedo, the nature of $\mathbf{L}$ is determined by the lighting source, while $\mathbf{A}$ is determined by the characteristics of the surface of object. A common assumption that $\mathbf{L}$ varies very slowly (i.e., $\mathbf{L}_x \approx 0$ and $\mathbf{L}_y \approx 0$) while $\mathbf{A}$ can change abruptly. Thus the hypothesis of Albedo can be exploited for generating lighting-insensitive features, and then the equations can be further approximated to $\boldsymbol{\Omega}_x = \mathbf{A}_x \times \mathbf{L}$ and $\boldsymbol{\Omega}_y = \mathbf{A}_y \times \mathbf{L}$. Therefore, $\boldsymbol{\Omega}_a = \boldsymbol{\Omega}_y./\boldsymbol{\Omega}_x \approx \mathbf{A}_y./\mathbf{A}_x$ is robust for light variations, where ./ is the matrix point division operator. Notice that the orientation feature $\boldsymbol{\Omega}_a$ has been used in gradientface [8], patterns of oriented edge magnitudes (POEM) [20], and IGO [3] methods for face recognition. However, they do not consider the spatial information and statistical structure of $\boldsymbol{\Omega}_x$ and $\boldsymbol{\Omega}_y$.

Assume that ∇_x and ∇_y are the x- and y-directional gradient operators, respectively

$$\nabla_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \quad \nabla_y = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

then $\boldsymbol{\Omega}_x$ and $\boldsymbol{\Omega}_y$ can be efficiently obtained via the local convolution operators $\boldsymbol{\Omega}_x = \boldsymbol{\Omega} * \nabla_x$ and $\boldsymbol{\Omega}_y = \boldsymbol{\Omega} * \nabla_y$. In practice, before generating the gradient features, the images are preprocessed by a Gaussian filter $\mathbf{G}(x, y, \sigma)$ to eliminate pixel noise, where the variance σ is set to 0.2 throughout this paper.

The second step of feature generation will be implemented by incorporating the local intensity order pattern (LIOP) [24], which is proposed to address the visual description and feature detection issues. It takes the order of elements in data vector into account and maps the data vector to its permutation space. Lei *et al.* [22] further extended the ordinal relationship into the gradient domain. The basic scheme of LIOP has been illustrated in Fig. 2. Given a d-dimensional vector $\mathbf{P} = [p_1, p_2, \ldots, p_d] \in \mathbb{R}^d$ and a possible permutation set $\boldsymbol{\Pi}$ of integers $\{1, 2, \ldots, d\}$, the mapping from $\mathbf{P}$ to $\boldsymbol{\Pi}$ is defined as follows. First, the elements in P are sorted in a nondescending order, i.e., $p_{i_1} \leq p_{i_2} \leq \cdots \leq p_{i_d}$. Second, the subscript list $i_1, i_2, \ldots, i_d$ is considered as the mapping result in set $\boldsymbol{\Pi}$ and is denoted using a unique scalar (LIOP code). To avoid ambiguity, $p_s \leq p_t$ is defined as if and only if: 1) $p_s < p_t$ or 2) $p_s = p_t$ and $s < t$. It is obvious that for a d-dimensional vector, there are $d!$ possible permutations.

For each pixel of an image $\boldsymbol{\Omega}$ that has two gradient response matrices $\boldsymbol{\Omega}_x$ and $\boldsymbol{\Omega}_y$, its neighbors are sampled and sorted in nondescending order. The order index is then mapped to its permutation space as in LIOP. Finally, the ordered patterns generated from $\boldsymbol{\Omega}_x$ and $\boldsymbol{\Omega}_y$ are obtained for further processing.

A detailed description for the instance in Fig. 2 is explained as follows. First, in the 2-D gradient domain, i.e., $\boldsymbol{\Omega}_x$ or $\boldsymbol{\Omega}_y$, the mean filter of 3×3 size is exploited for convolution. In particular, for the 3×3 window as shown in the figure, its mean value is 47. The same operation is applied to other 3×3 windows, and then we have the neighborhood data. In the second stage, for any pixel (denoted by star for emphasis), its four neighborhoods vector (20, 14, 68, and 47) is sampled and labeled by (1, 2, 3, and 4) in the clockwise sense. Then the four neighbors are sorted as index vector (2, 1, 4, and 3) in ascending order. Finally, we retrieve the index vector from the code book. We find that its position in the code book is eight, i.e., the eighth column, so we code the position of star as digital number 8. Notice that the code book can be previously saved in the memory/program for real-time retrieve and coding.

To preserve the spatial information in face image, a number of 2-D histogram features are extracted in local regions. Specifically, each 2-D image that has been locally ordered is divided into 8×8 blocks, and then the histogram is counted from each block respectively. In this way, histogram features are extracted from the spatial regions so that they can capture more representative and discriminant ability of facial images. These 2-D histogram features extracted from different local regions are vectorized and then finally concatenated to represent the whole face.

To exploit more complementary information, we adopt two different sampling ways, which have been shown in the dashed square in Fig. 2, and concatenate the feature vectors together. Notice that four neighbors are, respectively, sampled in $\boldsymbol{\Omega}_x$ and $\boldsymbol{\Omega}_y$ to save the computational cost.

We denote the histogram feature vector of local order codings by $\boldsymbol{\varphi}_i$ and the vectorized orientation feature vector by $\boldsymbol{\theta}_i$ for image $\boldsymbol{\Omega}_i$ from now on. We notice that $\boldsymbol{\theta}_i$ and $\boldsymbol{\varphi}_i$ may be heterogeneous and hence, we cannot directly concatenate them together for feature representation. Besides, the higher dimension of $\boldsymbol{\varphi}_i$ than that of $\boldsymbol{\theta}_i$ may bring over-domination in numerical computation. A natural way to deal with these problems is weighting the features by assigning different weights to each group

$$\boldsymbol{\xi}_i = \begin{bmatrix} \sqrt{\boldsymbol{\gamma}_{\boldsymbol{\theta}}}\boldsymbol{\theta}_i \\ \sqrt{\boldsymbol{\gamma}_{\boldsymbol{\varphi}}}\boldsymbol{\varphi}_i \end{bmatrix} \in \mathbb{R}^M \tag{2}$$

where $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_{\boldsymbol{\theta}}, \boldsymbol{\gamma}_{\boldsymbol{\varphi}}]^T \geq 0$, and it is called local order constrained gradient orientations (LOGO) in this paper.

The formulation of $\boldsymbol{\xi}_i$ provides spatial structure and local contrast in enhancing the discriminant capability of pure orientation features, thus it presents a natural regularization way to improve the performance of pure utilization of $\boldsymbol{\theta}_i$.

B. Enhanced Discriminant Subspace Learning

Due to the high dimension of feature augment in (2), dimension reduction is required for discriminant fusion and efficient classification. It is usually formulated as finding a low-dimensional and discriminant subspace, allowing larger margins between different classes as well as more compact representations within the same class. According to the graph embedding method [1], characterizing the separability between different classes and the compactness within the same class can be converted to and displayed by the weights w_{jk} and w'_{jk} between different samples. Assume that $\mathbf{V}$ is the projection matrix for dimension reduction, $\boldsymbol{\gamma}$ is the weight vector as shown in (2), then our objective function can be formulated as

$$\max_{\mathbf{V},\boldsymbol{\gamma}\geq 0} J(\mathbf{V},\boldsymbol{\gamma}) = \frac{\sum_{j,k=1}^{n} w_{jk}\left\|\mathbf{V}^T\boldsymbol{\xi}_j - \mathbf{V}^T\boldsymbol{\xi}_k\right\|_2^2}{\sum_{j,k=1}^{n} w'_{jk}\left\|\mathbf{V}^T\boldsymbol{\xi}_j - \mathbf{V}^T\boldsymbol{\xi}_k\right\|_2^2}. \tag{3}$$

Specific utilization of the weights is shown in Section IV.

However, (3) is different to traditional graph embedding method, as the non-negative vector $\boldsymbol{\gamma}$ is introduced into our formulation. The double unknowns $\mathbf{V} \in \mathbb{R}^{M\times h}$ and $\boldsymbol{\gamma}$ make the optimization problem difficult, here h is the size of discriminant subspace.

When the vector $\boldsymbol{\gamma}$ is fixed, the weighted feature vector $\boldsymbol{\xi}$ can be viewed as regular instance of vector-type, and the objective function is degenerated to Fisher's discriminant analysis so that the discriminant projection matrix $\mathbf{V}$ can be efficiently solved by the generated eigenvalue decomposition method.

