Sparse Tensor Discriminant Color Space for Face Verification

Su-Jing Wang, Jian Yang, *Member, IEEE,* Ming-Fang Sun, Xu-Jun Peng, Ming-Ming Sun and Chun-Guang Zhou*

***Abstract*—As one of the fundamental features, color provides useful information and plays an important role for face recognition. Generally, the choice of a color space is different for different visual tasks. How can a color space be sought for the specific face recognition problem? To address this problem, we propose a Sparse Tensor Discriminant Color Space (STDCS) model which represents a color image as a third-order tensor in this paper. The model can not only keep the underlying spatial structure of color images but also enhance robustness and give intuitionistic or semantic interpretation. STDCS transforms the eigenvalue problem to a series of regression problems. Then one spare color space transformation matrix and two sparse discriminant projection matrices are obtained by applying lasso or elastic net on the regression problems. The experiments on three color face databases, AR, Georgia Tech and LFW face databases, show that both the performance and the robustness of the proposed method outperform those of the state-of-the-art Tensor Discriminant Color Space (TDCS) model.**

***Index Terms*—Face recognition, Color images, Discriminant information, Tensor subspace, Sparse representation.**

I. Introduction

After decades of research and development, face recognition has attained considerable success in the field of personal identification and public security, such as crime and terrorist recognition. The various face recognition methods attract great interests from researchers. Some of them focus on how to extract the effective features from facial images. Principal Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [2] are two popular methods for feature extraction. To keep the spatial structure information of facial images, Two-Dimensional Principal Component Analysis (2D-PCA) [3] and Two-Dimensional Linear Discriminant Analysis (2D-LDA) [4] directly computed covariance (scatter) matrices from the image matrix. Recently, the feature extraction methods based on tensor is a hot topic. The methods can be divided into 2 categories. In one category [5][6][7], a high order tensor constructs a multilinear structure and models the multiple factors of facial variation (e.g., different user identities, various user postures and facial expressions, varying lights, etc.) using High-Order Singular Value Decomposition (HOSVD) [8][9][10]. In the other category[11][12][13][14], the conventional transformation methods (such as PCA, Singular Value Decomposition (SVD) and Locality Preserving Projections (LPP)[15]) are generalized to tensors. They treat a gray image as a 2nd-order tensor, a color image as a 3rd-order tensor.

Some of researchers focus on how to design novel classifiers for face recognition. Kumar *et al.* [16] created the first image search engine based entirely on faces. They divided the face into various regions and extracted the features from these regions. Then, Support Vector Machine (SVM) and Adaboost were used to classify attributes by various combinations of these regions. Wright *et al.* [17] developed a classifier based on sparse solution to enhance the performance of occluded face recognition. Schwartz *et al.* [18] developed a classifier based on Partial Least Squares (PLS) and discriminative tree to accelerate face identification for large data sets. To deal well with the large-scale and high dimensional data sets, Fan *et al.* [19]used the sample neighbors to effectively captures the structures of the data and may enhance the face recognition result. Zafeiriou *et al.* [20] used regularized kernel discriminant analysis to enhance face recognition and verification.

However, many methods only use gray-scale information of face images. Thus plenty of color information, which is useful for face recognition according to recent researches [21], [22], [23], [24], is lost. In [21], the experimental results showed that the recognition accuracy was improved if color information was available for PCA based methods. In [22], a RGB image whose size was $I_1 \times I_2$ was transformed into a $I \times 3$ matrix ($I = I_1 \times I_2$) and the 2D-PCA was applied on all transformed matrices to recognize the color face images. The experimental results showed that the accuracy can be improved by about 3% compared to the same method which was applied on the corresponding gray-scale images. It was demonstrated by Choi et al. [23] that the recognition performance can be significantly improved for low resolution face images (20 pixels or less) using facial color cue. Other researches [24] also reveal that different color spaces (such as RGB, PCA color space (PCS) and YIQ color space) provide better face

This work was supported by (1) the National Natural Science Foundation of China under Grant No. 60973092, 60903097, 61175023, (2) the Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China. This work was partially supported by the National Science Foundation of China under Grant No. 60973098,61005005 and the National Science Fund for Distinguished Young Scholars under Grant No. 61125305.

S.J Wang, M.F Sun and C.G Zhou are with the College of Computer Science and Technology, Jilin University, Changchun 130012, China. (e-mail:wangsj08@mails.jlu.edu.cn; sunmf09@mails.jlu.edu.cn; cgzhou@jlu.edu.cn).

J Yang is now with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China and also with the Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125. (e-mail: csjyang@njit.edu; csjyang@mail.njust.edu.cn).

X.J Peng is with Raytheon BBN technologies, Cambridge, MA, 02138, USA. (e-mail:xpeng@bbn.com).

M.M Sun is with the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail:sunmingming@gmail.com).

recognition performance than gray scale only.

In computer vision field, most color spaces can be defined by a transformation of the RGB color space, which is the most widely used color space. The linear transformation color spaces, such as the YUV and YIQ color space [25], are usually associated with the properties of some hardwares. While the nonlinear transformation color spaces, such as the HSV and L*a*b* color space, are generally related to the human vision system.

Usually, the R, G, and B component images in the RGB color space are correlated. Decorrelation among the components of component images helps reduce redundancy and is an important strategy to improve the accuracy of the subsequent recognition method [26]. Liu [27] proposed the uncorrelated color space (UCS), the independent color space (ICS), and the discriminating color space (DCS). Specifically, the UCS applies PCA to decorrelate the R, G, and B component images. The ICS and DCS further enhance the discriminating power for the subsequent recognition method by means of Independent Component Analysis (ICA) and LDA, respectively.

When an optimal color space is obtained, its effectiveness is evaluated by using a recognition method. This separate strategy cannot theoretically guarantee that the optimal color space is best for the subsequent recognition method, and therefore, cannot guarantee that the resulting face recognition system is optimal in performance. Color Image Discriminant (CID) model [28] is to seek a m optimal color space and an effective recognition method of color images in a unified framework. However, CID vectorizes each component image into a high dimensional vector. This results in the loss of spatial structure information of the component images. Moreover, CID also suffers from the small sample size problem. To overcome these drawbacks, Wang *et al.* [29] presented a Tensor Discriminant Color Space (TDCS) model which used a 3rd-order tensor to represent a color facial image.

Now, we review the above algorithms from the feature selection (or rather, the feature transformation) view. UCS, ICS and DCS are the results of feature transformations by using PCA, ICA and LDA on the RGB color components. CID used LDA not only on the RGB color components but also on image information. To keep more spatial structure information of the component images, TDCS use Discriminant Analysis with Tensor Representation (DATER) to transform the RGB color components and the image information.

Unfortunately, there is a lot of noise on real data. So, the above feature transformations can not obtain the best performance on real data. We can impose a sparse constraint on their object functions. The sparseness is a tradeoff between the optimal solution of their object functions and the noises.

Recently, the sparse feature transformation is one of the hottest topics. Many papers show that the sparse feature transformation methods can obtain better performance than their corresponding non-sparse methods in the real data. And these sparse methods can give an intuitionistic or semantic interpretation for the transformed subspace. Sparse PCA (SPCA) was first proposed in [30] by applying the least angle regression [31] and Elastic Net of ℓ_1-penalized regression [32] on regular principal components. In [33], Moghaddam et al. suggested a spectral bounds framework for sparse subspace learning and presented both exact and greedy algorithms for sparse LDA (SLDA). They also described the same framework for sparse PCA but they can only be applied to two-class problem [34]. In order to address this problem, Qiao et al. [35] extended the SPCA to obtain sparse discriminant vectors.

In this paper, we draw upon the insights from these approaches and explore a Sparse Tensor Discriminant Color Space (STDCS) model which is an extension of TDCS [29]. The aim of STDCS is to make two discriminant projection matrices U_1, U_2 and one color space transformation matrix U_3 sparse. Compared to the TDCS model, STDCS has two main advantages: 1) the intuitionistic or semantic interpretation and 2) the robustness not only for similarity measurement of images but also for noised images.

The rest of this paper is organized as follows: in Section II, we give the related definitions to tensor; in Section III, we will briefly review the TDCS model; which is followed by an introduction of the STDCS model in Section IV; in Section V, the experiments are conducted on three color face databases: AR, Georgia Tech and LFW face databases, and the results are covered in the same section which show that the efficiency and performance of STDCS are better than those of TDCS and CID; finally in Section VI, conclusions are drawn and several issues for the future works are described.

II. Tensor Fundamentals

A tensor is a multidimensional array. It is the higher-order generalization of scaler(zero-order tensor), vector(1st-order tensor), and matrix (2nd-order tensor). In this paper, lowercase italic letters (*a, b, ...*) denote scalars, bold lowercase letters (**a, b, ...**) denote vectors, bold uppercase letters (**A, B, ...**) denote matrices, and calligraphic uppercase letters ($\mathcal{A}$, $\mathcal{B}$, ...) denote tensors. The formal definition is given below[10]:

Definition 1. *The* order *of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is N. An element of $\mathcal{A}$ is denoted by $\mathcal{A}_{i_1 i_2 \ldots i_N}$ or $a_{i_1 i_2 \ldots i_N}$, where $1 \leq i_n \leq I_n$, $n = 1, 2, \ldots, N$.*

Definition 2. *The* n-mode vectors *of $\mathcal{A}$ are the I_n-dimensional vectors obtained from $\mathcal{A}$ by fixing every index but index i_n.*

Definition 3. *The* n-mode unfolding matrix *of $\mathcal{A}$, denoted by $(\mathcal{A})_{(n)} \in \mathbb{R}^{I_n \times (I_1 \ldots I_{n-1} I_{n+1} \ldots \times I_N)}$, contains the element $a_{i_1 \ldots i_N}$ at i_nth row and at jth column, where*

$$j = 1 + \sum_{k=1, k \neq n}^{N} (i_k - 1) J_k, \qquad \textit{with} \qquad J_k = \prod_{m=1, m \neq n}^{k-1} I_m. \tag{1}$$

We can generalize the product of two matrices to the product of a tensor and a matrix.

