Pattern Recognition Letters 107 (2018) 41–49

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Robust discriminative nonnegative dictionary learning for occluded face recognition

Weihua Ou[a,b,*], Xiao Luan[c], Jianping Gou[d], Quan Zhou[e,f], Wenjun Xiao[g], Xiangguang Xiong[a], Wu Zeng[h]

[a] School of Big Data and Computer Science, Guizhou Normal Univeristy, Guiyang, Guizhou Province, 550025, PR China
[b] Centre for Quantum Computation and Intelligent System, University of Technology Sydney, Broadway, NSW 2007, Australia
[c] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, PR China
[d] School of Computer Science and Telecommunication Engineering, JiangSu University, ZhenJiang, JiangSu, 212013, PR China
[e] National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing, PR China
[f] School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, PR China
[g] School of Physics and Electronics Science, Guizhou Normal University, Guiyang, China
[h] School of Electric and Electronic Engineering Institute, Wuhan Polytechnic University, Wuhan, PR China

ARTICLE INFO

Article history:
Available online 17 July 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Video surveillance
Face recognition
Occlusion
Nonnegative dictionary learning
Robust feature selection

ABSTRACT

Face recognition in real-world video surveillance needs to deal with a lot of challenges including low resolution, illumination variations, pose changes, occlusions and so on. Among them, occlusions are difficult and have not attracted enough attentions. To address this problem, in this paper, we propose a robust discriminative nonnegative dictionary learning method for occluded face recognition, which estimates the occlusions adaptively and selects the features robustly. Instead of modeling the reconstruction errors using a specific distribution, we estimate occlusions adaptively according to the reconstruction errors and learn different weights for different pixels during the iterative processing. To enhance discriminant ability of the dictionary, we constrain the low-dimensional representations of samples from the same class to be as close as possible and select the discriminative features robustly via $\ell_{2,1}$-norm. For the induced non-convex problem, we reformulate it into local convex optimization subproblem via utilizing the half-quadratic technique and propose new update rules. Extensive experiments are implemented on four benchmark datasets, and the experimental results demonstrate the effectiveness and robustness of the proposed method.

1. Introduction

Face recognition is one of the most active research areas in computer vision and achieves significant progress in recent years with the rapid development of deep learning technique [42,43]. Nowadays, existing methods can achieve satisfied performances in controlled conditions. However, in unconstrained environment, such as video surveillance, face recognition is still unsatisfactory. There are many challenges for the face recognition under the video surveillance, for example, complex illumination, pose changes [6–9], expression, low quality images and occlusions. Among them, occlusions are regarded as the most common and difficult one.

Generally, occlusions [5] can be categorized into following three types: accessories, self-occlusion, objects in front of faces, as shown in Fig. 1. Occlusions impose two kinds of difficulties for the face recognition system. First, the discriminative features are distorted and the intra-class variations are more large. Second, the types of occlusions in practical scenarios are unpredictable and the location, size and shape of the occlusions are unknown.

The intuitive idea for occluded face recognition is to detect the occlusions and then to recognize it on the unoccluded part. Based on this idea, [34] proposed to train a SVM classifier to detect the occluded region and then use the unoccluded area to match the gallery faces with the corresponding area. Similar works refer to [12,13,14]. It is noted that the performance is severely dependent on training samples. [26] proposed the skin color based mask to detect occluded area. However, it is difficult to detect the occlusions with the same color as faces utilizing skin color based mask. For example, the hands might be the same color as faces.

In the past decades, researchers have proposed many methods to deal with occlusions, which can be categorized into reconstruction-based method and local feature learning method.

* Corresponding author at: School of Big Data and Computer Science, Guizhou Normal Univeristy, Guiyang, Guizhou Province, 550025, PR China.
E-mail addresses: Weihua.Ou@uts.edu.au, ouweihuahust@gmail.com (W. Ou).

http://dx.doi.org/10.1016/j.patrec.2017.07.006
0167-8655/

Fig. 1. Some face samples with different occlusions from database TFWM. http://www.thefacewemake.org.

The reconstruction-based method treats occluded face recognition as a signal recovery problem, such as sparse representation based classification (SRC) [48]. However, SRC fails to large contiguous occlusions because the occlusions contain structure information. To remedy it, many occlusion dictionary based learning methods have been proposed. For example, [38] proposed structure constraints based occlusion dictionary learning method to deal with occluded face recognition. [3] proposed the extended SRC (ESRC) by utilizing the intra-class variation to construct the occlusion dictionary. ESRC can handle certain types of occlusions. Recently, [25] proposed kernel extended occlusions dictionary learning to improve the efficiency. Similar works can be referred to [4,21,47,51,59]. The common characteristic of those methods is to learn occlusions dictionary based on the assumptions that both of the non-occluded part and occluded part can be coded over the associated dictionary separately. However, they can only deal with the occlusions provided in the training stage and can not deal with unseen types of occlusions.

Local feature learning approaches are another effective way for occluded face recognition, which does not need the priors of occlusions. Those methods aim to extract features from local areas, and then use locally matching strategy to recognize the faces. A lot of local feature-based methods have been proposed to solve the occlusions problem in face recognition. For example, local subspace learning [33,45], distance measurement learning [44], multi-task sparse learning [31] and so on. Among them, nonnegative matrix factorization (NMF)[29] is one of effective and promising ways, which is consistent with the psychological intuition of combing parts to form a whole. Learning parts-based representation based on NMF provides a new way for robust face recognition under occlusions, and many related methods have been proposed in this direction, which will be reviewed in the next section. The existing local features methods are effective for the occluded face recognition. However, they can not select the discriminative features and few of them can tackle with outlier, which usually corresponding to the occlusions in face recognition.

In this paper, we propose a robust discriminative nonnegative dictionary learning method for occluded face recognition. To deal with the unpredictable occlusions, we estimate the occlusions adaptively according to the reconstruction errors. Instead of modeling the reconstruction errors using a specific distribution, different weights are learned for different pixels during the iterative processing. To enhance discriminant ability of the dictionary, we constrain the low-dimensional representations of samples from the same class to be as close as possible and select the discriminative features robustly via $\ell_{2,1}$-norm. To solve the induced non-convex problem, we utilize the half-quadratic optimization to formulate it into local convex subproblems and propose new update rules. Experimental results on four benchmark datasets demonstrate the learned low-dimensional representations are more robust to occlusions and large magnitude noises than that of the most existing methods.

The main contributions are summarized as follows:

- occlusions can be adaptively estimated according to the reconstruction errors without any priors for occlusions;
- the class mean regularization is proposed to enhance discriminant ability of the dictionary and the $\ell_{2,1}$-norm is utilized to select the features;
- experimental results on benchmark datasets demonstrate effectiveness of the proposed method.

The rest of paper is organized as follows. First, we review related works in Section 2. Then, we present the robust discriminative nonnegative dictionary learning (RDNDL) method in Section 3, and followed by the algorithm in Section 4. Finally, we show the experimental results in Section 5 and conclude this work in Section 6.

2. Related works

In this section, we first present non-negative matrix factorization (NMF) in Section 2.1, and then review the variants of NMF in Section 2.2. We denote matrix by uppercase letter X and denote vector by lowercase letter with right arrow $\vec{x}$. For matrix X, $X_{i,*}$ denotes the ith row and $X_{*,j}$ denotes the jth column, respectively.

