Pattern Recognition 45 (2012) 2912–2926

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Extracting non-negative basis images using pixel dispersion penalty

Wei-Shi Zheng [a,*], JianHuang Lai [a], Shengcai Liao [b], Ran He [c]

[a] School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China
[b] Department of Computer Science and Engineering, Michigan State University, USA
[c] Institute of Automation, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 30 September 2011
Received in revised form
19 January 2012
Accepted 30 January 2012
Available online 8 February 2012

Keywords:
Non-negative matrix factorization (NMF)
Non-negativity constraint
Spatially localized basis images
Feature extraction
Face image analysis

ABSTRACT

Non-negativity matrix factorization (NMF) and its variants have been explored in the last decade and are still attractive due to its ability of extracting non-negative basis images. However, most existing NMF based methods are not ready for encoding higher-order data information. One reason is that they do not directly/explicitly model structured data information during learning, and therefore the extracted basis images may not completely describe the "*parts*" in an image [1] very well. In order to solve this problem, the structured sparse NMF has been recently proposed in order to learn structured basis images. It however depends on some special prior knowledge, i.e. one needs to exhaustively define a set of structured patterns in advance. In this paper, we wish to perform structured sparsity learning as automatically as possible. To that end, we propose a pixel dispersion penalty (PDP), which effectively describes the spatial dispersion of pixels in an image without using any manually predefined structured patterns as constraints. In PDP, we consider each part-based feature pattern of an image as a cluster of non-zero pixels; that is the non-zero pixels of a local pattern should be spatially close to each other. Furthermore, by incorporating the proposed PDP, we develop a spatial non-negative matrix factorization (Spatial NMF) and a spatial non-negative component analysis (Spatial NCA). In Spatial NCA, the non-negativity constraint is only imposed on basis images and such constraint on coefficients is released, so both subtractive and additive combinations of non-negative basis images are allowed for reconstructing any images. Extensive experiments are conducted to validate the effectiveness of the proposed pixel dispersion penalty. We also experimentally show that Spatial NCA is more flexible for extracting non-negative basis images and obtains better and more stable performance.

1. Introduction

Extracting basis images using matrix factorization techniques is widely used in pattern recognition and computer vision. Given a set of N d-dimensional image data $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, N$, which are always assumed to be non-negative, matrix factorization aims to decompose a $d \times N$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ into a $d \times L$ basis matrix $\mathbf{W}$ and a $L \times N$ coefficient matrix $\mathbf{H}$ such that

$$\mathbf{X} \approx \mathbf{WH}, \tag{1}$$

where $L \ll \min(N,d)$ for learning a low dimensional subspace that describes main variations or distinctive factors of data. Principal component analysis (PCA) is one of the most popular matrix factorization techniques [2,3]. PCA learns a low dimensional subspace of data that preserves as much data information as possible. It is, however, always exploring holistic features as basis images and cannot explore spatially localized features, and thus the potential significant structured factors [4] cannot be explicitly explored.

The non-negativity matrix factorization (NMF) [5,6] is an effective and popular way to alleviate the above problem. NMF imposes non-negativity constraint on basis matrix $\mathbf{W}$ and coefficient matrix $\mathbf{H}$ at the same time. NMF aims to represent an image by additive combination of a set of non-negative basis images. This can sometimes effectively lead to part-based representation of the data. In order to make NMF support this property in diverse fields, many invariants, including using lasso penalty [7], imposing orthogonal penalty [8–10] and designing a criterion that measures the degree of sparsity [11], have also been reported in the last decade. One attraction of these methods is that they can experimentally retrieve sets or parts of variables as local patterns (e.g. eyes or mouth) in a face image, which are intuitively meaningful data structures and good for data analysis and understanding.

Although several typical penalty functions and constraints have been proposed in order to improve NMF for extracting sparser features, it still lacks of theoretical guarantee that the extracted non-negative basis images by the above methods can directly reflect the expected data structure information. It is because these constraints and penalty functions do not explicitly model the shapes or configuration of local patterns in an image.

* Corresponding author. Tel.: +86 020 84110175.
E-mail addresses: wszheng@ieee.org (W.-S. Zheng), stsljh@mail.sysu.edu.cn (J. Lai), scliao@msu.edu (S. Liao), rhe@nlpr.ia.ac.cn (R. He).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.01.022

Most of them are only concerned about the sparsity of the extracted basis images or the relationship between different basis images. A sparse basis image is conceptually different from a spatially localized basis image. A sparse basis image is mainly concerned about the number of its non-zero entries while no geometric constraints among those non-zero entries are specified. In comparison, a spatially localized basis image should be concerned more about the structured relationship between pixels (e.g. meaningful local patches). Therefore, the aforementioned related existing penalty functions cannot completely address a basic question raised by Mel [1] that how the extracted basis images are related to the concept "parts".

Recently, structured sparsity learning is introduced in order to quantify the correlation between variables (e.g. pixels in an image) in regression [12], classification [13], compressed sensing [14,15], and is also applied to NMF [4]. The main idea of the structured sparse NMF is to enforce some prior knowledge about structured information by formulating a structured regularization penalty added to the usual data reconstruction error term. The structured information differs for different applications and for example includes [16,17]: (1) shape of local rectangle patches (e.g. blocks as the set of axis-aligned half-spaces on a 2-dimensional grid across different sizes and scales); (2) shape of oblique local patterns with different angles in a plane; (3) a set of consecutive variables in a sequence; (4) group of variables (e.g. group of gene from the same pathway in gene analysis or group of dummy variables corresponding to the same factor in ANOVA factor analysis). However, in order to learn these structured sparse information, special prior knowledge about the structured information, e.g. group structure of variables or prior support patches in an image for different types of data, needs to be known and predefined manually. In order to realize this, one may need to provide much more exhaustive prior structured patterns during learning when using these structured sparsity learning methods. Hence, its successfulness is highly depending on the manual definition of structure in advance, and selection of prior information may have to be performed in order to reduce the complexity [17]. Moreover, it is still unknown how to define appropriate predefined structured patterns which are good for classification.

In this paper, we wish to learn structured sparse basis images in an image as automatically as possible without manually defining any special prior support patches. Our focus is on image analysis and we propose to consider the "parts" as the clusters of non-zero pixels in an image; that is all non-zero pixels of a part-based pattern in an image should be spatially close to each other. To this end, we devise a new spatially localized penalty function called the *pixel dispersion penalty*, which quantifies how pixels scatter spatially in an image.

Another contribution of this paper is that we will develop a spatial non-negativity matrix factorization (Spatial NMF) and a spatial non-negative component analysis (Spatial NCA), where both methods will incorporate the proposed pixel dispersion penalty and in particularly Spatial NCA allows both subtractive and additive combination of non-negative basis images at the same time. We find that when using the proposed pixel dispersion penalty, releasing non-negativity constraint on coefficient matrix in NMF will lead to better performance on image understanding and face recognition.

The rest of the paper is organized as follows. Section 2 first reviews some related work on extracting non-negative basis images using matrix factorization. The proposed pixel dispersion penalty is first detailed in Section 3 and utilized to develop new matrix factorization techniques in Section 4. Experiments are conducted for evaluating the proposed pixel dispersion penalty in Section 5. The paper is finally concluded in Section 6.

2. Review of related work

In this section, we mainly review related work of NMF for extracting sparse basis images (features). There are several other works on combining manifold learning and supervised learning with NMF [18–24] for classification. As manifold learning and supervised learning are not the main focus in this work, these works will not be covered in the following review and we will discuss them finally in the conclusion part.

Let $\mathbf{W}=(\mathbf{w}_1,\ldots,\mathbf{w}_L)$ and $\mathbf{H}=(\mathbf{h}_1,\ldots,\mathbf{h}_N)$, where L is the number of basis images and always much smaller than the dimension of data for a low-rank learning, and each $\mathbf{h}_i$ is the corresponding coding vector for each data sample $\mathbf{x}_i$. The non-negative matrix factorization (NMF) proposed by Lee and Seung [5,6] constrains the basis images and coding vectors to be non-negative and is therefore formulated as the following optimization problem:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{h}_i\in\mathbb{R}^L} \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{W}\mathbf{h}_i\|_F^2 \quad \text{s.t.} \quad \mathbf{W}\geq\mathbf{0},\ \mathbf{h}_i\geq\mathbf{0}. \tag{2}$$

Due to the use of additive combination of non-negative basis images, part-based basis images can be more easily extracted. However, NMF was not explicitly designed to extract sparse basis images, and it has been found that only imposing non-negativity constraint on both basis images and coding vectors sometimes does not sufficiently lead to extracting part-based basis images.

