Tensor Linear Regression and Its Application to Color Face Recognition

Quanxue Gao[1], Jiafeng Cheng[1]*, Deyan Xie[1], Pu Zhang[1], Wei Xia[1], Qianqian Wang[1]
[1] State Key Laboratory of Integrated Services Networks, Xidian University
Xi'an, 710071, China
qxgao@xidian.edu.cn; chengjf1208@163.com; qianqian174@foxmail.com

Abstract

Linear regression has achieved the promising preliminary results for face classification. But, most existing methods are incapable of tackling color images classification. The major reason is that they need to transform each color image to a vector or matrix, leading to the loss of multi-dimensional structure information embedded in color images. To address this problem, we study the tensor linear regression problem, and develop a novel tensor low-rank method, which utilizes tensor-Singular Value Decomposition (t-SVD) based tensor nuclear norm to emphasize the spatial structure embedded in color images. Applying it to color face classification, extensive experiments on three datasets demonstrate that our method is superior to several state-of-the-art methods.

1. Introduction

Face recognition is a biometric recognition technology based on human facial feature information for identification and has been one of the hot topics in pattern recognition and computer vision. Both feature extraction and classification are two of the most important problems for face recognition. We herein focus on classification. For classification, methods based on linear regression (LR) have achieved impressive results [3, 14, 2]. For example, Hoerl [7] used linear regression model for data classification and proposed ridge linear regression model which presents each probe data as a linear combination of all training samples. To improve performance, Naseem *et al*. developed a linear regression classifier (LRC) [14] for face recognition. In LRC, each probe face image can be described as a linear combination of the class-specific samples and classifies the probe image by minimizing the class-reconstruction error. In fact, several previous works, such as nearest feature line [10], nearest feature plane, and nearest feature space method [4] are all variants of LR based methods.

However, standard linear regression model always encounters the problem of over-fitting and does not make sense in many applications [7] when the coefficients are not limited. To tackle this problem, one of the most popular methods is to add l_1-norm based regularization on LR model, which is called Lasso. Applying it to face image classification, Wright *et al.* proposed a sparse representation based classification (SRC) [18]. Since the coefficients in SRC encode discriminative, SRC has achieved impressive performance in the experiments. Zhang *et al.* [23] analyzed the working mechanism of SRC and indicated that, compared with SRC, collaborative representation (CR) has competitive performance with significantly lower complexity. Then, CR based classification (CRC) is proposed for face recognition.

LRC, SRC and CRC employ l_1-norm or l_2-norm to measure the representation residual of error vector and have achieved impressive results when representation error follows a Gaussian or a Laplacian distribution. However, this distribution is very strict and cannot be satisfied in real-world face recognition [20, 17]. To improve robustness of LR, He *et al* proposed correntropy-based sparse representation (CESR) [6] by maximizing correntropy criterion for face recognition. Although motivations of the aforementioned methods are different, all of them use the one-dimensional pixel-based error model, in which the error on each pixel is characterized one by one, individually. Thus, they ignore spatial structure of representation error which is important for classification [22, 16].

Most existing works [22, 16] have demonstrated that nuclear norm can well characterize the whole structural information of data. Ma *et al.* integrated the rank minimization into sparse representation for dictionary learning and applied the model for face recognition [24]. Motivated by the fact that contiguous occlusion in image generally leads to low-rank representation error image, Qian *et al.* [15] proposed a model by adding a low-rank constraint of error image on Ridge regression. Yang *et al.* consolidated the work of Qian *et al* by incorporating nuclear norm based matrix regression (NMR) [19] which makes full use of the low-rank

*Corresponding author: Jiafeng Cheng (e-mail: chengjf1208@163.com).

structure.

The above mentioned methods are limited to handling 1-way (vector) or 2-way (matrix) data. In face recognition, the real face images are usually described by multi-dimensional way [13, 9]. For example, a color image is a 3-way object with column, row and color modes. However, existing models have to convert the multi-dimensional images into vectors or matrixes when dealing with multi-dimensional images. Thus, all of them fail to encode the multi-dimensional structure information embedded in color images. This motivates us to investigate how to exploit the multi-dimensional structure information.

Inspired by the following four observations that, First, a color probe image can be represented as a linear representation of a small number of dictionary atoms. Second, error image usually has low-rank structure due to the fact that the elements of error image are correlated in real applications. Third, the multi-dimensional structure information help improve robustness of regression models. Fourth, the recently proposed tensor-nuclear norm [12], which is based on tensor singular value decomposition (t-svd), is an effective convex relaxation of l_1-norm and well encodes discriminative information. We impose low-rank constraint of error image with tensor form in tensor linear regression, and present a new type of tensor nuclear norm (t-TNN) algorithm to optimize the objective function. Different form most existing tensor rank methods, t-SVD enjoys many similar properties as the matrix case and is more appropriate to describe the multi-dimensional information of color image. Thus, the complementary information among color channels can be explored more efficiently and thoroughly in our proposed tensor regression model. The main contributions of our method are summarized as follows:

- Our work extends matrix regression to tensor regression via proposing a new tensor regression model that is able to maintain more structural information without using the tensor-to-matrix or tensor-to-vectors converting step. Thus, our model well encodes multi-dimensional structure information embedded in color images

- In practice, contiguous occlusion generally leads to a low-rank error image. We use the whole structural information of an error image by minimizing the tensor nuclear norm to determine the regression coefficients. By doing it, the model becomes more robust against the the occlusion, disguise and so on.

2. Related work

2.1. Notations and preliminaries

For convenience, we summarize the notations used in our paper in Table 1. Given $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, denote by $\overline{\mathcal{A}}$ the discrete Fast Fourier transform (FFT) of tensor $\mathcal{A}$ along the third dimension i.e., $\overline{\mathcal{A}} = fft(\mathcal{A}, [], 3)$. Similarly, thus $\mathcal{A}$ can be obtained by inverse FFT (IFFT) of $\overline{\mathcal{A}}$ along the third dimension, i.e., $\mathcal{A} = ifft(\overline{\mathcal{A}}, [], 3)$.