The optimization difficulty concentrates on the computation of vector $\boldsymbol{\gamma}$. Fortunately, (3) can be transformed to its dual problem by the subspace theory. Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n]$, according to the representation theorem in subspace theory [25], the projection matrix $\mathbf{V}$ can be represented as a linear combination of columns in $\boldsymbol{\Xi}$, i.e., $\mathbf{V} = \boldsymbol{\Xi}\mathbf{U}$, where $\mathbf{U}$ is the coefficient matrix. As a result, we have $\mathbf{V}^T\boldsymbol{\Xi} = \mathbf{U}^T\boldsymbol{\Xi}^T\boldsymbol{\Xi}$.

Assume $\mathrm{Sim}(\mathbf{x}_1, \mathbf{x}_2)$ be any similarity function between feature vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, and

$$\mathbf{S}_j = \begin{bmatrix} \mathrm{Sim}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_j), & \mathrm{Sim}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_j) \\ \vdots & \vdots \\ \mathrm{Sim}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_j), & \mathrm{Sim}(\boldsymbol{\varphi}_n, \boldsymbol{\varphi}_j) \end{bmatrix} \tag{4}$$

we have

$$\boldsymbol{\xi}_j^T\boldsymbol{\xi}_k = \boldsymbol{\gamma}_\theta \mathrm{Sim}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_k) + \boldsymbol{\gamma}_\varphi \mathrm{Sim}(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_k) \tag{5}$$

and $\mathbf{V}^T\boldsymbol{\xi}_j = \mathbf{U}^T\mathbf{S}_j\boldsymbol{\gamma}$ by algebra operations. Then the function in (3) can be reformulated as

$$\max_{\mathbf{U},\boldsymbol{\gamma}\geq 0} J(\mathbf{U},\boldsymbol{\gamma}) = \frac{\sum_{j,k=1}^{n} w_{jk}\left\|\mathbf{U}^T\mathbf{S}_j\boldsymbol{\gamma} - \mathbf{U}^T\mathbf{S}_k\boldsymbol{\gamma}\right\|_2^2}{\sum_{j,k=1}^{n} w'_{jk}\left\|\mathbf{U}^T\mathbf{S}_j\boldsymbol{\gamma} - \mathbf{U}^T\mathbf{S}_k\boldsymbol{\gamma}\right\|_2^2}. \tag{6}$$

Now, the matrix-vector product $\mathbf{S}_j\boldsymbol{\gamma}$ can be viewed as new feature representation which is composed of several similarities, and determining the non-negative coefficient vector $\boldsymbol{\gamma}$ can be interpreted as finding appropriate weights for optimally dealing with LOGO. In particular, the possible over-domination phenomenon can be alleviated by the importance-based weighting method. By (5), we know the weights $\boldsymbol{\gamma}_\theta$ and $\boldsymbol{\gamma}_\varphi$ make a tradeoff between the two similarities $\mathrm{Sim}(\boldsymbol{\theta}_j, \boldsymbol{\theta}_k)$ and $\mathrm{Sim}(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_k)$.

Let $\mathbf{F}_w = \sum_{j,k} w'_{jk}(\mathbf{S}_j - \mathbf{S}_k)^T\mathbf{U}\mathbf{U}^T(\mathbf{S}_j - \mathbf{S}_k)$ and $\mathbf{F}_b = \sum_{j,k} w_{jk}(\mathbf{S}_j - \mathbf{S}_k)^T\mathbf{U}\mathbf{U}^T(\mathbf{S}_j - \mathbf{S}_k)$, the problem becomes

$$\min_{\boldsymbol{\gamma}} \boldsymbol{\gamma}^T\mathbf{F}_w\boldsymbol{\gamma} \quad \text{s.t.} \quad \boldsymbol{\gamma}^T\mathbf{F}_b\boldsymbol{\gamma} = 1, \text{ and } \boldsymbol{\gamma} \geq 0. \tag{7}$$

It is a nonconvex quadratically constrained quadratic programming problem, and its convex relaxation by adding an auxiliary matrix $\mathbf{B} = \boldsymbol{\gamma}\boldsymbol{\gamma}^T$ can be

$$\min_{\boldsymbol{\gamma},\mathbf{B}} \mathrm{tr}(\mathbf{F}_w\mathbf{B}) \quad \text{s.t.} \quad \mathrm{tr}(\mathbf{F}_b\mathbf{B}) = 1, \begin{bmatrix} 1 & \boldsymbol{\gamma}^T \\ \boldsymbol{\gamma} & \mathbf{B} \end{bmatrix} \geq 0, \boldsymbol{\gamma} \geq 0. \tag{8}$$

This is a semi-definite programming (SDP) relaxation of the nonconvex problem (7), and can be efficiently solved.

After obtaining the projection operator $\mathbf{V}$ or $\mathbf{U}$, we are ready to solve the out-of-sample problem as follows. For sample $\boldsymbol{\Omega}_t$, it is firstly transformed to LOGO feature $\boldsymbol{\xi}_t$ as shown in (2), and then embedded into the subspace by

$$\mathbf{z} = \mathbf{V}^T\boldsymbol{\xi}_t \quad \text{or} \quad \mathbf{U}^T\mathbf{S}_t\boldsymbol{\gamma} \tag{9}$$

where $\mathbf{S}_t$ is obtained via (4). Both formulations can be efficiently embedded into the next stage of label prediction.

We should pay attention to the similarity computation in subspace learning. The discontinuity which is usually brought by the image edges and local contrast-based features should be carefully treated. Li [26] systematically studied the discontinuity problem and defined a general discontinuity adaptive Markov random fields model. It shows that the fundamental difference between various models lines in the behavior of interaction within neighborhood instances, which is determined by the prior smoothness constraint encoded into the energy function.

One typical example of the AIF is the so-called correntropy in information theory and signal processing [27]. Formally, correntropy is defined as a generalized similarity between two arbitrary scalar random variables x and y, i.e., $\mathbb{V}_\sigma(x, y) = \mathbb{E}[\kappa_\sigma(x - y)]$. It is directly related to the probability of how similar two random variables are in a neighborhood of the joint space controlled by the bandwidth σ [27]. It can be immediately inferred that $\mathbb{V}_\sigma(x, y)$ is symmetric, positive, and bounded.

The correntropy of vectors $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^M$ is

$$\mathrm{Sim}(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^{M} \mathbb{E}\big[\kappa_\sigma(\mathbf{x}_m - \mathbf{y}_m)\big].$$

It has been used to obtain robust analysis and handle non-Gaussian noise [28]. It is worth noting that the AIF/correntropy is essentially different to general kernels. For a more detailed discussion, refer to Section III.

C. Classification via Sparse Regression

Recently, sparse representation-based classifiers have led to promising results in machine learning and pattern analysis [29], [30]. By the linear-core assumption, a test sample $\mathbf{y}$ can be sparsely represented as a linear combination $\boldsymbol{\chi}$ of the dictionary $\mathbf{D}$. Wright *et al.* [29] proposed the sparse representation classifier (SRC) for robust face recognition. An iterative reweighting method robust sparse coding (RSC) is developed by Yang *et al.* [31] for shrinking the influence of the partial occlusion. He *et al.* [28] proposed to handle the problem by their maximum correntropy model. In [32], we propose to use multiple loss measurements to improve robustness and discrimination of the regression model. Notice that these models only theoretically design for the simple scenario that training without occlusions while testing with occlusions, which is obviously restrictive in practice.

From now on, to avoid the notations confusion, the embedded low-dimensional representation of the training set is written as $\mathbf{Y}$, in which the ith atom (column) is $\mathbf{y}_i = \mathbf{V}^T\boldsymbol{\xi}_i = \mathbf{U}^T\mathbf{S}_i\boldsymbol{\gamma}$, and the embedded testing sample is denoted as $\mathbf{z}$ (the subscript is ignored).

The classification model is formulated via

$$\arg\min_{\boldsymbol{\chi}} \|\mathbf{Y}\boldsymbol{\chi} - \mathbf{z}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\chi}\|_0 = \nu \tag{10}$$

where $\|\boldsymbol{\chi}\|_0$ denotes the ℓ_0-norm of coefficient vector $\boldsymbol{\chi}$ and ν is the expected or predefined size of nonzero entries.

Fortunately, the objective function can be efficiently solved by the least angle regression (LARS), in which the least angle between two potentially correlated samples is exploited to search the optimally matched partners. The numerical optimization details can be referred to in [33].

One advantage by using of the least angle is its sequential and forward determination of the sparse coefficients, which will be favorable for choosing the size of the correlated samples or the active set. In our implementation, the size of the nonzero elements, ν, is fixed to be the number of training samples in each class.

