Front. Comput. Sci.
DOI 10.1007/s11704-015-4590-3

RESEARCH ARTICLE

Max-margin non-negative matrix factorization with flexible spatial constraints based on factor analysis

Dakun LIU, Xiaoyang TAN (✉)

Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract Non-negative matrix factorization (NMF) is a popular feature encoding method for image understanding due to its non-negative properties in representation, but the learnt basis images are not always local due to the lack of explicit constraints in its objective. Various algebraic or geometric local constraints are hence proposed to shape the behaviour of the original NMF. Such constraints are usually rigid in the sense that they have to be specified beforehand instead of learning from the data. In this paper, we propose a flexible spatial constraint method for NMF learning based on factor analysis. Particularly, to learn the local spatial structure of the images, we apply a series of transformations such as orthogonal rotation and thresholding to the factor loading matrix obtained through factor analysis. Then we map the transformed loading matrix into a Laplacian matrix and incorporate this into a max-margin non-negative matrix factorization framework as a penalty term, aiming to learn a representation space which is non-negative, discriminative and local-structure-preserving. We verify the feasibility and effectiveness of the proposed method on several real world datasets with encouraging results.

Keywords non-negative matrix factorization, factor analysis, loading matrix, flexible spatial constraints

1 Introduction

Dimensionality reduction or subspace representation plays an important role in data analysis and preprocessing. Principal component analysis (PCA) [1] as a basic method for dimensionality reduction and de-noising has been widely used in various disciplines due to its simplicity and efficiency. After the eigenface [2] was proposed, PCA becomes much more intuitive for understanding the image data. However, there are many limitations of PCA, such as the lack of encoding prior information of data and the simplicity of assumptions made for data noise and manifold. To address these issues, many variants of PCA were proposed, such as probabilistic PCA [3] and structured sparse PCA [4], etc. In computer vision, besides designing more complex models, we are interested in whether the learnt basis vectors are interpretable. Take face images for example, we may expect that each basis vector corresponds to a facial component such as nose or eyes, while the whole face should be the combination of these components.

Non-negative matrix factorization (NMF) [5] is a famous work in representation learning. The main difference between NMF and PCA lies in the former's non-negative constraints on basis matrix and coefficient matrix. It seems not to be a complex model, but it is quite intuitive and consistent with human cognition. Actually, NMF not only overcomes some of the flaws of eigenface, but also can extract local basis images from the data under some conditions. For this reason, NMF can be viewed as a parts-based representation [6] which is widely used in data mining and image analysis [7–12]. However, implicit local constraints in NMF is insufficient by themselves in practice. Prior knowledge of data or features is significant for a good representation learning and many im-

Received December 30, 2014; accepted July 1, 2015

E-mail: x.tan@nuaa.edu.cn

provements on NMF are inspired from the following two aspects.

Firstly, NMF is unsupervised, which means that the learnt features may not be good for classification. Actually, the objective of NMF is to optimally reconstruct the data under non-negative constraints, leading to a set of problem-independent features blind to the class labels. To increase the discriminativity of the learnt feature, one can either use Fisher discriminant criterion [1] or max-margin constraints [13]. Particularly, various discriminative NMF models are introduced based on the Fisher discriminant criterion [14–16] due to its well-understood theoretical properties and good performance in subspace representation. Accordingly, various methods for the optimization of Fisher NMF [17–20] and for the measurement of with-class and between-class matrix [15, 21] are also proposed. The max-margin constraints, on the other hand, are classical method in support vector machines (SVM) [1], but only recently is it embedded into the NMF framework [22, 23]. These works utilize hinge loss to measure the discriminative value of coefficient vectors but relax the non-negative constraints on basis vectors.

Secondly, there lacks an explicit local constraint for NMF. As a result, the local basis images are not unique [24] and sometimes the basis images are not so local [25]. In fact, the locality of basis vector is valuable for image processing — it is well-known that parts-based representation is robust to occlusion and illumination images [26–30]. Therefore, intensive efforts have been made to embed explicit and reasonable local structures into the basis images over the past decade. According to Ref. [31], the spatial constraints used by these works can be roughly divided into two categories: 1) structure-regularized sparsity or correlation, e.g., based on the correlation between features [32], sparseness regularization [25, 33] or Hyper-graph [34]; 2) structure-embedded Euclidean distance of pixels, e.g., using pixels density penalty to constrain the relationship between entries in basis images [35] or using L1-norm to measure the distance between adjacent pixels [36, 37].

Despite the partial success of the above methods, it is worth mentioning that the requirement of good features for discriminative power and locality plays quite different roles [38,39]. Taking face images for example, suppose that a basis image is about forehead, and this local structure is semantic and coincident with human cognition, but it may be useless if using it for gender classification, which makes the subspace dimension corresponding to this basis vector unhelpful. Representation with lots of useless dimensions is not efficient especially for the low-dimensional subspace. On the other hand, without locality constraints, it is not easy to integrate useful spatial structures into the model.

In this paper, we propose a flexible spatial constraint for max-margin non-negative matrix factorization based on factor analysis. One major advantage of this method is that we consider both the discriminative power and the locality of the features at the same time, which is barely the case in the previous research. Particularly, compared with the related NMF models, our model is superior in the following two aspects.

1) Representation is more robust and discriminative. Due to the locality constraint, prior knowledge about the data is integrated in basis vectors, yielding more interpretable representation. Furthermore, an extra basis vector (i.e., the projection) is utilized to undertake the discriminative constraint, such that the confliction goal between reconstruction and discriminant on coefficient vectors can be resolved (see Section 3 for details).

2) The spatial constraint is more flexible. Our spatial constraint is based on a generative model on variables. We learn independent local structures according to the correlation between variables and each latent variable, and embed these into basis images. Meanwhile, different data result in different correlations. So our spatial constraints are adaptive to data, which is more practical compared to the aforementioned rigid methods.

This paper is organized as follows. Related work on NMF is introduced in Section 2. The model and optimization method is described in Section 3. Extensive experiments on two face datasets are given in Section 4 and we conclude this paper in Section 5.

2 Related work

In this section, we review the NMF method and some of its variants with locality or discriminative constraints embedded.

Although the eigenface gives a nice explanation of PCA for face image data, there are negative entries in basis images which are difficult to explain. According to human cognition, all the things are composed of parts. Each part is non-negative and the whole thing is the non-negative combination of these parts. Based on this insight, Lee and Seung proposed a non-negative matrix factorization[1)]:

[1)] In fact, the essential of NMF was proposed much earlier in chemometrics [40] and was reinvented by a Finnish group of researchers in the middle of the 1990s [41]. Lee and Seung popularized it and gave it a better explanation as NMF.

$$\min_{\mathbf{B},\mathbf{h}_i} \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{B}\mathbf{h}_i||_2^2 \quad s.t.\ \ \mathbf{B} \geqslant 0, \mathbf{h}_i \geqslant 0, \tag{1}$$

where x_i is an image, $\mathbf{B}$ is the basis matrix with non-negative constraints, and $\mathbf{h}_i$ is the non-negative coefficient vector of x_i on $\mathbf{B}$. Due to the non-negative constraint, parts-based basis images can be extracted. This can explain, e.g., using a set of gray-scale images, which are data points lying in the positive orthant of feature space. Without loss of generality, some points may lie on the coordinate axes. The objective of NMF is then to find a set of vectors (i.e. basis vectors) in the positive orthant to optimally reconstruct these data points. As entries in coefficient vector are the weights for each basis vector, the basis vectors are mainly responsible for the structures of data points. So there must be zero entries in basis vectors to reconstruct those data points on the axes, and the localized basis images would emerge from that.

However, the non-negative constraint on basis matrix $\mathbf{B}$ and coefficient matrix $\mathbf{H}$ does not necessarily lead to parts-based basis images. In order to ensure the locality of basis images, various of extra constraints are proposed. Localized NMF (LNMF) [32] is one of the previous researches imposing explicitly local constraints onto the objective of NMF, focusing on the orthogonality between basis vectors and the expressiveness of coefficient vectors, i.e., $\|\mathbf{B}^T\mathbf{B}\|_1$ and $-trace(\mathbf{H}\mathbf{H}^T)$. Hoyer [25] proposed an indirect way to extract local basis images, by constraining the sparsity of basis vectors and coefficient vectors.