2.1. Nonnegative matrix factorization (NMF)

Given nonnegative matrix $X \in \mathbb{R}_+^{m\times n}$ whose columns are feature vectors, NMF decomposes X into product of non-negative basis matrix $W \in \mathbb{R}_+^{m\times r}$ and nonnegative coefficient matrix $H \in \mathbb{R}_+^{r\times n}$, i.e., $X \approx WH$. The objective function of NMF can be formulated as follows:

$$\min_{W,H} \| X - WH \|_F^2 \quad s.t. \;\; W \geq 0, H \geq 0. \tag{1}$$

Thus, the jth column feature vector $\vec{x}_j$ can be approximated by factorization $\vec{x}_j \approx W\vec{h}_j$, where $\vec{h}_j$ is the jth vector of coefficient matrix H. Usually, the projection subspace dimensionality r is lower than that of original m-dimensional subspace, i.e., $r \ll m$.

2.2. NMF variants

The standard NMF has been successful in many applications, however, it is unsupervised and is not suitable for classification. Many variants of NMF has been proposed to improve the discriminant and robustness.

Considering the local geometric structure of data, [1] proposed a graph-regularized NMF (GNMF), in which the geometric structure is described via k-NN graph. Under the graph embedding framework, [49] proposed non-negative embedding (NGE). NGE preserves the favorite similarities and unfavored similarities via intrinsic graph and penalty graph. Under the framework of patch alignment, [18] proposed a non-negative patch alignment (NPA). NPA shows that intrinsical differences of various NMFs are the patches that they build. More similar works can be referred to [23,39,52–54]. All those methods obtain satisfied performances in real applications. However, it is difficult to construct a reliable graph for noisy data, especially for occluded faces.

To improve discriminant of the nonnegative basis, various discriminative constraints on the nonnegative representation coefficients H have been imposed. Fox example, by imposing local sparsity constraints, local nonnegative matrix factorization (LNMF) obtained a localized part-based dictionary [30], which demonstrates that the local basis are more discriminative than that of NMF. Combining linear discriminant analysis and NMF, [57] proposed discriminative NMF, whose effectiveness is demonstrated in face verification. Exploiting the local geometric structure and label information, [16] proposed a manifold regularization and margin maximum NMF. Considering the multi-model distribution of data, [17] proposed a subclass discriminant NMF inspired by the clustering based discriminant analysis. Recently, [58] proposed a robust discriminative nonnegative matrix factorization, which learns each sub-class dictionary separately and constrains the cosine similarity between atoms from different classes.

To remove the effect of outliers and enhance the robustness of NMF, many robust NMFs have been proposed. For example, $\ell_{2,1}$-norm [24], ℓ_1-norm [15], earth mover's distance metric [40] and correntropy induced metric [10] are utilized to measure the reconstruction errors, respectively. Compared to ℓ_2-norm measure, those metrics achieve the better performance for outliers. However, the discriminant is not considered. Obviously, the occluded face images can be regarded as the outliers. In this paper, we proposed discriminative nonnegative dictionary learning method for occluded face recognition through simultaneously considering the reconstruction errors and the discriminant of features. Different from [58], we estimate the occlusions adaptively and select the features robustly via $\ell_{2,1}$-norm.

3. The proposed method

In the section, we present the objective function of our method, which includes class mean regularization term, robust feature selection term and reconstruction error term.

3.1. Class mean regularization

Given nonnegative face images dataset with c classes $X = [X^{(1)}, \ldots, X^{(c)}] \in \mathbb{R}_+^{m\times n}$, and the corresponding label matrix $Y = [Y^{(1)}, \ldots, Y^{(c)}] \in \{0,1\}^{c\times n}$, where each column $\vec{x}_j$ in X is a face image. We aim to seek discriminative nonnegative dictionary $W \in \mathbb{R}_+^{m\times r}$ that satisfies $X \approx WH$, where $H = [H^{(1)}, \ldots, H^{(c)}] \in \mathbb{R}_+^{r\times n}$ are the low-dimensional representations of X. As mention in the introduction, the basic NMF is unsupervised and is not suitable for classification.

For classification, the low-dimensional representations of samples in the same class should be as close as possible. Based on this idea, we propose class mean regularization. For the jth class samples $X^{(j)}$, we denote the corresponding low-dimensional representations as $H^{(j)} = [\vec{h}_1^{(j)}, \ldots, \vec{h}_{n_j}^{(j)}]$, where n_j is the sample number of the jth class and $\sum_{j=1}^{c} n_j = n$. To enhance the discriminant of low-dimensional representations, we encourage the representation of each sample as close as possible to the associated class mean below:

$$\min \sum_{j=1}^{c}\sum_{i=1}^{n_j} \left\| \vec{h}_i^{(j)} - \frac{1}{n_j}\sum_{i=1}^{n_j} \vec{h}_i^{(j)} \right\|_2^2 \tag{2}$$

For each term in the equation (2), we can formulate it as following using matrix trace,

$$\begin{aligned}
&\left\| \vec{h}_i^{(j)} - \frac{1}{n_j}\sum_{i=1}^{n_j} \vec{h}_i^{(j)} \right\|_2^2 \\
&= \mathrm{Tr}\left[\left(\vec{h}_i^{(j)} - \frac{1}{n_j}\sum_{i=1}^{n_j} \vec{h}_i^{(j)} \right)\left(\vec{h}_i^{(j)} - \frac{1}{n_j}\sum_{i=1}^{n_j} \vec{h}_i^{(j)} \right)^T \right] \\
&= \mathrm{Tr}\left[\left(\vec{h}_1^{(j)}, \ldots, \vec{h}_{n_j}^{(j)}\right) L_i^j \left(\vec{h}_1^{(j)}, \ldots, \vec{h}_{n_j}^{(j)}\right)^T \right] \\
&= \mathrm{Tr}\left[H^{(j)} L_i^j (H^{(j)})^T \right]
\end{aligned} \tag{3}$$

where Tr denotes the matrix trace, $L_i^j = \vec{v}_i \vec{v}_i^T, \vec{v}_i = [-\frac{1}{n_j}, \ldots, 1 - \frac{1}{n_j}, \ldots, -\frac{1}{n_j}]^T$ and the ith element of $\vec{v}_i$ is $1 - \frac{1}{n_j}$. Thus, formulation (2) can be rewritten as follows:

$$\begin{aligned}
&\sum_{j=1}^{c}\sum_{i=1}^{n_j} \left\| \vec{h}_i^{(j)} - \frac{1}{n_j}\sum_{i=1}^{n_j} \vec{h}_i^{(j)} \right\|_2^2 \\
&= \sum_{j=1}^{c} \mathrm{Tr}\left[H^{(j)} \left(\sum_{i=1}^{n_j} L_i^j \right) (H^{(j)})^T \right] \\
&= \mathrm{Tr}\left\{ \sum_{j=1}^{c} \left[H^{(j)} \left(L^j \right) (H^{(j)})^T \right] \right\} \\
&= \mathrm{Tr}\left(HLH^T \right)
\end{aligned} \tag{4}$$

where $L = diag\{L^1, \ldots, L^c\} \in \mathbb{R}^{n\times n}$ is a block diagonal matrix and $L^j = \sum_{i=1}^{n_j} L_i^j \in \mathbb{R}^{n_j \times n_j}$. The diagonal element in L^j equals to $1 - \frac{1}{n_j}$, while the non-diagonal element equals to $\frac{2-n_j}{n_j}$.