In the second stage for label prediction, we need to compute the reconstruction errors between $\mathbf{z}$ and the class-wise dictionary. The error is measured by the cosine function. Let $\boldsymbol{\chi}_k$ be the class-wise subvector of $\boldsymbol{\chi}$ and $\mathbf{Y}_k$ be the submatrix of $\mathbf{Y}$, both of them corresponding to the basis of class k. We find the identity of $\mathbf{z}$ via maximizing the reconstruction correlations

$$\text{Label}(\mathbf{z}) = \arg\max_k \text{Sim}\big(\mathbf{z}, \mathbf{Y}_k\boldsymbol{\chi}_k\big). \tag{11}$$

In the following sections of model selection and justification, the real performance of LOGO algorithm in face recognition is displayed. Through the sequential operations of discriminant dictionary building and the least angle-based sparse coding, our algorithm not only extracts discriminant features but also further filtrates the noisy components for imaging condition variations. Particularly, by simultaneously using the AIF metric and the subject-based reconstruction, the subject whose training samples illustrating the largest correlation (interaction) with $\mathbf{z}$ will automatically appear, and the corresponding label is used for identity prediction.

Algorithm 1 The Proposed LOGO Method

Input: Images $\{\boldsymbol{\Omega}_i\}_{i=1}^n$, Labels L, σ, ν, Test image $\boldsymbol{\Omega}_{n+1}$
Output: $\mathbf{U}$, $\boldsymbol{\gamma}$, predicted label of $\boldsymbol{\Omega}_{n+1}$.

1. Feature generation and augment;
   1.1 $\tilde{\boldsymbol{\Omega}}_i = \boldsymbol{\Omega}_i * \mathbf{G}(x, y, \sigma)$, $(i = 1, 2, \ldots, n+1)$;
   1.2 $\boldsymbol{\phi}_i = \tilde{\boldsymbol{\Omega}}_i * \nabla_x$, $\boldsymbol{\psi}_i = \tilde{\boldsymbol{\Omega}}_i * \nabla_y$;
   1.3 obtain $\boldsymbol{\varphi}_i$ via local order coding;
   1.4 obtain $\boldsymbol{\theta}_i$ via $\arctan(\boldsymbol{\psi}_i./\boldsymbol{\phi}_i)$ and vectorization;
   1.5 Return $\boldsymbol{\xi}_i$ as shown in (2);
2. Kernel fusion and model optimization;
   2.1 Calculate $\boldsymbol{\xi}_j^T\boldsymbol{\xi}_k$ via (5) and pool them into $\mathbf{S}_j$;
   2.2 Iteratively optimize $\mathbf{U}$ and $\boldsymbol{\gamma}$;
3. Feature embedding in the fused kernel space;
   3.1 $\mathbf{V} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n]\mathbf{U}$ is projection matrix;
   3.2 Dictionary $\mathbf{Y} = [\mathbf{V}^T\boldsymbol{\xi}_1, \mathbf{V}^T\boldsymbol{\xi}_2, \ldots, \mathbf{V}^T\boldsymbol{\xi}_n]$;
   3.3 $\mathbf{z}_{n+1} = \mathbf{U}^T\mathbf{S}_{n+1}\boldsymbol{\gamma}$ is the embedded vector of $\boldsymbol{\Omega}_{n+1}$;
4. Calculate sparse representation $\boldsymbol{\chi}$ of $\mathbf{z}_{n+1}$ via (10);
5. Labeling $\boldsymbol{\Omega}_{n+1}$ via $\boldsymbol{\chi}$ and (11);

A block diagram for the complete recognition system has been presented in Fig. 1, and the pseudocodes are presented in Algorithm 1.

The complexity of the algorithm focuses on three aspects: 1) similarity calculation; 2) feature extraction; and 3) classification. In detail, the computation of similarity values for every pair of vectors has a complexity of $\mathcal{O}(n^2)$. The alternative calculation of subspace $\mathbf{U}$ and weights $\boldsymbol{\gamma}$, by means of the generalized eigen-value decomposition and SDP, cost $\mathcal{O}(n^3)$ and polynomial time, respectively. Finally, the complexity of the sparse regression is $\mathcal{O}(n^3 + hn^2)$ [33] to obtain the identity coding vector.

III. Discussion

The relationships between our algorithm and some state-of-the-art methods, e.g., feature descriptors, similarities, and classifiers, are discussed in this section.

A. Feature Descriptors

The utilization of image gradient responses and mean filter in feature representation relates our method to the POEM operator [20], which is sequentially composed of gradient responses, mean filters, and local binary patterns.

More precisely, the POEM algorithm applies the self-similarity calculation approach from the local binary pattern (LBP)-based structure on the distribution of local edge through different orientations. To calculate the POEM codes for one pixel, the intensity values in the calculation of conventional LBP are replaced by gradient magnitudes, which are calculated by accumulating a local histogram of gradient directions over all pixels within a spatial patch (cell).

In a similar summarization approach, our new method can be sequentially realized by the gradient responses, mean filter, and LGOP descriptor. The histogram features are extracted for each local regions to enhance the robustness to noise and image rotations.

From these points of view, our method is distinct from the POEM operator on at least two aspects.

1) The LGOP features, instead of local binary patterns, are embedded in our algorithm. In fact, LBP encodes the ordinal relationship between the neighborhood samplings and the central one to obtain robust face representation. However, additional information like the difference among neighboring pixels, which may be helpful for face recognition, is ignored.
2) The histogram features are extracted in the final stage of our method to enhance the robustness to noise, rotations, and so on. However, in the POEM method, the histogram features are used for spatial magnitude accumulation in cell, thus they are extracted before LBP operations.

B. Correntropy Versus Radial Basis Kernel

The most important property of correntropy to this paper is that it is a second-order statistic of the mapped features. It can be proved by the knowledge in probability and statistics. Assume the dimension of the feature space is M and the mapping (linear or nonlinear) is $\boldsymbol{\zeta}(\mathbf{x}) = [\boldsymbol{\zeta}_1(\mathbf{x}), \boldsymbol{\zeta}_2(\mathbf{x}), \ldots, \boldsymbol{\zeta}_M(\mathbf{x})]^T$, the second-order statistic between $\boldsymbol{\zeta}(\mathbf{x})$ and $\boldsymbol{\zeta}(\mathbf{y})$ is expressed by the correlation matrix

$$\mathbf{R}_{xy} = \mathbb{E}[\boldsymbol{\zeta}(\mathbf{x})\boldsymbol{\zeta}^T(\mathbf{y})] = \begin{bmatrix} \mathbb{E}\big[\boldsymbol{\zeta}_1(\mathbf{x})\boldsymbol{\zeta}_1^T(\mathbf{y})\big] & \cdots & \mathbb{E}\big[\boldsymbol{\zeta}_1(\mathbf{x})\boldsymbol{\zeta}_M^T(\mathbf{y})\big] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\big[\boldsymbol{\zeta}_M(\mathbf{x})\boldsymbol{\zeta}_1^T(\mathbf{y})\big] & \cdots & \mathbb{E}\big[\boldsymbol{\zeta}_M(\mathbf{x})\boldsymbol{\zeta}_M^T(\mathbf{y})\big] \end{bmatrix}.$$

Meanwhile, $\mathbb{V}(\mathbf{x}, \mathbf{y}) = \mathbb{E}[\zeta^T(\mathbf{x})\zeta(\mathbf{y})] = \text{trace}(\mathbf{R}_{xy})$. The trace of $\mathbf{R}_{xy}$ is equal to the sum of the eigenvalues, which shows that $\mathbb{V}_\sigma(x, y)$ is a second-order statistic in the feature space induced by the Gaussian function. Specifically, it has been emphasized by [27] that correntropy is the trace (assuming centered data in reproducing kernel Hilbert space) of the cross variance operator defined in kernel methods.

The Gaussian kernel or radial basis function (RBF) on two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ is

$$\mathbf{K}_\sigma(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2}{2\sigma^2}\right)$$

where $\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2$ is the squared Euclidean distance between the two feature vectors and σ is a scale parameter.

Since the value of the RBF kernel decreases with distance and ranges between zero (in the limit) and one (when $\mathbf{x}_1 = \mathbf{x}_2$), it has a ready interpretation as a similarity measure. The feature space of the kernel has an infinite number of dimensions. For $\sigma = 1$, its expansion is

$$\mathbf{K}_1(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=0}^{\infty} \frac{\left(\mathbf{x}_1^T\mathbf{x}_2\right)^j}{j!} \exp\left(-\frac{||\mathbf{x}_1||_2^2}{2}\right) \exp\left(-\frac{||\mathbf{x}_2||_2^2}{2}\right).$$

From the definitions, we clearly observe that the RBF kernel is very different from the correntropy in high-dimensional space. Generally, global similarity that induced by the ℓ_2-norm $\|\mathbf{x}_1 - \mathbf{x}_2\|$ is emphasized by the RBF and in the reproducing kernel Hilbert space. However, the local interaction in each dimension is highlighted in the correntropy, which measures the similarity between $\mathbf{x}_1$ and $\mathbf{x}_2$ via the summarization of every pairwise interaction. In this view, the discontinuous problem in feature matching may be alleviated by correntropy.