The spatial location of pixels in a two-dimensional image is an important prior which is worth being exploited. Based on the assumption that pixels adjacent in the image should not be dispersed in basis images, Zheng et al. [35] proposed to combine a pixel dispersion penalty with NMF. The pixel dispersion penalty is measured as follows:

$$D(\mathbf{b}_i) = \mathbf{b}_i^T\{\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{x'=1}^{a}\sum_{y'=1}^{b} l([y,x],[y',x']) \times e_{y,x}e_{y',x'}^T\}\mathbf{b}_i, \tag{2}$$

where $[y, x]$ and $[y', x']$ are two coordinates in basis images $\mathbf{b}_i^{2D}$ respectively, $l([y, x], [y', x'])$ is the L1-norm distance measure, δ and $e_{y,x}$ are the indicator functions. Intuitively this term ensures that if two pixels are adjacent in the images, they should both be in the same local areas of basis images and hence the responses of them should be large. Otherwise, if the two pixels are far away from each other in the images, only one of them could be in the local areas at most.

Instead of using the L1-norm to measure the distance of all the pixels, Chen et al. [36] proposed a modified Neumann discretization [42] to penalize the correlation of adjacent pixels [36], which is described as follows,

$$\mathbf{D}_1^l = p_l \begin{pmatrix} -1 & 1 & & & 0 \\ & -1 & 1 & & \\ & & \cdots & & \\ & & & -1 & 1 \\ 0 & & & & -1 \end{pmatrix},$$

$$\mathbf{D}_2^l = p_l \begin{pmatrix} -1 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & & \cdots & & \\ & & 1 & -2 & 1 \\ 0 & & & 1 & -1 \end{pmatrix},$$

$$G_m^2(\mathbf{b}) = \sum_{l=1,2}\frac{1}{p_l}\sum_{\tau=1}^{p_l}\|\mathbf{D}_m^l\mathbf{b}^{(\tau,l)}\|^2, m = 1, 2, \tag{3}$$

where $G_m^2(\mathbf{b})$ is the penalty on the smoothness of the basis vector $\mathbf{b}$, $p_l, l = 1, 2$ is the dimension of row or column, and $\mathbf{b}^{(\tau,l)}$ is a sub-vector of $\mathbf{b}$ corresponding to either the τ-th row or the τ-th column of the rectangle. Note that all these work [35, 36] can be thought of as a rigid way to impose locality constraints since the spatial parameters have to be set beforehand. Instead, the information about spatial regularity of the data is learnt automatically in this work.

The information about the distribution structure of the whole data set can also be useful in data representation [43, 44]. Among others, discriminative NMF (DNMF) [15] is a typical supervised non-negative matrix factorization. The constraints are based on the Fisher discriminant criterion, i.e., the coefficient vectors of within-class should be close and those of between-class should be far away from each other. The within-class scatter S_w and the between-class scatter S_b are defined as follows:

$$\mathbf{S}_w = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{n_i}\sum_{j=1}^{n_i}(\mathbf{h}_j - \mathbf{u}_i)(\mathbf{h}_j - \mathbf{u}_i)^T, \tag{4}$$

$$\mathbf{S}_b = \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1}^{C}(\mathbf{u}_i - \mathbf{u}_j)(\mathbf{u}_i - \mathbf{u}_j)^T, \tag{5}$$

where C is the class number, n_i is the data number of class i, $\boldsymbol{u}_i$ is the mean value of class i.

Another type of supervised NMF was based on the max-margin constraint, i.e., max-margin semi-NMF (MNMF) [23]. The discriminant is based on the hinge loss of coefficient vectors. Meanwhile, the non-negative constraints on basis vectors is relaxed, so it is called semi-NMF. By contrast,

we do not make such relaxation, due to the fact that the non-negativeness is a natural property for some data types (such as face images considered here).

3 Max-margin non-negative matrix factorization with flexible spatial constraints

3.1 The motivations and the proposed model

In this section, we give details on how to impose both the locality structure and discriminative constraints simultaneously onto the NMF model.

Firstly, from the perspective of matrix factorization, data and dimension is equivalent in data matrix X, hence NMF could be regarded as not only data reconstruction with non-negative constraints but also feature reconstruction. Therefore, it is natural to use the method of Laplacian regularization (e.g., [45]) to encode the preferred correlation between features. Before giving the details of this, let us describe how and what kind of discriminative constraints can be incorporated into the model.

Secondly, although many works on discriminative NMF are based on the Fisher discriminant criterion, in our opinion, this strategy has one limitation which is difficult to deal with. That is, the discriminative constraints imposed on the coefficient vectors through fisher criterion are inherently conflicting with the original data reconstruction constraints on the same set of coefficient vectors by NMF objective. To understand this, we write the relationship between coefficient vectors and basis images as follows [5],

$$\mathbf{B}_{kl} = \mathbf{B}_{kl} \sum_j \frac{\mathbf{X}_{kj}\mathbf{H}_{lj}}{(\mathbf{BH})_{lj}}. \tag{6}$$

This equation shows that the coefficient vectors have a direct influence on the structure of basis images. It turns out that the discriminative constraints imposed on h make the model be in favor of basis images with some spatial locality [46]. Unfortunately, this conflicts with the goal of data reconstruction constrains of the original NMF model (Eq. (1)) which prefers a global basis vector.

One way to address this issue is to use an alternative project vector w to encode the discriminative information of the coefficient vectors instead of penalizing them through scatter matrix [15]. Actually this project vector can be thought of as an extra basis besides B, which allows us to impose locality constraints on the basis images of NMF without worrying about the aforementioned discriminativity-reconstruction dilemma.

With the above two aspects in mind, we propose the following max-margin non-negative matrix factorization with flexible spatial constraints:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{h}_i,(\mathbf{w},\mathbf{b})} \ & \frac{1}{2}\sum_i ||\mathbf{x}_i - \mathbf{B}\mathbf{h}_i||_2^2 + \frac{\alpha}{k} trace\left(\mathbf{B}^T\mathbf{L}\mathbf{B}\right) \\ & + \sum_i max(1 - y_i(\mathbf{w}^T\mathbf{h}_i + b), 0) \\ s.t. \quad & \mathbf{B} \geqslant 0, \mathbf{h}_i \geqslant 0, ||\mathbf{w}||_2^2 \leqslant 1, \end{aligned} \tag{7}$$

where $\boldsymbol{x}_i$ and $\mathbf{h}_i$ are respectively a data vector and its coefficient vector, $\mathbf{L}$ is the Laplacian matrix learnt for spatial constraints, α is the regularization parameter and k is the subspace dimension. Note that we use the hinge loss to maximize the function margins of coefficient vectors. This model bears some similarity with that of max-margin semi-NMF (MNMF) [23], but there are two major differences between these two:

1) MNMF relaxes the non-negative constraint on basis vectors and we do not. One reason for this is that for image data, all the features are naturally non-negative, e.g., the local structures of a face in basis images should also be non-negative. Hence in our opinion, preserving the non-negativeness of basis vectors is of significant importance.

2) There is no explicit local constraint on MNMF. Actually the relaxation of non-negativeness of basis vectors in MNMF makes locality of learned basis images even worse. By contrast, our model enjoys a flexible spatial constraint on the basis images, learnt automatically from the data based on factor analysis, as described below.

3.2 Learning flexible spatial constraints via factor analysis

In Section 2, we reviewed several methods to explicitly embed spatial structures of image data into the NMF model. However, these methods have one limitation in common: as the position of pixels in images are fixed for different data, the embedded spatial structures are essentially the same, independent of the content of images. Hence it is interesting to investigate alternative methods with flexible spatial constraints adaptive to the data.

The key idea of our method is as follows. We think of features as the realizations of a group of random variables, and hence the local structures is actually the correlations of these variables and we adopt the classic factor analysis (FA) method [47] to capture such correlations. The FA explains the observed covariance structures by assuming that the variables of data are generated by latent factors. Formally, it can be formulated as follows:

$$\begin{aligned} & \boldsymbol{x} = \boldsymbol{f} \times \boldsymbol{Z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \\ \Leftrightarrow \ & cov(\boldsymbol{X}) = \boldsymbol{F} \times \boldsymbol{F}^{\boldsymbol{T}} + \Lambda_{\varepsilon}, \end{aligned} \tag{8}$$

where $\boldsymbol{x}$ is a data vector, $\boldsymbol{\mu}$ is the mean vector of the data, $\boldsymbol{\varepsilon}$ is called uniqueness and Λ_ε is its covariance matrix, $\boldsymbol{f}$ is called factor loadings which is a column of loading matrix $\boldsymbol{F}$, while in the second equation, $cov(\boldsymbol{X})$ is the covariance matrix of data, and $\boldsymbol{Z}$ is called factors (or latent variables). The factor obeys $N(0, I)$ and the uniqueness obeys $N(0, \sigma^2)$.