3.2. Robust features selection based on $\ell_{2,1}$-norm

In classification, robust feature selection is very important for the extraction of meaningful features and elimination of noisy ones, especially for occluded face recognition because the face features are often contaminated by occlusions. Recently, sparsity regularization has been successful in dimensionality reduction, features selection and face recognition [19,35,41]. Compared to ℓ_1 norm, $\ell_{2,1}$ norm can select group features. Based on this, we use following robust feature selection function,

$$\|R^T H - Y\|_F^2 + \beta \|R\|_{2,1}, \tag{5}$$

where β is the regularization parameter, $R = [\vec{r}_1, \ldots, \vec{r}_c] \in \mathbb{R}^{r\times c}$ is the regression coefficient matrix, the $\ell_{2,1}$-norm regularization term $\|R\|_{2,1} = \sum_{i=1}^{r} \|R_{i,*}\|_2$ is to ensure R sparse in rows.

3.3. CIM-based reconstruction errors measurement

For occluded face recognition, it is very important to estimate the occlusions. Traditional ℓ_2-norm has nice mathematical properties, however, it is not the best choice for occluded face recognition. Because the errors induced in the occluded faces is much more complex and it is not suitable to model as a certain distribution.

Recently, the correntropy-induced metric (*CIM*) [32] has been very successful in face recognition [20], feature extraction [2,56] and non-negative matrix factorization [10]. The main idea is to approximate the unknown distribution adaptively via an iterative process. The formulation is defined as

$$CIM(\vec{x}, \vec{y}) = \left\{ k_\theta(0) - \frac{1}{m}\sum_{i=1}^{m} k_\theta[\vec{x}[i] - \vec{y}[i]] \right\}^{\frac{1}{2}},$$

where $\{\vec{x}, \vec{y}\} \in \mathbb{R}^m$ are the samples, and $k_\theta(\cdot)$ is a kernel function [22,55]. In this paper, we only consider Gaussian kernels, i.e., $g_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-x^2}{2\sigma^2}\right)$. The *CIM* assigns large weights for the elements with small errors, while small weights for the large errors which often correspond to the corruption or occlusions.

Utilizing the correntropy-induced metric to measure the reconstruction errors, we obtain the following reconstruction errors term:

$$\min_{W\geq 0,H\geq 0} \mathcal{J}(X,WH), \tag{6}$$

where $\mathcal{J}(X,WH)=\sum_{i,j}\left\{1-g_\sigma\left[X_{ij}-(WH)_{ij}\right]\right\}$.

3.4. Objective function

Combining (4)–(6), we obtain the final objective function for the proposed method as below:

$$\begin{aligned}&\min \mathcal{O}=\mathcal{J}(X,WH)+\alpha\|R^TH-Y\|_F^2+\beta\|R\|_{2,1}\\&+\lambda\mathrm{Tr}\left(HLH^T\right)\\&s.t. W\geq 0, H\geq 0\end{aligned} \tag{7}$$

where α, β, λ are regularization parameters which trade off different parts of the objective function.

4. Optimization

It is very difficult to directly minimize the objective function (7) with presence of the non-convex term. Fortunately, the half-quadratic technique [11] can deal with this problem efficiently by introducing additional auxiliary variables. The main idea of half-quadratic optimization is to reformulate the non-convex term as an augmented objective function in an enlarged parameter space, then finds the local optimum in the augmented parameter space by iterating between the auxiliary variables and optimized variables.

According to the half-quadratic optimization theory [36], we obtain the following equation for the non-convex term:

$$\mathcal{J}(X,WH)=\min_{P_{ij}\in\mathbb{R}_+}\left\{\sum_{i,j}P_{ij}\left[X_{ij}-(WH)_{ij}\right]^2+\varphi(P_{ij})\right\}, \tag{8}$$

where P_{ij} is the associated auxiliary variables and $\varphi(.)$ is the conjugate function of $g_\sigma(x)$.

Now, we can solve problem (7) by recursively optimizing one with the others fixed.

Optimize *P* for given *W, H* and *R*. Given W and H, problem (7) can be solved separately with respect to P_{ij}. Motivated by [50,56], for $E_{ij}=X_{ij}-(WH)_{ij}$, the estimation of P_{ij} is given by:

$$P_{ij}=g_\sigma\left(X_{ij}-(WH)_{ij}\right), \tag{9}$$

where $\sigma^2=\frac{1}{2mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[X_{ij}-(WH)_{ij}\right]^2$.

Optimize *W* for given *H, P* and *R*. Given H and P, the problem (7) can be solved by optimizing each row of W separately as follows,

$$\mathcal{L}(W,\Theta)=\sum_{i=1}^{m}(X_{i,*}-W_{i,*}H)A_i(X_{i,*}-W_{i,*}H)^T+\mathrm{Tr}(\Theta^TW),$$

where $A_i=diag(P_{i,*})\in\mathbb{R}^{n\times n}$, $\Theta=[\theta_{ik}]\in\mathbb{R}^{m\times r}$ is the Lagrange multipliers for the non-negative constraints $W\geq 0$. Setting the partial derivatives of $\frac{\partial\mathcal{L}(W,\Theta)}{\partial W_{ik}}$ to zero and utilizing the KKT conditions $\theta_{ik}W_{ik}=0$, we can get following equation for W_{ik},

$$\left(-2\left(X_{i,*}A_iH^T\right)_k+2\left(W_{i,*}HA_iH^T\right)_k+\theta_{ik}\right)W_{ik}=0.$$

Therefore, the update rule for W_{ik} can be described below:

$$W_{ik}=W_{ik}\frac{\left((X\odot P)H^T\right)_{ik}}{\left((WH)\odot P)H^T\right)_{ik}}. \tag{10}$$

Optimize *H* for given *W, R* and *P*. In this case, the problem (7) is equivalent to minimize following objective function:

$$\begin{aligned}\mathcal{L}(H,\Psi)=&\sum_{j=1}^{n}\left\{\vec{s}_j^TB_j\vec{s}_j\right\}+\lambda HLH^T\\&+\mathrm{Tr}\left(\Psi^TH+\alpha(R^TH-Y)(R^TH-Y)^T\right)\end{aligned} \tag{11}$$

where $\vec{s}_j=\left(X_{*,j}-WH_{*,j}\right)$, $B_j=diag(P_{*,j})\in\mathbb{R}^{m\times m}$, $\Psi=[\psi_{kj}]\in\mathbb{R}^{r\times n}$ is the Lagrange multipliers for the non-negative constraints $H\geq 0$. Setting the partial derivatives of $\frac{\partial\mathcal{L}(H,\Psi)}{\partial H_{kj}}$ to zero and utilizing the KKT conditions $\psi_{kj}H_{kj}=0$, we can get following equation for H_{kj},

$$\begin{aligned}&\left(-2\left(W^TB_jX_{*,j}\right)_k+2\left(W^TB_jWH_{*,j}\right)_k\right)H_{kj}+(\psi_{kj}\\&+2\lambda(HL)_{kj})H_{kj}+2\alpha(R(R^TH-Y))_{kj}H_{kj}=0\end{aligned} \tag{12}$$