C. Kernel-Based Sparse Representation Classifiers

Our recognition system can be viewed as a multikernel-based SRCs, in which the most typical methods include kernel-based SRC (KSRC) [34] and multiple kernel sparse representations (MKSR) [35]. Both KSRC and MKSR exploit the kernel-based sparse regression coefficients to complete the classification or feature representation tasks.

The objective function of KSRC is

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \mathbf{B}^T\mathbf{k}(\cdot, \mathbf{y}) = \mathbf{B}^T\mathbf{K}\boldsymbol{\alpha}$$

in which $\mathbf{B}$ is a discriminant projection matrix, $\mathbf{K}$ is the kernel matrix of the training set, $\mathbf{k}(\cdot, \mathbf{y})$ is the nonlinear mapped features of $\mathbf{y}$, and $\boldsymbol{\alpha}$ is the sparsest regression coefficients. In other words, after the dictionary learning/design in the nonlinear kernel space, the attained $\boldsymbol{\alpha}$ of $\mathbf{k}(\cdot, \mathbf{y})$ is directly used to classify $\mathbf{y}$. Obviously, the gradient order-based local features and the multiple similarities fusion in the kernel-based dimension reduction process make our method different from the KSRC. In particular, our method extends the single kernel function to the case of the multikernel or multicorrentropy, thus it can be favorable in dealing with image-based classification tasks.

The MKSR method is proposed by Thiagarajan *et al.* [35] to deal with complex visual recognition tasks, which is typical to adopt multiple features to describe different aspects of the images. The basic dimension reduction procedure in the multikernel space is similar to ours. However, besides feature generation, the kernel-based sparse representation solution approach and classifier rule are very different from our method. The multikernel coefficient vectors which are obtained by MKSR are used to learn the kernel-based multilevel dictionary. Then the updated kernel sparse representations are conversely used to compute the graph affinities instead of kernel-wise neighborhood. These two steps, i.e., sparse representation and affinities update, are alternatively optimized in an iterative manner. Finally, the support vector machine (SVM) classifier is directly used to complete the recognition task. These detailed analyses can clear up the relationships between LOGO and MKSR.

IV. Model Judgement and Analysis

In this section, some aspects of our algorithm including similarity function and classifier are evaluated.

The objective of discriminant subspace learning can be achieved by setting the weights w_{jk} and w'_{jk} as those in the Fisher discriminant analysis [1]

$$w'_{jk} = \begin{cases} 1/n_{\ell_j}, & \ell_j = \ell_k \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

$w_{jk} = (1/n) - w'_{jk}$, here n_{ℓ_j} is the number of samples labeled by ℓ_j. The numerator and the denominator characterizes the scatter for between-class and within-class, respectively.

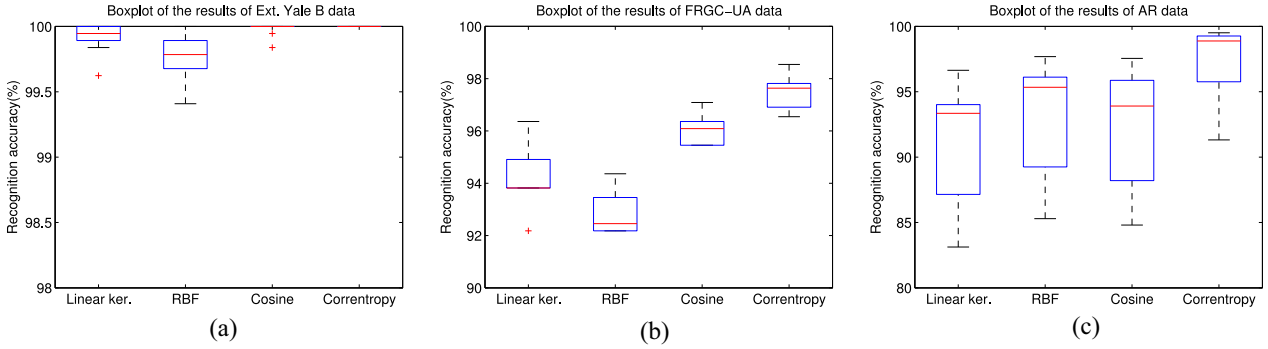


Fig. 3. Classification results by using different similarity functions in our dimension reduction stage. (a) Extended Yale B data. (b) FRGC uncontrolled illumination data. (c) AR data.

TABLE I
STATISTICS FOR THE EXPERIMENTAL DATA, WHERE n_c IS THE NUMBER OF CLASSES, n_s IS THE SIZE OF THE cTH CLASS

Database	n_c	n_s	Total size	Highlights
Ext. Yale B [36]	38	64	2432	Extreme Light
AR [37]	119	26	3094	Light & Occlusion
FRGC [38]	275	10	2750	Outdoor Light
LFW [39]	5749	–	13233	Uncontrolled condition

This configuration is used hereafter in our algorithm, and the number of projection vectors is set to $n_c - 1$, i.e., $h = n_c - 1$, in which n_c is the number of classes.

Some statistics of the image sets used in experiments are summarized in Table I. All images are aligned by the centers of eyes and mouth, according to the parameter values provided by the authors, and then resized to 56×46.

A. Similarities Comparison

The real performance of using different similarity functions are compared here. Four similarity functions, i.e., linear kernel, cosine, RBF, and correntropy, are respectively evaluated and then compared with each other. The hyper-parameter of correntropy is fixed to 1, while that of RBF is determined by fivefold cross-validation.

In the experiments with random sampling, the algorithm is independently evaluated 30 times, then the statistical tool boxplot is used to show the classification performance. On each box, the central mark in red is the median, the edges of the box are the 25th and 75th percentiles, and outliers are plotted individually.

For the extended Yale B set, ten samples are randomly chosen from each subject to build the training set, while the remainders are used for testing. The results obtained by using the four similarity functions are shown in Fig. 3(a). The results of RBF present a lower median accuracy while a larger variation than those of others. It even cannot exceed the results of the linear kernel, by which we infer that the hybrid lighting conditions destroy the basic assumption of Gaussian distribution. Conversely, the cosine function and correntropy present better results and the correntropy makes the extended Yale B data correctly classified.

For the uncontrolled face recognition grand challenge (FRGC) data, three samples of each subject are randomly chosen for training while the remainders are used for testing. The results are shown in Fig. 3(b). The median result of RBF is 92.5%, which is lower than 95.0% of the linear kernel. Both results of cosine and correntropy are higher than 96.0%, thus they display more competitive performance of the uncontrolled FRGC data. In particular, the median value of the correntropy is close to 98.0%, thus it outperforms others.

For the AR data, 14 samples in each person are randomly chosen to constitute the training set while the remainders are used for testing. From the boxplot as shown in Fig. 3(c), we can see that even though the variances of linear kernel and RBF are very close to each other, the median values of linear kernel and the cosine function are lower than those of Gaussian-based similarities. One possible reason of these contrasts may be attributed to the data distributions. Due to the mixture of lighting variations and continuous occlusions, the Gaussian-based similarities present much robustness in measuring the local structure and discriminant metric for subspace learning. In particular, the correntropy functions outperform the instance-based kernel functions in this experimental configuration.

Besides the random sampling-based experiments, the subset-based configurations which were designed by the authentic authors are also used to evaluate the algorithm's performance. The experimental results on the extended Yale B and AR databases are shown below.

For the extended Yale B data, subset 1 is used for training, and subsets 2–5 are used for testing. The lighting conditions become worse from subsets 2 to 5. For subset 2, the lighting condition is very close to the training set, all the results that using the four similarity functions can correctly recognize the whole testing set. For subset 3, the results of kernel-based similarity functions are lower than those of the correntropy-based similarities, both of which obtain 100% accuracy in this experimental setting. More evident contrasts are shown by the results of the remainder subsets. For subset 4, the linear kernel and RBF present their accuracies lower than 90%; however, both cosine and correntropy show their accuracies beyond 95% in Fig. 4(a). Moreover, the result of correntropy achieves to 98% for the testing data. The results on subset 5 display similar characteristics as subset 4 does. In particular, the correntropy outperforms other similarity functions and obtains the best result again.