To derive the correlation among variables, one can read them from loading matrix. Particularly, the variables corresponding to higher loadings in each $\boldsymbol{f}$ vector are generated by the same factor, and hence they can be thought of as correlated. In other words, each latent factor represents one locality structure of interest. However, the raw loading vectors yielded from factor analysis are not ideal for our purpose, so two further processing steps are needed.

3.2.1 Factor rotation

Usually more than one entry in each row of loading matrix has high loading values, and this feature overlapping would weaken the locality. Fortunately, due to the independence between latent variables, the loading matrix is rotation-invariant, and one can use this property to remedy this problem. In this work, we adopt the equamax method [47], which aims to find a orthogonal rotation matrix such that each variable is correlated with only one latent variable, as follows:

$$\begin{aligned} &\min_{\boldsymbol{R},\boldsymbol{T}} \sum_i \sum_j (\boldsymbol{R}_{ij}^2 - \mu_i)^2 \\ &s.t.\ \ \boldsymbol{R} = \boldsymbol{T}\boldsymbol{F}, \boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}, \end{aligned} \tag{9}$$

where $\boldsymbol{R}$ is the rotated loading matrix, $\boldsymbol{T}$ is the orthogonal rotation matrix to be found, and μ_i is the squared mean loading of each row of $\boldsymbol{F}$.

3.2.2 Thresholding

After rotation, each loading vector (the same size of the input image) encodes a particular spatial pattern of interest. We can refine it using a thresholding method to exclude some of loadings (e.g., those lower than the predefined threshold) out of the spatial pattern. However, choosing an appropriate threshold value is difficult in general. Here we adopt a k-means-based strategy with spatial information incorporated.

Specifically, we first choose the position corresponding to the highest loading in each loading vector as the anchor point, which can be viewed as the indicator of where the local structure is in a loading image. Then we construct a distance matrix based on the L2-norm, encoding the pairwise distance between the position of each pixel and the anchor point. The distance matrix is further constructed to be a similarity matrix $\boldsymbol{S}$ as follows,

$$S_{ij} = \begin{cases} \dfrac{1}{\|p_{i,j} - p_0\|_2}, & \text{if } p_{i,j} \neq p_0; \\ 1, & \text{otherwise,} \end{cases} \tag{10}$$

where $p_{i,j}$ is the coordinate of a pixel in the loading image, and p_0 is the coordinate of the anchor point.

To this end, we augment each loading vector with the similarity vector (i.e., the vectorized $\boldsymbol{S}$) obtained above side by side, and use the normal k-means algorithm to cluster these 2-dimensional data elements into two clusters. Finally we binarize them by setting all the data elements belonging to the first cluster (i.e., with higher loading values) to 1 and those in the other cluster 0. Note that there are other natural choices for clustering, e.g., a good alternative worth trying is the mean-shift clustering which finds the mode of distribution of values in a loading vector, but this alternative is beyond the scope of this work. Figure 1 illustrates the local structures (shown in the light area) found using this k-means-based thresholding method on the AR face dataset [48].

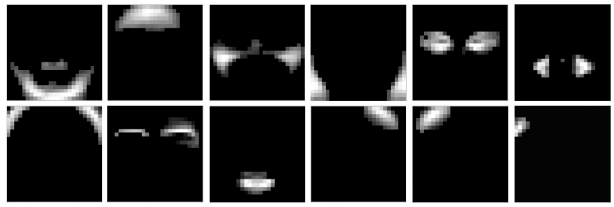


Fig. 1 Loading faces on the expression images of AR database, where the light areas are the local structures found

3.2.3 Laplacian reguarlization

As the final step to impose flexible spatial constraints, we need to map the transformed loading matrix to a laplacian matrix used for regularization (c.f., Eq. (7)). Particularly, a Laplacian matrix comes from a similarity matrix, that is,

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}, \tag{11}$$

where, $\boldsymbol{L}$ is the Laplacian matrix, $\boldsymbol{A}$ is a similarity matrix with element $\boldsymbol{A}_{ij}$ and $\boldsymbol{D}$ is a diagonal matrix,

$$D_{ij} = \begin{cases} \sum_k A_{ik}, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

In our case, each loading vector $\boldsymbol{f}_k$ represents a local structure of interest and we construct a separate similarity matrix A^k for each of them, which is as follows,

$$A_{ij}^k = \begin{cases} 1, & \text{if } f_{k,i} = f_{k,j} = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

where $f_{k,i}$ denotes the ith element of the vector $\boldsymbol{f}_k$. The equation means that any two elements of $\boldsymbol{f}_k$ in the same local structure are correlated and hence being similar to each other.

Suppose that we learn M latent factors from the data, we accordingly construct M similarity matrices $A^k, k = 1, 2, \ldots, M$, one for each (Eq. (13)). These are further mapped to M Laplacian matrices $L_k, k = 1, 2, \ldots, M$ via Eq. (11).

Our next task is to use these matrices to impose the spatial constrains onto the objective of NMF (c.f., Eq. (7)). Ideally, different spatial structures should be imposed on different basis images. Unfortunately, we cannot know which spatial prior should be applied to which basis image if no further information is available. In this work, we take a simple average pooling strategy and leave the more complex basis-specific constraints problem to the future. Particularly, here we assume that at least some of the basis images should meet the requirement of the locality constraints, hence the final Laplacian matrices L used for regularization can be calculated as $L = \frac{1}{M}\sum_k L_k$.

In our implementation, a weighted average pooling strategy is adopted,

$$\boldsymbol{L} = \sum_{k=1}^{M} \alpha_k \boldsymbol{L}_k, \tag{14}$$

where the weight α_k is defined as follows,

$$\alpha_k = \frac{d}{V_k \times R_k}, \tag{15}$$

where d is the dimension of loading factors, V_k is the total size of correlated features in the kth loading factor (i.e., the number of elements being "1" in that factor) while R_k is the number of connected regions. In other words, we prefer a locality constraint that has small number of correlated features and small number of connected regions.

To this end, our laplacian regularization used for flexible spatial constraints can be understood as follows,

$$tr\left(\mathbf{B}^T\mathbf{L}\mathbf{B}\right) = \sum_i \boldsymbol{b}_i^T \mathbf{L} \boldsymbol{b}_i, \tag{16}$$

$$= \sum_i \boldsymbol{b}_i^T (\sum_{k=1}^{M} \alpha_k \boldsymbol{L}_k) \boldsymbol{b}_i, \tag{17}$$

$$= \sum_k \alpha_k \, tr(\boldsymbol{B}^T \boldsymbol{L}_k \boldsymbol{B}). \tag{18}$$

In words, we prefer the basis images $\boldsymbol{B}$ learnt by NMF to satisfy every spatial constraint according to its quality.

The algorithm of learning Laplacian matrix is summarized in Algorithm 1.

Algorithm 1 Learning Laplacian matrix for spatial contraints

Input:

Training Data $\boldsymbol{X}$;

Steps:

1. Calculate loading matrix $\boldsymbol{F}$ (Eq. (8));
2. Perform factor rotation on $\boldsymbol{F}$ according to Eq. (9);
3. Normalize the rotated loading matrix (denoted as $\boldsymbol{F}_{rot}$);
4. Use k-means to cluster each column of $\boldsymbol{F}_{rot}$ and do the thresholding to find out spatial structures in each loading vector of $\boldsymbol{F}_{rot}$;
5. Calculate similarity matrix $\boldsymbol{A}^k$ according to Eq. (13);
6. Map $\boldsymbol{A}^k$ to Laplacian matrix $\boldsymbol{L}_k$ via Eq. (11);
7. Calculate the weights of each loading faces via Eq. (15);
8. Derive the final Laplacian matrix $\boldsymbol{L}$ via Eq. (14).

3.3 Optimization

In order to simplify the optimization, we transform our model to the following formulation:

$$\begin{aligned} \min_{\boldsymbol{B},\boldsymbol{h}_i,(\boldsymbol{w},b),\{\xi_i\}} \quad & \frac{1}{2}\sum_i \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{h}_i\|_2^2 + \frac{\alpha}{k} trace\left(\boldsymbol{B}^T\boldsymbol{L}\boldsymbol{B}\right) \\ & + \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_i \xi_i \\ s.t. \quad & \boldsymbol{B} \geqslant 0, \boldsymbol{h}_i \geqslant 0, \xi_i \geqslant 0, y_i(\boldsymbol{w}^T\boldsymbol{h}_i + b) \geqslant 1 - \xi_i, \\ & i = 1, \ldots, n. \end{aligned} \tag{19}$$

This objective can be alternatively optimized. Specifically, all the variables are divided into three groups: the coefficient matrix ($\boldsymbol{H}$), variables about max-margin projection ($\boldsymbol{w}, b, \xi_i$) and the basis matrix ($\boldsymbol{B}$). Two of them are fixed and the remaining one is optimized alternatively.