Table 1. Symbols and Meanings

symbol	meaning
$\mathcal{A}$	a tensor
$\mathbf{A}$	a matrix
$\mathbf{a}$	a vector
a	a scalar
a_{ijk}	(i, j, k)-th entry of $\mathcal{A}$
$\mathbf{A}^{(i)}$	the i-th frontal slice of $\mathcal{A}$
$\bar{\mathbf{A}}^{(i)}$	the i-th frontal slice of $\bar{\mathcal{A}}$

Definition 1 *[8] For a 3-way tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we denote the Frobenius norm as $\|\mathcal{A}\|_F = \sqrt{\sum_{ijk} |a_{ijk}|^2}$*

Definition 2 *[8] For a 3-way tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we denote $\overline{\mathbf{A}}$ as a block diagonal matrix with each block on diagonal as the frontal slice $\overline{\mathbf{A}}^{(i)}$ of $\overline{\mathcal{A}}$. $\overline{\mathbf{A}}$ has the following form:*

$$\overline{\mathbf{A}} = bdiag(\overline{\mathcal{A}}) = \begin{bmatrix} \overline{\mathbf{A}}^{(1)} & & & \\ & \overline{\mathbf{A}}^{(2)} & & \\ & & \ddots & \\ & & & \overline{\mathbf{A}}^{(n_3)} \end{bmatrix} \tag{1}$$

Definition 3 *[8] For a 3-way tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, its block circulant matrix is a matrix of $n_1 n_3 \times n_2 n_3$ having the following form:*

$$bcirc(\mathcal{A}) = \begin{bmatrix} \mathbf{A}^{(1)} & \mathbf{A}^{(n_3)} & ... & \mathbf{A}^{(2)} \\ \mathbf{A}^{(2)} & \mathbf{A}^{(1)} & ... & \mathbf{A}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}^{(n_3)} & \mathbf{A}^{(n_3-1)} & ... & \mathbf{A}^{(1)} \end{bmatrix} \tag{2}$$

Definition 4 *[8] For a tensor $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we have*

$$\begin{aligned} unfold(\mathcal{A}) &= \left[\mathbf{A}^{(1)}; \mathbf{A}^{(2)}; \cdots; \mathbf{A}^{(n_3)}\right] \\ fold(unfold(\mathcal{A})) &= \mathcal{A} \end{aligned} \tag{3}$$

Definition 5 *[8] (t-product) Let $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ and $\mathcal{B} \in \mathbb{R}^{n_2 \times l \times n_3}$, then the t-product between them is $\mathcal{E} \in \mathbb{R}^{n_1 \times l \times n_3}$, i.e.,*

$$\mathcal{E} = \mathcal{A} * \mathcal{B} = fold(bcirc(\mathcal{A}) \cdot unfold(\mathcal{B})) \tag{4}$$

t-product between $\mathcal{A}$ and $\mathcal{B}$ can be computed efficiently by

1. Calculate $\overline{\mathcal{A}} = fft(\mathcal{A}, [], 3)$ and $\overline{\mathcal{B}} = fft(\mathcal{B}, [], 3)$;

2. Multiply the each pair of the frontal slices of $\bar{\mathcal{A}}$ and $\bar{\mathcal{B}}$ to obtain $\bar{\mathcal{E}}$;

3. Calculate $\boldsymbol{\mathcal{E}} = ifft(\bar{\boldsymbol{\mathcal{E}}}, [], 3)$;

Meanwhile, note that the t-product reduces to the standard matrix-matrix product when $n_3 = 1$.

Definition 6 *[8] The conjugate transpose of a tensor $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is the tensor $\boldsymbol{\mathcal{A}}^T \in \mathbb{R}^{n_2 \times n_1 \times n_3}$ obtained by conjugate transposing each of the frontal slice and then reversing the order of transposed frontal slices 2 through n_3.*

Definition 7 *[8] A tensor is called f-diagonal if each of its frontal slices is diagonal matrix.*

Theorem 1 *[8] Block-circulant matrix can be block-diagonalized by*

$$(\mathbf{F_{n_3}} \otimes \mathbf{I_{n_1}}) \cdot bcirc(\boldsymbol{\mathcal{A}}) \cdot (\mathbf{F_{n_3}}^{\mathbf{-1}} \otimes \mathbf{I_{n_2}}) = \overline{\mathbf{A}} \tag{5}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{F}_{n_3}$ is the $n_3 \times n_3$ Discrete Fourier Transform (DFT) matrix, $\mathbf{I}_{n_1}$ and $\mathbf{I}_{n_2}$ denote $n_1 \times n_1$ and $n_2 \times n_2$ identity matrices, respectively.

Theorem 2 *[8](T-SVD). Let $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, then $\boldsymbol{\mathcal{A}}$ can be factored as*

$$\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^T \tag{6}$$

where $\boldsymbol{\mathcal{U}} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ and $\boldsymbol{\mathcal{V}} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ are orthogonal, $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a f-diagonal tensor.

Definition 8 *[12](tensor nuclear norm). Given $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, its nuclear norm is*

$$\begin{aligned} &\|\boldsymbol{\mathcal{A}}\|_{\circledast} = \textstyle\sum_{i=1}^{n_3} \left\| \overline{A}^{(i)} \right\|_* \\ &= \left\| (\boldsymbol{F}_{n_3} \otimes \boldsymbol{I}_{n_1}) \cdot bcirc(\boldsymbol{\mathcal{A}}) \cdot (\boldsymbol{F}_{n_3}^{-1} \otimes \boldsymbol{I}_{n_2}) \right\|_* \\ &= \|bcirc(\boldsymbol{\mathcal{A}})\|_* \end{aligned} \tag{7}$$

2.2. NMR

Given n image matrices $\mathbf{A}_1, ..., \mathbf{A}_n \in \mathbb{R}^{p \times q}$, then a probe $\mathbf{B} \in \mathbb{R}^{p \times q}$ is represented as

$$\begin{aligned} \mathbf{B} &= c_1\mathbf{A}_1 + c_2\mathbf{A}_2 +, ..., + c_n\mathbf{A}_n + \mathbf{E} \\ &\triangleq A(\mathbf{c}) + \mathbf{E} \end{aligned} \tag{8}$$

where, $c_1, c_2, ..., c_n$ is a set of representation coefficients, $A(\mathbf{c}) = c_1\mathbf{A}_1 + c_2\mathbf{A}_2 + ... + c_n\mathbf{A}_n$, and $\mathbf{E}$ is residual.

NMR [19] utilizes the minimal nuclear norm of representation error image as a criterion, and formulates the objective function as

$$\min_{\mathbf{c}} \|\mathbf{A}(\mathbf{c}) - \mathbf{B}\|_* + \frac{1}{2}\lambda \|\mathbf{c}\|_2^2 \tag{9}$$

Despite the promising classification results for face recognition in the presence of occlusion and illumination variations, NMR needs to transform each color image to a gray matrix, results in the loss of spatial structure information hidden in color image. Consequently, the performance of NMR is limited in color image classification. To address this problem, we extend NMR to the tensor form, and develop a novel tensor nuclear norm method in the following section.