By separating L into two parts, i.e., $L=L^+-L^-$, $L_{ij}^+=(|L_{ij}|+L_{ij})/2$, $L_{ij}^-=(|L_{ij}|-L_{ij})/2$, and with some simple calculus, equation (12) leads to the update rule for H_{kj}:

$$H_{kj}=H_{kj}\frac{\left(W^T(X\odot P)+\lambda HL^-+\alpha RY\right)_{kj}}{\left(W^T(WH\odot P)+\lambda HL^++\alpha RR^TH\right)_{kj}}. \tag{13}$$

Optimize *R* for given *W, H* and *P*. Motivated by [35], the derivative of objective function with respect to R is as follows:

$$\frac{\partial\mathcal{O}}{\partial R}=2\alpha H(R^TH-Y)^T+\beta DR \tag{14}$$

where D is a diagonal matrix with $D_{ii}=\frac{1}{2}\sqrt{R_{i,*}R_{i,*}^T+\epsilon}$. By setting (14) to zero, we obtain the following updating rule for regression coefficients R:

$$R=\left(HH^T+\frac{\beta}{2\alpha}D\right)^{-1}HY^T \tag{15}$$

The whole algorithm is summarized in Algorithm 1.

Fig. 2 shows the flowchart of the proposed method, which contains training and testing stage. In training stage, we use Algorithm 1 to learn discriminative dictionary W and regression matrix R. In testing stage, the low-dimensional representation $\vec{h}_t$ of each test sample $\vec{x}_t$ can be obtained on the learned dictionary W, then the predicted class label is given via computing the maximum regression value with the regression matrix. The test algorithm is summarized in Algorithm 2. Compared to the standard NMF, the computational complexity of proposed method in each step is $O(mnr+r^3)$ because the regression matrix is learned at the same time.

Algorithm 1 Robust discriminative nonnegative dictionary learning (RDNDL).

Input:
The non-negative dataset $X\in\mathbb{R}_+^{m\times n}$, label matrix Y, parameters α, β, λ, dimensionality r, iteration number nIter.
Output:
The non-negative dictionary $W\in\mathbb{R}_+^{m\times r}$ and the classification coefficients matrix $R\in\mathbb{R}^{r\times c}$.

1: Randomly initialize W, H and R.
2: **for** i = 1 : nIter **do**
3: Update P by equation (9);
4: Update W by equation (10);
5: Update H by equation (13);
6: Update R by equation (15).
7: **end for**

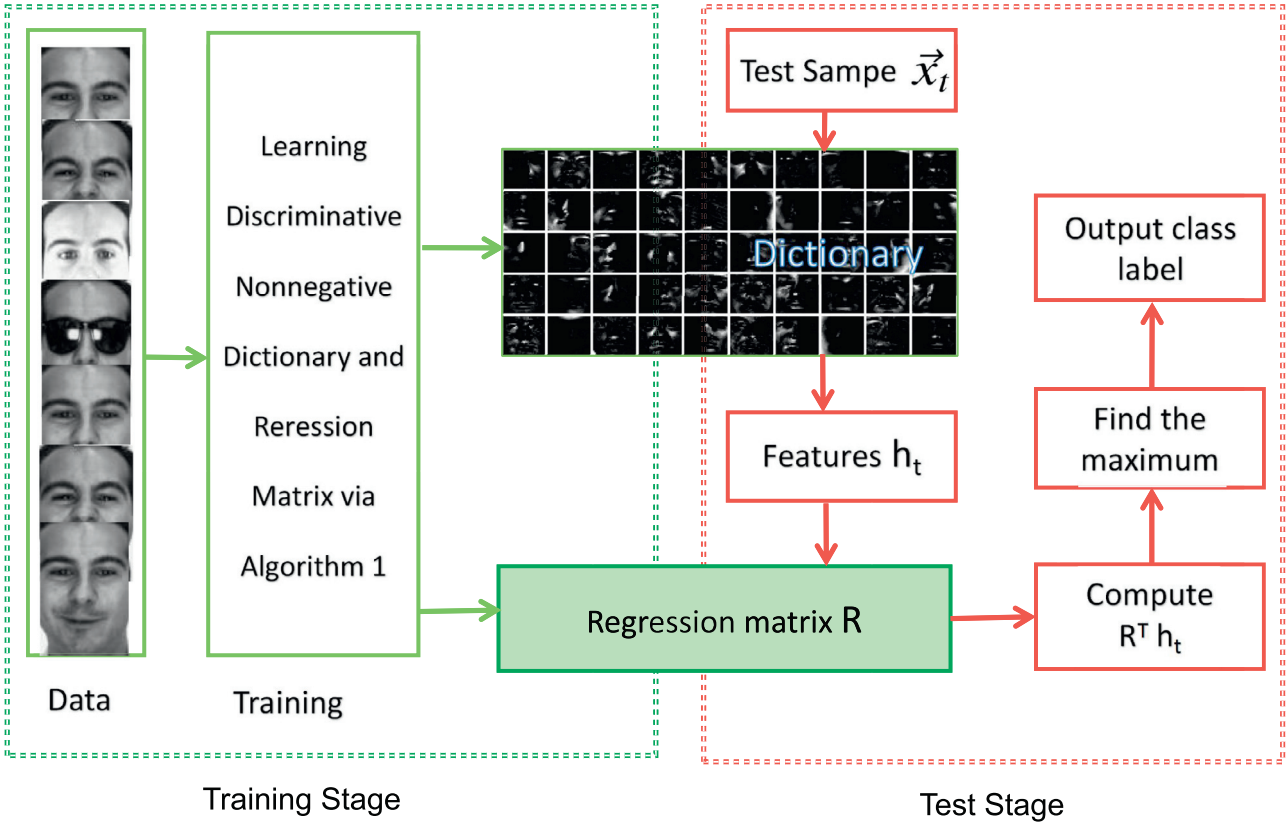

Fig. 2. The flowchart of our proposed method.

Algorithm 2 Occluded face recognition based on RDNDL.

Input:
The non-negative discriminative dictionary $W \in \mathbb{R}_+^{m\times r}$, test sample $\vec{x}_t \in \mathbb{R}_+^m$, classification coefficient matrix R, the iteration number nIter.

Output:
The predicted label of test sample $\vec{x}_t$.

1: Randomly initialize $\vec{h}_t$.
2: **for** i= 1 : nIter **do**
3: Update P by equation (9);
4: Update $\vec{h}_t$ by equation $\vec{h}_t \leftarrow \vec{h}_t \frac{(W^T(\vec{x}_t \odot P)}{(W^T(W\vec{h}_t \odot P)}$;
5: **end for**
6: Output $\vec{x}_t$ label: $l = arg\max_{\{1,2,\ldots,c\}}\{R^T\vec{h}_t\}$.