3.3.1 Update the projection vector and the coefficient matrix

As the basis matrix and the coefficient matrix can be initialized via the method of matrix factorization, the basis matrix and the coefficient matrix are fixed firstly. This is the inner alternative optimization, and can be done in two steps:

- Update the projection vector and slack variables. When the coefficient matrix and the basis matrix are fixed, the objective is simply a standard support vector machine:

$$\begin{aligned} \min_{(\boldsymbol{w},b),\{\xi_i\}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{j=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geqslant 0, y_i(\boldsymbol{w}^T\boldsymbol{h}_i + b) \geqslant 1 - \xi_i, i = 1, \ldots, n. \end{aligned} \tag{20}$$

- Update the coefficient matrix. When other variables are fixed, the optimization of the coefficient matrix is trans-

formed to quadratic programming:

$$\min_{\{\boldsymbol{h}_i\}} \frac{1}{2}\sum_i \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{h}_i\|_2^2$$
$$s.t.\ \ \boldsymbol{h}_i \geqslant 0, y_i(\boldsymbol{w}^T\boldsymbol{h}_i + b) \geqslant 1 - \xi_i, i = 1, \ldots, n, \quad (21)$$

and the objective can be decoupled further because of the independence of data:

$$\min_{\boldsymbol{h}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{B}\boldsymbol{h}\|_2^2$$
$$s.t.\ \ \boldsymbol{h} \geqslant 0, y(\boldsymbol{w}^T\boldsymbol{h} + b) \geqslant 1 - \xi. \quad (22)$$

The Lagrangian of above objective function is:

$$L = \|\boldsymbol{x} - \boldsymbol{B}\boldsymbol{h}\|^2 - \alpha^T\boldsymbol{h} - \beta[y(\boldsymbol{w}^T\boldsymbol{h} - b) - 1 + \xi], \alpha, \beta > 0, \quad (23)$$

where α and β are lagrangian multipliers, specifically α lagrangian multipliers vector.
Under the KKT conditions, we get:

$$\begin{cases} 2\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{h} - 2\boldsymbol{B}^T\boldsymbol{x} - \alpha - \beta y\boldsymbol{w} = \boldsymbol{0}, \\ \boldsymbol{1}^T\boldsymbol{h} = \boldsymbol{0}, \\ y(\boldsymbol{w}^T\boldsymbol{h} - b) - 1 + \xi = 0. \end{cases} \quad (24)$$

Transform Eq. (24) into matrix form and we get:

$$\begin{pmatrix} 2\boldsymbol{B}^T\boldsymbol{B} & \boldsymbol{1}^T & y(\boldsymbol{w}^T\boldsymbol{h} - b) - 1 + \xi \\ \boldsymbol{1}^T & \boldsymbol{0} & 0 \\ y\boldsymbol{w}^T & \boldsymbol{0} & 0 \end{pmatrix} \times \begin{pmatrix} \boldsymbol{h} \\ \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ yb + 1 - \xi \end{pmatrix}. \quad (25)$$

where $\boldsymbol{1}$ is a unit vector whose size is the same as $\boldsymbol{h}$, $\boldsymbol{0}$ is the zero vector. Then we can derive $\boldsymbol{h}$ by solving Eq. (25).

3.3.2 Update the basis matrix

The model is transformed to a non-negative matrix factorization with spatial constraint:

$$\min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{H}\|_F^2 + \frac{\alpha}{k} trace\left(\boldsymbol{B}^T\boldsymbol{L}\boldsymbol{B}\right)$$
$$s.t.\ \ \boldsymbol{B} \geqslant 0. \quad (26)$$

Due to the non-negative constraints, projected gradient descent method is adopted to solve it.

Specifically, the gradient of Eq. (26) is:

$$Grad = 2\boldsymbol{B}\boldsymbol{H}\boldsymbol{H}^T - 2\boldsymbol{X}^T\boldsymbol{H}^T - 2\alpha\boldsymbol{L}\boldsymbol{B}. \quad (27)$$

So we get $\boldsymbol{B}_{new} = \max(\boldsymbol{B} - \lambda Grad, 0)$, where λ is the step length.

The algorithm on max-margin non-negative matrix factorization with flexible spatial constraints based on factor analysis is summarized in Algorithm 2:

Algorithm 2 Max-margin non-negative matrix factorization with flexible spatial constraints based on factor analysis

Input:

Training Data and parameter: Data Matrix $\boldsymbol{X}$, the Laplacian matrix about the spatial structure $\boldsymbol{L}$, parameter of the regularization α;

Initialization: Initialize the basis matrix $\boldsymbol{B}^0$ and the coefficient matrix $\boldsymbol{H}^0$, let $t = 0$;

Steps:

Repeat

Let $s = 1$, $\boldsymbol{B} = \boldsymbol{B}^t$, $\boldsymbol{H}^s = \boldsymbol{H}^t$;

Repeat

Fix $\boldsymbol{B}$ and $\boldsymbol{H}^s$, estimate $(\boldsymbol{w}^{s+1}, b^{s+1})$ and $\{\xi_i^{s+1}\}$ via Eq. (20);

Fix $\boldsymbol{B}$, $(\boldsymbol{w}^{s+1}, b^{s+1})$ and $\{\xi_i^{s+1}\}$, estimate $\boldsymbol{H}^{s+1}$ via Eq. (22);

$s = s + 1$;

Until reaches the maximal iteration number;

Let $t = t + 1$, $\boldsymbol{H}^t = \boldsymbol{H}^s$, $\boldsymbol{w}^t = \boldsymbol{w}^s$, $b^t = b^s$ and $\xi_i^t = \xi_i^s$;

Learning the new basis matrix B^t via (**??**);

Until Eq. (19) converges or reaches the max. iter. number

Return: $\boldsymbol{B}^t$, $\boldsymbol{H}^t$, $\{(\boldsymbol{w}^t, b^t)\}$, $\{\xi_i^t\}$.

3.3.3 On the convergence of the algorithm

The objective is non-convex and usually we can only wish to obtain a sub-optimal solution with local minimum. In our implementation we initialize the algorithm using the method of Ref. [35]. In addition, as it is a triple alternative optimization and the results of inner alternative optimization are sensitive to those of the outer alternative optimization, the maximal iteration number of inner optimization is set to be much smaller than that of the outer loop. In experiments, we set the iteration number of inner alternative optimization to 30, and 100 for the outer optimization. Figure 2 illusrates a typical convergence curve of the objective function on the AR face dataset.

4 Experiments

In this section, in order to demonstrate the effectiveness of the learnt locality and discrimination, we give our experimental results on two publicly available face datasets, i.e., the AR face database [48] and the extended YaleB database [49] for face recognition, USPS digit database [50] for handwriting digit recognition and KTH [51] database for action recognition.

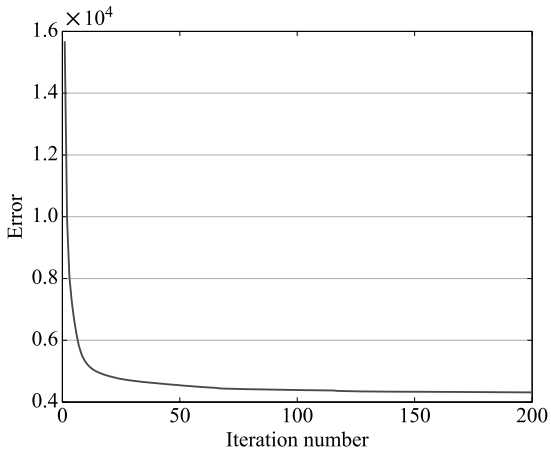


Fig. 2 Convergence curve of the proposed model on the AR dataset

4.1 Databases and settings

• AR database [48] There are 126 people composed by 70 males and 56 females in AR database. Every person has 26 face images, where the first 13 images are taken under four kinds of expression, three kinds of illumination and two kinds of occlusion. The remaining 13 images are taken in another time under the same condition. Here we use 100 people (50 males and 50 females) for experiments. All the images would be cropped and down sampled to 25×25 in experiments.

• Extended YaleB database [49] There are 16 128 face images of 28 individuals in extended YaleB. The images were taken under hine kinds of poses and 64 kinds of illumination conditions. For illumination face images, the light source direction with respect to the camera axis is at different degrees azimuth and degrees elevation. We use illumination images for experiments. All the images are down sampled to 20×25 in pixel in our experiments.