TABLE II
COMPARATIVE RESULTS(%) OF AR SET AND LABELLED FACES IN THE WILD (LFW) SET

	AR scarves images			AR sunglasses images			LFW images		
	Cosine	RBF	Correntropy	Cosine	RBF	Correntropy	Cosine	RBF	Correntropy
KSRC	60.00	94.33	91.67	72.33	92.33	95.62	50.67±1.58	83.70±1.69	82.50±1.63
LOGO	**86.67**	**93.67**	**93.67**	**95.00**	**93.67**	**97.00**	**52.33±1.47**	**85.28±1.55**	**86.78±1.50**

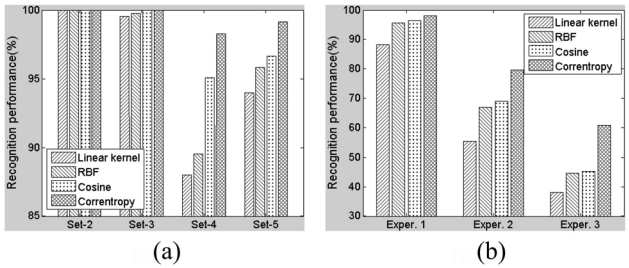


Fig. 4. Classification results by using different similarity functions. (a) Extended Yale B data. (b) Session-based AR data.

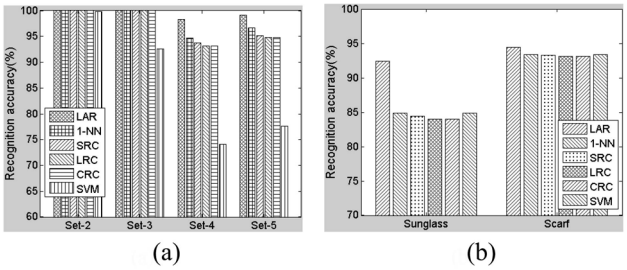


Fig. 5. Results by using different classifiers. (a) Extended Yale B data. (b) Occlusion-based AR data.

For the AR database, the frequently used settings [3], i.e., experiments 1–3, are exploited here to comparatively study the similarity functions. Specially, images 1–4 of session 1 for training, while:

1) in experiment 1, images 2–4 of session 2 for testing;
2) in experiment 2, images 5–7 of session 2 for testing;
3) in experiment 3, images 8–13 of session 2 for testing.

Which corresponds to the testing with expressions, illumination, and occlusions–illumination change, respectively. As the training instances and the testing instances are sampled from different sessions, these experimental configurations evaluate the generalization ability of algorithms.

In experiment 1, the result of linear kernel is obviously lower than those of other similarities. Fig. 4(b) shows that even though the results of other similarities are close to each other, the correntropy presents a higher accuracy therein. In experiment 2, the result of linear kernel has a large drop-off when compared with others. The result of RBF is close to that of cosine, but both of them have large accuracy gaps to the result of correntropy. The results of experiment 3 display similar characteristics, and all the similarities have lower accuracies than those of experiment 2.

In the experiments that are conducted above, we observe that correntropy outperforms the kernels and cosine in face recognition, no matter how complex the imaging conditions or the testing environments are. Just for this reason, the correntropy will be used in the following sections to evaluate other aspects of our new algorithm and then compare the performance with some state-of-the-art methods.

B. Classifiers Comparison

To illustrate the classification performance of the proposed classifier, some state-of-the-art classifiers, e.g., 1NN, SVM, SRC, and linear regression classification (LRC), are exploited to finish the feature matching task by using the same features. In particular, we also compare our results with those of KSRC method, which is the special case of $\boldsymbol{\gamma}_\theta = \boldsymbol{\gamma}_\varphi = 1$.

The results on the extended Yale B data are shown in Fig. 5(a). As we stated before, the lighting condition of subset 2 is very close to subset 1, thus all the classifiers can completely classify the images in subset 2. For other subsets, the recognition rates present very similar characteristics. Within the compared classifiers, SVM cannot achieve the high accuracies as others do. Specifically, the results of both subsets 4 and 5 are below 80%. However, our proposed classification rule can automatically choose the most correlated instances in its sparse representation procedure, thus it outperforms others in all the experiments.

Fig. 5(b) presents the results on the two special experiments, in which the occluded images, i.e., by sunglasses and scarves, are used for testing, but all the remainders are used for training. For the sunglasses occluded images, we can see the least angle-based criterion has a large improvement in classification accuracy, thus it obviously outperforms all the other classifiers. For the scarves occluded images, although the recognition rates are not significantly different, our new classifier correctly classify 94.46% testing instances thus still outperforms others.

In comparisons between our method and KSRC, we notice that the superiority of weighting optimization is more evidently presented by the learning procedure with complex data. For the data sets of controlled lighting conditions, the final classification accuracies of the two methods are almost the same. Therefore, only the results on the AR set and LFW data are presented as follows.

For the AR data, we use the frontal images without expression and lighting variations in session 1 to learn the discriminant subspace, and then use the occluded images in session 2 for testing. The results are presented in Table II. In the experiments with scarf-occlusions, when KSRC is used, the result of cosine function, RBF, and correntropy is 60.00%, 92.33%, and 94.64%, respectively. However, when our weighting procedure is applied to classification, the result is 86.67%, 93.67%, and 93.67%, respectively. In the KSRC-based experiments of sunglass occlusions, the result of cosine, RBF, and correntropy is 72.33%, 92.33%, and 95.62%, respectively. However, the result is increased to 95.00%, 93.67%, and 97.00% by our

TABLE III
RESULTS(%) OF EXTENDED YALE B DATA USING RANDOM SAMPLING, WHERE n_0 IS THE NUMBER OF TRAINING SAMPLES OF EACH SUBJECT

n_0	LDA [12]	Gabor [16]	LTV [40]	LBP [17]	Gradface [8]	DFD [7]	LGOP [22]	IGO [3]	LOGO
5	70.99±5.49	92.99±4.53	80.16±4.21	88.07±4.17	91.89±2.29	88.87±4.48	94.33±3.38	94.61±4.55	**99.85±0.24**
10	81.28±5.14	97.78±1.77	88.85±3.89	94.24±2.85	95.15±0.96	95.09±1.97	98.73±0.98	99.76±0.17	**99.99±0.03**

TABLE IV
RESULTS(%) OF EXTENDED YALE B DATA USING DIFFERENT SUBSETS

subset	LDA [12]	Gabor [16]	LTV [40]	LBP [17]	Gradface [8]	DFD [7]	LGOP [22]	IGO [3]	LOGO
2	100	100	100	100	96.05	100	99.78	99.78	**100**
3	95.60	95.60	99.34	99.78	93.41	96.04	96.70	99.56	**100**
4	34.41	35.17	69.58	80.42	66.92	71.71	62.59	67.30	**95.06**
5	6.72	6.58	34.31	76.19	61.34	67.93	67.93	63.03	**96.64**

weighting approach. Therefore, the superiority of weighting is experimentally validated by the AR data.

For the LFW data, we use the standard test protocol to compare the results between different similarity functions. The reported results shown in Table II are the average accuracies and their variations in tenfold cross-validation. When the KSRC is used to obtain the discriminant subspace and classification, the result of the three similarities is 50.67%, 83.70%, and 82.50%, respectively. Nevertheless, when our weighting approach is used to simultaneously optimize the discriminant subspace, we can see the average result is 52.33%, 85.28%, and 86.78%, respectively.

Therefore, both results that obtained by using the AR and LFW sets validate the superiority of weighting approach. In other words, the weighted regularization in computing the fused similarities indeed plays an important role in feature normalization and over-dominance alleviation.

C. Computational Time and Feature Dimensions

The computational time of both feature learning and classification is shown here. Particularly, the feature learning methods include linear discriminant analysis (LDA) [12], kernel discriminant analysis (KDA) [41], Gabor [16], logarithmic total variation (LTV) [40], LBP [17], discriminant face descriptor (DFD) [7], LGOP, and IGO-LDA, while classifiers include 1NN, SVM, SRC, LRC [42], collaborative representation classification (CRC) [43], RSC, correntropy-based sparse representation (CESR), and LARS. The total 570 training images are randomly chosen from extended Yale B set.

The final features that will be input to the classifiers are summarized as follows. All the discriminant analysis-based feature extractors, i.e., LDA, KDA, Gabor, IGO, and LOGO, preserve $n_c - 1$ features for classification, where n_c is the number of classes. LTV and gradientface generate the same feature sizes as the original image, namely 2576. LBP divides the image into 8×6 blocks and then extracts the local binary coding-based histograms, thus it generates 12 288 features to compute the χ^2 kernel values. For DFD and LGOP, we use the default parameter settings in the local coding stage, and then reduce the feature dimensions to 400 by the whitened principal component analysis [20].