In order to make NMF extract sparser basis images in diverse applications, several constraints or penalty functions are also combined in order to make the extracted basis images less overlapped or sparse. Li et al. united three penalty functions together and derived the following criterion based on the divergence distance [8]:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{H}\in\mathbb{R}^{L\times N}} \left\{\sum_{ij}\mathbf{X}_{ij}\log\frac{\mathbf{X}_{ij}}{(\mathbf{W}\mathbf{H})_{ij}}-\mathbf{X}_{ij}+(\mathbf{W}\mathbf{H})_{ij}+\alpha\|\mathbf{W}^T\mathbf{W}\|_1-\beta\,trace(\mathbf{H}\mathbf{H}^T)\right\}$$
$$\text{s.t.} \quad \mathbf{W}\geq\mathbf{0},\ \mathbf{H}\geq\mathbf{0},\ \alpha,\beta\geq 0. \tag{3}$$

The penalty $\|\mathbf{W}^T\mathbf{W}\|_1$ at least includes the orthogonal penalty between the columns of $\mathbf{W}$ which would reduce the redundancy between basis images, and maximizing the penalty $trace(\mathbf{H}\mathbf{H}^T)$ would enforce the algorithm to achieve maximum expressiveness of basis images $\mathbf{w}_i$ [8].

Hoyer designed a sparsity measurement that computes the sparsity degree [11] of each basis image $\mathbf{w}_i$ directly as follows:

$$G(\mathbf{w}_i)=\frac{\sqrt{d}-(\sum_j|\mathbf{w}_i(j)|/\sqrt{\sum_j\mathbf{w}_i(j)^2})}{\sqrt{d}-1}. \tag{4}$$

By applying the same sparsity measurement to each coding vector $\mathbf{h}_i$ as well, a NMF with sparsity constraint (NMFnc) is formulated in [11]. Apart from the above two variants, Pascual-Montano et al. modified the NMF criterion by introducing a smooth matrix $\mathbf{S}$ [25] as follows:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{h}_i\in\mathbb{R}^L} \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{W}\mathbf{S}\mathbf{h}_i\|_F^2 \quad \text{s.t.} \quad \mathbf{W}\geq\mathbf{0},\ \mathbf{h}_i\geq\mathbf{0}. \tag{5}$$

where

$$\mathbf{S}=(1-\theta)\mathbf{I}+\frac{\theta}{L}\mathbf{1}\mathbf{1}^T,\ \theta\in[0,1]. \tag{6}$$

The smooth matrix is designed in order to balance the sparsity of the basis images $\mathbf{w}_i$ and coding vectors $\mathbf{h}_j$.

In addition, work in [7] aimed to extract more effective sparse non-negative basis images by means of sparse coding learning using lasso penalty, and a related efficient algorithm was developed by Lee et al. [26]. However, as suggested in [25], for learning a low rank basis

matrix $\mathbf{W}$, it is still hard to learn part-based and spatially localized basis images from real data without directly penalizing them. There is also work to combine non-negativity constraint, lasso penalty and orthogonal penalty together by Zass and Shashua [9]. However, the developed sparse non-negative PCA is computationally expensive as it is a fourth order optimization problem.

Although the above constraints and penalty functions have been widely recognized, they are not directly and explicitly designed for exploring the structured information in an image. The structured information is useful and significant. It is because variables or parameters are not always completely independent, and they may be correlated or grouped together in terms of some kind of structured information. In order to overcome this problem, several recent works attempted to impose group structure between variables, for example incorporating some prior support patch information to quantify these variables jointly [4,12,27]. Particularly, for extracting structured sparse features in an image, Jenatton et al. proposed to use different sizes and scales of rectangle blocks to group the pixels in an image and applied this penalty to non-negativity matrix factorization [4]. Therefore, the learning model is formulated as follows[1]:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{h}_i\in\mathbb{R}^L} \quad \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{W}\mathbf{h}_i\|^2+\lambda\sum_{j=1}^{L}\Omega_w(\mathbf{w}_j)$$
$$\text{s.t.} \quad \forall i,\ \Omega_h(\mathbf{h}_i)\le 1\ \&\ \mathbf{W}\ge\mathbf{0},\ \mathbf{h}_i\ge\mathbf{0}. \tag{7}$$

where $\Omega_w(\mathbf{w})=\sum_{G\in\mathcal{G}}\|d^G\circ\mathbf{w}\|_2$, $\mathcal{G}$ is a pre-specified subset of power set of $\{1,\ldots,n\}$, d^G is a n-dimensional vector and performs as a filter function [4], and $\Omega_h(\mathbf{h}_i)$ is some kind of constraint on coding vector $\mathbf{h}_i$. More general formulation can be found in [4]. However, these data structured information needs to be known a priori as aforementioned in the introduction.

Compared to the above existing works, the novelty and new developments of this work include:

1. We propose a pixel dispersion penalty in order to describe the spatial structured relations between pixels in an image without using any prior structured patterns as constraints.
2. By applying the pixel dispersion penalty, two new developments are presented. That is, (1) we apply our proposed pixel dispersion penalty to non-negative matrix factorization and develop Spatial NMF, and (2) we particularly develop spatial non-negative component analysis (Spatial NCA). In Spatial NCA, the non-negativity constraint is only imposed on basis images $\mathbf{w}_i$ and such constraint on codings $\mathbf{h}_j$ is released. An optimization algorithm is also developed accordingly.

3. The proposed pixel dispersion penalty

We wish to learn a low rank basis matrix $\mathbf{W}$, each column of which (i.e. each basis image $\mathbf{w}_i$) describes a spatially localized part of an image. We say a basis image $\mathbf{w}_i$ is spatially localized if the non-zero pixels of the basis image are spatially and locally non-dispersive, i.e. those non-zero pixels should be clustered and close to its center. Suppose the image we are concerned is b pixels in height and a pixels in width. To measure the dispersion degree of non-zero pixels in each basis $\mathbf{w}_i$, we propose the following criterion $D(\mathbf{w}_i)$:

$$D(\mathbf{w}_i)=\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{x'=1}^{a}\sum_{y'=1}^{b}\delta(\mathbf{w}_i^{2D}(y,x)\ne 0)\times\delta(\mathbf{w}_i^{2D}(y',x')\ne 0)\times l([y,x],[y',x']), \tag{8}$$

where $\mathbf{w}_i^{2D}$ ($\in\mathbb{R}^{b\times a}$) is the corresponding matrix form of the basis image vector $\mathbf{w}_i$, $\delta(true)=1$ and 0 otherwise, and l is an association function between two coordinate vectors $[y,x]$ and $[y',x']$ in an image and measures the distance between them. In this paper, we use the following association function:

$$l([y,x],[y',x'])=|y-y'|+|x-x'|. \tag{9}$$

The larger the $D(\mathbf{w}_i)$ is the more dispersive the non-zero pixels are.

The measurement Eq. (8), however, is a non-convex function of $\mathbf{w}_i$, which would be hard for optimizing $\mathbf{w}_i$ in an analytic way. In order to make the modeling of dispersion degree more tractable for optimization, we further develop the following weighted dispersion degree modeling:

$$D(\mathbf{w}_i)=\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{x'=1}^{a}\sum_{y'=1}^{b}|\mathbf{w}_i^{2D}(y,x)|\,|\mathbf{w}_i^{2D}(y',x')|\,l([y,x],[y',x']). \tag{10}$$

To investigate the rationale of the above modeling, we let $d_{y,x}(\Delta y,\Delta x)=l([y,x],[y+\Delta y,x+\Delta x])$. As shown in Fig. 1, $d_{y,x}(\Delta y,\Delta x)$ is a special high-pass filter, and the penalty function $D(\mathbf{w}_i)$ is actually a weighted combination of a set of special high-pass filter's responses as follows:

$$D(\mathbf{w}_i)=\sum_{x=1}^{a}\sum_{y=1}^{b}|\mathbf{w}_i^{2D}(y,x)|\times\left\{\sum_{x'=1}^{a}\sum_{y'=1}^{b}|\mathbf{w}_i^{2D}(y',x')|\,d_{y,x}(y'-y,x'-x)\right\}. \tag{11}$$

The high-pass filter then enlarges the effect of any points which are away from the corresponding center $[y,x]$. Therefore, minimizing the weighted dispersion $D(\mathbf{w}_i)$ would suppress the scenario that two disjoint pixels at $[y,x]$ and $[y',x']$ which are far away from each other and both have higher weights $|\mathbf{w}_i^{2D}(y,x)|$ and $|\mathbf{w}_i^{2D}(y',x')|$ (i.e. higher pixel values), due to the non-negativity of $|\mathbf{w}_i^{2D}(y,x)|$. This encourages the algorithm to learn spatially localized patches in an image.