• USPS digit database [50] The dataset contains 9 298 images of 16×16 recording handwriting digits from 0 to 9, as shown in Fig. 3. Traditionally, it has been used in a splitting of 7 291 images for training and 2 007 images for testing. However, these two sets are actually collected in slightly different ways leading that the images in the test set are much harder than those in the training set [52]. In order to get rid of the impacts on local structures from outliers, we concatenate both sets, randomly reshuffel the images and divide images of each class into three parts. Two-thirds of images in each class are for training and the rest part for test.

• KTH action database [51] This video database contains six kinds of human actions(boxing, hand waving, hand clapping, jogging, running and walking) performed by 25 subjects in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. In our experiment, we "naively" chosen (not time-scaled) nine frames from each scenarios of the subjects. In order to align the frames better, we extract bounding boxes of size 80×64 around the objects. Part of derived images are shown in Fig. 4.

Fig. 3 Samples in USPS digit database

Fig. 4 Samples in KTH action database

For performance comparison, we use the original NMF algorithm and its several variants, including LNMF [32], NMF with sparseness constraints (NMFsc) [25], Spatial NMF [35], Spatial non-negative component analysis (Spatial NCA) [35], max-margin semi-NMF (MNMF) [22] and Fisher non-negative matrix factrization (FNMF) [14].

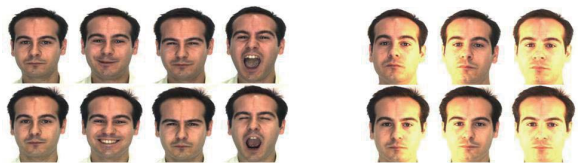
In classification, we use SVM as the classifier. Since the experiments are multi-class classification, one-versus-all criterion is used. The labels of test coefficient vectors are predicted by max-win. Additionally, since FNMF is not suitable for multi-class classification, all-vs-all strategy is specifically used for FNMF.

Parameters of these methods are set to be same in both databases, and described as follows: α in our method is set to 0.7; the structure constraints on basis images and coefficient vectors in LNMF are 0.1; the sparsity degrees of the basis vectors and coefficients are both 0.5 in NMFsc; the param-

eter for pixel dispersion penalty is 0.1 in spatial NMF and spatial NCA; and the parameter for S_w and S_b is 1 in FNMF.

4.2 Face recognition under illumination variations

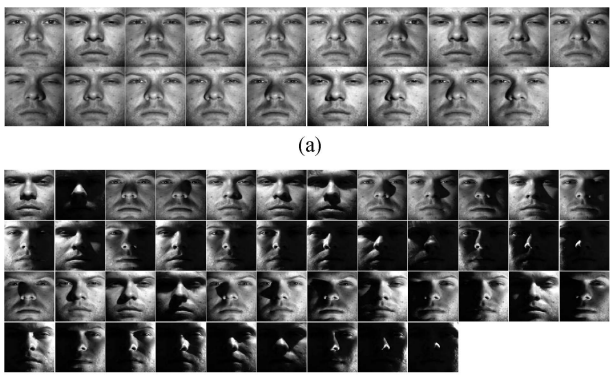
The first series of experiments are about face recognition under illumination variations, which is one of the main challenges for face system designers. Specifically, on the AR database, we randomly choose six face images from each subject with expression variations (but without any lighting changes) for training and six faces with lighting changes for testing. Figure 5 demonstrates some samples of one subject for training and test, respectively.



(a) (b)

Fig. 5 Samples for (a) training and (b) test in the AR database

While on the extended YaleB database, two groups of face images with different lighting conditions are constructed respectively for training and test. Particularly, we choose 19 illumination images whose light source direction with respect to the camera axis is less than 35 at degrees azimuth and elevation for training, while the remaining 45 faces with different lighting conditions for testing. Figure 6 illustrates some face images on this dataset.



(a)

(b)

Fig. 6 Samples for (a) training and (b) test in the extended YaleB database

Figure 7 gives mean accuracies (of ten experiments in total) of various methods on the two databases as a function of different subspace dimensions. It can be seen from the figures that the proposed method consistently performs better than the compared methods on both databases. Particularly, we have the following observations.

1) Compared with two NMF variants with spatial constraints, i.e., Spatial NMF [35] and Spatial NCA [35]: The accuracy of our method is about 8.0% higher than these two on the AR lighting subset (94.6% vs. 85.0% and 85.9%), while on the extended YaleB lighting set, our method is also superior to them by over 7.0% in terms of accuracy. This reveals that although explicit spatial constraints are useful, e.g., on the AR lighting subset, the accuracy of the NMF baseline algorithm is about 79.7%, which is much worse than both Spatial NMF and Spatial NCA, our method with flexible spatial constraints effectively improves the performance.
2) Compared with the two NMF variants with discriminative constraints, i.e., Fisher non-negative matrix factrization (FNMF) [14] and max-margin semi-NMF (MNMF) [22]: The figure shows that exploiting supervision information is helpful to improve the discriminative performance of the learnt basis, e.g, on the AR subset, the methods of FNMF and MNMF respectively improve the performance from 79.7% of the baseline to 88.67% and 86.67%, while our method outperforms both discriminatively trained NMF variants by 6.0%. Additionally, one can see that the MNMF works worse than both our method and FNMF, one possible explanation is that it drops the non-negative constraints on the basis, which are important for non-negative data type such as image data.
3) Compared with other NMF variants, i.e., LNMF [32] and NMF with sparseness constraints (NMFsc): Although both LNMF and NMFsc methods improve performance on both datsets over the original NMF algorithm, such improvement is limited (less than 3.0% on both datasets) compared to other variants of NMF. Note that both methods impose very general constraints on the NMF model, e.g., basis orthogonality or sparsity, without explicitly exploiting useful domain knowledge (e.g., spatial structure for images) or side information (e.g, data labels). On the other hand, our model takes both factors into account and achieves the best performance among the compared ones.

4.3 Face recognition with partial occlusions

Due to variety of occlusion size and randomness of occluded parts, recognizing the identity of occluded face images is another challenging issue on face recognition. In AR database, the occlusion is a scarf or glasses in Fig. 8(a). Since there are no occlusion images in extended YaleB database, partial oc-

clusion is simulated by randomly removing some rectangular patches from test samples (c.f., Fig. 8(b)). The size of occlusion patches are 20% of the image size in extended YaleB database.

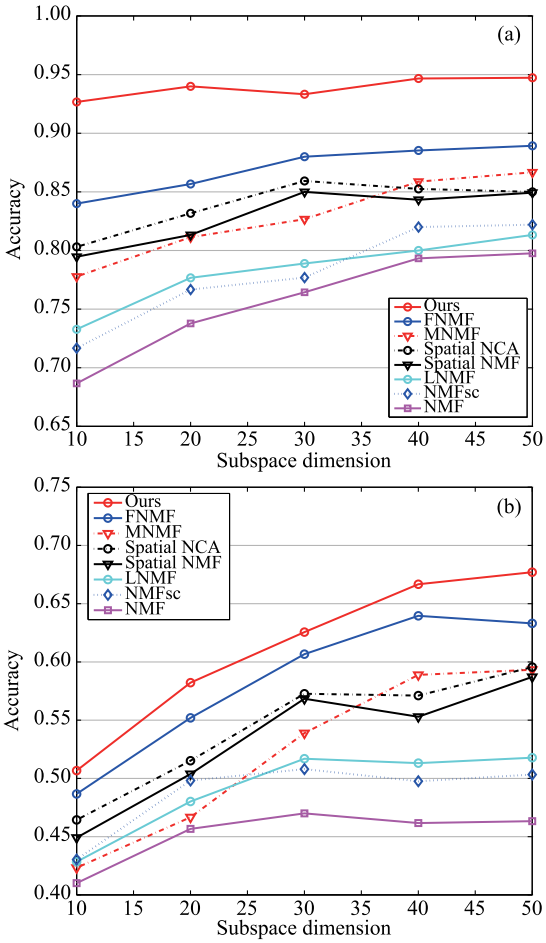


Fig. 7 Performance comparison of our method with various methods. (a) Mean accuracy on the AR database; (b) Mean accuracy on the extended YaleB database

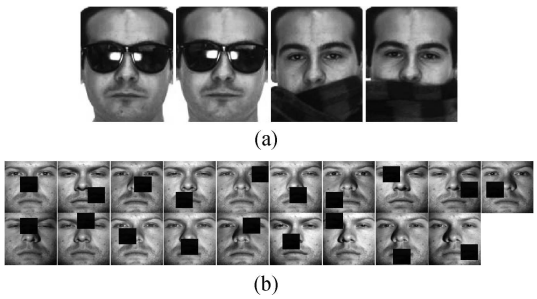


Fig. 8 Samples for test in (a) AR and (b) extended YaleB database

Mean accuracies on different subspace dimensions are demonstrated in Figs. 9(a) and 9(b). Similar to face recognition on illumination, our method on partial occlusions outperforms compared methods. Particularly, for the two NMF variants with spatial constraints, the accuracies on AR occlusion subset are 69.1% and 72.6%, while our accuracy is 80%. On the extended YaleB occlusion set, our method is superior to them by about 10.0%. Meanwhile, compared with NMF variants with discriminative constraints, the improvements of our method are about 5% and 8% on the two subsets, respectively. Due to the deficiency of discriminative and spatial constraint, the performance of other NMF variants are not comparable to that of discriminative NMF and spatial constrained NMF. Overall, compared with the performances on the AR subsets, all the methods have better accuracies on extended YaleB occlusion set due to the simplicity of occlusion .