The computational time of each training method is shown in Table V. As a classical dimension reduction method, LDA takes 0.84 s to compute the discriminant subspace by directly using the pixel images. When the Gabor filters are used to generate multiscale features, the training time is increased to 60.16 s. Although gradientface does not need to attain discriminant subspace for dimension reduction, the gradient features require 9.36 s to prepare for the cosine-based distances. Due to the time-spending ℓ_1-based total variation optimization, LTV uses more time than 40 min to compute the robust features. IGO method takes 16.88 s to complete the cosine similarity calculation and discriminant analysis. The computational time of LBP, DFD, and LGOP is 119.19, 512.38, and 348.30 s, respectively. As our LOGO method enhances the LGOP features by using the correntropy as the similarity metric and then using the multiple kernel-based subspace learning procedure, it needs 455.96 s in training.

To measure the test time of the classifiers, we input the LOGO features into different classifiers, and then report the computational time of classifying one test instance. We think the same input features can provide a fair condition in comparing the classification efficiency. The results are shown in Table VI.

Theoretically, the 1NN classifier takes little time in classification because it only need to compute the simple Euclidean distances without any optimization. The practical computational time is 2.18e-4 s, which is obviously less than those of other classifiers. Except for SVM, of which the test time is 3.23e-2 s, the remainder classifiers can be viewed as the same category because all of them are regression models of different regularization terms. The efficiency of SRC, LRC, and CRC is close to each other, as the time is 11.05e-2, 10.78e-2, and 17.85e-2, respectively. As RSC is motivated by dealing with the occlusion-based data representation problem, the extended Yale B data make the optimization of RSC very simple, and thus it only needs 4.35e-2 s to classify the test instance. In contrast, CESR adopts half-quadratic optimization approach to solve the coefficient vector and the time is 32.10e-2 s. Notice that the correlation-based LARS only takes 6.84e-2 s in test, thus it is more efficient than CRC and CESR.

V. COMPARATIVE RESULTS

In this section, we compare LOGO with some state-of-the-art methods, which include descriptor-based methods, e.g., Gabor-LDA, LBP, IGO-LDA, and the regression-based classifiers such as SRC, LRC, RSC, and CESR.

TABLE V
TRAINING TIME (SECOND) OF THE COMPARED METHODS. THE TOTAL 570 TRAINING IMAGES OF RESOLUTION 56*46 ARE RANDOMLY CHOSEN FROM THE EXTENDED YALE B SET

LDA	KDA	Gabor	LTV	LBP
0.84	0.98	60.16	2809.78	119.19
Gradface	DFD	LGOP	IGO	LOGO
9.36	512.38	348.30	16.88	455.96

TABLE VI
TESTING TIME (SECOND) OF THE CLASSIFIERS. THE TESTING IMAGE IS RANDOMLY CHOSEN FROM THE EXTENDED YALE B SET

1NN	SVM	SRC	LRC
2.18e-4	3.23e-2	11.05e-2	10.78e-2
CRC	RSC	CESR	LARS
17.85e-2	4.35e-2	32.10e-2	6.84e-2

A. Results for Extended Yale B Data

We make two groups of experiments for different test tasks, in which the first group is based on random sampling and the other is based on different lighting subsets.

1) Testing With Random Sampling: First, ten samples from each subject are randomly selected to constitute the training set, and the remainders are used for testing. We present the average recognition rates in Table III. As the pixel-based data are used for subspace learning, the average accuracy of LDA is only close to 71%. By the total variation transform and discriminant learning, LTV and LBP obtain their respective accuracy as 81.06% and 88.07%. The results of Gabor and DFD are 93.0% and 88.07%, respectively, thus they further improve the recognition performance. For the gradient-based feature descriptors, the result of gradientface is 91.89% which is lower than those of LGOP, IGO, and LOGO methods. Due to the local order-based feature fusion, our LOGO correctly classify 99.85% of the testing instances, thus outperforms other methods.

When the training size for each subject is increased to ten, all the methods enhance their discrimination capacities which are reflected by their recognition accuracies. The detailed results are also shown in Table III. Although the results of LGOP and IGO achieve to 98.73% and 99.76%, respectively, our LOGO method can correctly classify almost all the samples as its accuracy is stably close to 100%.

2) Testing With Preassigned Subsets: In another experiment setting, the extended Yale B data are divided into five nonoverlapping subsets according to the original data constitution. Then subset 1 is used for training, and other sets are used for subsequent testing. We present the test results in Table IV. The illumination condition becomes more extreme as the indexes of subset increase, all these methods except for gradientface and IGO-LDA obtain 100% accuracies for subset 2. However, when subset 3 is used for testing, only our LOGO keep the 100% accuracy, which is closely followed by the LTV, LBP, and IGO methods. For subset 4, LOGO obtains a result close to 95.06%, which is followed by the accuracy 80.42% of LBP. All the remainder methods can only obtain lower recognition rates than 80%. When subset 5 is used for testing, the results of LDA and Gabor-LDA methods only achieve 6.72% and 6.58%, which indicate that both the pixel features and Gabor filters cannot provide robust features for the extremely distributed lighting variations. The phrase angle-based methods, i.e., gradientface and IGO-LDA, improve the recognition accuracies by different ranges, but their results are still inferior to that of LBP, which correctly recognizes 76.2% samples in the test set. Notice that the result of LOGO is 96.64%, thus it significantly outperforms others even though with these extreme illuminations.

B. Results for FRGC Data

The outdoor lighting subset of FRGC data is exploited here for evaluating the recognition performance. The uncontrolled images were taken in varying illumination conditions, e.g., hallways, atriums, or outside. We conduct two groups of experiments for different evaluation purposes.

In the first group, five controlled lighting images of each person compose the whole data, and three of which are randomly chosen for training and the remainders for testing. This is a relative easy recognition task due to its controlled lighting conditions, and the results have been shown in Table VII. Except for Gabor and LTV, each of the other methods can obtain a recognition rate higher than 95%. The best performance is obtained by LGOP, which correctly classify all the testing samples. Notice that the results of DFD and LOGO are 99.78% and 99.56%, respectively, thus they are very close to that of LGOP. The result of LTV is only 89.78%, it indicates that the total variation operator cannot capture the discriminant and structured information within the neighborhood regions. On the contrary, the gradient-based descriptors, such as orientations and local order codes, provide very important and structured patterns in feature representation.

In the second group of experiments, three of five uncontrolled lighting images of each person are randomly chosen for training and the remainders for testing. Due to the more complex lighting image mechanism in the outdoor environment, it presents a relatively difficult feature extraction and matching task. As shown in Table VII, the top three results are obtained by the LGOP, IGO, and LOGO methods, of which the recognition rate is 94.05%, 96.36%, and 97.66%, respectively. Several conclusions can be summarized from the results. First, due to the more complex lighting variations than that of the controlled data, all the methods decrease their recognition rates by different degrees. Second, except for IGO and LOGO, other methods are sensitive to the uncontrolled lighting condition. Moreover, we find the subspace learning-based methods, i.e., IGO and LOGO, outperform the gradientface model. It indicates that the supervised subspace is indeed important for discriminant and compact feature representation.

C. Results for AR Data

We make three groups of experiments as follows.

1) Testing With Random Sampling: The first group of experiments is based on random sampling, and the objective focuses on training with mixed instances, which may be captured in different periods and with continuous occlusions. The recognition task incorporating occluded samples within

TABLE VII
RESULTS(%) OF FRGC DATA USING RANDOM SAMPLING

subset	LDA [12]	Gabor [16]	LTV [40]	LBP [17]	Gradface [8]	DFD [7]	LGOP [22]	IGO [3]	LOGO
Controlled	98.85±0.66	94.03±2.47	89.78±1.52	98.42±1.72	96.54±1.64	99.78±0.25	100.0±0.00	98.71±1.23	**99.56±0.28**
Uncontrolled	81.36±2.00	80.80±1.63	72.41±1.32	74.05±1.23	75.17±2.03	84.75±2.74	94.05±1.15	96.36±0.65	**97.66±0.62**

TABLE VIII
RESULTS(%) OF AR DATA USING RANDOM SAMPLING, WHERE n_0 IS THE NUMBER OF TRAINING SAMPLES OF EACH SUBJECT

n_0	LDA [12]	Gabor [16]	LTV [40]	LBP [17]	Gradface [8]	DFD [7]	LGOP [22]	IGO [3]	LOGO
10	74.37±5.47	84.02±7.74	52.77±7.66	79.82±4.71	84.70±4.37	81.16±4.57	92.74±2.97	87.17±3.66	**93.24±3.31**
14	72.33±4.55	92.81±4.22	62.21±7.50	83.86±3.50	89.17±3.72	86.31±4.64	96.16±2.53	93.95±4.93	**96.22±3.49**

TABLE IX
RESULTS(%) OF AR DATA USING OCCLUDED FACES IN DIFFERENT SESSIONS

occlusion	LDA [12]	Gabor [16]	LBP [17]	SRC [29]	RSC [31]	CESR [28]	Gradface [8]	DFD [7]	LGOP [22]	IGO [3]	LOGO
Scarves	11.67	47.67	76.67	12.70	72.70	12.67	66.67	75.33	91.33	42.33	**93.67**
Sunglasses	31.33	44.00	87.67	57.30	80.30	63.00	83.33	67.33	91.67	59.67	**97.00**

Fig. 6. Demo of sunglasses and scarves occlusions.

training set has a mixed blessing. On one hand, the occluded samples will lead to a large variation and make biased statistical computation, then increase the difficulty of classification. On the other hand, occlusion can be viewed as a typical noise, as shown in [44], so introducing noisy samples into training set is equivalent to make a Tikhonov regularization on the learning system, thus the generalization ability of the obtained subspace can be enhanced for out-of-sample problem.