3. The classification of nuclear norm based tensor regression model and its ADMM algorithm

3.1. Nuclear norm based tensor linear regression

Given a set of training samples $\boldsymbol{\mathcal{A}}_{1,1}$,...,$\boldsymbol{\mathcal{A}}_{1,p_1}$,..., $\boldsymbol{\mathcal{A}}_{N,1}$,...,$\boldsymbol{\mathcal{A}}_{N,p_N} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. The set of training samples has N classes and the i-th class has p_i samples. For a new test sample $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, it can be represented linearly using $\boldsymbol{\mathcal{A}}_{i,m} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ as the following equation:

$$\begin{aligned} \boldsymbol{\mathcal{X}} &= x_{1,1}\boldsymbol{\mathcal{A}}_{1,1} + ... + x_{1,p_1}\boldsymbol{\mathcal{A}}_{1,p_1} + ... \\ &+ x_{N,1}\boldsymbol{\mathcal{A}}_{N,1} + ... + x_{N,p_N}\boldsymbol{\mathcal{A}}_{N,p_N} + \boldsymbol{\mathcal{E}} \end{aligned} \tag{10}$$

where, $x_{1,1}, ..., x_{1,p_1}, ..., x_{N,1}, ..., x_{N,p_N}$ is a set of representation coefficients and $\boldsymbol{\mathcal{E}}$ is the representation residual. We have the following linear mapping for convenience.

$$\begin{aligned} &\boldsymbol{\mathcal{A}}(\mathbf{x}) = x_{1,1}\boldsymbol{\mathcal{A}}_{1,1} + ... + x_{1,p_1}\boldsymbol{\mathcal{A}}_{1,p_1} + ... + \\ &x_{N,1}\boldsymbol{\mathcal{A}}_{N,1} + ... + x_{N,p_N}\boldsymbol{\mathcal{A}}_{N,p_N} \end{aligned} \tag{11}$$

So, Eq. (10) becomes

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{A}}(\mathbf{x}) + \boldsymbol{\mathcal{E}} \tag{12}$$

For the probe image, ideally, the representation of a set of training samples is closed to the probe as possible in regression analysis, it may lead to the structural low-rank of the error image. Meanwhile, in more general cases, the illumination such as shadows and the occlusion such as sunglass and a scarf yield a low-rank error image for grayscale images. So we extend the low-rank of representation residual image matrix to the low-rank of representation residual image tensor as to make full use of the low-rank structural information. We formulate the optimal problem as:

$$\begin{aligned} &\min_{\mathbf{x}, \boldsymbol{\mathcal{E}}} \|\boldsymbol{\mathcal{E}}\|_{\circledast} \\ &st. \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{A}}(\mathbf{x}) + \boldsymbol{\mathcal{E}} \end{aligned} \tag{13}$$

To avoid over-fitting, we need to add the limit on regression coefficients. Borrowing the idea of the Ridge regression, we add a similar regularization term to the objective function. Finally, we obtain the following model:

$$\begin{aligned} &\min_{\mathbf{x}, \boldsymbol{\mathcal{E}}} \|\boldsymbol{\mathcal{E}}\|_{\circledast} + \lambda \|\mathbf{x}\|_2^2 \\ &st. \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{A}}(\mathbf{x}) + \boldsymbol{\mathcal{E}} \end{aligned} \tag{14}$$

3.2. Optimization

Due to the operations of tensor are different from matrix, in order to solve Eq. (14) and find the optimal solution, we given the following Lemma and Theorem.

Lemma 1 *[1] Let $\mathbf{Y} = \mathbf{U}_Y * \mathbf{D}_Y * \mathbf{V}_Y^T$ be the SVD of $\mathbf{Y} \in \mathbb{R}^{m \times n}$, $\tau > 0$, for the following optimization problem:*

$$\arg\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{X} - \mathbf{Y}\|_F^2 + \tau \|\mathbf{X}\|_* \tag{15}$$

Then, the optimal solution of the model (15) is

$$\mathrm{D}_\tau[\mathbf{Y}] = \mathbf{U}_Y P_\tau(\mathbf{Y}) \mathbf{V}_Y^T \tag{16}$$

where,

$$P_\tau(\mathbf{Y}) = diag(\gamma_1, \gamma_2, \cdots, \gamma_l)$$
$$\gamma_i = sign(\sigma_i(\mathbf{Y})) \max(\sigma_i(\mathbf{Y}) - \tau, 0)$$

$\sigma_i(\mathbf{Y})$ *denotes the* i *largest singular value of* $\mathbf{Y}$.

Theorem 3 *For* $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, *let the t-SVD of* $\boldsymbol{\mathcal{A}}$ *be* $\boldsymbol{\mathcal{A}} = \boldsymbol{\mathcal{U}} * \boldsymbol{\mathcal{S}} * \boldsymbol{\mathcal{V}}^T$.

For a minimization problem:

$$\arg\min_{\boldsymbol{\mathcal{X}}} \frac{1}{2} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}\|_F^2 + \tau \|\boldsymbol{\mathcal{X}}\|_{\circledast} \tag{17}$$

Then, the optimal solution of Eq. (17) can be expressed by the operator defined as:

$$\boldsymbol{\mathcal{X}} = \mathrm{D}_\tau(\boldsymbol{\mathcal{A}}) = \boldsymbol{\mathcal{U}} * ifft(P_\tau(\overline{\boldsymbol{\mathcal{A}}})) * \boldsymbol{\mathcal{V}}^T \tag{18}$$

where $P_\tau(\overline{\boldsymbol{\mathcal{A}}})$ *is a tensor and* $P_\tau(\overline{\mathbf{A}}^{(i)})$ *is the* i*-th frontal slice of* $P_\tau(\overline{\boldsymbol{\mathcal{A}}})$.

Proof:
In Fourier domain, the optimization problem of Eq. (17) can be reformulated as:

$$\overline{\boldsymbol{\mathcal{X}}}^* = \arg\min_{\overline{\boldsymbol{\mathcal{X}}}} \frac{1}{2} \left\|\overline{\boldsymbol{\mathcal{X}}} - \overline{\boldsymbol{\mathcal{A}}}\right\|_F^2 + \tau \|bdiag(\boldsymbol{\mathcal{X}})\|_* \tag{19}$$

According to the character of F-norm and the character of nuclear norm of block diagonal matrix, we can get the following equation further:

$$\overline{\boldsymbol{\mathcal{X}}}^* = \arg\min_{\overline{\boldsymbol{\mathcal{X}}}} \sum_{i=1}^{n_3} \frac{1}{2} \left\|\overline{\mathbf{X}}^{(i)} - \overline{\mathbf{A}}^{(i)}\right\|_F^2 + \tau \left\|\overline{\mathbf{X}}^{(i)}\right\|_* \tag{20}$$

where $\overline{\mathbf{X}}^{(i)}$ and $\overline{\mathbf{A}}^{(i)}$ are the i the frontal slice of $\overline{\boldsymbol{\mathcal{X}}}$ and $\overline{\boldsymbol{\mathcal{A}}}$ respectively. Thus all $\overline{\mathbf{X}}^{(i)}$ are independent, so are $\overline{\mathbf{A}}^{(i)}$. Then Eq. (20) can be divided into n_3 subproblems.