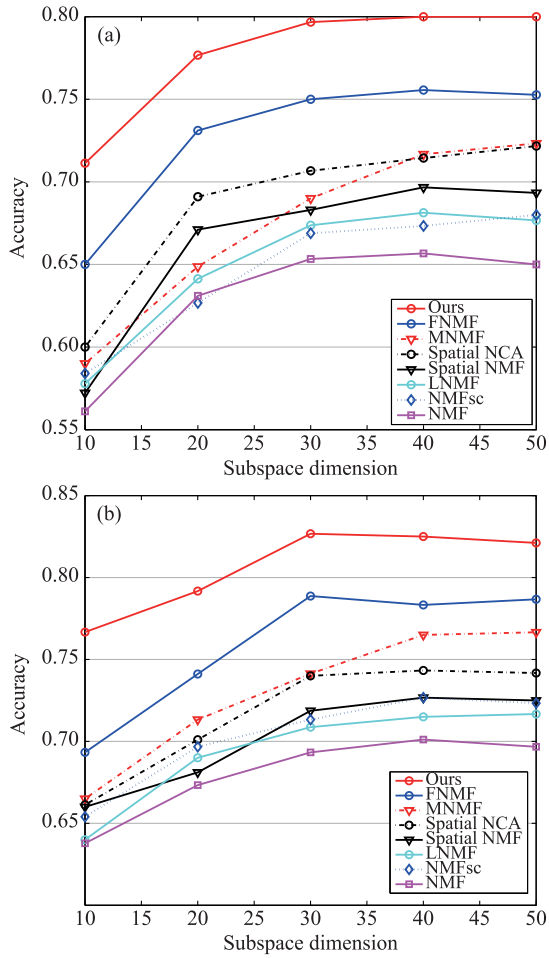


Fig. 9 Mean accuracy in (a) AR database and (b) extended YaleB database on occlusion images

4.4 Digit recognition

Besides face recognition, digit recognition is another common application to check the robustness of representation. The complexity mainly comes from the variants of digits written by different people and similarities between some specifical digits such as 2 and 7, 1 and 7, etc. Hence, the

robustness of representation is more dependent on spatial constraints and local structures. Meanwhile, each digit has its own label, and the discriminative power is another vital property for good representation. However, some variants of NMF in comparison lack these considerations. Specifically, LNMF and NMFsc do not have spatial constraints, while spatial NMF is not so flexible. In this series of experiments, comparisons are mainly focused on NMF, Spatial NCA, FNMF, MNMF and ours.

Tabel 1 gives the results of accuracies in USPS databas. It can be seen that the proposed method performs the best among the compared methods consistently over various dimensions of projected subspace. This reals that imposing both spatial constraints and discriminative constraints are important for this task. Figure 10 shows the basis images learnt by various methods. One can see that the basis images of our method effectively capture the most discriminative local structures of handwritten digits, while the local structures in basis images of MNMF, for example, are difficult to understand, due to the lack of non-negative constraints.

Table 1 Accuracies in USPS database/%

Dim.	NMF	Spatial NCA	FNMF	MNMF	Ours
10	63.62	81.71	88.71	88.11	91.83
20	66.54	89.73	89.91	91.73	93.64
30	69.32	91.58	90.60	93.97	97.00
40	71.88	92.70	94.37	92.30	96.19
50	70.35	93.22	93.50	92.37	94.82

4.5 Action recognition

In this section, we show how to use middle-level descriptors to improve the robustness of our method against misalignment, which could be a major challenge for any algorithm that tries to exploit the spatial layout of the objects. Particularly, we conduct our experiments on the KTH action database [51]. As shown in Fig. 4, the manually cropped images are not aligned well to each other, which may seriously influence the performance.

To deal with this issue, we use the HoG features [53] as the feature descriptor. The HoG features are extracted by creating a non-overlapping spatial grid with size of 8×8, which improves its tolerance against misalignment. However, the original HoG representation is essentially built on concatenated histograms over each cells, without preserving the spatial information of the images. Therefore, instead of concatenating those local histograms at each cell, we transform each image into nine feature maps according to the nine orientational bins per histogram in the Hog space. For an input image with size of $H \times W$ in pixels, the size of the feature map is $\lceil \frac{H}{8} \rceil \times \lceil \frac{W}{8} \rceil$. Then, we run our algorithm separately on each of the nine feature maps and fuse the results at the decision level.

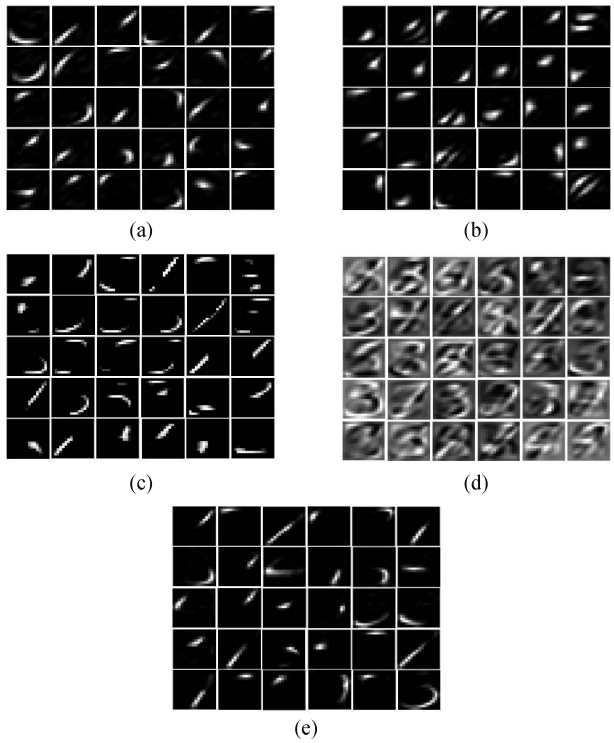


Fig. 10 The basis images of different models in USPS digit database. (a) NMF; (b) Spatial NCA; (c) FNMF; (d) MNMF; (e) ours

Tabel 2 gives the results of accuracies in KTH database. It can be seen that compared to the gray-scaled features, working on the HoG features significantly improves the performance despite of the misalignment on this dataset.

Table 2 Accuracies in KTH database

Method	NMF	MNMF	Ours(gray)	Ours(HoG)
Accuracy/%	67.1	71.6	28.7	74.2

4.6 Discussions

In this section, we give some discussions on details of the proposed method.

4.6.1 The contribution of each component

The proposed method contains several components, mainly about non-negative constraints and spatial constraints. Figure 11 illustrates the effect of removing each of these two constraints in turn while leaving the remaining in place (the comparison is thus against our full model) on the lighting datasets of AR and extended Yale B, respectively. In general, each of constraints is beneficial, but most performance loss occurs if

both constraints are removed. This indicates that both of these constraints are necessary in practice.