Definition 4. *The* n-mode product *of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{U}$, is an $(I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N)$-tensor of which the entries are given by:*

$$(\mathcal{A} \times_n \mathbf{U})_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} \stackrel{def}{=} \sum_{i_n} a_{i_1 i_2 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} u_{j_n i_n}. \tag{2}$$

Definition 5. *The* scalar product *of two tensors* $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$*, denoted by* $\langle \mathcal{A}, \mathcal{B} \rangle$*, is defined in a straightforward way as* $\langle \mathcal{A}, \mathcal{B} \rangle \stackrel{def}{=} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_N} a_{i_1 i_2 \ldots 1_N} b_{i_1 i_2 \ldots 1_N}$*. The* Frobenius norm *of a tensor* $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ *is then defined as* $\|\mathcal{A}\|_F \stackrel{def}{=} \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle}$

Form the definition of the n-mode unfolding matrix, we have

$$\|\mathcal{A}\|_F = \|(\mathbf{A})_{(n)}\|_F \tag{3}$$

By using tensor decomposition, any tensor $\mathcal{A}$ can be expressed as the product

$$\mathcal{A} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_N \mathbf{U}_N \tag{4}$$

where $\mathbf{U}_n$, $n = 1, 2, \ldots, N$, is an orthonormal matrix and contains the ordered principal components for the nth mode. $\mathcal{C}$ is called the *core tensor*. Unfolding the above equation, we have

$$\mathbf{A}_{(n)} = \mathbf{U}_n \mathbf{C}_{(n)} (\mathbf{U}_N \otimes \ldots \otimes \mathbf{U}_{n+1} \otimes \mathbf{U}_{n-1} \otimes \ldots \otimes \mathbf{U}_1)^T \tag{5}$$

where operator $\otimes$ is the Kronecker product of the matrices.

III. Overview of Tensor discriminant color space model

In this section, we overview the Tensor Discriminant Color Space (TDCS) model. In TDCS, a color image is naturally represented by a 3rd-order tensor. The 1-mode of a tensor is the height of an image, the 2-mode of a tensor is the width of an image and the 3-mode of tensor is the color space of an image. For instance, a RGB image with size $I_1 \times I_2$ is represented as a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, where $I_3 = 3$. The 3-mode of $\mathcal{A}$ is the color variable in the RGB color space which has 3 components corresponding to $\mathbf{R}$, $\mathbf{G}$ and $\mathbf{B}$ in RGB space.

Assuming C is the number of individuals, $\mathcal{A}_i^c$ is the ith color face image of cth individual, and M_c is the number of color face images of cth individual, where $M = M_1 + M_2 + \ldots + M_C$. the TDCS algorithm seeks two discriminant projection matrices $\mathbf{U}_1 \in \mathbb{R}^{I_1 \times L_1}$, $\mathbf{U}_2 \in \mathbb{R}^{I_2 \times L_2}$ and a color space transformation matrix $\mathbf{U}_3 \in \mathbb{R}^{I_3 \times L_3}$ (usually $L_1 < I_1$, $L_2 < I_2$ and $L_3 \leq I_3$) for transformation

$$\begin{aligned} &\mathcal{D}_i^c = \mathcal{A}_i^c \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \times_3 \mathbf{U}_3^T, \\ &i = 1, 2, \ldots, M_c, \quad c = 1, 2, \ldots, C. \end{aligned} \tag{6}$$

which ensures that the projected tensors from the same individual are distributed as close as possible, while the projected tensors from different individuals are distributed as far as possible.

The mean image of the cth individual is defined by:

$$\overline{\mathcal{A}}^c = \frac{1}{M_c} \sum_{i=1}^{M_c} \mathcal{A}_i^c \tag{7}$$

and the mean image of all individuals is defined by:

$$\overline{\mathcal{A}} = \frac{1}{C} \sum_{c=1}^{C} \overline{\mathcal{A}}^c \tag{8}$$

The between-class scatter of color images is defined as:

$$\Psi_b(\mathcal{A}) = \sum_{c=1}^{C} \|\overline{\mathcal{A}}^c - \overline{\mathcal{A}}\|_F^2 \tag{9}$$

and within-class scatter of color images is defined as:

$$\Psi_w(\mathcal{A}) = \sum_{c=1}^{C} \sum_{i=1}^{M_c} \|\mathcal{A}_i^c - \overline{\mathcal{A}}^c\|_F^2 \tag{10}$$

A reasonable idea is to maximize the between-class scatter of projected tensors $\Psi_b(\mathcal{D})$ and to minimize the within-class scatter of projected tensors $\Psi_w(\mathcal{D})$. Then TDCS criterion can be defined as follows:

$$\max J(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3) = \frac{\Psi_b(\mathcal{D})}{\Psi_w(\mathcal{D})} \tag{11}$$

Here, three matrices $\mathbf{U}_n$ need to be simultaneously updated for achieving the optimal solution of the criterion function J. We can define n-mode between-class scatter matrix $\mathbf{S}_b^{(n)}$ and n-mode within-class scatter matrix $\mathbf{S}_w^{(n)}$ as:

$$\mathbf{S}_b^{(n)} = \sum_{c=1}^{C} \left(\overline{\mathbf{A}}_{(n)}^c - \overline{\mathbf{A}}_{(n)} \right) \tilde{\mathbf{U}}_n \tilde{\mathbf{U}}_n^T \left(\overline{\mathbf{A}}_{(n)}^c - \overline{\mathbf{A}}_{(n)} \right)^T \tag{12}$$

and

$$\mathbf{S}_w^{(n)} = \sum_{c=1}^{C} \sum_{i=1}^{M_c} \left(\mathbf{A}_{i(n)}^c - \overline{\mathbf{A}}_{(n)}^c \right) \tilde{\mathbf{U}}_n \tilde{\mathbf{U}}_n^T \left(\mathbf{A}_{i(n)}^c - \overline{\mathbf{A}}_{(n)}^c \right)^T \tag{13}$$

where $\tilde{\mathbf{U}}_n = \mathbf{U}_N \otimes \ldots \otimes \mathbf{U}_{n+1} \otimes \mathbf{U}_{n-1} \otimes \ldots \otimes \mathbf{U}_1$, $n = 1, 2, \ldots, N$ and $N = 3$.

Then, the between-class scatter of the projected tensors $\Psi_b(\mathcal{D})$ and the within-class scatter of the projected tensors $\Psi_w(\mathcal{D})$ can be rewritten as follows:

$$\Psi_b(\mathcal{D}) = tr\left(\mathbf{U}_n^T \mathbf{S}_b^{(n)} \mathbf{U}_n \right) \tag{14}$$

and

$$\Psi_w(\mathcal{D}) = tr\left(\mathbf{U}_n^T \mathbf{S}_w^{(n)} \mathbf{U}_n \right) \tag{15}$$

So, given all other projection matrices $\mathbf{U}_1, \ldots, \mathbf{U}_{n-1}, \mathbf{U}_{n+1}, \ldots, \mathbf{U}_N$, the TDCS criterion can be written as follow:

$$\max \frac{tr\left(\mathbf{U}_n^T \mathbf{S}_b^{(n)} \mathbf{U}_n \right)}{tr\left(\mathbf{U}_n^T \mathbf{S}_w^{(n)} \mathbf{U}_n \right)} \tag{16}$$

According to Rayleigh quotient, Eq. (16) is maximized if and only if the matrix $\mathbf{U}_n$ consists of L_n generalized eigenvectors, which are corresponding to the largest L_n generalized eigenvalues of the matrix pencil $(\mathbf{S}_b^{(n)}, \mathbf{S}_w^{(n)})$, which satisfies:

$$\mathbf{S}_b^{(n)} \mathbf{v} = \lambda \mathbf{S}_w^{(n)} \mathbf{v} \tag{17}$$

Since $\mathbf{S}_b^{(n)}$ and $\mathbf{S}_w^{(n)}$ are dependent on $\mathbf{U}_1, \ldots, \mathbf{U}_{n-1}, \mathbf{U}_{n+1}, \ldots, \mathbf{U}_N$, we can see that the optimization of $\mathbf{U}_n$ depends on the projections of other modes. An iterative procedure can be constructed to maximize Eq. (11). For details, please refer to our previous work [29].

IV. Sparse tensor discriminant color space model

In this section, we discuss how to model the Sparse Tensor Discriminant Color Space (STDCS). The same symbols described and defined in Section III are re-used. The aim of STDCS is not only to maximize Eq. (11) but also to make three matrices $\mathbf{U}_n$ sparse. Here, sparsity means that $\mathbf{U}_n$ has only a small number of nonzero elements or it has lots of zero elements. Therefore, the criterion function of STDCS is defined as:

$$\begin{aligned} &\max J(\mathbf{U}_1,\mathbf{U}_2,\mathbf{U}_3)=\frac{\Psi_b(\mathcal{D})}{\Psi_w(\mathcal{D})} \\ &\text{subject to} \quad Card(\mathbf{U}_n)<K_n, \quad n=1,2,3 \end{aligned} \tag{18}$$

where $Card(\cdot)$ denotes the number of nonzero elements of $\mathbf{U}_n$. The only difference between Eq. (18) and Eq. (11) is a sparseness constraint imposed in Eq. (18). We convert the generalized eigenvalue problem (16) to a regression problem and then apply penalized least squares with an ℓ_1 penalty. We combine all color face images $\mathcal{A}_i^c \in \mathbb{R}^{I_1\times I_2\times 3}$ into a fourth-order tensor $\mathcal{A}\in\mathbb{R}^{I_1\times I_2\times 3\times M}$. To facilitate the subsequent discussion, two tensors $\mathcal{H}_w\in\mathbb{R}^{I_1\times I_2\times 3\times M}$ and $\mathcal{H}_b\in\mathbb{R}^{I_1\times I_2\times 3\times M}$ are introduced as follows:

$$\mathcal{H}_w=\left\{\begin{array}{c}(\mathcal{A}_1^1-\overline{\mathcal{A}}^1)\\ \vdots\\ (\mathcal{A}_{M_1}^1-\overline{\mathcal{A}}^1)\\ (\mathcal{A}_1^2-\overline{\mathcal{A}}^2)\\ \vdots\\ (\mathcal{A}_{M_2}^2-\overline{\mathcal{A}}^2)\\ \vdots\\ (\mathcal{A}_1^C-\overline{\mathcal{A}}^C)\\ \vdots\\ (\mathcal{A}_{M_C}^C-\overline{\mathcal{A}}^C)\end{array}\right\} \quad\text{and}\quad \mathcal{H}_b=\left\{\begin{array}{c}(\overline{\mathcal{A}}^1-\overline{\mathcal{A}})\\ \vdots\\ (\overline{\mathcal{A}}^1-\overline{\mathcal{A}})\\ (\overline{\mathcal{A}}^2-\overline{\mathcal{A}})\\ \vdots\\ (\overline{\mathcal{A}}^2-\overline{\mathcal{A}})\\ \vdots\\ (\overline{\mathcal{A}}^C-\overline{\mathcal{A}})\\ \vdots\\ (\overline{\mathcal{A}}^C-\overline{\mathcal{A}})\end{array}\right\} \tag{19}$$

where $\{\cdot\}$ denotes the combination of M Nth order tensors into a $(N+1)$th order tensor. Further, $\mathcal{H}_b$ can be reduced to a lower dimension tensor $\mathcal{H}_b'\in\mathbb{R}^{I_1\times I_2\times 3\times C}$ according to:

$$\mathcal{H}_b'=\left\{\begin{array}{c}(\overline{\mathcal{A}}^1-\overline{\mathcal{A}})\\ (\overline{\mathcal{A}}^2-\overline{\mathcal{A}})\\ \vdots\\ (\overline{\mathcal{A}}^C-\overline{\mathcal{A}})\end{array}\right\} \tag{20}$$

Theorem 1. *Given $N-1$ projection matrices $\mathbf{U}_1,\ldots,\mathbf{U}_{n-1},\mathbf{U}_{n+1},\ldots,\mathbf{U}_N$, let*

$$\mathcal{G}_w=\mathcal{H}_w\times_1\mathbf{U}_1\ldots\times_{n-1}\mathbf{U}_{n-1}\times_{n+1}\mathbf{U}_{n+1}\ldots\times_N\mathbf{U}_N, \tag{21}$$

$$\mathcal{G}_b=\mathcal{H}_b\times_1\mathbf{U}_1\ldots\times_{n-1}\mathbf{U}_{n-1}\times_{n+1}\mathbf{U}_{n+1}\ldots\times_N\mathbf{U}_N \tag{22}$$

and

$$\mathcal{G}_b'=\mathcal{H}_b'\times_1\mathbf{U}_1\ldots\times_{n-1}\mathbf{U}_{n-1}\times_{n+1}\mathbf{U}_{n+1}\ldots\times_N\mathbf{U}_N \tag{23}$$

Then, $\mathbf{S}_w^{(n)}=\mathbf{G}_{w(n)}\mathbf{G}_{w(n)}^T$, $\mathbf{S}_b^{(n)}=\mathbf{G}_{b(n)}\mathbf{G}_{b(n)}^T$ and $\mathbf{S}_b^{(n)}=\mathbf{G}_{b(n)}'\mathbf{G}_{b(n)}'^T$

Proof: Apply n-mode unfolding on Eq. (22):

$$\mathbf{G}_{b(n)}'=(\mathcal{H}_b\times_1\mathbf{U}_1\ldots\times_{n-1}\mathbf{U}_{n-1}\times_{n+1}\mathbf{U}_{n+1}\ldots\times_N\mathbf{U}_N)_{(n)} \tag{24}$$

For c-th class,

$$\begin{aligned}&((\overline{\mathcal{A}}^c-\overline{\mathcal{A}})\times_1\mathbf{U}_1\ldots\times_{n-1}\mathbf{U}_{n-1}\times_{n+1}\mathbf{U}_{n+1}\ldots\times_N\mathbf{U}_N)_{(n)}\\ &=(\overline{\mathcal{A}^c}-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n\end{aligned} \tag{25}$$

Obviously, $\mathbf{G}_{b(n)}'$ can be rewritten as:

$$\mathbf{G}_{b(n)}'=[(\overline{\mathcal{A}}^1-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n,(\overline{\mathcal{A}}^2-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n,\ldots,(\overline{\mathcal{A}}^C-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n] \tag{26}$$

So

$$\begin{aligned}&\mathbf{G}_{b(n)}'\mathbf{G}_{b(n)}'^T=\\ &[(\overline{\mathcal{A}}^1-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n,(\overline{\mathcal{A}}^2-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n,\ldots,(\overline{\mathcal{A}}^C-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n]\\ &\cdot\left[\begin{array}{c}\widetilde{\mathbf{U}}_n^T\cdot(\overline{\mathcal{A}}^1-\overline{\mathcal{A}})_{(n)}^T\\ \widetilde{\mathbf{U}}_n^T\cdot(\overline{\mathcal{A}}^2-\overline{\mathcal{A}})_{(n)}^T\\ \vdots\\ \widetilde{\mathbf{U}}_n^T\cdot(\overline{\mathcal{A}}^C-\overline{\mathcal{A}})_{(n)}^T\end{array}\right]\\ &=\sum_{c=1}^{C}(\overline{\mathcal{A}}^c-\overline{\mathcal{A}})_{(n)}\cdot\widetilde{\mathbf{U}}_n\cdot\widetilde{\mathbf{U}}_n^T\cdot(\overline{\mathcal{A}}^c-\overline{\mathcal{A}})_{(n)}^T\\ &=\sum_{c=1}^{C}(\overline{\mathcal{A}}_{(n)}^c-\overline{\mathcal{A}}_{(n)})\cdot\widetilde{\mathbf{U}}_n\cdot\widetilde{\mathbf{U}}_n^T\cdot(\overline{\mathcal{A}}_{(n)}^c-\overline{\mathcal{A}}_{(n)})^T\\ &=\mathbf{S}_b^{(n)}\end{aligned} \tag{27}$$

Similarly, we can prove $\mathbf{S}_w^{(n)}=\mathbf{G}_{w(n)}\mathbf{G}_{w(n)}^T$ and $\mathbf{S}_b^{(n)}=\mathbf{G}_{b(n)}\mathbf{G}_{b(n)}^T$. ■

Theorem 2. *$\mathbf{S}_w^{(n)}$ is positive definite and its Cholesky decomposition is denoted as $\mathbf{S}_w^{(n)}=\mathbf{R}_{w(n)}^T\mathbf{R}_{w(n)}$, where $\mathbf{R}_{w(n)}\in\mathbb{R}^{I_n\times I_n}$ is an upper triangular matrix. $\mathbf{v}_1,\ldots,\mathbf{v}_{L_n}$ are eigenvectors of Eq. (17) which correspond to the L_n largest eigenvalues. $\mathbf{A}=[\alpha_1,\ldots,\alpha_{L_n}]$ and $\mathbf{B}=[\beta_1,\ldots,\beta_{L_n}]$ ($\mathbf{A}\in\mathbb{R}^{I_n\times L_n}$). For $\lambda>0$, then $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are the solutions of the following problem:*

$$\begin{aligned}&\min_{\mathbf{A},\mathbf{B}}\sum_{i=1}^{\widetilde{I}_n}\|\mathbf{R}_{w(n)}^{-T}\mathbf{g}_i-\mathbf{A}\mathbf{B}^T\mathbf{g}_i\|^2-\lambda\sum_{j=1}^{L_n}\beta_j^T\mathbf{S}_w^{(n)}\beta_j\\ &\text{subject to}\quad \mathbf{A}^T\mathbf{A}=\mathbf{I}\end{aligned} \tag{28}$$

where $\widetilde{I}_n=I_1\times\ldots,\times I_{n-1}\times I(n+1)\times\ldots\times I_N\times M$ and $\mathbf{g}_i$ is the ith row of $\mathbf{G}_{b(n)}$. Then $\hat{\beta}_j$ spans the same linear space as $\mathbf{v}_j$, where $j=1,\ldots,\boldsymbol{L}_n$

Proof: The proof is similar to Theorem 1 in [35]. ■

According to Theorem 2, the generalized eigenvalue of Eq. (17) is transformed to the regression problem of Eq. (28). The regression problem (28) has two variables $\mathbf{A},\mathbf{B}$ that need to be optimized simultaneously. It can be solved by iteratively minimizing over $\mathbf{A}$ and $\mathbf{B}$.

Given a fixed $\mathbf{B}$, the second term in Eq (28) is constant. Its first term can be rewritten as:

$$\begin{aligned}
&\sum_{i=1}^{\widetilde{I}_n} \|\mathbf{R}_{w(n)}^{-T}\mathbf{g}_i - \mathbf{A}\mathbf{B}^T\mathbf{g}_i\|^2 \\
&= \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1} - \mathbf{G}_{b(n)}\mathbf{B}\mathbf{A}^T\|^2 \\
&= tr[(\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1} - \mathbf{G}_{b(n)}\mathbf{B}\mathbf{A}^T)(\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1} - \mathbf{G}_{b(n)}\mathbf{B}\mathbf{A}^T)^T] \\
&= tr(\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\mathbf{R}_{w(n)}^{-T}\mathbf{G}_{b(n)}^T + \mathbf{G}_{b(n)}\mathbf{B}\mathbf{B}^T\mathbf{G}_{b(n)}^T) \\
&- 2tr(\mathbf{B}^T\mathbf{G}_{b(n)}^T\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\mathbf{A})
\end{aligned} \tag{29}$$

The second term in this equation is constant. Therefore, we only need to maximize the $tr(\mathbf{B}^T\mathbf{G}_{b(n)}^T\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\mathbf{A})$ subject to the constraint $\mathbf{A}^T\mathbf{A} = \mathbf{I}$. The solution is obtained by computing the singular value decomposition:

$$\mathbf{R}_{w(n)}^{-T}(\mathbf{G}_{b(n)}^T\mathbf{G}_{b(n)})\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{30}$$

Given a fixed $\mathbf{B}$, the solution of Eq. (28) is $\hat{\mathbf{A}} = \mathbf{U}\mathbf{V}^T$ according to Theorem 4 in [30].