For the i-th ($i = 1, 2, \cdots, n_3$) subproblem, we have

$$\overline{\mathbf{X}}^{(i)*} = \arg\min_{\overline{\mathbf{X}}^{(i)}} \frac{1}{2} \left\|\overline{\mathbf{X}}^{(i)} - \overline{\mathbf{A}}^{(i)}\right\|_F^2 + \tau \left\|\overline{\mathbf{X}}^{(i)}\right\|_* \tag{21}$$

According to Lemma 1, the solution $\overline{\mathbf{X}}^{(i)*}$ of Eq. (21) is $\mathrm{D}_\tau(\overline{\mathbf{A}}^{(i)}) = \overline{\mathbf{U}}^{(i)} P_\tau(\overline{\mathbf{A}}^{(i)}) {\overline{\mathbf{V}}^{(i)}}^T$ and it is the i-th frontal slice of $\overline{\boldsymbol{\mathcal{X}}}^*$.

According to Definition 5, we can easily get

$$\boldsymbol{\mathcal{X}}^* = \mathrm{D}_\tau(\boldsymbol{\mathcal{A}}) = \boldsymbol{\mathcal{U}} * ifft(P_\tau(\overline{\boldsymbol{\mathcal{A}}})) * \boldsymbol{\mathcal{V}}^T \tag{22}$$

where $\boldsymbol{\mathcal{U}} = ifft(\overline{\boldsymbol{\mathcal{U}}}, [], 3)$ and $\overline{\mathbf{U}}^{(i)}$ is the i-th frontal slice of $\overline{\boldsymbol{\mathcal{U}}}$, $\boldsymbol{\mathcal{V}} = ifft(\overline{\boldsymbol{\mathcal{V}}}, [], 3)$ and $\overline{\mathbf{V}}^{(i)}$ is the i-th frontal slice of $\overline{\boldsymbol{\mathcal{V}}}$.

Our objective function in Eq. (14) simultaneously learns the residual error $\boldsymbol{\mathcal{E}}$ and the coefficients $\mathbf{x}$. Each of them can be solved efficiently by fixing the other. The alternating direction method of multipliers or the augmented Lagrange multipliers (ALM) method is an efficient solver for our problem [11]. Eq. (14) can be solved by minimizing the following unconstrained ALM problem:

$$\min_{\mathbf{x}, \boldsymbol{\mathcal{E}}} \|\boldsymbol{\mathcal{E}}\|_{\circledast} + \lambda \|\mathbf{x}\|_2^2 + \frac{\mu}{2} \left\|\boldsymbol{\mathcal{E}} - \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{A}}(\mathbf{x}) + \frac{\boldsymbol{\mathcal{Y}}}{\mu}\right\|_F^2 \tag{23}$$

where μ is a positive scalar, $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is an estimate of the Lagrange multiplier, $\|\cdot\|_{\circledast}$ denotes the tensor nuclear norm, $\|\cdot\|_2$ denotes the vector l_2 norm, $\|\cdot\|_F$ denotes the tensor Frobenius norm. We separate our problem into the following subproblems.

1. **x-subproblem**: To update $\mathbf{x}$, we solve the following optimization problem by fixing the other variables

   $$\arg\min_{\mathbf{x}} \tfrac{\mu_k}{2} \left\|\boldsymbol{\mathcal{E}}_k - \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{A}}(\mathbf{x}) - \tfrac{\boldsymbol{\mathcal{Y}}_k}{\mu_k}\right\|_F^2 + \lambda \|\mathbf{x}\|_2^2 \tag{24}$$

   For Eq. (24), We denote $\mathbf{e}_k = vect(\boldsymbol{\mathcal{E}}_k)$, $\mathbf{z} = vect(\boldsymbol{\mathcal{X}})$, $\mathbf{y}_k = vect(\boldsymbol{\mathcal{Y}}_k)$, $\mathbf{D} = (\ vect(\boldsymbol{\mathcal{A}}_{1,1}),\ vect(\boldsymbol{\mathcal{A}}_{1,2}),\ \ldots\ , vect(\boldsymbol{\mathcal{A}}_{N,p_N})\ )$, where $vect(\cdot)$ is notation which transforms tensor into vector. This transformation is reasonable according to the definition of F-norm. So Eq. (24) becomes the following optimization problem:

   $$\arg\min_{\mathbf{x}} \tfrac{\mu_k}{2} \left\|\mathbf{e}_k - \mathbf{z} + \mathbf{D}\mathbf{x} - \tfrac{\mathbf{y}_k}{\mu_k}\right\|_F^2 + \lambda \|\mathbf{x}\|_2^2 \tag{25}$$

   taking the derivative of Eq. (25) w.r.t $\mathbf{x}$ and setting the derivative to zero. We have the solution as the following:

   $$\begin{aligned} \mathbf{x}_{k+1} = (\mu_k \mathbf{D}^T \mathbf{D} + 2\lambda \mathrm{I})^{-1} * \mu_k \\ * (\mathbf{D}^T \mathbf{z} + \mathbf{D}^T \tfrac{\mathbf{y}_k}{\mu_k} - \mathbf{D}^T \mathbf{e}_k) \end{aligned} \tag{26}$$

2. $\boldsymbol{\mathcal{E}}$ **-subproblem**: To update $\boldsymbol{\mathcal{E}}$, we solve the following optimization problem by fixing the other variables

   $$\begin{aligned} \arg\min_{\boldsymbol{\mathcal{E}}} \tfrac{1}{2} \left\|\boldsymbol{\mathcal{E}} - (\boldsymbol{\mathcal{X}} + \mu_k^{-1} \boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}))\right\|_F^2 \\ + \tfrac{1}{\mu_k} \|\boldsymbol{\mathcal{E}}\|_{\circledast} \end{aligned} \tag{27}$$

   According to Theorem 3, the optimal solution of Eq. (27) is

   $$\mathrm{D}_{\mu_k^{-1}}(\boldsymbol{\mathcal{X}} + \mu_k^{-1} \boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1})) \tag{28}$$

The optimization process of solving Eq. (14) is summarized in Algorithm 1.

3.3. Classification

Similar to the strategy of NMR, we use the training samples of all classes to form the set of regressors. Let $\boldsymbol{\mathcal{A}}_{1,1}$,...,$\boldsymbol{\mathcal{A}}_{1,p_1}$,..., $\boldsymbol{\mathcal{A}}_{N,1}$,...,$\boldsymbol{\mathcal{A}}_{N,p_N}$ be training sample images of all classes. For a given test image $\boldsymbol{\mathcal{X}}$, we use all training samples to represent it and obtain the representation coefficient vector by solving Eq. (14) via Algorithm 1 and obtain the optimal solution.