The results using 10 and 14 samples/class for training are shown in Table VIII. The result of LTV in the two experiments is 52.77% and 62.21%, respectively, thus they are obviously lower than those of other methods. However, due to the stronger distinguish-ability of Gabor and gradient filters, the Gabor and gradient-based descriptors obtain better results than those of LTV, LDA, and LBP. When 14 samples are randomly chosen from each subject for training, for example, the recognition rate of Gabor-based method is 92.81%, while the result of LGOP, IGO, and LOGO is 96.16%, 93.95%, and 96.22%, respectively. With concern of the gradientface method, which corresponds to the accuracy of 89.17%, we can see the local order coding plays an important role in classifying the samples, and therefore LGOP and LOGO obtain the top results.

2) Test by Continuously Occluded Images: In the second group of experiments, we focus on the performance in presence of contiguous occlusion, which is inarguably one of the most challenging paradigm in robust face recognition. The AR database consists of two modes of contiguous occlusion, i.e., sunglasses and scarves. Fig. 6 reflects these two scenarios for two sessions, and the second and the fifth images are with lighting variations.

As in [3], [29], and [31], a subset (with only illumination and expression changes) that contains 50 male and 50 female subjects was chosen from the original AR database in this experiment. For each subject, the seven images from session 1 were used for training, with other seven images from session 2 for testing. The detailed recognition accuracies are listed in Table IX. Due to this special configuration, the data in training set and testing set can be viewed as heterogeneous, thus the recognition task is different and the feature representation should be more carefully designed.

For the experiments that tested by sunglasses occlusions. The pure pixel-based subspace learning and feature matching methods are not reasonably effective, thus LDA can only classify 13.6% of the testing samples. Although Gabor and LBP transformations integrate local contrasts in the structured coding stage, the result is only 22.69% and 91.74%, respectively. By incorporating the gradient orientation features, LGOP and gradientface obtain their recognition rates of 83.19% and 86.27%, respectively. Further improvements are displayed by the sparse regression-based classifiers with robust similarity metrics. The result of SRC, RSC, and CESR is 87.0%, 60.5%, and 74.86%, respectively. As we can see, the best result is obtained by 92.44% of our LOGO method, which is closely followed by LBP. By comparing with other results as shown in Table IX, e.g., CESR, IGO, and LGOP, we attribute the remarkable performance to the ensemble of structured local order patterns and robust correntropy for discriminant metric learning.

For the scarves occluded images, the result of SRC and RSC is 59.5% and 57.0%, respectively, which are lower than 89.92% of CESR. It indicates the importance of maximum correntropy metric in sparse representation-based classification rules. By learning of the locality importance, DFD slightly promotes the result of LBP from 78.99% to 80.81%. By using the gradient orientation features, the results of gradientface and IGO are 85.57% and 71.15%, respectively, which are consistently lower than 92.30% of LGOP and 94.46% of LOGO. It shows that the local order coded descriptors indeed improve the recognition accuracy of the simple orientation-based features. Moreover, even though the supervised learning approach are respectively applied to IGO and LOGO, the superiority of local order patterns is still displayed by the LOGO algorithm.

TABLE X
MEAN ACCURACY (%) ON LFW DATABASE

Method	IGO	Gradientface	LARK	APEM	DML-eigSIFT	MRF-MLBP	LBP multi-shot	DFD	LGOP	LOGO
Accuracy	66.02±1.09	67.38±1.72	72.23±0.50	84.04±1.20	81.27±2.30	79.08±0.14	85.17±0.61	83.13±0.50	83.35±1.13	**86.78±1.50**

Fig. 7. Some examples of one person in LFW set.

D. Face Verification Results for LFW Data

LFW database [39] contains 13 233 images of 5749 people downloaded from the Web, which is designed for totally unconstrained face recognition with dramatic variations of pose, illumination, expression, misalignment, occlusion, and so on. Some images are shown in Fig. 7.

The specified LFW evaluation protocols are used for face verification. The database is divided into ten disjoint splits, which contain different identities. In the unrestricted protocol, the training information is provided as simply the names of the people in each split, thus one can formulate as many match and mismatch pairs as one desires, from people within each split. The testing set is the 600 predefined image pairs in the remaining split, where 300 are positive pairs portraying the same person and the remaining 300 are negative pairs portraying different people.

The images are resized to 50×50 for reducing the time cost. Notice that there are more than 10 000 training samples of more than 5000 people in each experiment, but most of people have only one or two samples. To obtain more representative features, the persons who have the more than five images are chosen to learn the discriminant subspace.

Besides the methods including gradientface, IGO, LGOP, and DFD, some benchmark and related methods, e.g., multi-resolution LBP in Markov random field (MRF-MLBP) [45], LBP multishot [46], locally adaptive regression kernel (LARK) [47], adaptive probabilistic elastic matching (APEM) [48], distance metric learning (DML)-eigSIFT [49], information theoretic-based discriminant metric learning (IDML) [50], and V_1-like multiple kernel learning (MKL) [51] are also exploited here for a broader comparison. Notice that the IDML and DML-eigSIFT are metric learning methods. LARK and V_1-like MKL emphasize the kernel-based regression models. The MRF-MLBP, LBP multishot, and APEM can be attributed into the LBP or scale-invariant feature transform (SIFT)-based local descriptor fusion methods. We present the results (mean±std) of the compared methods in Table X, which shows that the result of LOGO is 86.78±1.50. It indicates that the proposed new features not only beat the traditional gradient-based descriptors, but also outperform the local feature fusion-based kernel learning models. Fig. 8 displays the receiver operating characteristic curve (ROCs), and it demonstrates the superiority of our LOGO in the LFW verification task.

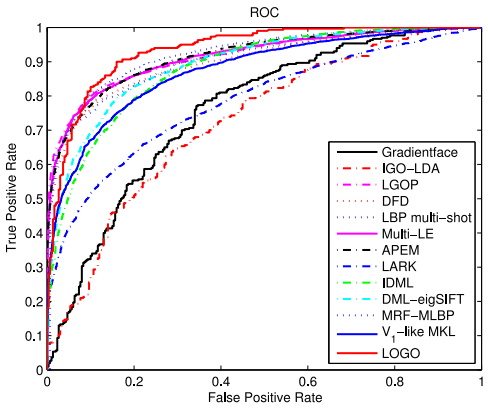


Fig. 8. ROC over view 2 on the LFW database.

It has been recently reported that the best accuracy of the LFW validation task has exceeded 99%, which was obtained by deep learning method [23] with large-scale data. It should be noted that the great improvement of validation accuracy is reported under a different test protocol [23], which exploits lots of foreign-aided data with more complex image variations to assist the training process. Pure comparison of recognition accuracies but in different test protocols is not suitable, and thus it is not further discussed in this paper.

VI. CONCLUSION

This paper proposes a enhanced IGO descriptor based on discriminant subspace learning. The novelty concentrates on the local order-based feature coding and the correntropy-based similarity in sparse representation classification. Along this way, the gradient filters are used to describe local contrasts within neighboring pixel points, thus the proposed feature descriptor enhances the local textures and further discovers intrinsic structure of facial images. Two kinds of similarity functions, i.e., kernel and correntropy, are used to measure the closeness of each pair of instances, and then present a comparative study. Experimental results show that the feature representation method achieves competitive performance under complex conditions, including extreme illumination and occlusion variations. How to improve the domain adaptation ability of the algorithm is our future work.

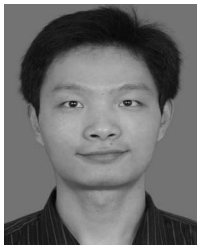
REFERENCES

[1] S. C. Yan *et al.*, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[2] Y. L. Chen and C. T. Hsu, "Multilinear graph embedding: Representation and regularization for images," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 741–754, Feb. 2014.