Given a fixed $\mathbf{A}$, let $\mathbf{A}_\perp$ be an orthogonal matrix such that $[\mathbf{A}; \mathbf{A}_\perp]$ is $I_n \times I_n$ orthogonal, where $[\mathbf{A}; \mathbf{A}_\perp]$ means to concatenate matrices $\mathbf{A}$ and $\mathbf{A}_\perp$ along rows. This is feasible since $\mathbf{A}$ has orthonormal columns. Thus the first term in Eq. (28) can be rewritten as following:

$$\begin{aligned}
&\sum_{i=1}^{\widetilde{I}_n} \|\mathbf{R}_{w(n)}^{-T}\mathbf{g}_i - \mathbf{A}\mathbf{B}^T\mathbf{g}_i\|^2 \\
&= \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1} - \mathbf{G}_{b(n)}\mathbf{B}\mathbf{A}^T\|^2 \\
&= \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}[\mathbf{A}; \mathbf{A}_\perp] - \mathbf{G}_{b(n)}\mathbf{B}\mathbf{A}^T[\mathbf{A}; \mathbf{A}_\perp]\|^2 \\
&= \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\mathbf{A} - \mathbf{G}_{b(n)}\mathbf{B}\|^2 + \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\mathbf{A}_\perp\|^2 \\
&= \sum_{j=1}^{L_n} \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\alpha_j - \mathbf{G}_{b(n)}\beta_j\|^2 + \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\mathbf{A}_\perp\|^2
\end{aligned} \tag{31}$$

Given a fixed $\mathbf{A}$, therefore, the solution of Eq (28) is $\mathbf{B} = [\beta_1, \ldots, \beta_{L_n}]$, where β_j can be obtained by solving the following ridge regression problem:

$$\min_{\beta_j} \|\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\alpha_j - \mathbf{G}_{b(n)}\beta_j\|^2 + \lambda\beta_j^T\mathbf{S}_w^{(n)}\beta \tag{32}$$

Here, L_n ridge regression problems are independent to each other. In summary, we transform the generalized eigenvalue problem (17) to L_n ridge regression problems (32). However,the ridge regression does not provide a sparse solution. In order to get the sparse solutions, one can use lasso regression [36] on β_j by using ℓ_1 norm. Fig. 1 shows that the sparse solution can be found by solving a ℓ_1 norm problem but not by solving a traditional ℓ_2 norm problem.

We denote $\widetilde{\mathbf{y}}_j = (\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\alpha_j, 0_{1\times I_n})^T$ and $\widetilde{\mathbf{W}} = (\mathbf{G}_{b(n)}^T, \sqrt{\lambda}\mathbf{R}_{w(n)}^T)^T$, the ridge regression problems (32) are equal to the following lasso regression problems:

$$\min_{\beta_j} \|\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{W}}\beta_j\|^2 + \lambda_1\|\beta_j\|_1 \tag{33}$$

where λ_1 is the ℓ_1 norm tuning parameter. When λ_1 is large enough, some elements in β_j will be shrunk to zero. However,

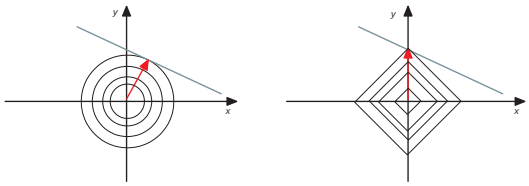


Fig. 1. the sparse solution can be found by solving a ℓ_1 norm problem but not by solving a traditional ℓ_2 norm problem. The x component of solution of ℓ_1 norm is zero.

the lasso has several shortages as pointed out in [32]. For instance, the number of extracted features by the lasso is limited by the number of samples. To address the shortage, the elastic net [32] generalizes the lasso by combining both the ℓ_1 norm and ℓ_2 norm as the penalty. The lasso regression problems (33) can be written as the following elastic net problems:

$$\min_{\beta_j} \|\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{W}}\beta_j\|^2 + \lambda_{1,j}\|\beta_j\|_1 + \lambda_2\|\beta_j\|^2 \tag{34}$$

where λ_2 is the ℓ_2 norm tuning parameter. When $\lambda_2 = 0$, elastic net is degraded to lasso regression.

Now, given all other sparse projection matrices $\mathbf{U}_1, \ldots, \mathbf{U}_{n-1}, \mathbf{U}_{n+1}, \ldots, \mathbf{U}_N$, the sparse solution $\mathbf{U}_n$ can be obtained by solving the L_n elastic net problems (34). Since the $\mathbf{G}_{b(n)}$ and $\mathbf{R}_{w(n)}$ depend on $\mathbf{U}_1, \ldots, \mathbf{U}_{n-1}, \mathbf{U}_{n+1}, \ldots, \mathbf{U}_N$, it can be seen that the optimization of $\mathbf{U}_n$ depends on the projections in other modes. An iterative procedure can be constructed to maximize Eq. (18). The pseudocode of the proposed method is summarized in Algorithm 1.

V. Experiments and Results

A. Database

We conducted the experiments on three well-known color face databases: AR[37], Georgia Tech face databases[1] and Labeled Faces in the Wild (LFW) face databases[2].

The AR database contains over 4,000 color face images of 126 people (70 male and 56 female). A subset of 100 people (50 male and 50 female) were selected in our experiment. The selected images were frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). All images were taken under strictly controlled conditions. No restrictions on wear (clothes, glasses, etc.), make-up, hair style, etc. were imposed to participants. The same pictures were taken for each individual in two separate sessions which were apart for two weeks. 26 images of each individual were selected in our experiment. All images were cropped into 32×32 pixels. The sample images of one individual from the AR database are shown in Fig. 2, where Fig. 2(a)-Fig. 2(m) are from the first session used as the training set, and Fig. 2(n)-Fig. 2(z) are from the second session used for testing purpose.

Georgia Tech face database contains images of 50 individuals which were taken in two or three sessions at different times. For each individual in this database, 15 color JPEG images

[1] http://www.anefian.com/research/face_reco.htm
[2] http://vis-www.cs.umass.edu/lfw/

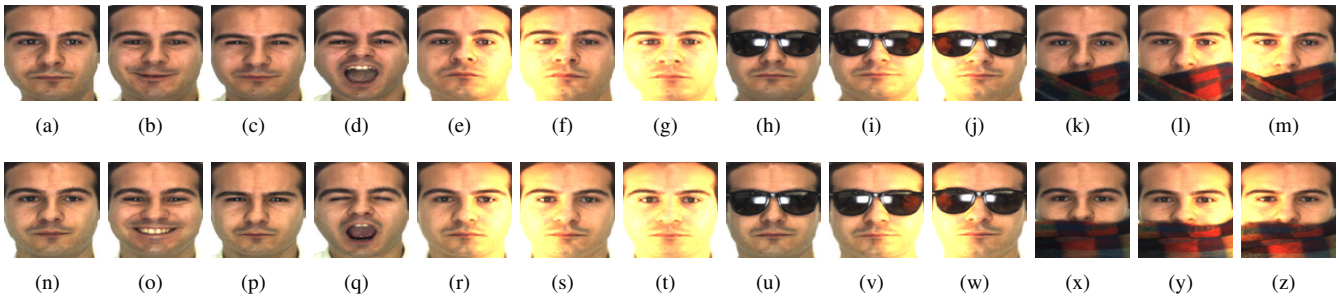


Fig. 2. Sample images of one individual from the AR database.

Algorithm 1 STDCS

INPUT: a set of M labeled tensor samples $\mathcal{A}_i^c$, $i = 1, 2, \ldots, M_c$, $c = 1, 2, \ldots, C$, the number of reduced dimensions L_n, $n = 1, 2, 3$ and sparse tuning parameters $\lambda, \lambda_1, \lambda_2$.