Moreover, by Eq. (11), we find that $D(\mathbf{w}_i)$ in some aspect can be viewed as a special weighted ℓ_1-norm function. However, the difference is that the weight $\{\sum_{x'=1}^{a}\sum_{y'=1}^{b}|\mathbf{w}_i^{2D}(y',x')|d_{y,x}(y'-y,x'-x)\}$ for each entry $|\mathbf{w}_i^{2D}(y,x)|$ has incorporated the pixel spatial information surrounding the pixel at $[y,x]$ in an image, and this makes the pixel dispersion penalty be a second-order function.

Nevertheless, without using any manually predefined structured patterns as constraints, the spatially localized basis images can be favored by the pixel dispersion penalty. The experiments will show the learned basis images $\mathbf{w}_i$ are also sparse.

We aim to extract non-negative basis images in this work. Therefore, we finally present the following special pixel dispersion penalty for computation.

3.1. The pixel dispersion penalty for non-negative $\mathbf{w}_i$

Let $\mathbf{e}_{y,x}$($\in\mathbb{R}^d$) be the indicator vector such that

$$\mathbf{e}_{y,x}(j)=\begin{cases}1 & j=(x-1)\times b+y,\\ 0 & \text{otherwise},\end{cases} \tag{12}$$

where b is the height of an image. Note that $\mathbf{w}_i^{2D}(y,x)=\mathbf{w}_i^T\mathbf{e}_{y,x}$. When $\mathbf{w}_i$ is non-negative, Eq. (10) becomes

$$D(\mathbf{w}_i)=\mathbf{w}_i^T\left\{\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{x'=1}^{a}\sum_{y'=1}^{b}l([y,x],[y',x'])\times\mathbf{e}_{y,x}\mathbf{e}_{y',x'}^T\right\}\mathbf{w}_i=\mathbf{w}_i^T\mathbf{E}_i\mathbf{w}_i, \tag{13}$$

[1] Please note that Eq. (7) is slightly different from Jenatton's because each sample is formulated as a row vector in [4] while it is formulated as a column vector in this work.

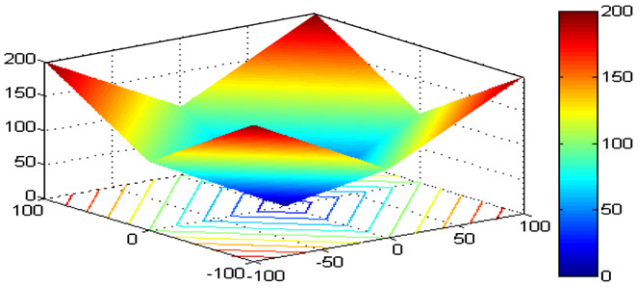

Fig. 1. Illustration of the filter function $d_{y,x}(\Delta y,\Delta x)$ in Eq. (11).

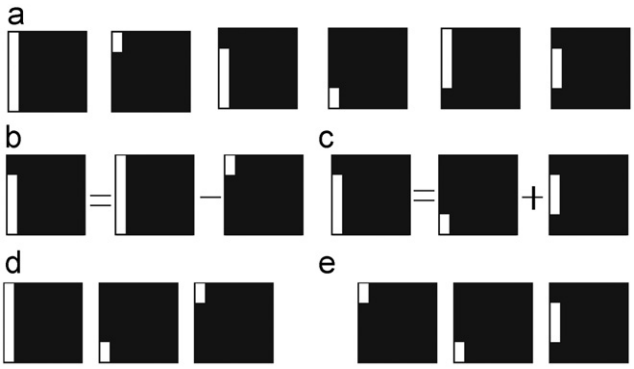

Fig. 2. Demonstration of NCA: (a) six image samples; (b) and (c) are two examples of image construction; (d) and (e) are two alternative basis images that can be learned by NCA from (a).

where

$$\mathbf{E}_l = \sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{x'=1}^{a}\sum_{y'=1}^{b} l([y,x],[y',x'])\mathbf{e}_{y,x}\mathbf{e}_{y',x'}^T. \tag{14}$$

We call $\mathbf{E}_l$ the dispersion kernel matrix in this paper.

4. Low-rank matrix factorization with pixel dispersion penalty

4.1. Spatial non-negative matrix factorization

In order to extract non-negative structured local patterns of the data, we are now incorporating the pixel dispersion penalty to develop a new penalized NMF-based matrix factorization as follows:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{h}_i\in\mathbb{R}^L} \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{W}\mathbf{h}_i\|_F^2+\frac{\lambda}{L}\,trace(\mathbf{W}^T\mathbf{E}_l\mathbf{W})$$
$$\text{s.t.}\quad \mathbf{W}\geq\mathbf{0},\mathbf{0}\leq\mathbf{h}_i\leq c_0, \tag{15}$$

where $\lambda\geq 0$, $L\ll d$, and c_0 is a simple positive constant bound parameter which removes the scale effect during minimization of $trace(\mathbf{W}^T\mathbf{E}_l\mathbf{W})$. Note that the condition $L\ll d$ is necessary for learning a low-rank basis matrix $\mathbf{W}$ and also avoiding trivial solution if $L\geq d$. We call the above model as spatial non-negative matrix factorization (Spatial NMF).

In this paper, we particularly focus on developing a special spatially localized semi non-negative matrix factorization method, in which the basis images are non-negative and no non-negativity constraint is imposed on coefficients. We call such kind of matrix factorization as spatial non-negative component analysis (Spatial NCA). The next section will detail the motivation and model.

4.2. Non-negative component analysis with pixel dispersion penalty

Although the usefulness of non-negative basis images in image understanding and representation has been recognized, it may still be hard for an object to be completely represented by additive combination of a few spatially localized non-negative basis images. To this end, there are recent attempts to partially release the non-negativity constraint in NMF in [28,29] and our early work [10]. Compared to [10,28,29], we first in the following give more detailed analysis about the release of non-negativity constraint on the coefficient part.

Since the non-negativity constraint on coefficients is released, the subtraction between non-negative basis images is allowed. In particular, this enables the algorithm to realize the scenario that each pattern can be represented by removing and adding spatially localized basis images on a few other basis images. Such a kind of flexibility can be shown in Fig. 2. By allowing subtraction between positive basis images the algorithm can probably find basis images in Fig. 2(d) and (e); but NMF can only find Fig. 2(e). Hence a much more flexible way to explore spatially localized basis images can be possible if the non-negativity constraint on coefficients is released. From another point of view, it is the fact that by using matrix factorization techniques such as PCA, each pattern $\mathbf{x}_i$ is approximately reconstructed by a set of basis images $\{\mathbf{w}_j\}_{j=1}^L$ as follows:

$$\mathbf{x}_i\approx\sum_{j=1}^{L}h_{ji}\times\mathbf{w}_j. \tag{16}$$

Let $\mathbf{w}_j^+=\max(\mathbf{w}_j,0)$ and $\mathbf{w}_j^-=-\min(\mathbf{w}_j,0)$. Then

$$\mathbf{x}_i\approx\sum_{j=1}^{L}h_{ji}\times\mathbf{w}_j^+-h_{ji}\times\mathbf{w}_j^-. \tag{17}$$

So each data $\mathbf{x}_i$ is able to be represented by combination of non-negative basis images, where subtractive and additive combinations exist simultaneously. Note that as subtraction between an two positive basis images is allowed, we are also able to analyze any data with negative entries. For a more general matrix factorization problem, we can generalize the above reconstruction process as follows:

$$\mathbf{x}_i\approx\sum_{j=1}^{L}h_{ji}\times\mathbf{w}_j,\quad \mathbf{w}_j\geq 0 \text{ for any } j. \tag{18}$$

where the basis matrix $\mathbf{W}=[\mathbf{w}_1,\ldots,\mathbf{w}_L]$ and coefficient matrix $\mathbf{H}=(h_{ji})=[\mathbf{h}_1,\ldots,\mathbf{h}_N]$ are learned by

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{h}_i\in\mathbb{R}^L} \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{W}\mathbf{h}_i\|_F^2$$
$$\text{s.t.}\quad \mathbf{W}\geq\mathbf{0}. \tag{19}$$

The above model as *non-negative component analysis* (NCA), and compared to the NMF criterion only the constraint $\mathbf{h}_i\geq\mathbf{0}$ in Eq. (2) has been removed.