Let δ_i: $\mathbb{R}^n \to \mathbb{R}^n$ be the characteristic function that selects the coefficients associated with the i-th class. For $\mathbf{x} \in \mathbb{R}^n$, $\delta_i(\mathbf{x})$ is a vector whose only nonzero entries are the entries in $\mathbf{x}$ that are associated with Class i. Using the coefficients associated with the i-th class, one can get the reconstruction of $\boldsymbol{\mathcal{X}}$ in Class i as $\hat{\boldsymbol{\mathcal{X}}}_i = \boldsymbol{\mathcal{A}}(\delta_i(\mathbf{x}))$. The corresponding class reconstruction error is defined by

$$d_i(\boldsymbol{\mathcal{X}}) = \left\|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}_i\right\|_F^2 \tag{29}$$

The decision rule is defined as: if $d_l = \min\limits_i d_i(\boldsymbol{\mathcal{X}})$, then $\boldsymbol{\mathcal{X}}$ is assigned to Class l.

```
Algorithm 1: Solving Eq. (14) by ADMM
Initialize: $\boldsymbol{\mathcal{E}}_0 = \boldsymbol{\mathcal{Y}}_0 = \mathbf{x}_0 = 0$, $\rho = 1.1$, $\mu_0 = 1e-3$,
$\mu_{\max} = 1e10$, $\varepsilon = 1e-8$
whlie not converged do
    1. update $\mathbf{x}_{k+1}$ by
        $\mathbf{x}_{k+1} = \arg\min\limits_{\mathbf{x}} \frac{\mu_k}{2}\left\|\boldsymbol{\mathcal{E}}_k - \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{A}}(\mathbf{x}) - \frac{\boldsymbol{\mathcal{Y}}_k}{\mu_k}\right\|_F^2 + \lambda\|\mathbf{x}\|_2^2$
    2. update$\boldsymbol{\mathcal{E}}_{k+1}$ by
        $\boldsymbol{\mathcal{E}}_{k+1} = \arg\min\limits_{\boldsymbol{\mathcal{E}}} \frac{1}{2}\|\boldsymbol{\mathcal{E}} - (\boldsymbol{\mathcal{X}}+\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}))\|_F^2 + \frac{1}{\mu_k}\|\boldsymbol{\mathcal{E}}\|_{\circledast}$
    3.$\boldsymbol{\mathcal{Y}}_{k+1} = \boldsymbol{\mathcal{Y}}_k + \mu_k(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1})$
    4.update $\mu_{k+1}$ by $\mu_{k+1} = \min(\rho\mu_k, \mu_{max})$
    5.check the convergence conditions
        $\|\boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{A}}(\mathbf{x}_k)\|_\infty \leq \varepsilon$,
        $\|\boldsymbol{\mathcal{E}}_{k+1} - \boldsymbol{\mathcal{E}}_k\|_\infty \leq \varepsilon$
        $\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}\|_\infty \leq \varepsilon$
end while
6. Output $\boldsymbol{\mathcal{E}}$, $\mathbf{x}$
```

3.4. Convergence Analysis

Corollary 1 *The sequence $\{\boldsymbol{\mathcal{Y}}_k\}$ is generated by Algorithm 1 is bounded.*

Proof: Based on the Lagrange multiplier updating method in step of Algorithm1, we have:

$$\begin{aligned}
&\|\boldsymbol{\mathcal{Y}}_{k+1}\|_F = \|\boldsymbol{\mathcal{Y}}_k + \mu_k(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1})\|_F \\
&\quad = \mu_k\left\|\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k + \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}\right\|_F \\
&\quad = \mu_k \times \tfrac{1}{\sqrt{n_3}}\left\|bdiag(\overline{\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k + \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}})\right\|_F \\
&\quad = \mu_k \times \tfrac{1}{\sqrt{n_3}}\left\|bdiag(\overline{\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k + \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1})})- \right. \\
&\qquad\qquad \left. bdiag(\overline{\boldsymbol{\mathcal{E}}_{k+1}})\right\|_F
\end{aligned} \tag{30}$$

Denote by $\boldsymbol{\mathcal{U}}_k * \boldsymbol{\mathcal{S}}_k * \boldsymbol{\mathcal{V}}_k^T$ the SVD of the tensor $\boldsymbol{\mathcal{X}}+\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1})$ in the $(k+1)$-th iteration. Based on the solution of tensor nuclear norm, we have

$$\boldsymbol{\mathcal{E}}_{k+1}{=}\boldsymbol{\mathcal{U}}_k * ifft(P_\tau(\overline{\boldsymbol{\mathcal{X}}+\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1})})) * \boldsymbol{\mathcal{V}}_k^T \tag{31}$$

According to Theorem 2, we have

$$\begin{aligned}
&\|\boldsymbol{\mathcal{Y}}_{k+1}\|_F = \mu_k \times \tfrac{1}{\sqrt{n_3}}\left\|bdiag(\overline{\boldsymbol{\mathcal{U}}_k}) \bullet (bdiag(\overline{\boldsymbol{\mathcal{S}}_k})\right. \\
&\left.-bdiag(P_\tau(\overline{\boldsymbol{\mathcal{X}}+\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1})}))){\bullet}bdiag(\overline{\boldsymbol{\mathcal{V}}_k}^T)\right\|_F \\
&{=}\mu_k \times \tfrac{1}{\sqrt{n_3}}\left\|bdiag(\overline{\boldsymbol{\mathcal{S}}_k})- \right. \\
&\qquad\qquad \left. bdiag(P_\tau(\overline{\boldsymbol{\mathcal{X}}+\mu_k^{-1}\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1})}))\right\|_F \\
&\leq \mu_k \times \tfrac{1}{\sqrt{n_3}}\sqrt{\sum_{j=1}^{n_3}\sum_i\left(\tfrac{1}{\mu_k}\right)^2}
\end{aligned} \tag{32}$$

Thus, $\{\boldsymbol{\mathcal{Y}}_k\}$ is bounded.

Corollary 2 *The sequences $\{\mathbf{x}_k\}$ and $\{\boldsymbol{\mathcal{E}}_k\}$is generated by Algorithm 1 is bounded.*

Proof: To annlyze the boundedness of $\Gamma(\boldsymbol{\mathcal{E}}_{k+1}, \mathbf{x}_{k+1}, \boldsymbol{\mathcal{Y}}_k, \mu_k)$, first we can see the following inequality hold becase in step 1 and step 2 we have achieved the globally optimal solutions of the $\boldsymbol{\mathcal{E}}$ and $\mathbf{x}$ subproblems:

$$\Gamma(\boldsymbol{\mathcal{E}}_{k+1}, \mathbf{x}_{k+1}, \boldsymbol{\mathcal{Y}}_k, \mu_k) \leq \Gamma(\boldsymbol{\mathcal{E}}_k, \mathbf{x}_k, \boldsymbol{\mathcal{Y}}_k, \mu_k) \tag{33}$$

We update $\boldsymbol{\mathcal{Y}}$ according to step 3, then we have:

$$\boldsymbol{\mathcal{Y}}_{k+1} = \boldsymbol{\mathcal{Y}}_k + \mu_k(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}) \tag{34}$$