OUTPUT: a set of projected tensors $\mathcal{D}_i^c$, two sparse discriminant projection matrices $\mathbf{U}_1 \in \mathbb{R}^{I_1 \times L_1}$, $\mathbf{U}_2 \in \mathbb{R}^{I_2 \times L_2}$ and a sparse color space transformation matrix $\mathbf{U}_3 \in \mathbb{R}^{I_3 \times L_3}$

Algorithm:

Initialize $\mathbf{U}_n$ with a set of identity matrices;
Calculate the mean image of the cth individual $\overline{\mathcal{A}}^c$ and the mean image of all individuals $\overline{\mathcal{A}}$ by Eq. (7) and Eq. (8);
Calculate $\mathcal{H}_w$ and $\mathcal{H}_b$ by using Eq. (19);
repeat
  for $n = 1$ to 3 **do**
    Calculate $\mathcal{G}_w$ and $\mathcal{G}_b$ by Eq. (21) and Eq. (22);
    Calculate the mode-n unfolding $\mathbf{G}_{b(n)}$ and $\mathbf{G}_{w(n)}$;
    Calculate the upper triangular matrix $\mathbf{R}_{w(n)} \in \mathbb{R}^{I_n \times I_n}$ from the Cholesky decomposition of $\mathbf{G}_{w(n)}^T \mathbf{G}_{w(n)}$ such that $\mathbf{G}_{w(n)}^T \mathbf{G}_{w(n)} = \mathbf{R}_{w(n)}^T \mathbf{R}_{w(n)}$;
    Initialize $\mathbf{A}$ as an identity matrix;
    Calculate $\widetilde{\mathbf{W}} = (\mathbf{G}_{b(n)}^T, \sqrt{\lambda}\mathbf{R}_{w(n)}^T)^T$;
    repeat
      for $j = 1$ to L_n **do**
        Calculate $\widetilde{\mathbf{y}}_j = (\mathbf{G}_{b(n)}\mathbf{R}_{w(n)}^{-1}\alpha_j, 0_{1\times I_n})^T$;
        Solve $\min_{\beta_j} \|\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{W}}\beta_j\|^2 + \lambda_{1,j}\|\beta_j\|_1 + \lambda_2\|\beta_j\|^2$ by using the elastic net;
      end for
      Calculate $\mathbf{B} = [\beta_1, \ldots, \beta_{L_n}]$;
      Calculate $\mathbf{R}_{w(n)}^{-T}(\mathbf{G}_{b(n)}^T \mathbf{G}_{b(n)})\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ by using SVD;
      Calculate $\mathbf{A} = \mathbf{U}\mathbf{V}^T$;
    until $norm(\mathbf{B}_{k'+1} - \mathbf{B}_{k'}) < \epsilon_1$
  end for
  Calculate $\mathbf{U}_n = \mathbf{B}$;
  Calculate J_{k+1} by Eq. (11);
until $|J_{k+1} - J_k| < \epsilon$
Compute a set of projected tensors $\mathcal{D}_i^c$ by Eq. (6);

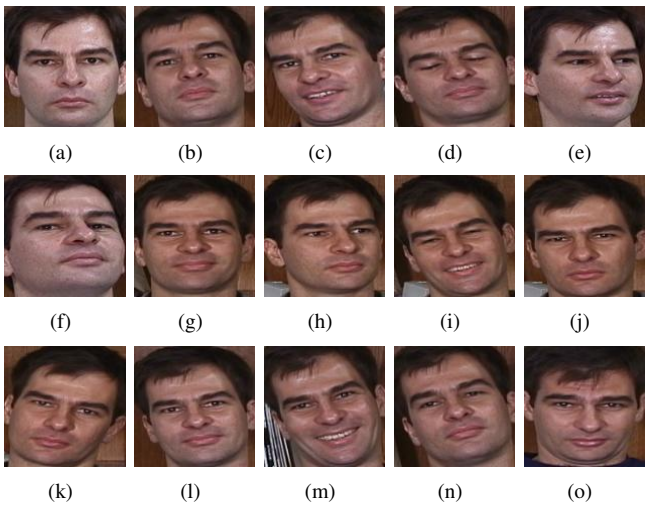


Fig. 3. Sample images of one individual from the Georgia Tech database (unaligned head images).

were captured with clutter backgrounds. The resolution of these images was 640×480 pixels and the size of face within these images was around 150×150. Faces illustrated in these images may be frontal and/or tilted with different expressions, illuminations and scales. Each image was manually cropped and resized to 32×32 pixels. The sample images of one individual from the Georgia Tech database are showed in Fig. 3.

LFW database is designed for studying the problem of unconstrained face recognition. It contains more than 13,000 images of faces collected from the web. Each face has been labeled with the name of the person pictured. 1680 of the people pictured have two or more distinct photos in the data set. The only constraint on these faces is that they were detected by the Viola-Jones face detector. In our experiments, we choose 1,251 images from 86 people pictured have 11-20 images. Each image was manually cropped and resized to 32×32 pixels. The sample images of one individual from the LFW database are showed in Fig. 4.

B. Experiment setting

For the purpose of evaluating the performance of STDCS, we used *face verification rate* as the criteria. The FERET Verification Testing Protocol [38] recommends using the Receiver Operating Characteristic (ROC) curves to depict the relations between the Face Verification Rate (FVR) and the

Fig. 4. Sample images of one individual from the LFW database.

False Accept Rate (FAR). In order to get better performance, in our experiments, the score matrices were generated by the Manhattan distance and Euclidean distance, respectively. The ROC curves were plotted by using the Statistical Learning Toolbox[3] according to the obtained score matrix. For tensor operations, we used the tensor toolbox developed by Bader and Kolda in MATLAB^{TM}[39]. All experiments on the Microsoft Windows XP 64-bits version machine with 2.66 GHz Intel CPU and 16 GB memory.

In our experiments, we will discuss the robustness of algorithms from the following two aspects:

- The performance of algorithms is insensitive to the measures (Manhattan distance and Euclidean distance).
- The performance of algorithms is insensitive to the aligned faces, occluded faces and noised faces.

C. Experiments and results on the AR database

In this experiment, we trained STDCS, TDCS and CID by using 7 un-occluded color face images of each individual from the first session in AR database and tested them using the corresponding images in the second session. The convergence threshold ϵ was set as 0.1 and $\mathbf{x}_1$ was initialized as $[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T$. In STDCS, we set $\lambda = 1000$ and $\lambda_2 = 10^{-6}$. For the parameter λ_1, we used another strategy to tune the sparseness. The number of non-zero elements in each column of three sparse projection matrices were 10, 10 and 2. Meanwhile, we carried out LDA and 2D-LDA on corresponding gray images. Because there were 100 individuals in the AR database, only 99 discriminant projection basis vectors were extracted in LDA and CID. For 2D-LDA, TDCS and STDCS, the two numbers of the reduced spatial dimensions both are 10. The score matrices were generated by Manhattan distance and Euclidean distance, respectively. The ROC curves of the five methods are shown in Fig. 5. The results indicate that the performance of TDCS with Euclidean distance is slightly better than that of

[3]The **slverifyroc** function in the Statistical Learning Toolbox can only plot the ROC curve illustrating the relations of the false reject rate versus the FAR. We modified it to depict the relations between the FVR and the FAR.

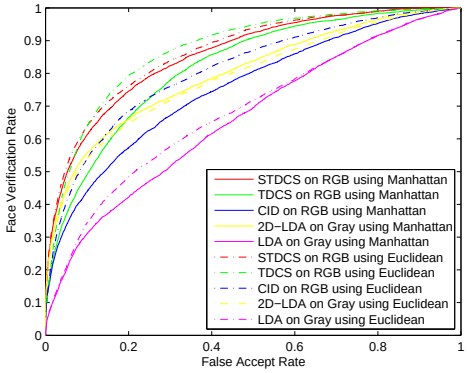


Fig. 5. ROC curves of STDCS, TDCS, CID, 2D-LDA and LDA on the un-occluded facial images of AR database.

STDCS with Manhattan distance. However, the space between two curves of STDCS is narrower than the space between two curves of TDCS. This shows that STDCS is more robust to the measures than TDCS. The curves in Fig. 5 also show that the 20 discriminant projection basis vectors in STDCS and TDCS contain more discriminant information than 99 ones in CID. It is derived from the fact that some discriminant information is thrown away in the PCA step of CID model.

In order to investigate the robustness to noise, we used all images including the occluded facial images in the first session to train the models. All images in the second session were used for testing. In this case, we got three color space transformation matrices:

$$\mathbf{X} = \begin{bmatrix} 0.4126 & -0.2107 & -0.5558 \\ -0.0261 & -0.4683 & 1.0536 \\ 1.0000 & 0.9739 & -0.5524 \end{bmatrix}, \tag{35}$$

$$\mathbf{U}_3 = \begin{bmatrix} 0.1267 & -0.2084 & 0.3358 \\ -0.2128 & -0.4168 & -0.7897 \\ 0.9689 & 0.8848 & 0.5134 \end{bmatrix} \tag{36}$$

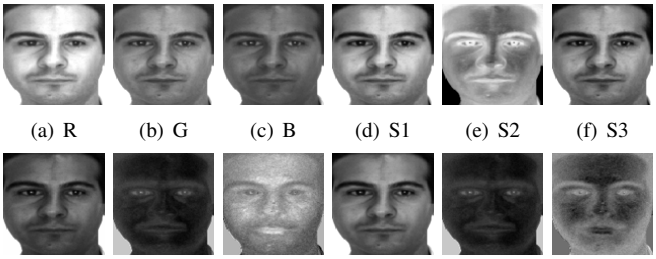
and

$$\mathbf{U}_3^{sparse} = \begin{bmatrix} 0.8270 & 0 & 0.2287 \\ 0.5622 & -0.9975 & 0.9735 \\ 0 & -0.0705 & 0 \end{bmatrix} \tag{37}$$

There are two non-zero elements in each column of $\mathbf{U}_3^{sparse}$. Using these three matrices, we got three color components $\mathbf{D}^1$, $\mathbf{D}^2$, $\mathbf{D}^3$ of CID; three color components $\mathbf{T}^1$, $\mathbf{T}^2$, $\mathbf{T}^3$ of TDCS and three color components $\mathbf{S}^1$, $\mathbf{S}^2$, $\mathbf{S}^3$ of STDCS. These components are illustrated in Fig. 8. Compared to CID color space and TDCS color space, STDCS color space is more intuitionistic, i.e. the color component images of STDCS look more like real faces. Although $\mathbf{S}^1$ and $\mathbf{S}^3$ are similar to RGB space, the influence of light on R component in RGB is decreased in STDCS by the linear combination of R and G components. The ROC curves are illustrated in Fig. 9, where STDCS with Manhattan distance obtains the best performance. Comparing with Fig. 5, STDCS obtains the best performance for the occluded facial images. This also verify the STDCS robustness to the occluded facial images.

In order to further investigate the robustness to noise, some images were randomly selected and occluded with a

Fig. 6. noise face images on the AR database.