Recent work in [10,28,29] has shown that although NCA can explore non-negative basis images in a more flexible way, NCA itself is hard to learn spatially localized basis images. In this work, we wish to incorporate the proposed pixel dispersion penalty into NCA in order to learn spatially localized structured patterns.

By imposing the proposed pixel dispersion penalty (Eq. (13)) on the basis images, a *spatial non-negative component analysis* (Spatial NCA) can be formulated by weighting the reconstruction

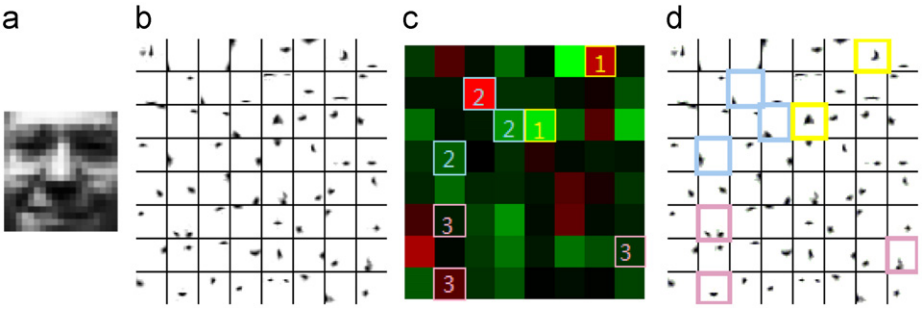

Fig. 3. (c) is the corresponding coefficients for the basis images in (b) learned by Spatial NCA in order to reconstruct (a); (d) shows examples of simultaneous additive and subtractive combination of parts, where basis images of the same group are marked using the same color bounding box and the corresponding group numbers are also shown in (c). The green ones in (c) are positive coefficients, the red ones are negative coefficient and the dark ones are towards zero. In each basis image in (b) and (d), white pixels denote (almost) zero gray values and black ones denote positive gray values (the roles of white and black pixels there are different from those in (a) due to the traditional use for visualization). (a) Example, (b) Basis Images, (c) Coefficients and (d) Combination Examples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

error and the pixel dispersion penalty function as follows:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{h}_i\in\mathbb{R}^L} f_{\mathbf{E}_l}(\mathbf{W},\mathbf{H})$$
$$\text{s.t.}\quad \mathbf{W}\geq\mathbf{0},\ |\mathbf{h}_i(j)|\leq c_0, \tag{20}$$

where $\mathbf{h}_i(j)$ is the jth entry of coding vector $\mathbf{h}_i$ and

$$f_{\mathbf{E}_l}(\mathbf{W},\mathbf{H})=\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i-\mathbf{W}\mathbf{h}_i\|_F^2+\frac{\lambda}{L}\,trace(\mathbf{W}^T\mathbf{E}_l\mathbf{W}),\ \lambda\geq 0. \tag{21}$$

Specially, when the input data $\mathbf{x}_i$ are non-negative (e.g. images represented by pixel values), as the extracted basis images are non-negative, so the criterion can implicitly ensure that the learned coefficients $\mathbf{h}_i(j)$ should not be too small negative value as it is not good for minimizing the reconstruction error in Eq. (20). In view of this, we can relax the above criterion for $\mathbf{x}_i\geq 0$ as follows:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times L},\mathbf{h}_i\in\mathbb{R}^L} f_{\mathbf{E}_l}(\mathbf{W},\mathbf{H})$$
$$\text{s.t. } \mathbf{W}\geq\mathbf{0},\ \mathbf{h}_i(j)\leq c_0. \tag{22}$$

An example on reconstructing a face image by simultaneous additive and subtractive combination of spatially localized basis images learned by Spatial NCA is shown in Fig. 3, where some operations appear near the nose and mouth.

4.3. Optimization algorithm: a projected gradient method

We now mainly develop an optimization algorithm for Spatial NCA (Eq. (20)). The following developed optimization algorithm can be easily applied to Spatial NMF (Eq. (15)) with slight modifications.

As Criterion (20) is convex for each variable respectively but not for all variables jointly, it is not straightforward to compute a globally optimal solution. To solve this problem, we seek optimal basis matrix $\mathbf{W}$ and coefficient matrix $\mathbf{H}$ in an alternating update manner. Among the alternating techniques, the multiplicative rule is popular for the NMF-based matrix factorization techniques and also recently used for a different semi non-negative algorithm where non-negativity is imposed on coefficient rather than basis images. However, due to the bound constraint in Eq. (20), the multiplicative rule may not be easily applicable. Recently, Lin has demonstrated that for NMF the projected gradient update method is a more efficient technique as compared to the multiplicative update rule [30] and related projection techniques has also been used recently in [4]. We in this paper adopt this strategy and develop a projected gradient update based alternating method for computing an optimal solution for the proposed model in Eq. (20).

Algorithm 1. Learning the Spatial NCA model.

Data: Data matrix $\mathbf{X}$, number of basis images L
begin
  Initialization of $\mathbf{W}$ and $\mathbf{H}$ (see text);
  Formulation of the dispersion kernel matrix $\mathbf{E}_l$ in Eq. (14);
  while *stopping criterion not reached* **do**
    Update $\mathbf{W}$ by Eq. (23);
    Update $\mathbf{H}$ by Eq. (24);
  end
end
Output: Basic Matrix $\mathbf{W}$ and Coefficient Matrix $\mathbf{H}$

In this work, the alternating procedure for learning Spatial NCA is shown in Algorithm 1. More specifically, the alternating procedure at each iteration consists of the following two steps:

1. *Update of basis images*:

$$\mathbf{W}(i,j)\leftarrow\max\left\{0,\mathbf{W}(i,j)-\eta_1\times\frac{\partial f_{\mathbf{E}_l}(\mathbf{W},\mathbf{H})}{\partial\mathbf{W}}(i,j)\right\}$$
$$\text{where}\ \frac{\partial f_{\mathbf{E}_l}(\mathbf{W},\mathbf{H})}{\partial\mathbf{W}}=\frac{2}{N}(\mathbf{W}\mathbf{H}\mathbf{H}^T-\mathbf{X}\mathbf{H}^T)+\frac{2\lambda}{L}\mathbf{E}_l\mathbf{W}. \tag{23}$$

2. *Update of coefficients*:

$$\mathbf{H}(i,j)\leftarrow\min\left\{c_0,\max\left\{-c_0,\mathbf{H}(i,j)-\eta_2\times\frac{\partial f_{\mathbf{E}_l}(\mathbf{W},\mathbf{H})}{\partial\mathbf{H}}(i,j)\right\}\right\}$$
$$\text{where}\ \frac{\partial f_{\mathbf{E}_l}(\mathbf{W},\mathbf{H})}{\partial\mathbf{H}}=\frac{2}{N}(\mathbf{W}^T\mathbf{W}\mathbf{H}-\mathbf{W}^T\mathbf{X}). \tag{24}$$

The η_1 and η_2 in Eqs. (23) and (24) are the step lengths in gradient decent and can be adaptively determined as similarly done in [30].

The alternating procedure will repeat the above two steps until convergence. The alternating update procedure ensures the value of the objective function in Criterion (20) decreases after each update [30] and can be finally terminated if the difference between the last two updated criterion values is lower than some tolerance value (e.g. 10^{-6} used in our experiments).

Initialization. Initialized values for $\mathbf{W}$ and $\mathbf{H}$ are necessary for the above alternating procedure. While it is still an open/unsolved issue on investigating the best optimization method for alternating algorithms (e.g. the proposed method and many other non-convex methods), rather than using random initialization,

inspired by Eq. (17), we initialize these two values using a PCA-based method. The motivation is we still wish those basis images are informative to some extent, albeit extraction of spatially localized/sparse basis images. Such an idea is also embedded in the proposed criteria and existing related methods (see Section 2). This inspires a way to initialize the basis matrix using informative vectors. While PCA is a frequently used technique for extracting most informative basis images from a set of images, the extracted basis images, in general, do not satisfy the non-negativity. In view of this, we present the following way to extract related non-negative basis images which are learned using PCA-based technique.