[3] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2454–2466, Dec. 2012.

[4] J. Chen and J. Yang, "Robust subspace segmentation via low-rank representation," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1432–1445, Aug. 2014.
[5] M. Kan, D. Xu, S. Shan, W. Li, and X. Chen, "Learning prototype hyperplanes for face recognition in the wild," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3310–3316, Aug. 2013.
[6] H. C. Chi, M. A. Tahir, J. Kittler, and M. Pietikainen, "Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1164–1177, May 2013.
[7] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning discriminant face descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 289–302, Feb. 2014.
[8] T. Zhang, Y. Y. Tang, B. Fang, Z. Shang, and X. Liu, "Face recognition under varying illumination using gradientfaces," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2599–2606, Nov. 2009.
[9] M. D. Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust face recognition for uncontrolled pose and illumination changes," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 1, pp. 149–163, Jan. 2013.
[10] Y. Xu *et al.*, "Data uncertainty in face recognition," *IEEE Trans. Cybern.*, vol. 44, no. 10, pp. 1950–1961, Oct. 2014.
[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[12] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[13] D.-Q. Dai and P. C. Yuen, "Face recognition by regularized discriminant analysis," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 4, pp. 1080–1085, Aug. 2007.
[14] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.
[15] C. X. Ren, D. Q. Dai, X. X. Li, and Z. R. Lai, "Band-reweighted Gabor kernel embedding for face image representation and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 725–740, Feb. 2014.
[16] C. J. Liu and H. Wechsler, "Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.
[17] 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, Dec. 2006.
[18] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.
[19] Z. Lei, S. Liao, M. Pietikainen, and S. Z. Li, "Face recognition by exploring information jointly in space, scale and orientation," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 247–256, Jan. 2011.
[20] N.-S. Vu and A. Caplier, "Enhanced patterns of oriented edge magnitudes for face recognition and image matching," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1352–1365, Mar. 2012.
[21] D. Huang, C. Zhu, Y. H. Wang, and L. M. Chen, "HSOG: A novel local image descriptor based on histograms of the second-order gradients," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4680–4695, Nov. 2014.
[22] Z. Lei, D. Yi, and S. Z. Li, "Local gradient order pattern for face representation and recognition," in *Proc. ICPR*, Stockholm, Sweden, Aug. 2014, pp. 387–392.
[23] Y. Sun, Y. Cheng, X. G. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," *Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 1988–1996.
[24] Z. Wang, B. Fan, and F. Wu, "Local intensity order pattern for feature description," in *Proc. ICCV*, Barcelona, Spain, Nov. 2011, pp. 603–610.
[25] X. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Process. Mag.*, vol. 28, no. 2, pp. 16–26, Mar. 2011.
[26] S. Z. Li, "On discontinuity-adaptive smoothness priors in computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 6, pp. 576–586, Jun. 1995.
[27] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.
[28] R. He, W.-S. Zheng, and B.-G. Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.
[29] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[30] R. Diao, F. Chao, T. Peng, N. Snooke, and Q. Shen, "Feature selection inspired classifier ensemble reduction," *IEEE Trans. Cybern.*, vol. 44, no. 8, pp. 1259–1268, Aug. 2014.
[31] M. Yang, D. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *Proc. CVPR*, Providence, RI, USA, Jun. 2011, pp. 625–632.
[32] Z.-R. Lai, D.-Q. Dai, C.-X. Ren, and K.-K. Huang, "Discriminative and compact coding for robust face recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1900–1912, Sep. 2015.
[33] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.
[34] L. Zhang *et al.*, "Kernel sparse representation-based classifier," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1684–1695, Apr. 2012.
[35] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Multiple kernel sparse representations for supervised and unsupervised learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2905–2915, Jul. 2014.
[36] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.
[37] A. Martínez and R. Benavente, "The AR face database," Univ. Autónoma de Barcelona, Barcelona, Spain, CVC Tech. Rep. 24, Jun. 1998.
[38] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE CVPR*, vol. 1. San Diego, CA, USA, 2005, pp. 947–954.
[39] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, Oct. 2007.
[40] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. S. Huang, "Total variation models for variable lighting face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1519–1524, Sep. 2006.
[41] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller, "Fisher discriminant analysis with kernels," in *Proc. IEEE Int. Workshop Neural Netw. Signal Process.*, vol. 6. Madison, WI, USA, Aug. 1999, pp. 41–48.
[42] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, Nov. 2010.
[43] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. ICCV*, Barcelona, Spain, Nov. 2011, pp. 471–478.
[44] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Comput.*, vol. 7, no. 1, pp. 108–116, 1995.
[45] S. R. Arashloo and J. Kittler, "Efficient processing of MRFs for unconstrained-pose face recognition," in *Int. Conf. Biometrics Theor., Appl. Syst.*, Arlington, VA, USA, 2013, pp. 1–8.
[46] Y. Taigman, L. Wolf, and T. Hassner, "Multiple one-shots for utilizing class label information." in *Proc. BMVC*, London, U.K., Sep. 2009, pp. 1–12.
[47] H. J. Seo and P. Milanfar, "Face verification using the LARK representation," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 4, pp. 1275–1286, Dec. 2011.
[48] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE CVPR*, Portland, OR, USA, 2013, pp. 3499–3506.
[49] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1–26, 2012.
[50] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE CVPR*, Kyoto, Japan, 2009, pp. 498–505.
[51] N. Pinto, J. J. DiCarlo, and D. D. Cox, "How far can you get with a modern face recognition test set using only simple features?" in *Proc. IEEE CVPR*, Miami, FL, USA, 2009, pp. 2591–2598.



Chuan-Xian Ren received the B.S. degree in mathematics from Fuyang University, Fuyang, China, in 2005, and the Ph.D. degree in applied mathematics from Sun Yat-Sen University, Guangzhou, China, in 2010.

From 2010 to 2011, he was a Senior Research Associate with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong. He is currently an Assistant Professor with the Faculty of Mathematics and Computational Science, Sun Yat-Sen University. He was elected as a candidate for the "Thousand-Hundred-Ten" Talents Program of Guangdong Province in 2014. His current research interests include image processing, pattern analysis, and machine learning.

Zhen Lei (M'11) received the B.S. degree in automation from the University of Science and Technology of China, Hefei, China, in 2005, and the Ph.D. degree in automation from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010.

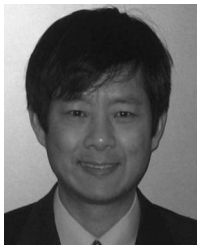
He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. He has published over 90 papers in international journals and conferences. His current research interests include computer vision, pattern recognition, image processing, and face recognition in particular.

Dr. Lei served as an Area Chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015.

Dao-Qing Dai (M'07) received the B.Sc. degree from Hunan Normal University, Changsha, China, in 1983, the M.Sc. degree from Sun Yat-Sen University, Guangzhou, China, in 1986, and the Ph.D. degree from Wuhan University, Wuhan, China, in 1990, all in mathematics.

From 1998 to 1999, he was an Alexander von Humboldt Research Fellow with Free University, Berlin, Germany. He is currently a Professor with the Faculty of Mathematics and Computing, Sun Yat-Sen University, Guangzhou. His current research interests include image processing, wavelet analysis, face recognition, and bioinformatics. He has authored or co-authored over 100 refereed technical papers.

Dr. Dai was a recipient of the Outstanding Research Achievements in Mathematics Award from the International Society for Analysis, Applications, and Computation, Fukuoka, Japan, in 1999. He served as a Program Co-Chair of Sinobiometrics in 2004 and a Program Committee Member for several international conferences.



Stan Z. Li (F'09) received the B.Eng. degree in radio technology from Hunan University, Changsha, China, the M.Eng. in radio technology degree from the National University of Defense Technology, Changsha, and the Ph.D. degree in pattern recognition from Surrey University, Surrey, U.K.

He is currently a Professor and the Director of the Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was an Associate Professor with Nanyang Technological University, Singapore. He was a Researcher with Microsoft Research Asia from 2000 to 2004. He has published over 200 papers in international journals and conferences, and authored and edited eight books. His current research interests include pattern recognition and machine learning, image and vision processing, face recognition, biometrics, and intelligent video surveillance.

Mr. Li was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is acting as the Editor-in-Chief for the *Encyclopedia of Biometrics*. He served as the Program Co-Chair for the International Conference on Biometrics in 2007, 2009, and 2015, and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was elevated to IEEE Fellow for his contributions to the fields of face recognition, pattern recognition, and computer vision.