(a) R (b) G (c) B (d) S1 (e) S2 (f) S3

(g) T1 (h) T2 (i) T3 (j) D1 (k) D2 (l) D3

Fig. 8. Illustration of R, G, and B color components and the various components generated by CID, TDCS and STDCS on the AR face database.

rectangular noise consisting of random black and white dots whose size was at least 128 pixels, located at a random position. The manner of forming rectangle noise is similar to that in [40]. Fig. 6 shows typical examples of noised images. Fig. 7(a) shows the ROC curves of STDCS and TDCS by adding noise to 20%, 40%, 60% and 80% samples of both the training set and the testing set. Fig. 7(b) and Fig. 7(c) show the ROC curves of the two algorithms by adding noise on the training set and the testing set, respectively. From the figures, the performances of STDCS with 80% noised samples are better than those of TDCS with 20% noised samples in the three cases. It is interesting that the shapes of ROC curves are changed in the case of adding noise on the testing set.

For the intuition, we enhanced the sparse constraint and got a sparse color sparse color transformation matrix as following:

$$\mathbf{U}_3^{sparse'} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{38}$$

Intuitively, the effect of B component differs from the other two. In order to verify this difference, 2D-LDA is implemented on R, G and B component images, respectively. In Fig. 10, three ROC curves are illustrated from which we can see that the curves of R and G components are similar, while the curve of B component is quite different.

D. Experiments and results on the Georgia Tech face database

Georgia Tech face database is more complex than AR database, because it contains various pose faces with different expressions on cluttered background. In this experiment, we used the first 8 images of each individual as the training set and the remaining images as the testing set. The CID, TDCS and STDCS models were trained and we got three color space transformation matrices:

$$\mathbf{X} = \begin{bmatrix} -1.0000 & 0.4894 & 0.4076 \\ 0.8473 & 0.3595 & -1.0134 \\ -0.2767 & -1.0401 & 0.5332 \end{bmatrix}, \tag{39}$$

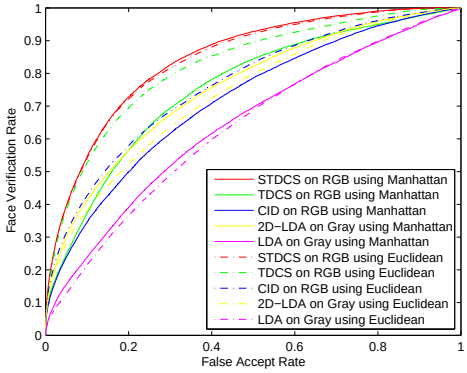


Fig. 9. ROC curves of STDCS, TDCS, CID, 2D-LDA and LDA on AR face database.

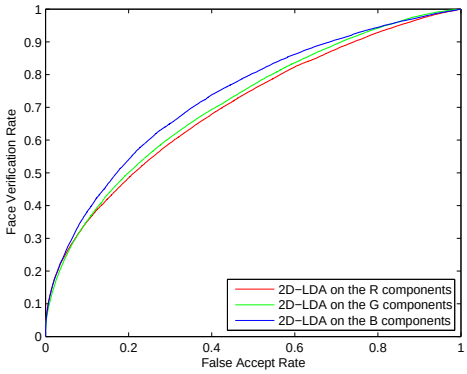


Fig. 10. ROC curves of 2D-LDA on R, G and B components of AR face database.

$$\mathbf{U}_3 = \begin{bmatrix} 0.1067 & -0.3004 & 0.6192 \\ 0.7589 & 0.7798 & -0.7852 \\ -0.6424 & -0.5492 & 0.0080 \end{bmatrix} \tag{40}$$

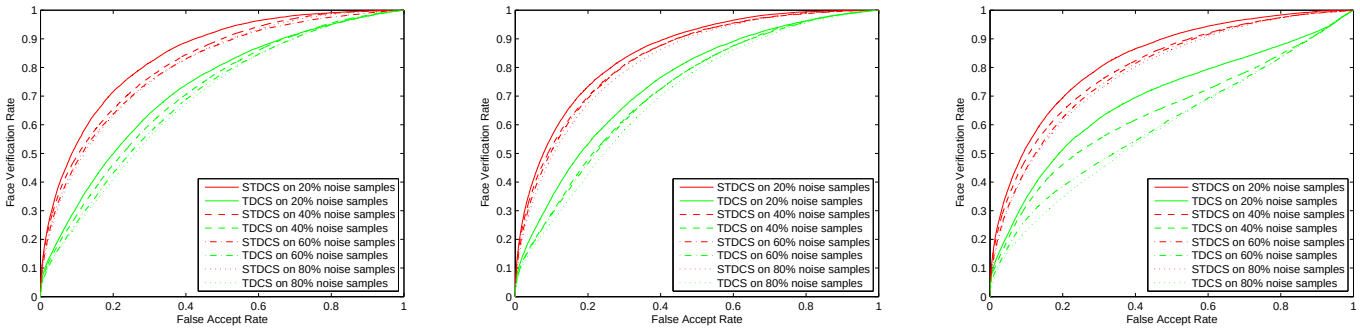
and

$$\mathbf{U}_3^{sparse} = \begin{bmatrix} 0.9051 & 0 & -0.7533 \\ -0.4252 & -0.7941 & 0 \\ 0 & 0.6077 & 0.6576 \end{bmatrix} \tag{41}$$

These three matrices are not the same as Eq. (35), Eq. (36) and Eq. (37) due to the different training sets. Using these three matrices, we got three color components $\mathbf{D}^1$, $\mathbf{D}^2$, $\mathbf{D}^3$ of CID; three color components $\mathbf{T}^1$, $\mathbf{T}^2$, $\mathbf{T}^3$ of TDCS and three color components $\mathbf{S}^1$, $\mathbf{S}^2$, $\mathbf{S}^3$ of STDCS. These components are illustrated in Fig. 11. In order to investigate the semantic interpretation of each color component, we set the number of non-zero elements in each column of the sparse color space transformation matrix as one. The matrix is obtained as:

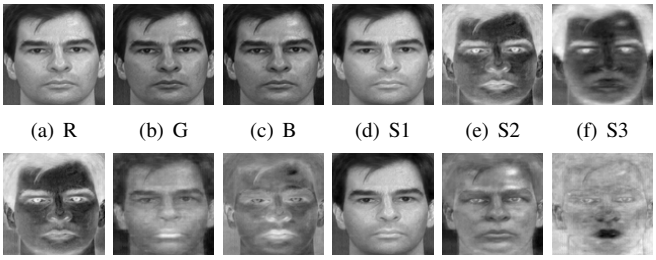
$$\mathbf{U}_3^{sparse'} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} \tag{42}$$

Fig. 12 illustrates three color components $\mathbf{S}^1$, $\mathbf{S}^2$, $\mathbf{S}^3$ in STDCS. The three components $\mathbf{S}^1$, $\mathbf{S}^2$, $\mathbf{S}^3$ come from R, B and negative B components in RGB color space. From above analysis, the combination of R, B components from RGB space is most effective for facial recognition on Georgia



(a) noise on both training set and testing set (b) noise on training set (c) noise on testing set

Fig. 7. ROC curves of STDCS and TDCS the noised facial images of AR database.



(a) R (b) G (c) B (d) S1 (e) S2 (f) S3

(g) T1 (h) T2 (i) T3 (j) D1 (k) D2 (l) D3

Fig. 11. Illustration of R, G, and B color components and the various components generated by CID, TDCS and STDCS on the Georgia Tech face database.

(a) S1 (b) S2 (c) S3

Fig. 12. Illustration of 3 color components $\mathbf{S}^1$, $\mathbf{S}^2$, $\mathbf{S}^3$ in STDCS on the Georgia Tech face database.

Tech face database. The same conclusion can also be drawn from the matrix $\mathbf{U}_3^{sparse'}$. In order to verify this, TDCS was conducted on three different combinations of (R,B), (R,G) and (G,B), respectively. The score matrix was generated by using Euclidean distance. The ROC curves are illustrated in Fig. 13, where the ROC curve of the combination of (R,B) shows better performance than other two combinations.

To 50 individuals in the Georgia Tech database, 49 discriminant projection basis vectors were extracted using LDA and CID. For 2D-LDA, TDCS and STDCS, the two numbers of the reduced spatial dimensions were both 10. In this experiment, the score matrices were generated by using Manhattan distance and Euclidean distance, respectively. In Fig. 14, Manhattan distance and Euclidean distance are denoted by solid lines and dash-dot lines. The results indicate that STDCS has the best performance compared to other four algorithms. For TDCS, the performance of using Euclidean distance is better than that of using Manhattan distance. While for STDCS, the

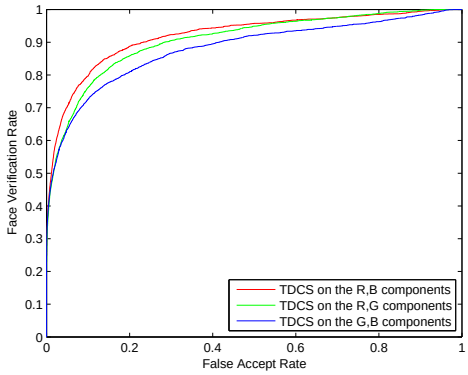


Fig. 13. ROC curves of TDCS on (R,B) (R,G) and (G,B) combinations.

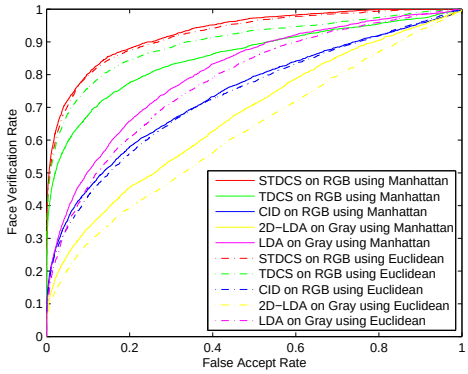


Fig. 14. ROC curves of TDCS, CID, 2D-LDA and LDA on the Georgia Tech face database.

performance of using Euclidean distance is worse than that of using Manhattan distance. Even though the score matrix was generated by the Euclidean distance, the performance of STDCS is better than that of TDCS. From this figure, we can also see that the space between two curves of STDCS is narrower than that between two curves of TDCS. This shows that the STDCS is more insensitive to similarity measurement of images than TDCS.