Further, $\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1} = \mu_k^{-1}(\boldsymbol{\mathcal{Y}}_{k+1} - \boldsymbol{\mathcal{Y}}_k)$
There is,

$$\begin{aligned}
\Gamma(\boldsymbol{\mathcal{E}}_k, \mathbf{x}_k, \boldsymbol{\mathcal{Y}}_k, \mu_k) &{=}\Gamma(\boldsymbol{\mathcal{E}}_k, \mathbf{x}_k, \boldsymbol{\mathcal{Y}}_{k-1}, \mu_{k-1}) \\
&+ \tfrac{\mu_k - \mu_{k-1}}{2}\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_k) - \boldsymbol{\mathcal{E}}_k\|_F^2 \\
&+ \langle\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{Y}}_{k-1}, \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_k) - \boldsymbol{\mathcal{E}}_k\rangle \\
&= \Gamma(\boldsymbol{\mathcal{E}}_k, \mathbf{x}_k, \boldsymbol{\mathcal{Y}}_{k-1}, \mu_{k-1}) \\
&+ \tfrac{\mu_k - \mu_{k-1}}{2}\left\|\mu_{k-1}^{-1}(\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{Y}}_{k-1})\right\|_F^2 \\
&+ \left\langle\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{Y}}_{k-1}, \mu_{k-1}^{-1}(\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{Y}}_{k-1})\right\rangle \\
&= \Gamma(\boldsymbol{\mathcal{E}}_k, \mathbf{x}_k, \boldsymbol{\mathcal{Y}}_{k-1}, \mu_{k-1}) \\
&+ \tfrac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2}\|\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{Y}}_{k-1}\|_F^2
\end{aligned} \tag{35}$$

Denote by Θ the bound of $\|\boldsymbol{\mathcal{Y}}_k - \boldsymbol{\mathcal{Y}}_{k-1}\|_F^2$ for k ($k = 1, ..., \infty$). We have:

$$\begin{aligned}\Gamma(\boldsymbol{\mathcal{E}}_{k+1}, \mathbf{x}_{k+1}, \boldsymbol{\mathcal{Y}}_k, \mu_k) \leq \Gamma(\boldsymbol{\mathcal{E}}_1, \mathbf{x}_1, \boldsymbol{\mathcal{Y}}_0, \mu_0) \\ +\Theta \sum_{k=1}^{\infty} \tfrac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2}\end{aligned} \tag{36}$$

Since the penalty parameter $\{\mu_k\}$ satisfies $\sum\limits_{k=1}^{\infty} \frac{\mu_{k+1}}{\mu_k^2} < +\infty$, we have:

$$\sum_{k=1}^{\infty} \frac{\mu_k + \mu_{k-1}}{2\mu_{k-1}^2} \leq \sum_{k=1}^{\infty} \frac{\mu_k}{\mu_{k-1}^2} < +\infty \tag{37}$$

Thus, we know that $\Gamma(\boldsymbol{\mathcal{E}}_{k+1}, \mathbf{x}_{k+1}, \boldsymbol{\mathcal{Y}}_k, \mu_k)$ is also upper bounded. The boundedness of $\{\boldsymbol{\mathcal{E}}_k\}$ and $\{\mathbf{x}_k\}$ can be easily deduced as follows:

$$\begin{aligned}&\|\boldsymbol{\mathcal{E}}_k\|_{\circledast} + \|\mathbf{x}_k\|_2^2 \\ &= \Gamma(\boldsymbol{\mathcal{E}}_k, \mathbf{x}_k, \boldsymbol{\mathcal{Y}}_{k-1}, \mu_{k-1}) + \tfrac{\mu_{k-1}}{2}(\tfrac{1}{\mu_{k-1}^2} \|\boldsymbol{\mathcal{Y}}_{k-1}\|_F^2 \\ &- \left\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_k) - \boldsymbol{\mathcal{E}}_k + \tfrac{1}{\mu_{k-1}}\boldsymbol{\mathcal{Y}}_{k-1}\right\|_F^2) \\ &= \Gamma(\boldsymbol{\mathcal{E}}_k, \mathbf{x}_k, \boldsymbol{\mathcal{Y}}_{k-1}, \mu_{k-1}) - \tfrac{1}{2\mu_{k-1}}(\|\boldsymbol{\mathcal{Y}}_k\|_F^2 - \|\boldsymbol{\mathcal{Y}}_{k-1}\|_F^2)\end{aligned} \tag{38}$$

Because $\Gamma(\boldsymbol{\mathcal{E}}_{k+1}, \mathbf{x}_{k+1}, \boldsymbol{\mathcal{Y}}_k, \mu_k)$ and $\{\boldsymbol{\mathcal{Y}}_k\}$ is bounded , in addition, $\|\boldsymbol{\mathcal{E}}_k\|_{\circledast}$ and $\|\mathbf{x}_k\|_2^2$ all are nonnegative, thus $\{\mathbf{x}_k\}$ and $\{\boldsymbol{\mathcal{E}}_k\}$ generated by the propose algorithm are all bounded.

Theorem 4 *The sequence* $\{\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}\}$ *generated by Algorithm 1 satisfy:*

$$\lim_{k\to\infty} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}\|_{\infty} = 0 \tag{39}$$

Proof: From corollary 1 and corollary 2, we know $\{\mathbf{x}_k\}$, $\{\boldsymbol{\mathcal{E}}_k\}$ and $\{\boldsymbol{\mathcal{Y}}_k\}$ generated by the propose algorithm are all bounded. There exists at least one accumulation point for $\{\mathbf{x}_k, \boldsymbol{\mathcal{E}}_k, \boldsymbol{\mathcal{Y}}_k\}$. Specifically, we have:

$$\begin{aligned}&\lim_{k\to\infty} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}\|_F \\ &= \lim_{k\to\infty} \tfrac{1}{\mu_k}\|\boldsymbol{\mathcal{Y}}_{k+1} - \boldsymbol{\mathcal{Y}}_k\|_F = 0\end{aligned} \tag{40}$$

According to the equivalence property of between norms, we have

$$\lim_{k\to\infty} \|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{A}}(\mathbf{x}_{k+1}) - \boldsymbol{\mathcal{E}}_{k+1}\|_{\infty} = 0 \tag{41}$$

4. Experimental results

Extensive experiments were carried out to illustrate the efficacy of the proposed approach. Essentially, three standard color databases, i.e. the AR, the Georgia Tech and the FET have been addressed. We compare our model with NMR, LRC, SRC, and CRC. (1) NMR [19]: Learning the regression coefficients by incorporating nuclear norm based matrix regression for a probe; (2) LRC [14]: Learning the regression coefficients of each class by linear regression for a probe and divide it into the class whit the minimum reconstruction error; (3) SRC [18]: It is a general classification algorithm for object recognition based on a sparse representation computed by l_1-minization; (4) CRC [23]: Learning the coefficients by collaborative representing based on ridge regression.