5. Experiments

In this section, we conduct experiments to evaluate the effectiveness and robustness of the proposed method on CMU-PIE, extended Yale b, AR and Movie Trailer Face datasets. In Section 5.1, we present the datasets, and followed by the compared methods in Section 5.2. Then, we show the experimental results on CMU-PIE and extended Yale b in Section 5.3. Finally, we present the experimental results on the AR and Movie Trailer Face datasets in Section 5.4.

5.1. Datasets

CMU-PIE There are 41,368 images under 68 persons with 13 different poses, 43 different illumination, and 4 different expression in the CMU-PIE dataset. In our experiment, we chose 42 images at pose 27 for each person at different illumination conditions with resolution 32×32. There are 2856 images in all.

Extended Yale b This dataset contains 161,289 images from 38 persons under 9 poses and 64 different illuminations with resolution 32×32. In our experiment, we select about 64 images with the frontal pose under different illumination for each person. There are 2414 images in all.

AR This dataset contains 4000 color frontal images taken from 126 subjects at two separate sessions with different occlusion (sunglasses and scarf), illumination variation and facial expression. In our experiment, we choose a subset containing 50 man and 50 women with 26 images for each subject. We have about 2600 images in all.

Movie trailer face dataset This face video dataset [37] contains 101 movie trailers from YouTube in the 2010 release year that contained celebrities present in the supplemented PublicFig+10 [28] dataset. In total, PubFig+10 consists of 34,522 images and Movie Trailer Face Dataset has 4485 face tracks. We learn dictionary on the PubFig+10 and then test on the Movie Trainler Face Dataset. More details, please refer to [37].

To evaluate the robustness of the proposed method, we add salt&pepper noises and random white block for CMU-PIE and extended Yale b dataset, respectively. For salt&pepper noises, the noises level is varied from {5%, 10%, 20%, 30%, 40%, 50%}, and the block size for the random block is varied from $\{6\times 6, 8\times 8, 10\times 10, 12\times 12, 14\times 14, 16\times 16\}$. Some samples are shown in Fig. 3. Similar to [17], we randomly select half of the images from each person for training, and keep the rest as testing samples.

5.2. Compared methods

We use PCA and NMF as the baselines, and compare with other different nonnegative matrix factorization methods. The details are shown as below:

Fig. 3. Sample images are selected from the CMU-PIE, extended Yale b, and AR datasets, respectively. The first four rows are selected from the CMU-PIE and extended Yale b, respectively. The original images are shown in the left-most column. The upper row for each dataset shows the images corrupted by salt & pepper noises with levels varying from {5%, 10%, 20%, 30%, 40%, 50%}. The lower row for each dataset shows the images occluded by a random white block varying in size from $\{6\times 6, 8\times 8, 10\times 10, 12\times 12, 14\times 14, 16\times 16\}$. The last two rows are selected from AR dataset.

- PCA [46]: unsupervised dimensionality reduction method
- SRC [48]: Sparse representation based classification
- NMF [29]: unsupervised nonnegative dimensionality reduction method
- GNMF [1]: supervised nonnegative matrix factorization with graph regularization
- L_{21}-NMF [27]: NMF combining with $\ell_{2,1}$-loss function
- CIM-NMF [10]: NMF combining with CIM error measurement
- NDLA [18]: NMF combining with local nonnegative discriminant analysis
- our method: NMF combining with CIM error measurement, class mean regularization and robust features selection.

The subspace dimension r ranges from {100, 150, 200, 250, 300}. For PCA, the nearest neighbor rule is used for classification. For the other NMF-based methods, we adopt the algorithm proposed in [40] to test. For GNMF, the neighbor size k is set to 5 and λ is set to 0.001 according to [1]. For NDLA, we set $k_1 = k_2 = 6$, $\lambda = 0.0001$ and utilize multiplicative updating rule by separating the whole alignment matrix L into two parts according to [18]. For our method, we set $\alpha = \lambda = 0.001, \beta = 0.1$ empirically for all the experiments. The maximum iteration number nIter is set to 300 for both training and testing.

5.3. Experimental results on CMU-PIE and extended Yale b

We implement two kinds of experiments on the CMU-PIE and extended Yale b datasets. For the first, the noise levels or the occlusions sizes are fixed and the subspace dimensionality r is varied from 100 to 300, while the other is to set the subspace dimensionality $r = 200$ and to change the noise levels or occlusions sizes.

Table 1
Recognition rates (%) with increase of subspace dimension r on CMU-PIE dataset under fixed noises level. (Salt&Pepper noises level 40%).

Method	Dimensionality					
	100	150	200	250	300	**Avg.**
PCA	66.25	66.58	71.71	72.83	74.02	70.28
NMF	89.92	91.39	91.52	92.58	91.95	91.47
GNMF	91.53	91.95	89.71	90.97	93.28	91.49
L_{21}-NMF	90.55	90.83	90.41	90.13	89.29	90.24
CIM-NMF	95.66	96.22	97.97	98.11	99.23	97.44
NDLA	91.60	91.53	91.57	90.62	91.32	91.33
ours	**95.98**	**98.18**	**98.81**	**99.65**	**99.93**	**98.51**

Table 2
Recognition rates (%) with increase of subspace dimension r on CMU-PIE dataset under fixed occlusions size (occlusion size 10×10).

Method	Dimensionality					
	100	150	200	250	300	**Avg.**
PCA	78.50	79.13	79.13	80.67	81.51	79.79
NMF	85.78	86.62	84.94	85.29	84.59	85.44
GNMF	84.80	86.20	85.57	85.78	86.48	85.77
L_{21}-NMF	84.94	86.69	85.64	86.27	87.04	86.12
CIM-NMF	96.44	97.32	97.25	97.55	97.83	97.28
NDLA	88.94	90.90	90.06	91.11	90.90	90.38
ours	**97.14**	**99.10**	**99.12**	**99.32**	**99.79**	**98.80**

Table 3
Recognition rates (%) with increase of subspace dimension r on extended Yale b dataset under fixed noises level (Salt&Pepper noises level 40%).

Method	Dimensionality					
	100	150	200	250	300	**Avg.**
PCA	31.64	34.73	37.98	39.65	39.98	36.80
NMF	67.78	73.12	74.12	75.21	75.63	73.17
GNMF	69.03	74.37	75.16	77.31	76.42	74.46
L_{21}-NMF	70.45	74.21	73.46	77.30	76.04	74.29
CIM-NMF	67.78	70.70	72.12	73.13	76.79	72.10
NDLA	68.28	73.79	76.96	76.80	77.13	74.59
ours	**71.09**	**79.83**	**81.72**	**86.75**	**87.19**	**81.32**

Table 4
Recognition rates (%) with increase of subspace dimension r on extended Yale b dataset under fixed occlusions size (occlusion size 10×10).