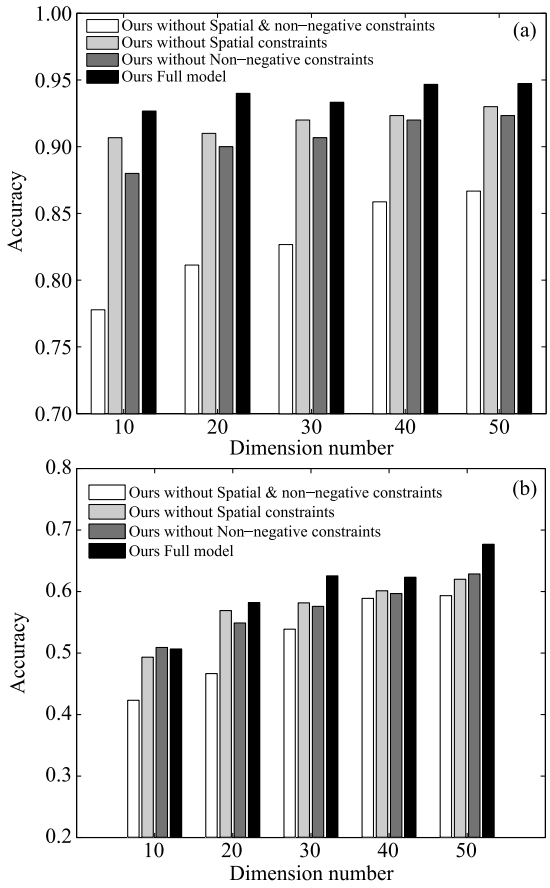


Fig. 11 Influence of the individual constraint of our model on the lighting datasets of (a) AR database and (b) extended Yale B database

4.6.2 Quality of basis images in face recognition

The aim of representation learning is to learn a new efficient representation of data. As the most relevant part to representation, basis vectors are crucial to the system performance. There are two main aspects reflecting the quality of basis vectors. One is the locality of basis images. If local basis images are semantically meaningful, the corresponding representation would be more interpretable. Furthermore, locality is vital to the robustness of representation. The other one is related to the reconstruction error of basis images, which reflects how much characteristic of data is inherited to basis vectors.

To this end, we use three kinds of criteria to measure the quality of basis images: 1) Pixel based mean squared error (PMSE). PMSE is defined as mean squared error (MSE) divided by the pixel number to measure the mean reconstruction error for each pixel. Therefore the smaller PMSE is, the better it is for reconstruction; 2) Sparseness degree(SD) [25]. SD measures the average number of zero entries in basis vectors. Sparseness is the prerequisite of locality but not all sparse basis images have local structures; 3) Pixel dispersion degree (PDD) [35]. PDD measures the spatial compactness of variables in basis vectors. The smaller PDD becomes, The more compact the response regions are in basis images. Note that locality requires that the neighboring entries in basis images should be compactly distributed, and hence the density of high intensity variables in basis vectors could indirectly reflect the locality.

Additionally, one can jointly use SD and PDD to measure the locality of basis vectors. Specifically, a big SD value and a small PDD value usually indicate scattered structures in basis images, while a big PDD value and a small SD value could mean a big connected region in basis images. Additionally, in order to merge all these quality measurements, we define quality of basis (QoB) as QoB = normalized PMSE + normalized PDD − normalized SD, and the smaller the better.

Considering the negativity of basis images of MNMF, its quality is not measured here. Meanwhile, since the basis images of Spatial NCA and Spatial NMF are quite similar, we choose Spatial NCA as the representation of NMF with spatial constraint because of its better performance. Since it is difficult to obtain locality and good reconstruction jointly in the low-dimensional subspace, experiments in this part are conducted on the subspace dimensions of 10 and 30, respectively. Tables 3 and 4 give the results of basis images in AR database and extended YoleB database, and Figs. 12 and 13 demonstrate all the basis images learnt by the compared methods.

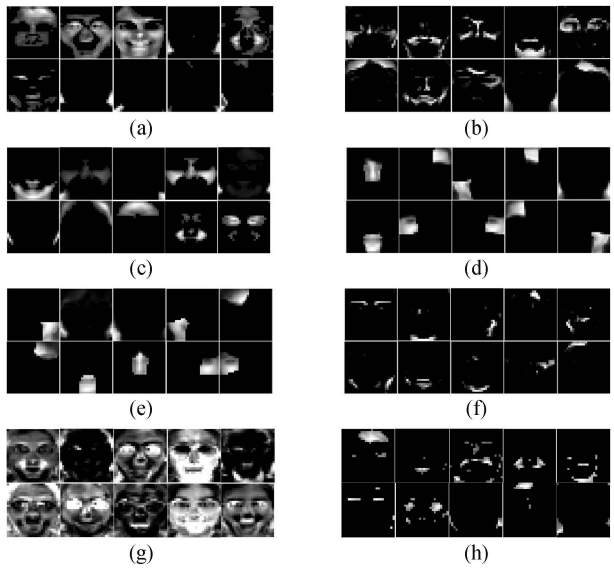


Fig. 12 The basis images of different models in AR database: (a) NMF, (b) LNMF, (c) NMFsc, (d) Spatial NCA, (e) Spatial NMF, (f) FNMF, (g) MNMF, (h) ours

Table 3 Analysis on basis images in AR database

Method	NMF	LNMF	Spatial NCA	FNMF	Ours
PMSE	**125**	1 532.7	1 003.8	1 450.31	363.1
SD	0.282 3	0.63	0.548	**0.672 7**	0.61
PDD/10^6	8.22	32.2	**3.72**	33.7	29.33
QoB	0.034	0.793 2	**–0.024 3**	0.713 8	0.06

Table 4 Analysis on basis images in extended YaleB database

Method	NMF	LNMF	Spatial NCA	FNMF	Ours
PMSE	**317.18**	1 242.1	1 046.95	1 251.71	511.4
SD	0.350 6	**0.645 3**	0.583 4	0.603 3	0.480 8
PDD/10^7	8.19	4.45	**0.745**	5.10	1.85
QoB	0.500 5	0.385 8	0.061 1	0.462 6	**0.006 0**

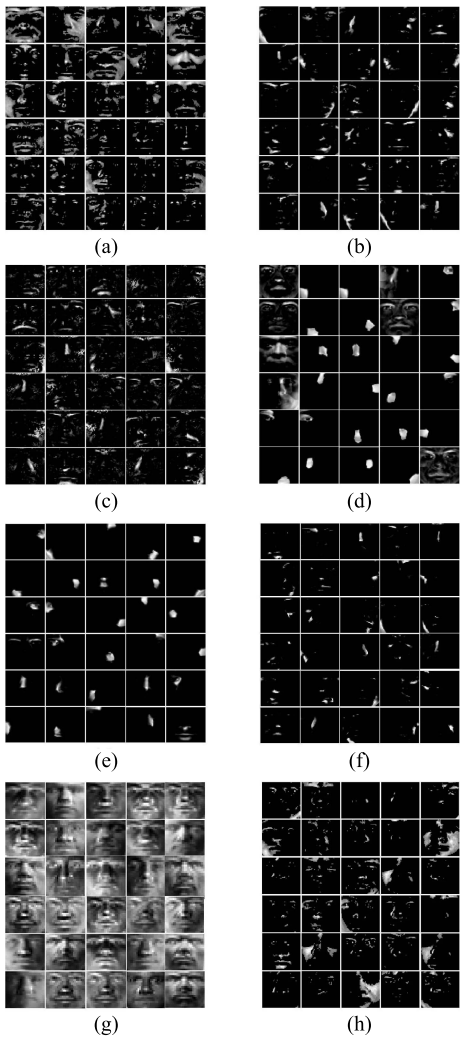


Fig. 13 The basis images of different models in extended YaleB database. (a) NMF; (b) LNMF; (c) NMFsc; (d) Spatial NCA; (e) Spatial NMF; (f) FNMF; (g) MNMF; (h) ours

Several observations can be made from these results:

1) Spatial constraint is an efficient way to derive local basis images. To measure the locality, we could use the difference of normalized PDD and normalized SD. It is found that the NMF variants with spatial constraints have better locality. Specifically, the locality values of these methods (i.e.,NMF, LNMF, Spatial NCA, FNMF, and ours) are respectively 0.15, 0.11, –0.67, 0, –0.04 on AR database, and 1, –0.5, –0.7, –0.31, –0.35 on extended YaleB database. From this, we see that spatial NCA always has the best locality value on the two databases, while our method ranks the second and third best on the two databases respectively.
2) Flexibility of spatial constraints leads to self-adaptive basis images. For Spatial NCA and Spatial NMF, the spatial constraints are the same on the two face databases. This explains why their basis images are similar to each other on both databases. Instead, our method learns different local structures on different datasets.
3) Non-negative constraints on basis images are important for face images. As there are negative entries in the basis vectors of MNMF, the local structures are immersed in the images and the locality is unable to measure. especially on the Extended YaleB database, the local structures of face images are nearly vanished.