More specifically, to initialize L basis images, the L largest principal component vectors $\mathbf{q}_i$, $i=1,\ldots,L$ are first extracted by PCA and then the following $2L$ non-negative basis images are computed:

$$\mathbf{q}_i^+ = \max(\mathbf{q},0),\ \mathbf{q}_i^- = -\min(\mathbf{q},0). \tag{25}$$

Without loss of generality, we assume that all these non-negative basis images are non-zero; otherwise, the zero basis images are removed first. Let $\{\tilde{\mathbf{q}}_i\} = \{\mathbf{q}_i^+/\|\mathbf{q}_i^+\|_2\} \bigcup \{\mathbf{q}_i^-/\|\mathbf{q}_i^-\|_2\}$. Then we select L non-negative basis images from $\{\tilde{\mathbf{q}}_i\}$ through the following steps:

- The first component is selected such that it has the maximal correlation to the mean of training data $\mathbf{u}$, i.e.

$$\mathbf{w}_1 = \arg\max_{\tilde{\mathbf{q}}_i} \tilde{\mathbf{q}}_i^T \mathbf{u}. \tag{26}$$

- Let $Q_i = \{\tilde{\mathbf{q}}_j\}_{j=1}^{2L} - \{\mathbf{w}_j\}_{j=1}^{i}$. Then for $i \geq 1$, the $i+1$ component which has the lowest correlation to the already selected basis images is selected by

$$\mathbf{w}_{i+1} = \arg\min_{\tilde{\mathbf{q}} \in Q_i} \sum_{s=1}^{i} \tilde{\mathbf{q}}^T \mathbf{w}_s. \tag{27}$$

  This procedure repeats until all the rest $L-1$ non-negative basis images are selected.

After initializing the basis matrix $\mathbf{W}$, we initialize the coefficient matrix $\mathbf{H}$ simply by $\mathbf{H} = \mathbf{W}^T\mathbf{X}$.

5. Experiments

The section is to demonstrate the effectiveness of the proposed pixel dispersion penalty for extracting spatially localized basis images in image understanding and face recognition.

5.1. Datasets and experiment setting

5.1.1. Datasets

Five datasets were selected for evaluation. Fig. 4 shows some images from these datasets. These five datasets are introduced as follows:

- *Swimmer*. The Swimmer dataset is always used for evaluation, as the ground truth decomposition [31] is known for this dataset, i.e. a group of images can be completely represented using a few non-overlapping basis images. Swimmer consists of 256 images of size 32×32. Each image is constituted by 5 parts from the 17 distinct non-overlapping basis images (as shown in Fig. 5(a)), i.e., a centered invariant part called *torso* of 12 pixels and four *limbs* of 6 pixels appear in one of the 4 positions.
- *CBCL*. The CBCL used in [5,25] consists of 2429 frontal face images of resolution 19×19. It is widely used as a standard dataset to evaluate matrix factorization algorithms. Note that, no preprocessing is performed for CBCL here, while some preprocessing, such as removing mean, clipping etc, were first applied by Lee et al. [5] and Hoyer [11]. Without these preprocessing, the different performances of the compared methods are solely due to the differences of these methods. In the experiment, 64 basis images were learned for CBCL.
- *GB2312*. GB2312 is a Chinese character dataset, which consists of the most frequently used 3755 simple Chinese characters. They were centered and normalized with resolution 20×20. Fig. 4(c) shows some examples of them. Each local part of a Chinese character is always constructed by eight strokes: Dian, Shu, Heng, Pie, Duan Pie, Na, Ti, and Gou. These strokes are shown in Fig. 4(d). In our experiments, we aim to evaluate how

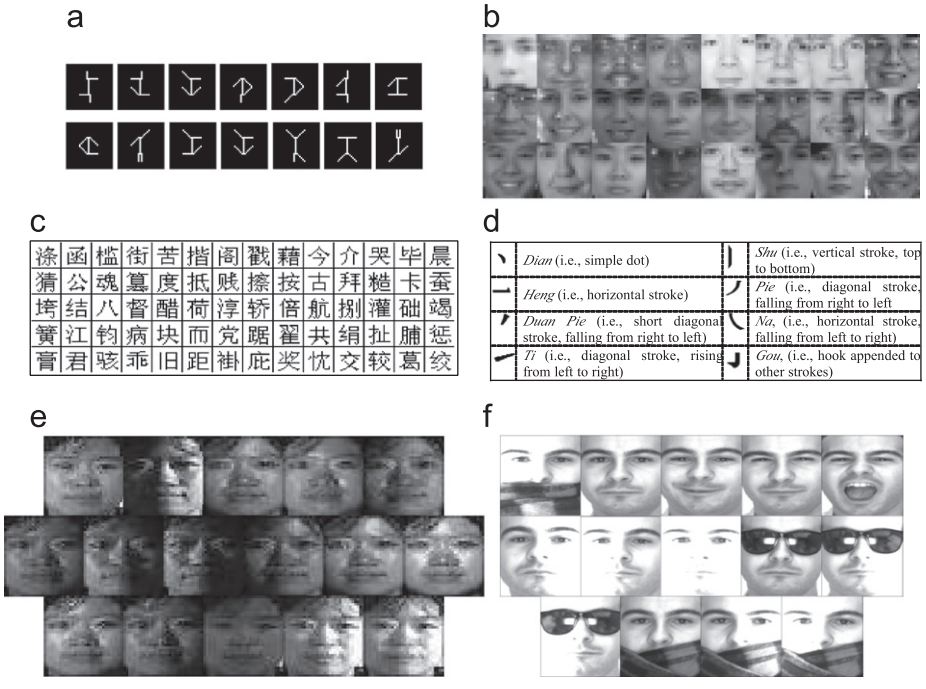

Fig. 4. Examples of images in (a) Swimmer, (b) CBCL, (c) GB2312, (d) basic strokes in Chinese characters, (e) CMU and (f) AR databases.

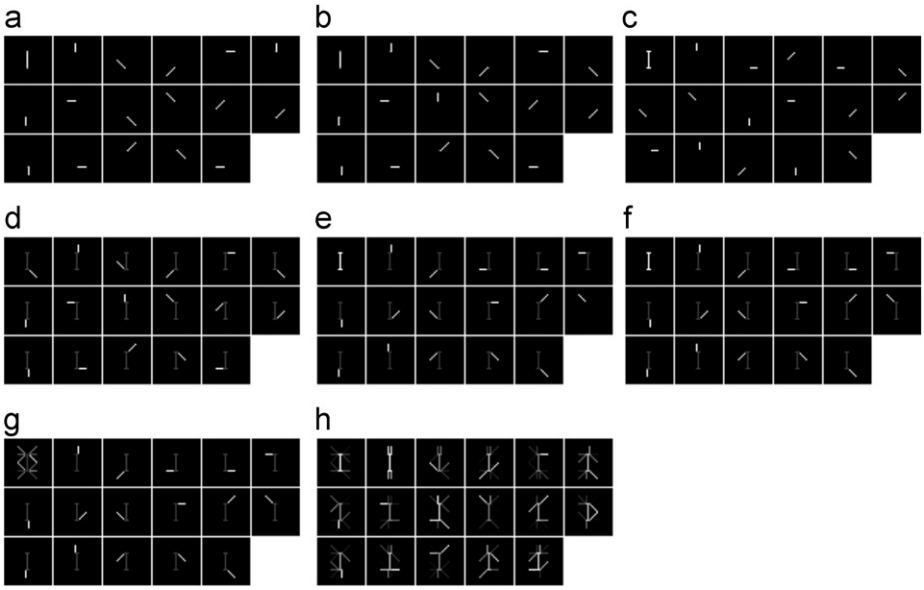

Fig. 5. Illustration of experiment on Swimmer. Black pixels are for (almost) zero entries and white pixels are for positive ones. (a) Spatial NCA, (b) Spatial NMF, (c) SSNMF, (d) Sparse NMF, (e) LNMF, (f) nsNMF, (g) NMF and (h) NCA.

well these strokes are extracted by different methods. In the experiment, 100 basis images were learned for GB2312.

- *CMU*. A subset from the CMU PIE dataset [32] was selected. It consists of 2924 illuminated frontal face images from 68 people, where 43 images for each person under different illumination conditions were captured both indoor and outdoor. All images were resized to 32×32 due to the computational issue as similarly done in [20,22,33].
- *AR*. The AR data set [34] used in this paper consists of 3094 images from 119 people, and each person has 26 images with illumination variation, expression variation, or occlusion (with or without glasses/scarf). All images were resized to 32×32 due to the computational issue.

5.1.2. Methods for evaluation

The main objective of this paper is to validate the proposed pixel dispersion penalty (PDP) for extracting non-negative and spatially localized basis images. Hence, we mainly compare our proposed two methods Spatial NMF and Spatial NCA with NMF [6], localized non-negative matrix factorization (LNMF) [8], non-smooth non-negative matrix factorization and (nsNMF) [25], sparse coding based NMF (Sparse NMF) [7] and structured sparse NMF (SSNMF) [4].