Method	Dimensionality					
	100	150	200	250	300	**Avg.**
PCA	39.07	44.24	44.41	44.99	45.08	43.56
NMF	65.53	68.78	71.45	71.54	71.67	69.80
GNMF	66.53	69.62	72.12	69.12	70.03	69.48
L_{21}-NMF	63.94	67.70	71.78	71.50	71.73	69.33
CIM-NMF	69.28	72.95	74.21	76.29	78.46	74.24
NDLA	74.58	73.33	74.17	77.92	77.50	75.50
ours	**75.54**	**81.75**	**84.64**	**84.75**	**87.98**	**82.93**

The experimental results of the first group are shown at from Tables 1–4. Obviously, the performances of all the methods on the PIE dataset are better than those on the extended Yale b dataset. This might be the illumination variations on the extended Yale b dataset is more complex than that of PIE dataset. On the PIE dataset, the proposed method RDNDL obtains the best performance, and the CIM-NMF gets the second best performance. This demonstrates that the CIM-induced measurement is effective for the occlusions and corruption. However, on the extended Yale b dataset, the performances of CIM-NMF are worse than that of NDLA, while the proposed method still achieves the best performances. Those verifies the robust features selection used in our

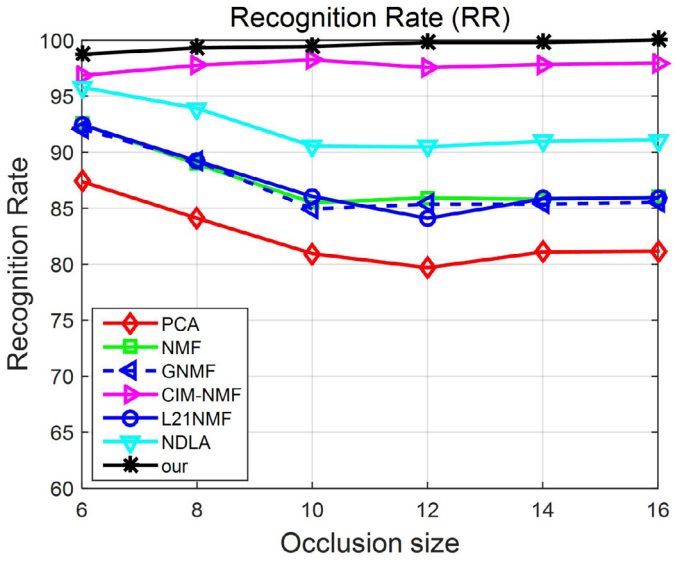

Fig. 4. Recognition rates of different methods on the CMU-PIE dataset with the occlusion size varied from 6×6 to 16×16.

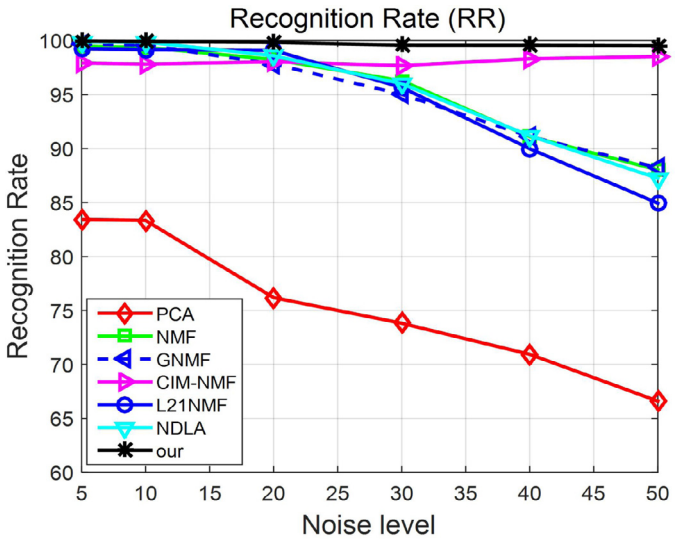

Fig. 5. Recognition rates of different methods on the CMU-PIE dataset with the noise levels varied from 5% to 50%.

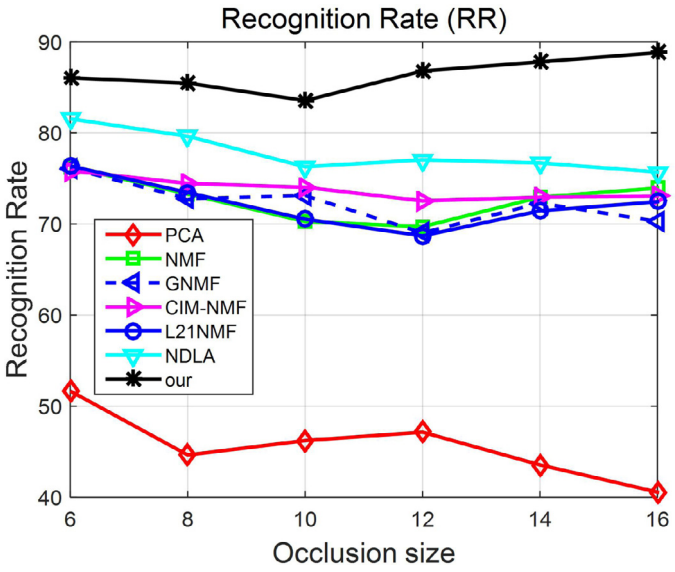

Fig. 6. Recognition rates of different methods on the extended Yale b dataset with the occlusion size varied from 6×6 to 16×16.

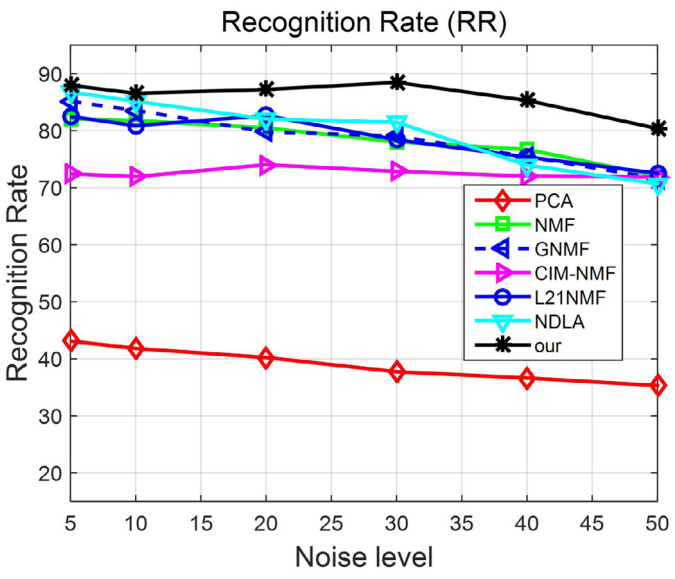

Fig. 7. Recognition rates of different methods on the extended Yale b dataset with the noises level varied from 5% to 50%.

method can efficiently improve the recognition rates for the complex illumination variations.

The second group experimental results are shown at from Figs. 4–7. From Figs. 4 and 5, we can see that the performances of the proposed method and CIM-NMF are better than that of the others, even with increase of the noise level and the occlusion size on the CMU-PIE. Compared to CIM-NMF, our method achieved the best performances. On the extended Yale b dataset, the recognition rate of CIM-NMF is lower than that of NDLA, while our methods obtained the best performances at both occlusion and salt&pepper noise, which are shown in Figs. 6 and 7. This demonstrates combining the CIM-induced measurement and the robust feature selection can improve the robustness. In all, GNMF and L_{21}NMF achieved nearly the same performances on all those different scenarios.