5 Conclusion

In this paper, a max-margin non-negative matrix factorization method with flexible spatial constraints is proposed. One of the major advantages of this method is that the spatial constraints for non-negative matrix factorization are learnt flexibly from the data. The spatial relationship is derived from the statistical relationship and spatial distance between features. We show that the learnt spatial constraints are adaptable to the data and could be used for some middle-level descriptors, not just gray-scaled images. The second advantage of this model is that it combines the local constraints and discriminative constraints jointly. Due to the local constraints, prior knowledge about the data could be integrated into basis vectors, yielding more interpretable representation. Meanwhile, based on the max-margin criterion, an extra basis vector is learnt to undertake the discriminative constraint, which effectively resolves the confliction goal between low reconstruction error and high discriminative power on coefficient vectors, leading to better locality of basis vectors and smaller reconstruction error. The feasibility and effectiveness of the proposed method is verified extensively on several real world datasets.

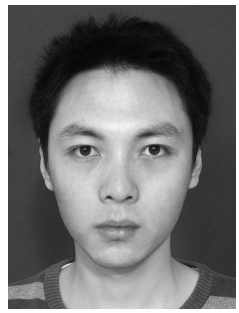
Acknowledgements The work was financed by the National Natural Science Foundation of China (Grant No. 61073112), the National Science Foundation of Jiangsu Province (BK2012793), and the Doctoral Fund of Ministry of Education of China (20123218110033).

References

1. Bishop C M, Nasrabadi N M. Pattern recognition and machine learning. volume 1. springer New York, 2006
2. Turk M, Pentland A. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 1991, 3(1): 71–86
3. Tipping M E, Bishop C M. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999, 61(3): 611–622
4. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics, 2006, 15(2): 265–286
5. Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788–791
6. Seung D, Lee L. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems, 2001, 13: 556–562
7. Ross D A, Zemel R S. Learning parts-based representations of data. The Journal of Machine Learning Research, 2006, 7: 2369–2397
8. Lemme A, Reinhart R F, Steil J J. Online learning and generalization of parts-based image representations by non-negative sparse autoencoders. Neural Networks, 2012, 33: 194–203
9. Wang S, Uchida S, Liwicki M, Feng Y. Part-based methods for handwritten digit recognition. Frontiers of Computer Science, 2013, 7(4): 514–525
10. Zhang Y, Chen L, Jia J, Zhao Z. Multi-focus image fusion based on non-negative matrix factorization and difference images. Signal Processing, 2014, 105: 84–97
11. Du H, Hu Q, Zhang X, Hou Y. Image feature extraction via graph embedding regularized projective non-negative matrix factorization. In: Pattern Recognition, 196–209. Springer, 2014
12. Wu Y, Shen B, Ling H. Visual tracking via online nonnegative matrix factorization. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(3): 374–383
13. Wang X, Wang B, Bai X, Liu W, Tu Z. Max-margin multiple-instance dictionary learning. In: Proceedings of the 30th International Conference on Machine Learning. 2013, 846–854
14. Wang Y, Jia Y. Fisher non-negative matrix factorization for learning local features. In: Proceedings of Asian Conference on Computer Vision. 2004, 27–30
15. Zafeiriou S, Tefas A, Buciu I, Pitas I. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. IEEE Transactions on Neural Networks, 2006, 17(3): 683–695
16. Li X, Fukui K. Fisher non-negative matrix factorization with pairwise weighting. In: Proceedings of MVA. 2007, 380–383
17. Kotsia I, Zafeiriou S, Pitas I. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. IEEE Transactions on Information Forensics and Security, 2007, 2(3): 588–595
18. Nieto O, Jehan T. Convex non-negative matrix factorization for automatic music structure identification. In: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013, 236–240
19. Huang K, Sidiropoulos N D, Swami A. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Transactions on Signal Processing, 2014, 62(1): 211–224
20. Yanez F, Bach F. Primal-dual algorithms for non-negative matrix factorization with the kullback-leibler divergence. arXiv preprint arXiv:1412.1788, 2014
21. Wang J J Y, Gao X. Max–min distance nonnegative matrix factorization. Neural Networks, 2015, 61: 75–84
22. Kumar B, Kotsia I, Patras I. Max-margin non-negative matrix factorization. Image and Vision Computing, 2012, 30(4): 279–291
23. Kumar V B, Patras I, Kotsia I. Max-margin semi-nmf. In: Proceedings of the British Machine Vision Conference. 2011
24. Donoho D, Stodden V. When does non-negative matrix factorization give a correct decomposition into parts? In: Proceedings of the Neural Information Processing Systems Conference. 2003, 1141–1148
25. Hoyer P O. Non-negative matrix factorization with sparseness constraints. The Journal of Machine Learning Research, 2004, 5: 1457–1469
26. Tan X, Triggs B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing, 2010, 19(6): 1635–1650
27. Wang Y, Liu J, Tang X. Robust 3d face recognition by local shape difference boosting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(10): 1858–1870
28. Wang X, Ling H, Xu X. Parts-based face super-resolution via nonnegative matrix factorization. Computers & Electrical Engineering, 2014, 40(8): 130–141
29. Sharma G, Jurie F, Pérez P. Epml: Expanded parts based metric learning for occlusion robust face verification. In: Proceedings of Asian Conference on Computer Vision. 2014, 1–15
30. Tang Z, Zhang X, Zhang S. Robust perceptual image hashing based on ring partition and nmf. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(3): 711–724
31. Tian Q, Chen S, Tan X. Comparative study among three strategies of incorporating spatial structures to ordinal image regression. Neurocomputing, 2014, 136: 152–161
32. Li S Z, Hou X W, Zhang H J, Cheng Q S. Learning spatially localized, parts-based representation. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2001, I–207
33. Jiang B, Zhao H, Tang J, Luo B. A sparse nonnegative matrix factorization technique for graph matching problems. Pattern Recognition, 2014, 47(2): 736–747
34. Zeng K, Yu J, Li C, You J, Jin T. Image clustering by hyper-graph regularized non-negative matrix factorization. Neurocomputing, 2014, 138: 209–217
35. Zheng W S, Lai J, Liao S, He R. Extracting non-negative basis images using pixel dispersion penalty. Pattern Recognition, 2012, 45(8): 2912–2926
36. Chen X, Li C, Cai D. Spatially correlated nonnegative matrix factorization for image analysis. In: Intelligent Science and Intelligent Data

Engineering, 148–157. Springer, 2013

37. Chen X, Li C, Liu H, Cai D. Spatially correlated nonnegative matrix factorization. Neurocomputing, 2014, 139: 15–21
38. Wu J, Qu W, Hu H, Li Z, Xu Y, Tao Y. A discriminative spatial bag-of-word scheme with distinct patch. In: Proceedings of the 2014 International Conference on Audio, Language and Image Processing. 2014, 266–271
39. Mu Y, Ding W, Tao D. Local discriminative distance metrics ensemble learning. Pattern Recognition, 2013, 46(8): 2337–2349
40. Lawton W H, Sylvestre E A. Self modeling curve resolution. Technometrics, 1971, 13(3): 617–633
41. Paatero P, Tapper U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics, 1994, 5(2): 111–126
42. Chen X, Tong Z, Liu H, Cai D. Metric learning with two-dimensional smoothness for visual analysis. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012, 2533–2538
43. Cai D, He X, Wu X, Han J. Non-negative matrix factorization on manifold. In: Proceedings of the Eighth IEEE International Conference on Data Mining. 2008, 63–72
44. Cai D, He X, Han J, Huang T S. Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1548–1560
45. Ando R K, Zhang T. Learning on graph with laplacian regularization. Advances in Neural Information Processing Systems, 2007, 19: 25
46. Fidler S, Skocaj D, Leonardis A. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(3): 337–350
47. Basilevsky A T. Statistical factor analysis and related methods: theory and applications. volume 418. John Wiley & Sons, 2009
48. Martínez A M, Kak A C. Pca versus lda. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(2): 228–233
49. Georghiades A S, Belhumeur P N, Kriegman D J. From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 643–660
50. Hull J J. A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1994, 16(5): 550–554
51. Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition. 2004, 32–36
52. Hido S, Tsuboi Y, Kashima H, Sugiyama M, Kanamori T. Inlier-based outlier detection via direct density ratio estimation. In: Proceedings of the Eighth IEEE International Conference on Data Mining. 2008, 223–232
53. Dalal BN. T. Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005, 886–893



Dakun Liu received his BS and MS in applied mathematics in 2006 and 2009, respectively. He is now a PhD candidate at Nanjing University of Aeronautics and Astronautics, China. His research interests include computer vision, machine learning, etc.

Xiaoyang Tan received his PhD in machine learning from Nanjing University, China in 2005. He is a professor and PhD supervisor at Nanjing University of Aeronautics and Astronautics, China. His research interests include computer vision, machine learning, etc.