For comparison in our experiments, we will report the best results of the compared methods NMF, LNMF, nsNMF, Sparse NMF and SSNMF on each data set. For image understanding the best visual results of basis images learned by these compared methods are presented, and for face recognition the best recognition rates are reported. Note that, we tuned the θ in nsNMF(Eq. (5)) in $[0:0.1:1]$, the sparse parameter in Sparse NMF in $[0:0.1:1]$, and the sparse parameter in SSNMF in $2^{(-10-100\cdot\eta)}$, $\eta \in [0:0.1:1]$ set in a similar form to [4] and its online code.[2]

For the proposed methods Spatial NMF and Spatial NCA, unless otherwise stated, we have the following setting. For experiments on image understanding, the parameter λ in these two proposed methods is set in order to make the proposed algorithms generate basis images that are with similar sparsity degree to the best results of SSNMF, so that the comparison can be done more fairly at almost the same sparsity degree of features; that is $\lambda = 0.1$ for CBCL and $\lambda = 1$ for GB2312 for the two proposed methods. For face recognition, the parameter λ in our two proposed methods is always set to 1. The effect of the parameter in these two proposed methods will be investigated in Section 5.5.

All basis and coefficient matrices in all the iterative methods evaluated in our experiments were initialized by the same PCA-based initialization technique as described in Section 4.2 and the maximum number of iteration is 500.

5.2. Experiments on image understandings

We in the following compare all related methods on both synthetic and real-world datasets.

5.2.1. Experiments on synthetic dataset

We first present the comparison results on the Swimmer data, in which all images can be exactly reconstructed by 17 ground truth image parts. An algorithm is good for Swimmer dataset if it can learn the exact 17 ground truth basis images, where the exact basis images are shown in Fig. 5(a).

Since we know the ground truth decomposition, we report the best results that are the most similar to the ground truth for all other compared methods in this experiment, while we fixed the parameter λ in Spatial NMF and Spatial NCA to be 0.1. We find that performance of Spatial NMF and Spatial NCA are almost the same by setting the parameter value around 0.1, e.g. 0.2 or 0.3.

As shown in Fig. 5, we find that Spatial NMF and Spatial NCA are able to learn the 17 ground truth basis images. Compared to Spatial NMF and Spatial NCA, NMF always preserves the shadow torso (i.e. the centered vertical segment) in its extracted basis images. In comparison, LNMF, nsNMF and Sparse NMF still preserve the torso in each basis images for Swimmer. This shows the usefulness of the proposed pixel dispersion penalty that imposes the spatial relationship between pixels as a constraint in an image directly, while no similar constraint has been considered in LNMF, nsNMF and Sparse NMF.

We also note that SSNMF can also successfully extract the ground truth basis images. Similar to Spatial NMF and Spatial NCA, SSNMF is also a structured sparsity learning method. On one

[2] http://www.di.ens.fr/~jenatton/

hand, this validates the effectiveness of structured sparsity learning used by our proposed methods and SSNMF; on the other hand, it suggests that Spatial NMF and Spatial NCA, which are based on the pixel dispersion penalty and do not depend on pre-specified structured patterns, can also perform as good as SSNMF. Note that the structured sparsity learning in SSNMF is highly depending on the special prior knowledge used to formulate the penalty; in comparison, pixel dispersion penalty is more unsupervised and thus less complicated.

5.2.2. Experiments on real-world dataset

The CBCL and GB2312 were used for experiments here. The CBCL is used to see how local facial basis images can be investigated and the Chinese character set is used to investigate the frequently used basic strokes in Chinese character images.

Besides comparing visual basis images learned by different methods, three criteria including mean square error (MSE), the normalized absolute overlap degree (AOD) [10] and the sparsity degree (SD) (Eq. (4)) were also used for evaluation. The MSE is to evaluate whether the extracted spatially localized basis images are informative, AOD measures the redundancy between basis images, and SD measures the sparsity of each basis image. More specifically, the AOD is defined by

$$AOD(\mathbf{W}) = \frac{1}{L(L-1)} \sum_{r=1}^{L} \sum_{1,r' \neq r}^{L} \widehat{\mathbf{w}}_r^T \widehat{\mathbf{w}}_{r'}. \tag{28}$$

where $\widehat{\mathbf{w}}_r(i) = |\mathbf{w}_r(i)| / \sum_{j=1} |\mathbf{w}_r(j)|$.

We wish to see that a better localized feature extraction method is featured with low AOD values, large SD values and a reasonable MSE value. The lower the AOD is the less overlap between extracted basis images will be; the larger the SD is the sparser the extracted basis images are. Though minimizing reconstruction error is not our main concerns in this work, MSE is still a necessary criterion to show whether the extracted basis images are informative, as sparse basis images are the ones we preferred only if it can explore structured localized basis images and these images are also good at describing images.

We first report the visual results in Figs. 6 and 7. As shown, by using the pixel dispersion penalty, the basis images extracted by Spatial NMF and Spatial NCA are more spatially localized and less overlapping, resulting in more clear and meaningful localized facial basis images and strokes investigated in these two datasets respectively. Compared to NMF, nsNMF and Sparse NMF which are popular non-negative matrix factorization methods, Spatial NMF and Spatial NCA apparently are able to extract less overlapping basis images, as less shadows are observed in the basis images. Although Sparse NMF finds some interesting radicals in Chinese characters on data set GB2312, they are not the basic strokes we aim to investigate. Also the extracted part-based features can be less effective compared to Spatial NMF and Spatial NCA, as Sparse NMF achieves obviously higher MSE than Spatial NCA (see Table 1) on GB2312. Compared to LNMF, we will show later that the reconstruction ability of LNMF is much unsatisfactory than our proposed methods (see Table 1), albeit obtaining similar visual results. From the figures, especially Fig. 6, the main differences between SSNMF and Spatial NMF/Spatial NCA are two-fold: (1) SSNMF sometimes preserves more shadows around localized feature in basis images (see Fig. 6(c)); (2) SSNMF extracts more pairwise features (e.g. two eyes, two eyebrows as shown in Fig. 6(c)), while Spatial NMF and Spatial NCA are good at extracting isolated localized features. It is probably because the basis images extracted by SSNMF are regularized by pre-specified structured patterns, which might include pairwise ones; in comparison, the pixel dispersion penalty used in Spatial NMF and Spatial NCA aims to extract basis images where non-zero pixels are clustered together locally, therefore prone to extract isolated localized features. For some applications, the isolated and localized features can be more robust for recognition under illumination and occlusion, and this would be shown in the next section.

In order to quantify the visual results, we also report the absolute overlapping rate between the extracted basis images and the sparsity degree of the basis images in Table 1. These two types of results validate the observations from the visual images shown in Figs. 6 and 7. That is by using the pixel dispersion penalty, less

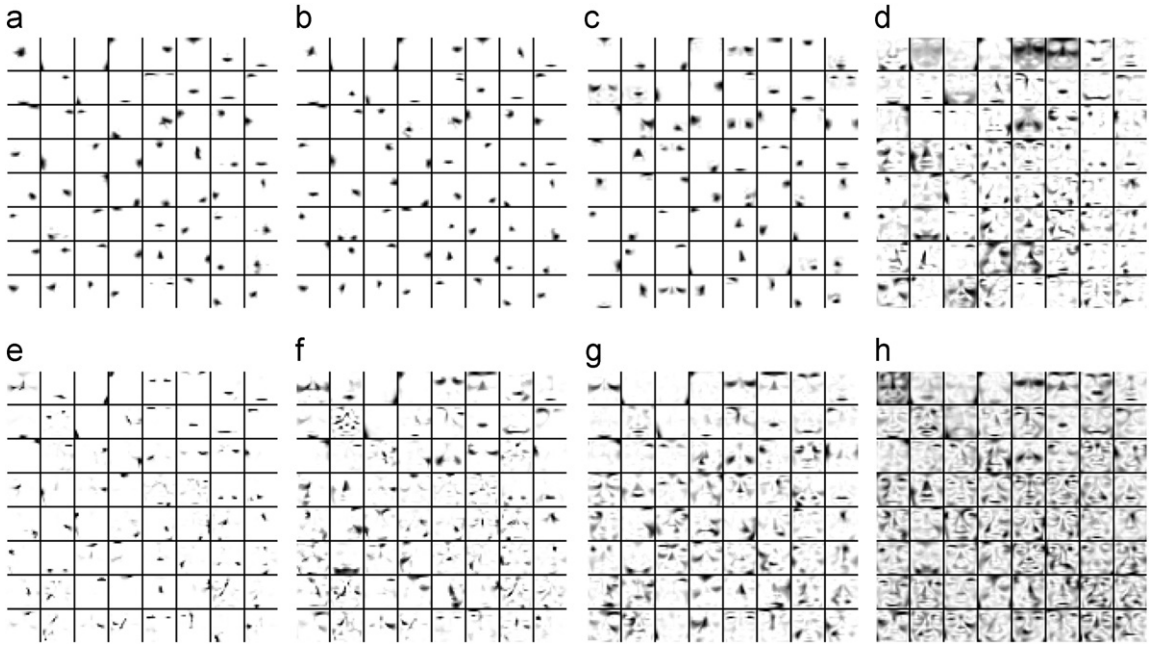

Fig. 6. Illustration of experiment on CBCL. In each basis image, white pixels denote (almost) zero gray values and black ones denote positive gray values (the roles of white and black pixels here are different from those in Fig. 5 due to the traditional use for visualization). (a) Spatial NCA, (b) Spatial NMF, (c) SSNMF, (d) Sparse NMF, (e) LNMF, (f) nsNMF, (g) NMF and (h) NCA.