The learned basis of different methods on the extended Yale b dataset are shown in Fig. 8. It can be seen that the learned basis of NMF, GNMF, L_{21}-NMF and NDLA include occlusions, which are bounded by the red boxes. It means that those approaches are not robust to occlusions because they regard occlusions as parts of the faces. However, the CIM-NMF and our method are not affected by the occlusions. Compared to CIM-NMF, the basis learned by our method are more sparse than that of the CIM-NMF. Among them, the basis learned by NDLA are the most sparsest one although they are not robust to occlusions than that of CIM-NMF and our method.

5.4. Results on AR and movie trailer face datasets

For AR dataset, we set the dimensionality $r = 200$ and present the results of all the methods on the Table 5. Compared to CMU-PIE and extended Yale b dataset, the occlusions of sunglass and scarf in the AR dataset are realistic and the occlusion area are more large. From the Table 5, we can see the performance of PCA falls sharply, the best result is only 18.72%. The performances of NMF, GNMF and L_{21}-NMF are also lower than that of the other two datasets. However, NDLA and our methods still achieved over 92%. Compared to NDLA, our method obtained the best performances.

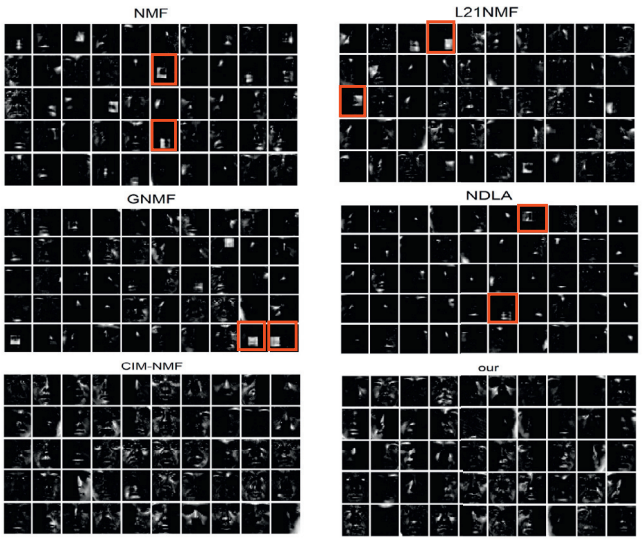

Fig. 8. The learned basis of different methods on the extended Yale b dataset with occlusion size 14×14. Only 50 atoms are shown for each method.

Table 5
Recognition rates (%) of different methods on the AR dataset with dimensionality $r = 200$.

Method	Occlusion	
	Sunglass	Scarf
PCA	18.72	15.62
NMF	67.47	63.46
GNMF	67.27	67.37
L_{21}-NMF	70.07	66.27
CIM-NMF	69.37	61.96
NDLA	93.29	92.09
ours	**95.20**	**94.17**

Table 6
Results (%) of different methods on the movie trailer face dataset.

Method	Results	
	AP (%)	Recall (%)
NN	9.53	0.00
SVM	50.06	9.69
L_2	36.16	0.00
SRC (Voting)	54.88	23.47
MSSRC	58.70	30.23
ours	59.31	30.19

For the move trailer face dataset, according to the settings in [37], we use Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Gabor wavelets to represent each face track after eye alignment. All the descriptors were scaled to unit norm with 1536 dimensions after using Principle Components Analysis (PCA). Combining with three different types descriptors, we obtain 4608 dimensional features for each frame. We set the dimensionality $r = 300$ for our method and show the precision and recall rates of different methods in Table 6, where the results of NN, SVM, L_2, SRC and MSSRC are cited directly from [37]. It can seen that our method achieved comparable performance to the state-of-the-art.

6. Conclusion

In this paper, we propose a robust discriminative nonnegative dictionary leaning method for occluded face recognition. Robustness and discriminant of the dictionary are considered simultaneously. For robustness, the occlusions are estimated via iterative update rule. To the discriminative, we constrain the low-dimensional representations from the same class to be close to the associated class mean and select features via $\ell_{2,1}$-norm. Experimental results demonstrate that the learned representations are more robust to large continuous occlusions than that of the most existing methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61402122, 61461008, 61502067, 61401228, 61502208), the Natural Science Foundation of Guizhou Province (No.[2017]1130,[2014]2124), the 2014 PhD. Recruitment Program of Guizhou Normal University, the China Scholarship Council (No. 201508525007), Foundation of Guizhou Educational Department (KY[2016]027), the Outstanding Innovation Talents of Science and Technology Award Scheme of Education Department in Guizhou Province (QianJiao KY[2015]487), Chongqing Research Program of Application Foundation and Advanced Technology (cstc2015jcyjA40013), Natural Science Foundation of Jiangsu Province (No. BK20150522), China Postdoctoral Science Foundation (No. 2015M581841), and Postdoctoral Science Foundation of Jiangsu Province (No. 1501019A), and the Priority Academic Program Development of Jiangsu Higer Education Institutions, and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology.

References

[1] D. Cai, X. He, J. Han, T. Huang, Graph regularized nonnegative matrix factorization for data representation, TPAMI 33 (8) (2011) 1548–1560.
[2] S. Yu, Z. Cao, X. Jiang, Robust linear discriminant analysis with a Laplacian assumption on projection distribution, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2567–2571.
[3] W. Deng, J. Hu, J. Guo, Extended src: undersampled face recognition via intraclass variant dictionary, IEEE Trans. Pattern Anal. Mach. Intell. 34 (9) (2012) 1864–1870.
[4] W. Deng, J. Hu, J. Guo, In defense of sparsity based face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 399–406.
[5] C. Ding, J. Choi, D. Tao, L.S. Davis, Multi-directional multi-level dual-cross patterns for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2016) 518–531.
[6] C. Ding, D. Tao, Robust face recognition via multimodal deep face representation, IEEE Trans. Multimed. 17 (11) (2015) 2049–2058.
[7] C. Ding, D. Tao, A comprehensive survey on pose-invariant face recognition, ACM TIST 7 (3) (2016) 1–42.
[8] C. Ding, D. Tao, Pose-invariant face recognition with homography-based normalization, Pattern Recognit. 66 (2017) 144–152.
[9] C. Ding, C. Xu, D. Tao, Multi-task pose-invariant face recognition, IEEE Trans. Image Process. 24 (3) (2015) 980–993.
[10] L. Du, X. Li, Y. Shen, Robust nonnegative matrix factorization via half-quadratic minimization, in: ICDM, IEEE, 2012, pp. 201–210.
[11] D. Geman, G. Reynolds, Constrained restoration and the recovery of discontinuities, TPAMI 14 (3) (1992) 367–383.
[12] B. Gu, V.S. Sheng, A robust regularization path algorithm for ν-support vector classification, IEEE Trans. Neural Netw. Learn. Syst. 28 (5) (2017) 1241–1248.
[13] B. Gu, V.S. Sheng, K.Y. Tay, W. Romano, S. Li, Incremental support vector learning for ordinal regression, IEEE Trans. Neural Netw. Learn. Syst. 26 (7) (2015) 1403–1416.
[14] B. Gu, X. Sun, V.S. Sheng, Structural minimax probability machine, IEEE Trans. Neural Netw. Learn. Syst. 2017 (2017) DOI: 10.1109/TNNLS.2016.2544779.
[15] N. Guan, D. Tao, Z. Luo, J. Shawe-Taylor, Mahnmf: manhattan non-negative matrix factorization, Mach. Learn. 1 (5) (2012) 11–43.
[16] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, IEEE Trans. Image Process. 20 (7) (2011) 2030–2048.
[17] S. Nikitidis, A. Tefas, N. Nikolaidis, I. Pitas, Subclass discriminant nonnegative matrix factorization for facial image analysis, Pattern Recognit. 45 (12) (2012) 4080–4091.
[18] N. Guan, D. Tao, Z. Luo, B. Yuan, Non-negative patch alignment framework, TNN 22 (8) (2011) 1218–1230.
[19] R. He, T. Tan, L. Wang, W.-S. Zheng, $\ell_{2,1}$ regularized correntropy for robust feature selection, in: CVPR, IEEE, 2012, pp. 2504–2511.
[20] R. He, W. Zheng, B. Hu, Maximum correntropy criterion for robust face recognition, TPAMI 33 (8) (2011) 1561–1576.