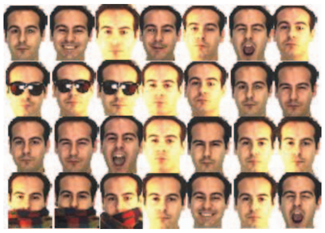

Figure 1. The data for experiments, the first row is the testing data without shelter, the second row is the testing data with shelter, the third row is the training data without shelter, the fourth row is the training data with shelter.

4.1. AR Database

The AR database consists of 126 (70 men and 56 women) subjects. Each of them has 26 images including 14 unobstructed images and 12 occluded (Sunglasses and scarf) images. We conduct two different types of experiments on the AR database. To simulate the real situation, each type of experiments consists of four cases: 1. train samples and test samples are all clean; 2. test samples are contaminated and train samples are clean; 3. test samples are clean and train samples are contaminated; 4. train samples and test samples are all contaminated.

4.1.1 Recognition with natural occlusion

In the first experiment, 7 images randomly chosen form 14 unobstructed images of per subject are used for training data without shelter, the other are used for testing data without shelter. Additional, we choose 7 images randomly which include 4 unobstructed images and 3 occluded images (Sunglasses or scarf) to use for training data with shelter. We choose 7 images randomly which include 4 unobstructed images and 3 occluded images (Sunglasses and scarf) to use for testing data with shelter. These data for experiments are shown in fig. 1. We use the four experiments to test the performance of the proposed model. We conduct face recognition tests and show the recognition rates of LRC, CRC, SRC, NMR and our model in Table 2.

In this set of experiments, generally speaking, the contiguous occlusion caused by sunglass and scarf leads to a low-rank error image. It is more reasonable to use nuclear norm to measure the representation error. NMR takes the

Table 2. Recognition accuracy rate for natural occlusion on the AR database

ACC(%)	1	2	3	4
LRC	0.9548	0.9286	0.8905	0.8536
CRC	0.9678	**0.9583**	0.9214	0.8833
SRC	0.9798	0.9036	0.9583	0.9274
NMR	0.9714	0.8964	0.9548	0.9202
OURS	**0.9917**	0.9429	**0.9833**	**0.9774**

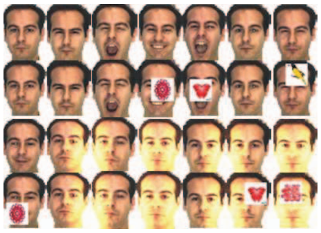

Figure 2. The data for experiments, the first row is the testing data without shelter, the second row is the testing data with shelter, the third row is the training data without shelter, the fourth row is the training data with shelter

low-rank into account while LRC and CRC treat the contiguous occlusion as a random noise point. So the result of NMR is better than RLC and CRC. Meanwhile compared with the NMR, our model is significantly superior, because our model considers the depth information. Our model is lower than SRC in the second experiment, but is better than SRC in the other experiment, which shows that our model is more stable, compare to SRC.

4.1.2 Recognition with artificial occlusion

In the second type experiments, 7 images randomly chosen from 14 unobstructed images of per subject were used for training without shelter, while the other are used for testing without shelter. The testing data with shelter are generated by this way where three images chosen from testing data without shelters randomly are corrupted by a randomly located square block images and the rest of testing data without shelters are unchanged. We can obtain the training data with shelter via corrupting any three images from the training data without shelter by a randomly located square block image and keep the rest of training data without shelters to be unchanged. These data for experiments are shown in fig. 2. We conduct face recognition tests and show the recognition rates of LRC, CRC, SRC, NMR and our model in Table 3.

The difference between these experiments and the previous experiments is that we added an additional occlusion block to the unoccluded images, but our model performs the best among all methods. Particularly, our models is significantly superior to other methods in those cases where the training data or the testing data include occlusion blocks. It shows that our model is more robust to occlusion blocks.

Table 3. Recognition accuracy rate for artificial occlusion on the AR database

ACC(%)	1	2	3	4
LRC	0.9667	0.9214	0.8452	0.8202
CRC	0.9702	0.9607	0.8976	0.9119
SRC	0.9690	0.9179	0.9619	0.9310
NMR	0.9643	0.9405	0.9583	0.9286
OURS	**0.9893**	**0.9821**	**0.9786**	**0.9667**

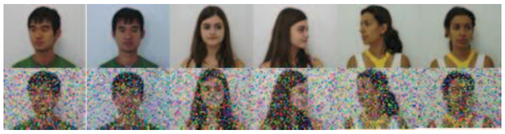

Figure 3. The data for experiments, the first row is the samples without noise, the second row is the samples with noise

Table 4. Accuracy for artificial occlusion on the first sub-database

ACC(%)	1	2	3	4
LRC	0.9300	0.8800	0.8833	0.8667
CRC	0.9066	0.8900	0.8633	0.8700
SRC	0.8967	0.8433	0.8833	0.8433
NMR	0.9033	0.8533	0.8730	0.8670
OURS	**0.9733**	**0.9233**	**0.9300**	**0.8933**

4.2. FEI Database

FEI database includes four sub-databases. Per sub-database consists of 50 subjects and each of them has 14 images. We perform the same experiments on these four sub-databasea to verify the performance of our model respectively. For each sub-database, 8 images chosen randomly from each subject are used for the training data without noise, while the remaining 6 images served as the testing data without noise. In addition, we randomly add noise like salt noise to any three images of the training without noise samples to form the training data with noise. We also randomly add noise like salt noise to any three images of the testing without noise samples to form the testing data with noise. To simulate the real situation, experiment consists of four cases: 1. train samples and test samples are all clean; 2. test samples are contaminated and train samples are clean; 3. test samples are clean and train samples are contaminated; 4. train samples and test samples are all contaminated. These data for experiments are shown in fig. 3. Finally, we put the four sub-databases together and do the same experiments as the sub-databases above. Tables show a detailed comparison of our model with LRC, CRC, SRC, NMR.