In order to investigate robustness of models, all images in the Georgia Tech database were manually aligned (two eyes

Fig. 15. Sample images for one individual of the Georgia Tech database (aligned facial images).

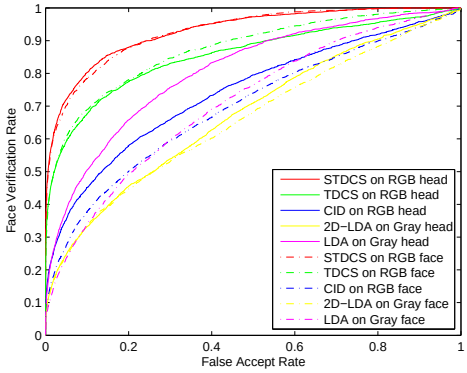


Fig. 16. ROC curves of TDCS, CID, 2D-LDA and LDA on the Georgia Tech face database.

Fig. 17. noise face images on the Georgia Tech face database.

were used for alignment), cropped, and then re-sized to 32×32 pixels. To the cropped images as shown in Fig. 15, we retained as much of the facial region as possible, by eliminating most of the non-facial regions. The experiments with the same setting described above were conducted on them. The score matrix is generated by using Manhattan distance. The results of aligned facial images and unaligned head images are plotted and compared in Fig. 16. The solid lines denote the ROC curves on the unaligned head images and the dash-dot lines denote the ROC curves on the aligned facial images. Generally, the performance on the aligned facial images should be better than the performance on the unaligned head image. However, image vector based methods, such as CID and LDA, achieve opposite results where unaligned image provides better performance than aligned image. This is due to the fact that vectorization causes loss of the spatial structure information of images. Comparing three color space models, we can also see that the performances of STDCS and TDCS are better than that of CID. Furthermore, the margin between two curves of STDCS (or TDCS) is narrower than the margin between two curves of CID. Among them, the two ROC curves of STDCS are almost overlapped. This indicates that STDCS is more robust than TDCS and CID for the color images. This also shows that ℓ_1 norm is more robust than ℓ_2 norm [40].

In order to further investigate the robustness to noise, the noises are added on the Georgia Tech face database by using the same manner. Fig. 17 shows typical examples of noised images. Fig. 18 shows the ROC curves of the two algorithms in the three cases. The similar conclusions are drawn from the figure. It is interesting that the performance of STDCS with 60% noised training samples is the best.

E. Experiments and results on the LFW face database

LFW database is designed for studying the problem of unconstrained face recognition. From Fig. 4, we can see that the skin color of the same person is different due to various cameras. In this experiment, we randomly selected $\lfloor p/2 \rfloor$ images of each person (the person has p images) as the training set and the remaining images as the testing set. The CID, TDCS and STDCS models were trained and we got three color space transformation matrices:

$$\mathbf{X} = \begin{bmatrix} 1.0000 & -1.0089 & -0.3600 \\ 0.3166 & 0.2805 & 0.9981 \\ -0.2236 & 0.9078 & -0.6621 \end{bmatrix}, \tag{43}$$

$$\mathbf{U}_3 = \begin{bmatrix} -0.9103 & 0.6923 & 0.1885 \\ 0.4085 & -0.7043 & -0.7688 \\ -0.0673 & -0.1571 & 0.6111 \end{bmatrix} \tag{44}$$

and

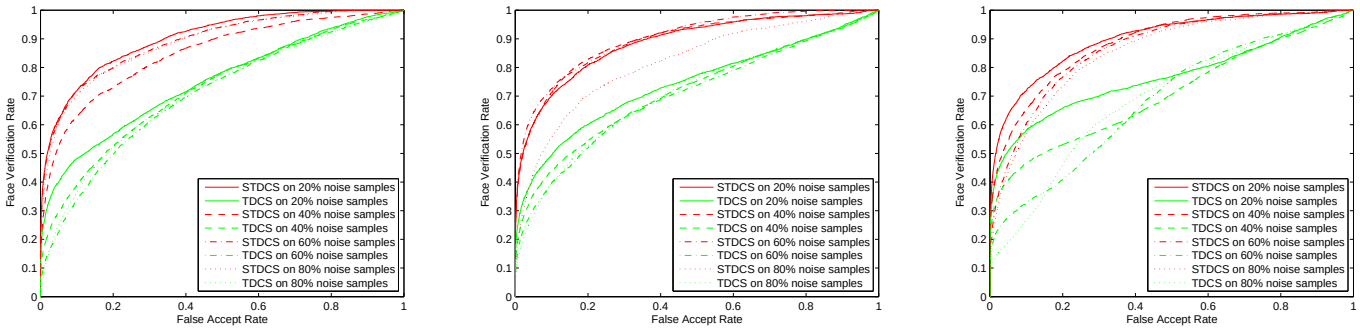
$$\mathbf{U}_3^{sparse} = \begin{bmatrix} 0 & 0.6733 & -0.9736 \\ -0.4744 & -0.7394 & 0 \\ 0.8803 & 0 & 0.2281 \end{bmatrix} \tag{45}$$

Using these three matrices, we got three color components $\mathbf{D}^1$, $\mathbf{D}^2$, $\mathbf{D}^3$ of CID; three color components $\mathbf{T}^1$, $\mathbf{T}^2$, $\mathbf{T}^3$ of TDCS and three color components $\mathbf{S}^1$, $\mathbf{S}^2$, $\mathbf{S}^3$ of STDCS. These components are illustrated in Fig. 19. Intuitively, the components of STDCS are more clear than those of TDCS and CID. In order to further investigate the semantic interpretation of each color component, we set the number of non-zero elements in each column of the sparse color space transformation matrix as one. The matrix is obtained as:

$$\mathbf{U}_3^{sparse'} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{46}$$

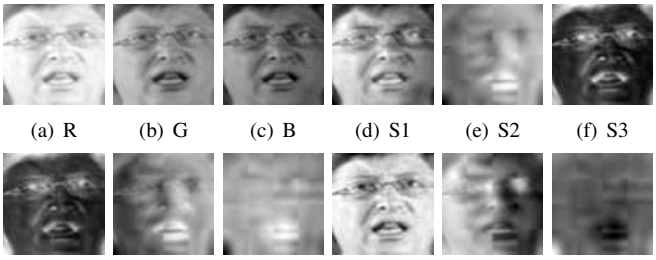
From the above equation, we can see that each component plays the same role. By the same way, 2D-LDA is implemented on R, G and B components to verify the Intuition. In Fig. 20, three ROC curves are illustrated from which we can see that the curves of R G and B components are similar.

In the same way, STDCS, TDCS, CID, 2D-LDA and LDA are conducted on the LFW face database. Due to 86 individuals in the LFW database, LDA and CID only extract 85



(a) noise on both training set and testing set (b) noise on training set (c) noise on testing set

Fig. 18. ROC curves of STDCS and TDCS the noised facial images of the Georgia Tech face database.



(a) R (b) G (c) B (d) S1 (e) S2 (f) S3

(g) T1 (h) T2 (i) T3 (j) D1 (k) D2 (l) D3

Fig. 19. Illustration of R, G, and B color components and the various components generated by CID, TDCS and STDCS on the LFW face database.

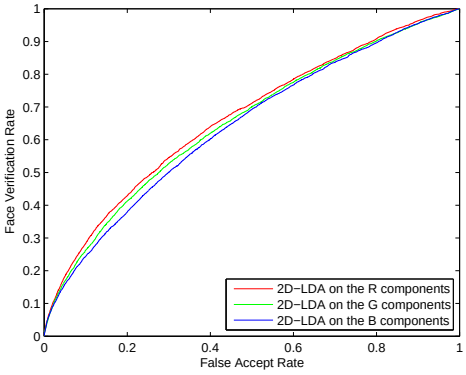


Fig. 20. ROC curves of 2D-LDA on R, G and B components of LFW face database.

discriminant projection basis vectors. Other parameters are the same with previous section. The results indicate that STDCS has the best performance compared to other four algorithms from Fig. 21. Comparing with TDCS, two curves of STDCS are overlapped. This shows that the STDCS is more insensitive to similarity measurement of images than TDCS. we can also see that the curves of the algorithms for gray images are almost consistent and their performance are the worst. It reveals that gray information is not enough for the discrimination of color images in this more complex case.

Similar noised experiments are conducted on LFW database. The results are illustrated in Fig. 22. Similar conclusions are drawn from the figure.

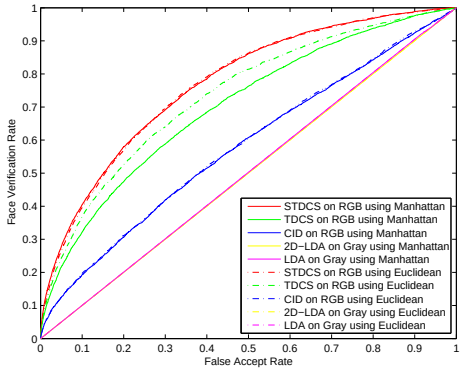


Fig. 21. ROC curves of STDCS, TDCS, CID, 2D-LDA and LDA on the LFW face database.

VI. Conclusion

In this paper, we present a new color space model which is named as the Sparse Tensor Discriminant Color Space (STDCS). By learning from training samples, the proposed model optimizes one sparse color space transformation matrix and two sparse discriminant projection matrices simultaneously. The experiments on the AR, Georgia Tech and LFW color face databases are systematically performed and analyzed. The experimental results reveal a number of interesting remarks:

1) STDCS model can give an intuitionistic or semantic interpretation.
2) STDCS is more robust not only for similarity measurement of images but also for image alignments.