[21] Z. He, S. Yi, Y. Cheung, X. You, Y.Y. Tang, Robust object tracking via key patch sparse representation, IEEE Trans. Cybern. 47 (2) (2017) 354–364.
[22] R. Herbrich, Learning Kernel Classifiers: Theory and Algorithms, MIT Press, 2001.
[23] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Trans. Ind. Electron. 62 (6) (2015) 3742–3751.
[24] J. Huang, F. Nie, H. Huang, C. Ding, Robust manifold nonnegative matrix factorization, ACM TKDD 8 (3) (2014) 11.
[25] K.-K. Huang, D.-Q. Dai, C.-X. Ren, Z.-R. Lai, Learning kernel extended dictionary for face recognition, IEEE Trans. Neural Netw. Learn. Syst. 99 (2016) 1–13.
[26] H. Jia, A.M. Martinez, Support vector machines in face recognition with occlusions, in: CVPR, IEEE, 2009, pp. 136–141.
[27] D. Kong, C. Ding, H. Huang, Robust nonnegative matrix factorization using l21-norm, in: Proceedings of the 20th International Conference on Information and Knowledge Management, ACM, 2011, pp. 673–682.
[28] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 365–372.
[29] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.
[30] S.Z. Li, X.W. Hou, H.J. Zhang, Q.S. Cheng, Learning spatially localized, parts-based representation, in: CVPR, IEEE, 2001, pp. 1–8.
[31] S. Liao, A.K. Jain, S.Z. Li, Partial face recognition: alignment-free approach, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1193–1205.
[32] W. Liu, P. Pokharel, J. Principe, Correntropy: properties and applications in non-gaussian signal processing, IEEE Trans. Sig. Process. 55 (11) (2007) 5286–5298.
[33] A.M. Martínez, Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class, IEEE Trans. Pattern Anal. Mach. Intell. 24 (6) (2002) 748–763.
[34] R. Min, A. Hadid, J.-L. Dugelay, Improving the recognition of faces occluded by facial accessories, in: Automatic Face & Gesture Recognition and Workshops, IEEE, 2011, pp. 442–447.
[35] F. Nie, H. Huang, X. Cai, C.H. Ding, Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization, in: Advances in Neural Information Processing Systems, 2010, pp. 1813–1821.
[36] M. Nikolova, R. Chan, The equivalence of half-quadratic minimization and the gradient linearization iteration, TIP 16 (6) (2007) 1623–1627.
[37] E.G. Ortiz, A. Wright, M. Shah, Face recognition in movie trailers via mean sequence sparse representation-based classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3531–3538.
[38] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, Z. Zhu, Robust face recognition via occlusion dictionary learning, Pattern Recognit. 47 (4) (2014) 1559–1572.
[39] W. Ou, S. Yu, G. Li, J. Lu, K. Zhang, G. Xie, Multi-view non-negative matrix factorization by patch alignment framework with view consistency, Neurocomputing 204 (2016) 116–124.
[40] R. Sandler, M. Lindenbaum, Nonnegative matrix factorization with earth mover's distance metric for image analysis, TPAMI 33 (8) (2011) 1590–1602.
[41] A. Singh, R. Pokharel, J. Principe, The c-loss function for pattern classification, Pattern Recognit. 47 (1) (2014) 441–453.
[42] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: Advances in Neural Information Processing Systems, 2014b, pp. 1988–1996.
[43] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1891–1898.
[44] X. Tan, S. Chen, Z.-H. Zhou, J. Liu, Face recognition under occlusions and variant expressions with partial similarity, IEEE Trans. Inf. Forensics Secur. 4 (2) (2009) 217–230.
[45] X. Tan, S. Chen, Z.-H. Zhou, F. Zhang, Recognizing partially occluded, expression variant faces from single training image per person with som and soft k-nn ensemble, IEEE Trans. Neural Netw. 16 (4) (2005) 875–886.
[46] M. Turk, A. Pentland, Face recognition using eigenfaces, in: CVPR, IEEE, 1991, pp. 586–591.
[47] Y. Wen, W. Liu, M. Yang, Y. Fu, Y. Xiang, R. Hu, Structured occlusion coding for robust face recognition, Neurocomputing 178 (2016) 11–24.
[48] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[49] J. Yang, S. Yang, Y. Fu, X. Li, T. Huang, Non-negative graph embedding, in: CVPR, IEEE, 2008, pp. 1–8.
[50] M. Yang, L. Zhang, J. Yang, D. Zhang, Regularized robust coding for face recognition, TIP 22 (5) (2013) 1753–1766.
[51] S. Yi, Z. Lai, Z. He, Y.-m. Cheung, Y. Liu, Joint sparse principal component analysis, Pattern Recognit. 61 (2017) 524–536.
[52] X. You, W. Ou, C.L.P. Chen, Q. Li, Z. Zhu, Y. Tang, Robust nonnegative patch alignment for dimensionality reduction, IEEE Trans. Neural Netw. Learn. Syst. 26 (11) (2015) 2760–2774.
[53] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multi-view features for image re-ranking, IEEE Trans. Multimed. 16 (1) (2014) 159–168.
[54] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. 99 (2016) 1–11.
[55] S. Yu, X. You, K. Zhao, W. Ou, Y. Tang, Kernel normalized mixed-norm algorithm for system identification, in: Neural Networks (IJCNN), 2015 International Joint Conference on, IEEE, 2015, pp. 1–6.
[56] X. Yuan, B. Hu, Robust feature extraction via information theoretic learning, in: ICML, ACM, 2009, pp. 1193–1200.
[57] S. Zafeiriou, A. Tefas, I. Buciu, I. Pitas, Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, IEEE Trans. Neural Netw. 17 (3) (2006) 683–695.
[58] R. Zhang, Z. Hu, G. Pan, Y. Wang, Robust discriminative non-negative matrix factorization, Neurocomputing 173 (2016) 552–561.
[59] Z.-Q. Zhao, Y.-m. Cheung, H. Hu, X. Wu, Corrupted and occluded face recognition via cooperative sparse representation, Pattern Recognit. 56 (2016) 77–87.