From Table 4, Table 5, Table 6, Table 7, Table 8, we can see our method is generally better than other methods. From Table 6, our model is lower than SRC in the second experiment. SRC is based on assumption that the representation errors are pixel-wise and are of independent identical dis-

Table 5. Accuracy for artificial occlusion on the second sub-database

ACC(%)	1	2	3	4
LRC	0.9233	0.8733	0.8533	0.8467
CRC	0.9433	0.8933	0.9200	0.8733
SRC	0.9367	0.8900	0.8900	0.8633
NMR	0.9267	0.9033	0.9000	0.8600
OURS	**0.9800**	**0.9600**	**0.9467**	**0.9333**

Table 6. Accuracy for artificial occlusion on the third sub-database

ACC(%)	1	2	3	4
LRC	0.8933	0.8667	0.8367	0.8333
CRC	0.9133	0.8733	0.8833	0.8400
SRC	0.9100	**0.9667**	0.8600	0.8133
NMR	0.9200	0.8700	0.8767	0.8300
OURS	**0.9633**	0.9367	**0.9400**	**0.9200**

Table 7. Accuracy for artificial occlusion on the fourth sub-database

ACC(%)	1	2	3	4
LRC	0.9267	0.9000	0.8933	0.8700
CRC	0.9366	0.8766	0.8966	0.8433
SRC	0.9267	0.8833	0.8700	0.8333
NMR	0.9367	0.9000	0.8933	0.8567
OURS	**0.9667**	**0.9433**	**0.9467**	**0.9300**

Table 8. Accuracy for artificial occlusion on the FEI database

ACC(%)	1	2	3	4
LRC	0.8692	0.8100	0.7925	0.7608
CRC	0.8517	0.7550	0.7141	0.6417
SRC	0.8567	0.6608	0.7708	0.6142
NMR	0.8108	0.6717	0.7308	0.6092
OURS	**0.9250**	**0.8150**	**0.8508**	**0.7608**

tribution. The representation error image is not a low-rank when pixels of testing sample are contaminated. But our model is better than SRC in the other experiments. It shows that our model is more robust to noise, compared with SRC. From Table 8, our model has the same recognition rate with RLC when the testing data and the training data are all contaminated, but our model is significant superior to RLC in other cases. Meanwhile, as the number of categories increases, the performance of all the methods is dramatically degraded especially NMR, but our model is more stable.

4.3. Georgia Tech Database

The Georgia Tech Database consists of 50 subjects with 15 images per subject. It characterizes several variations such as pose, expression, cluttered background, and illumination. 8 images chosen randomly from each subject are used for the training data without noise, while the remaining 7 images served as the testing data without noise. In addition, we randomly add noise like salt noise to any three images of the testing samples to form the testing data with noise and randomly add noise like salt noise to any three images of the training samples to form the training data with noise. These data for experiments are shown in fig. 4. To simulate the real situation, experiment consists of two cases: 1. train samples and test samples are all clean; 2. test samples are contaminated and train samples are clean. Table 9 shows a detailed comparison of our model with LRC, CRC, SRC, NMR.

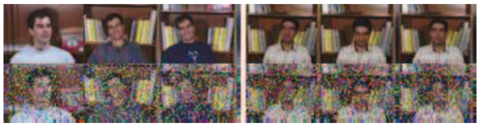

Figure 4. The data for experiments, the first row is the samples without noise, the second row is the samples with noise

Table 9. Accuracy rate for on the Georgia Tech Database

ACC(%)	LRC	CRC	SRC	NMR	**OURS**
1	0.9829	0.9743	0.9743	0.9829	**0.9886**
2	0.9657	0.9371	**0.9868**	0.9771	0.9857

From the Table 9, it is clear that our model performs the best in the first experiment. However, our model performs slightly worse than SRC in the second experiment. NMR and our model all take the whole structural information into account when training images or testing images are in the existence of occlusion, but our model is much better than the NMR in terms of the performance. It shows that our model can retain more structural information of the picture by using the color images directly, which is useful for the task like face recognition.

5. Conclusion

We present a tensor nuclear norm based linear regression for color face recognition and develop an efficient algorithm to solve optimal solution. Our model avoids transforming each color image to a matrix or a vector, thus the multi-dimensional structure information, which is embedded in color image, can be well preserved. Extensive experiments on several standard databases indicate that our model is superior to most state-of-art classifier methods. In the future work, we will consider how to integrate deep learning, which has achieved impressive results [5, 21], and our proposed classifier method into unified framework to further improve classification accuracy

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61773302, Innovation Fund of Xidian University No.10221150004 and the Fundamental Research Funds for the Central Universities.

References

[1] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[2] Xiao Cai, Chris Ding, Feiping Nie, and Heng Huang. On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1124–1132, 2013.

[3] Xiujuan Chai, Shiguang Shan, Xilin Chen, and Wen Gao. Locally linear regression for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 16(7):1716–1725, 2007.

[4] Jen Tzung Chien, Chia Chen Wu, Jen Tzung Chien, and Chia Chen Wu. Discriminant waveletfaces and nearest feature classifiers for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1644–1649, 2002.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[6] Ran He, Wei Shi Zheng, and Bao Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1561–1576, 2011.

[7] Arthure. Hoerl and Robertw. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[8] Misha E. Kilmer and Carla D. Martin. Factorization strategies for third-order tensors. *Linear Algebra and Its Applications*, 435(3):641–658, 2011.

[9] Yuan-Cheng Lee, Jiancong Chen, Ching Wei Tseng, and Shang-Hong Lai. Accurate and robust face recognition from rgb-d images with a deep learning approach. In *The British Machine Vision Conference*, pages 123–124, 2016.

[10] S. Z Li and J Lu. Face recognition using the nearest feature line method. *IEEE Transactions on Neural Networks*, 10(2):439–43, 1999.

[11] Zhouchen Lin, Minming Chen, Leqin Wu, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *Eprint Arxiv*, 9, 2010.

[12] Canyi Lu, Jiashi Feng, Yudong Chen, Liu Wei, and Shuicheng Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(10.1109/TPAMI.2019.2891760):1–1, 2019.

[13] Ze Lu, Xudong Jiang, and Alex Kot. Color space construction by optimizing luminance and chrominance components for face recognition. *Pattern Recognition*, 83:456–468, 2018.

[14] Imran Naseem, Roberto Togneri, and Mohammed Bennamoun. Linear regression for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):2106–2112, 2010.

[15] Jianjun Qian, Jian Yang, Fanglong Zhang, and Zhouchen Lin. Robust low-rank regularized regression for face recognition with occlusion. *Pattern Recognition*, 48(10):3145–3159, 2015.

[16] Qianqian Wang, Fang Chen, Quanxue Gao, Xinbo Gao, and Feiping Nie. On the schatten norm for matrix based subspace learning and classification. *Neurocomputing*, 216:192–199, 2016.

[17] Xing Wang, Meng Yang, and Linlin Shen. Structured regularized robust coding for face recognition. *IEEE Transactions on Image Processing*, 22(5):1753–1766, 2013.

[18] 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*, 31(2):210–227, 2009.

[19] Jian Yang, Lei Luo, Jianjun Qian, Ying Tai, Fanlong Zhang, and Yong Xu. Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):156–171, 2017.

[20] Meng Yang, Lei Zhang, Jian Yang, and D. Zhang. Robust sparse coding for face recognition. In *Computer Vision and Pattern Recognition*, pages 625–632, 2011.

[21] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1290–1299, 2017.

[22] Debing Zhang. Matrix completion by truncated nuclear norm regularization. In *Computer Vision and Pattern Recognition*, pages 2192–2199, 2012.

[23] Lei Zhang and Meng Yang. Sparse representation or collaborative representation: Which helps face recognition? In *International Conference on Computer Vision*, pages 471–478, 2012.

[24] Wen Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *Computer Vision and Pattern Recognition*, pages 2586–2593, 2012.