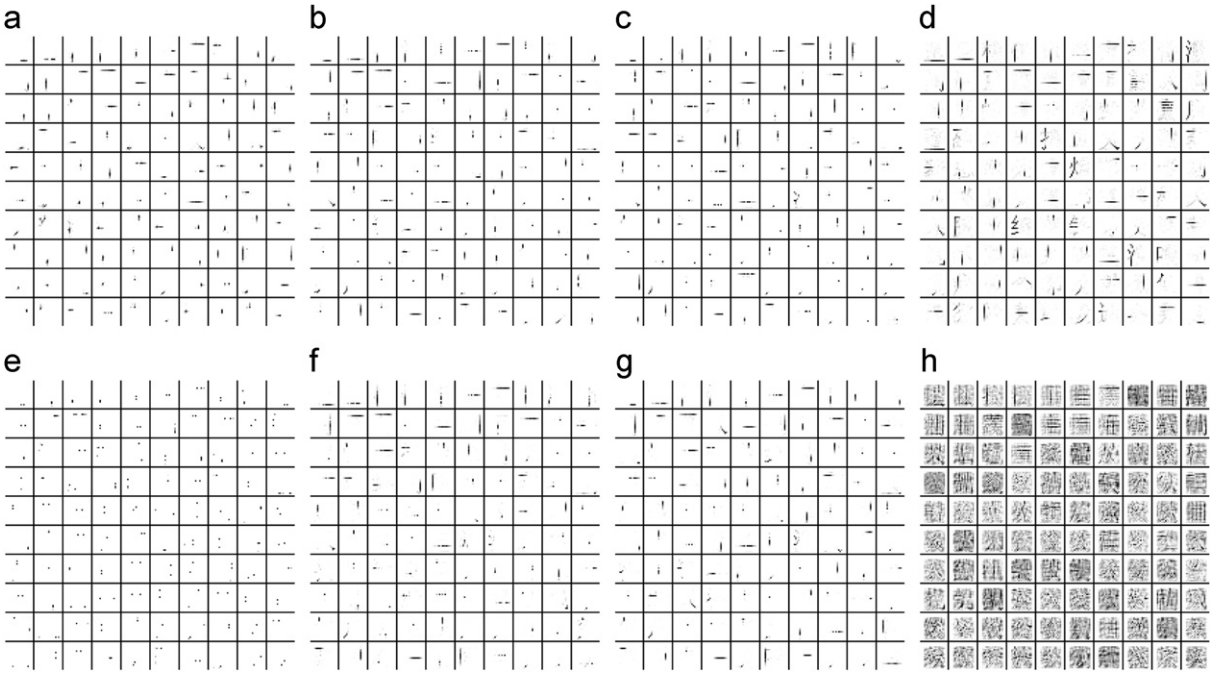

Fig. 7. Illustration of experiment on GB2312. In each basis image, white pixels denote (almost) zero gray values and black ones denote positive gray values. (a) Spatial NCA, (b) Spatial NMF, (c) SSNMF, (d) Sparse NMF, (e) LNMF, (f) nsNMF, (g) NMF and (h) NCA.

Table 1
Comparison among non-negativity based methods for image understanding.

Database	Criterion	NMF	LNMF	nsNMF	Sparse NMF	SSNMF	Spatial NMF	Spatial NCA
CBCL	SD (Eq. (4))	0.6080	0.8496	0.7108	0.5915	0.79717	0.8561	0.8448
	AOD (Eq. (28))	0.1717	0.0259	0.1042	0.1873	0.0458	0.0217	0.0269
	MSE	0.5833	4.9371	2.0561	1.51	0.6046	0.7121	0.6955
GB2312	SD (Eq. (4))	0.9378	0.9750	0.9437	0.9065	0.9495	0.9573	0.9355
	AOD (Eq. (28))	0.0116	2.1421e−005	0.0095	0.0251	0.0076	0.0056	0.0131
	MSE	14.3415	21.4311	15.2772	15.62	14.707	15.3095	12.8823

overlapping (smaller AOD), more spatially localized and sparser basis images (larger SD) are learned. According to the MSE values shown in Table 1, Spatial NCA and Spatial NMF are able to extract equally sparse but much more informative basis images as compared to the ones extracted by LNMF. Compared to NMF, nsNMF and Sparse NMF, the two proposed methods extract more spatially localized and sparser features while keeping reasonable MSE values. Although compared to SSNMF, Spatial NMF and Spatial NCA achieve a little higher MSE values, minimizing the reconstruction error is not the main objective of this work. A little higher MSE is the price one has to pay when using the pixel dispersion penalty for effectively extracting a few spatially localized basis images. From another point of view, this also suggests a tradeoff between locality/sparsity and the quality of data reconstruction. Pixel dispersion penalty will help interpret part-based representation of images, but at the same time the MSE is higher, because the basis images are less overlapped and thus more sparse, resulting in relatively less information used for reconstruction.

5.3. Experiments on face recognition

For recognition, we first applied each method to extract basis image matrix $\mathbf{W}$ on each data set. Then, any input image vector $\mathbf{x}$ will be transformed to $(\mathbf{W}^T\mathbf{W}+\gamma\mathbf{I})^{-1}\mathbf{W}^T\mathbf{x}$ as similarly done in [8], where $\mathbf{W}$ is the basis matrix and the identity matrix $\mathbf{I}$ is to avoid the singularity problem (γ is small, e.g. 10^{-6} in our experiments).

Table 2
Recognition comparison: classification rate (%): $L=100$.

Database	NMF	LNMF	nsNMF	Sparse NMF	SSNMF	Spatial NMF	Spatial NCA
CMU	77.1	76.92	77.20	77.36	78.61	78.99	**79.06**
AR	57.90	61.84	61.39	59.99	60.19	62.15	**63.39**

For recognition, we extracted L basis images, where $L \in \{100, 200\}$. We then applied Linear Discriminant Analysis [2], a popular technique in face recognition, for learning a discriminant subspace that implicitly selects or combines the extracted basis images. Finally the nearest neighbor classifier was used to classify testing samples in the discriminant feature space.

The CMU and AR datasets were used for recognition against variations such as illumination, occlusion and expression. In our experiments, 3 and 6 images were randomly selected for CMU and AR respectively for training, and the rest were used for testing. This procedure was repeated 10 times for each method on each dataset, and the average recognition rate is reported.

During the experiments, we fixed the parameter λ in Spatial NMF and Spatial NCA to be 1, and for comparison we report the best results of the compared methods. The comparison results are reported in Tables 2 and 3 with respect to different numbers of basis images. As shown, the proposed Spatial NMF and Spatial NCA, especially the latter one, outperform the others on CMU and

AR datasets when the number of extracted basis images is 100; when more basis images are learned, e.g. 200, our proposed methods still perform better overall than other methods on AR dataset, yielding about 1−2 higher recognition rate. Note that the two proposed methods do not achieve superior performance on CMU when $L=200$, because there are less classes (people) in CMU than AR and it is possible that some spatially localized basis images extracted could be noisy if more basis images are learned. Note that this finding can be applied to all other methods as they all obtain lower performance on CMU when more basis images are extracted. In addition, we here fixed the parameter λ to 1 in our two proposed methods, while reporting the best results of other methods for comparison. We show later it is possible to improve the performance of the proposed methods if larger λ is used. Nevertheless, the pixel dispersion penalty is an effective penalty function for learning non-negative basis images.

5.4. Further comparison

In previous sections, we show that among the two proposed methods, Spatial NCA is better than Spatial NMF. The difference between these two methods is that there is no non-negativity constraint on coefficients in Spatial NCA while there is in Spatial NMF. This implies that by releasing the non-negativity constraint on coefficients, more effective non-negative basis images can be learned.