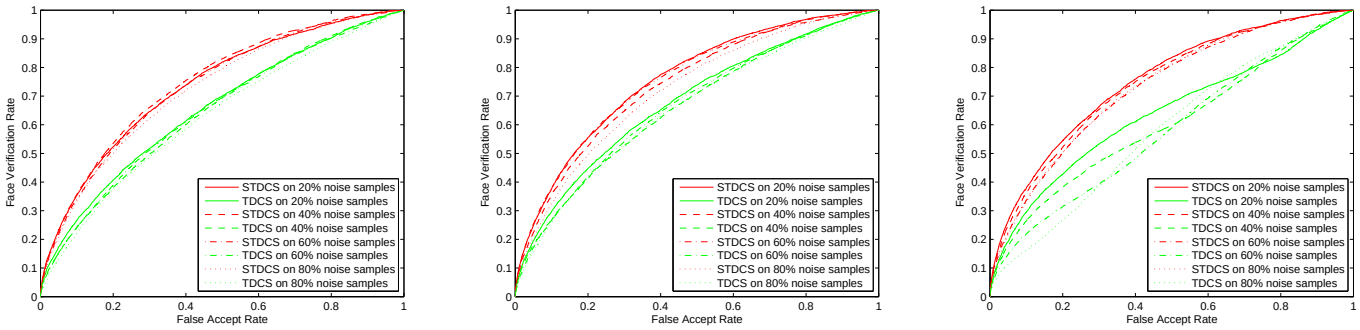
Our future work will be on the theoretical analysis of convergence of the algorithm.

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[3] J. Yang, D. Zhang, A. F. Frangi, and J. Y. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.



(a) noise on both training set and testing set (b) noise on training set (c) noise on testing set

Fig. 22. ROC curves of STDCS and TDCS the noised facial images of LFW database.

[4] M. Li and B. Z. Yuan, "2D-LDA: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527–532, 2005.

[5] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*. London, UK: Springer-Verlag, 2002, pp. 447–460.

[6] S. Park and M. Savvides, "Individual kernel tensor-subspaces for robust face recognition: A computationally efficient tensor framework without requiring mode factorization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 5, pp. 1156–1166, 2007.

[7] S.-J. Wang, C.-G. Zhou, Y.-H. Chen, X.-J. Peng, H.-L. Chen, G. Wang, and X. Liu, "A novel face recognition method based on sub-pattern and tensor," *Neurocomputing*, vol. 74, no. 17, pp. 3553–3564, 2011.

[8] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(r1,r2,...,r-n) approximation of higher-order tensors," *Siam Journal On Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.

[9] ——, "A multilinear singular value decomposition," *Siam Journal On Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[10] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *Siam Review*, vol. 51, no. 3, pp. 455–500, 2009.

[11] J. Ye, "Generalized low rank approximations of matrices," *Machine Learning*, vol. 61, no. 1, pp. 167–191, 2005.

[12] X. He, D. Cai, and P. Niyogi, "Tensor subspace analysis," in *In Advances in Neural Information Processing Systems 18 (NIPS)*. MIT Press, 2005.

[13] H. P. Lu, N. P. Konstantinos, and A. N. Venetsanopoulos, "MPCA: Multilinear principal component analysis of tensor objects," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 18–39, 2008.

[14] S.-J. Wang, C.-G. Zhou, N. Zhang, X.-J. Peng, Y.-H. Chen, and X. Liu, "Face recognition using second-order discriminant tensor subspace analysis," *Neurocomputing*, vol. 74, no. 12-13, pp. 2142–2156, Jun. 2011.

[15] X. F. He and P. Niyogi, "Locality preserving projections," *Advances In Neural Information Processing Systems 16*, vol. 16, pp. 153–160, 2004.

[16] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *Proceedings of the 10th European Conference on Computer Vision: Part IV*. Citeseer, 2008, pp. 340–353.

[17] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[18] W. Schwartz, H. Guo, and L. Davis, "A robust and scalable approach to face identification," *Computer Vision–ECCV 2010*, pp. 476–489, 2010.

[19] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *Neural Networks, IEEE Transactions on*, vol. 22, no. 7, pp. 1119–1132, 2011.

[20] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 23, no. 99, pp. 526–534, 2012.

[21] L. Torres, J. Reutter, and L. Lorente, "The importance of the color information in face recognition," in *Proceedings. 1999 International Conference on Image Processing, 1999. ICIP 99.*, vol. 3. IEEE, 1999, pp. 627–631.

[22] C. Wang, B. Yin, X. Bai, and Y. Sun, "Color face recognition based on 2DPCA," in *19th International Conference on Pattern Recognition, 2008. ICPR 2008.*, 2008, pp. 1–4.

[23] J. Choi, Y. Ro, and K. Plataniotis, "Color face recognition for degraded face images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 5, pp. 1217–1230, 2009.

[24] M. Villegas, R. Paredes, A. Juan, and E. Vidal, "Face verification on color images using local features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08*. IEEE, 2008, pp. 1–6.

[25] W. H. Buchsbaum, *Color TV Servicing*, 3rd ed. Prentice-Hall, 1975.

[26] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Pr, 1990.

[27] C. Liu, "Learning the uncorrelated, independent, and discriminating color spaces for face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 213–222, 2008.

[28] J. Yang and C. Liu, "Color image discriminant models and algorithms for face recognition," *IEEE Transactions on Neural Networks*, vol. 19, no. 12, pp. 2088–2098, 2008.

[29] S.-J. Wang, J. Yang, N. Zhang, and C.-G. Zhou, "Tensor discriminant color space for face recognition," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2490–2501, 2011.

[30] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[31] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[32] H. Zou and T. Hastie, "Regression shrinkage and selection via the elastic net, with applications to microarrays," *JR Statist. Soc. B*, 2004.

[33] B. Moghaddam, Y. Weiss, and S. Avidan, "Spectral bounds for sparse PCA: Exact and greedy algorithms," *Advances in Neural Information Processing Systems*, vol. 18, p. 915, 2006.

[34] ——, "Generalized spectral bounds for sparse LDA," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 641–648.

[35] Z. Qiao, L. Zhou, and J. Z. Huang, "Sparse linear discriminant analysis with applications to high dimensional low sample size data," *IAENG International Journal of Applied Mathematics*, vol. 39, 2009.

[36] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag, 2001.

[37] A. Martinez and R. Benavente, "The AR face database," *Univ. Purdue, CVC Tech. Rep*, vol. 24, 1998.

[38] H. Moon and P. Phillips, "The FERET verification testing protocol for face recognition algorithms," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 48–53.

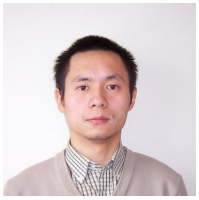
[39] B. Bader and T.G.Kolda, "Tensor toolbox version 2.3, copyright 2009, sandia national laboratories, http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/."

[40] N. Kwak, "Principal component analysis based on L1-norm maximization," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 1672–1680, 2008.

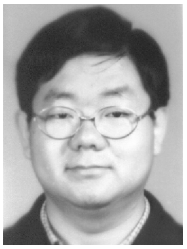
Su-Jing Wang received the Master's degree from the Software College of Jilin University, Changchun, China, in 2007. From September 2008, he is pursuing to the Ph.D. degree at the College of Computer Science and Technology of Jilin University. He has published more than 30 scientific papers. He is One of Ten Selectees of the Doctoral Consortium at International Joint Conference on Biometrics 2011. He was called as *Chinese Hawkin* by the Xinhua News Agency. His research was published in IEEE Transactions on Image Processing, Neurocomputing, etc. His current research interests include pattern recognition, computer vision and machine learning. He also reviews for several top journals, such as IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Neural Networks and Learning Systems. For details, please refer to his homepage http://sujingwang.name.

Jian Yang (M'08) received the BS degree in mathematics from the Xuzhou Normal University in 1995. He received the MS degree in applied mathematics from the Changsha Railway University in 1998 and the Ph.D. degree from the Nanjing University of Science and Technology(NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a postdoctoral researcher at the University of Zaragoza, and in the same year, he was awarded the RyC program Research Fellowship sponsored by the Spanish Ministry of Science and Technology. From 2004 to 2006, he was a postdoctoral fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a postdoctoral fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a professor in the School of Computer Science and Technology of NUST. He is the author of more than 50 scientific papers in pattern recognition and computer vision. His research interests include pattern recognition, computer vision and machine learning. Currently, he is an associate editor of Pattern Recognition Letters and IEEE Transactions on Neural Networks and Learning Systems, respectively.

Ming-Fang Sun received the B.E degree in Computer Science and Technology from Jilin University, Changchun, China, in 2009. Currently pursuing her Master's degree at the College of Computer Science and Technology of Jilin University. Her research includes face recognition and human action recognition.



Xu-Jun Peng obtained his Ph.d from department of computer science and engineering at the state university of New York at Buffalo. Currently, he is a research scientist with Raytheon BBN technologies. His research interests include Machine Learning, Image Processing and Document Analysis.



Ming-Ming Sun received the B.S. degree in mathematics from Xinjiang University, Urumqi, China, in 2002, and the Ph.D. degree in pattern recognition and intelligence systems from the Department of Computer Science, Nanjing University of Science and Technology (NUST), Nanjing, China, in 2007. He is currently a Lecturer with the School of Computer Science and Technology, NUST. His current research interests include pattern recognition, machine learning and image processing.

Chun-Guang Zhou PhD, professor, PhD supervisor, Dean of Institute of Computer Science of Jilin University. He is Jilin-province-management Expert, Highly Qualified Expert of Jilin Province, One-hundred Science-Technique elite of Changchun. And he is awarded the Governmental Subsidy from the State Department. He has many pluralities of national and international academic organizations. His research interests include related theories, models and algorithms of artificial neural networks, fuzzy systems and evolutionary computations, and applications of machine taste and smell, image manipulation, commercial intelligence, modern logistic, bioinformatics, and biometric identification based on computational intelligence. he has published over 168 papers in Journals and conferences and he published 1 academic book.