In this section, we further show that the superiority of Spatial PCA is partially due to the use of the proposed pixel dispersion penalty. In order to show that, we additionally compare Spatial NCA with SSNCA, where SSNCA is to impose the structured sparsity in [4] onto NCA. The sparse parameter in SSNCA was tuned in the same way as the one for SSNMF and the best results were reported. For Spatial NCA, we fixed the parameter λ in Spatial NCA to 0.1 for experiments on CBCL and GB2312 and 1 for all recognition experiments. Please note that we set the parameter in Spatial NCA to 0.1 here for GB2312 because it generates basis images that have the most similar sparse degree to the best ones learned by SSNCA. The results are shown in Fig. 8, Tables 4 and 5. From these results, we have:

- Spatial NCA can extract much sparser and less overlapping basis images (i.e. smaller AOD and larger SD). This can be investigated from Fig. 8 and Table 4.
- Spatial NCA overall outperforms SSNCA for recognition, especially on AR dataset. This is shown in Table 5.

Hence, the proposed pixel dispersion penalty plays an important role in Spatial NCA and performs more effectively than the structured sparsity constrain in SSNCA, especially from the recognition aspect.

5.5. More discussion on Spatial NCA and Spatial NMF

5.5.1. The effect of the parameter λ

We finally investigate the effect of the parameter λ in Spatial NMF and Spatial NCA which indicates the importance of the pixel dispersion penalty. In previous section, the importance parameter λ is actually fixed to be 0.1 in most of the cases for experiments on image understanding and always 1 for recognition. We are now varying the value of this parameter and see its effect on the performance of the two proposed methods. For image analysis, we varied the parameter value from 0.1 to 1 in Tables 6 and 7; for face recognition, we varied the parameter in [0.01 : 0.01 : 0.09 0.1: 0.1 : 0.9 1 : 0.5 : 10] in Figs. 9 and 10.

From these results, we can find that:

Table 3
Recognition comparison: classification rate (%): $L=200$.

Database	NMF	LNMF	nsNMF	Sparse NMF	SSNMF	Spatial NMF	Spatial NCA
CMU	75.53	75.79	**77.49**	75.53	74.93	74.88	74.69
AR	70.69	69.65	70.79	70.91	70.36	68.64	**72.76**

Table 4
Image understanding: SSNCA vs. Spatial NCA (see text in Section 5.4 for the setting of Spatial NCA here).

Database	Methods	SD (Eq. (4))	AOD (Eq. (28))	MSE
CBCL	SSNCA	0.7552	0.0679	0.5330
	Spatial NCA	0.8448	0.0269	0.6955
GB2312	SSNCA	0.8175	0.0835	10.831
	Spatial NCA	0.8433	0.0653	11.5268

Table 5
Recognition comparison: SSNCA vs. Spatial NCA (%).

Database	$L=100$		$L=200$	
	SSNCA	Spatial NCA	SSNCA	Spatial NCA
CMU	78.38	**79.06**	**77.58**	74.69
AR	58.69	**63.39**	69.61	**72.76**

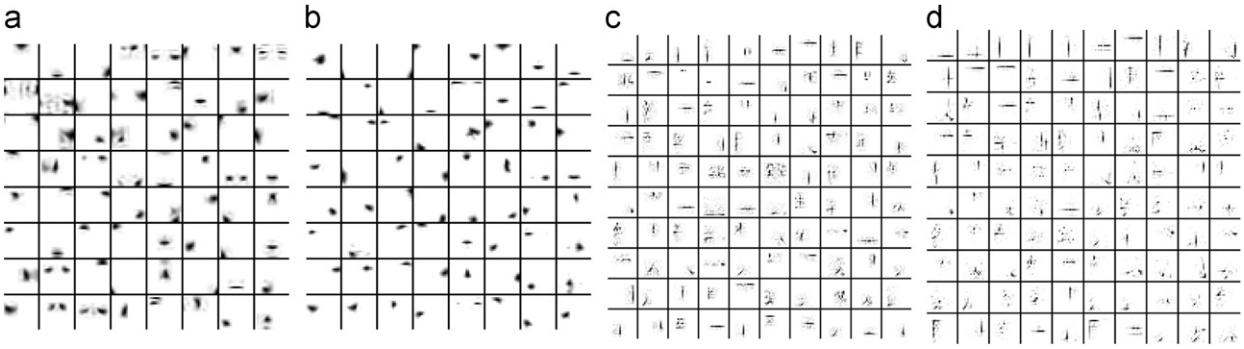

Fig. 8. Illustration of experiment on CBCL: SSNCA vs. Spatial NCA (see text in Section 5.4 for the setting of Spatial NCA here). (a) SSNCA, (b) Spatial NCA, (c) SSNCA and (d) Spatial NCA.

Table 6
Evaluation of pixel dispersion penalty on CBCL: AOD (Average Overlap Degree in Eq. (28)), SD (Sparsity Degree in Eq. (4)), MSE.

Criterion	Method	λ				
		0.1	0.3	0.5	0.7	0.9
AOD	Spatial NMF	0.0217	0.0141	0.0113	0.0090	0.0074
	Spatial NCA	0.0269	0.0163	0.0134	0.0109	0.0097
SD	Spatial NMF	0.8561	0.8779	0.8864	0.8935	0.8987
	Spatial NCA	0.8448	0.8729	0.8811	0.8880	0.8916
MSE	Spatial NMF	0.7121	0.8816	0.9502	1.0649	1.1078
	Spatial NCA	0.6955	0.8047	0.8879	0.9681	1.0155

Table 7
Evaluation of pixel dispersion penalty on GB2312: AOD, SD, MSE.

Criterion	Method	λ				
		0.1	0.3	0.5	0.7	0.9
AOD	Spatial NMF	0.0088	0.0073	0.0066	0.0062	0.0057
	Spatial NCA	0.0653	0.0280	0.0197	0.0161	0.0137
SD	Spatial NMF	0.9456	0.9509	0.9534	0.9546	0.9572
	Spatial NCA	0.8433	0.9032	0.9207	0.9289	0.9342
MSE	Spatial NMF	14.6579	14.8326	14.8908	14.8406	15.1489
	Spatial NCA	11.5268	12.1136	12.4202	12.6183	12.8025

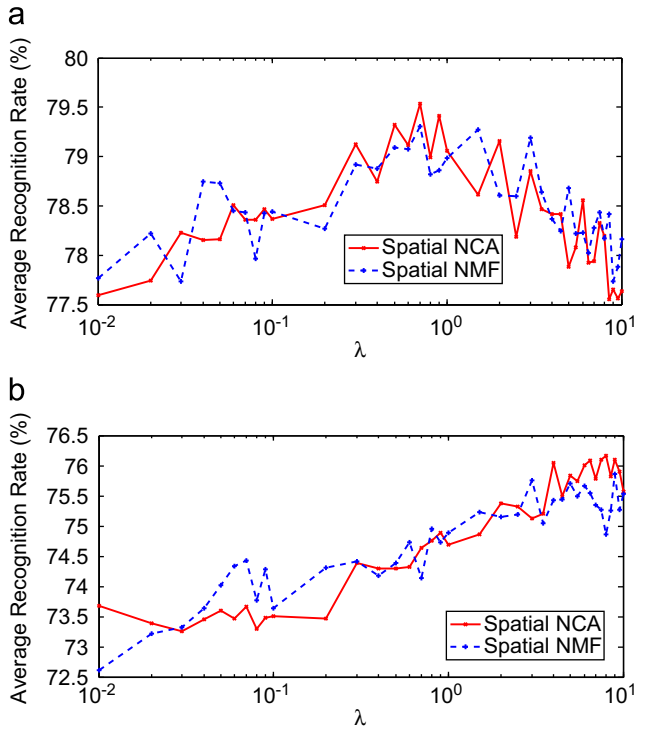

Fig. 9. Illustration of the effect of PDP for recognition on CMU. The x-axis is on logarithm scale. (a) $L=100$ and (b) $L=200$.

- The sparse degree and overlapping degree of the basis images extracted by Spatial NMF and Spatial NCA are always almost the same, while Spatial NCA always has a lower MSE performance.
- In most of the cases, much better recognition performance can be obtained by setting a larger λ value (>1), for example $\lambda=3$ for Spatial NMF and $\lambda=4$ for Spatial NCA on CMU when $L=200$. This indicates much better classification results can be obtained by the proposed Spatial NMF and Spatial NCA when $\lambda>1$.
- The Spatial NCA performs better and more robust than Spatial NMF. This further validates the comparison analysis between these two methods in previous sections.

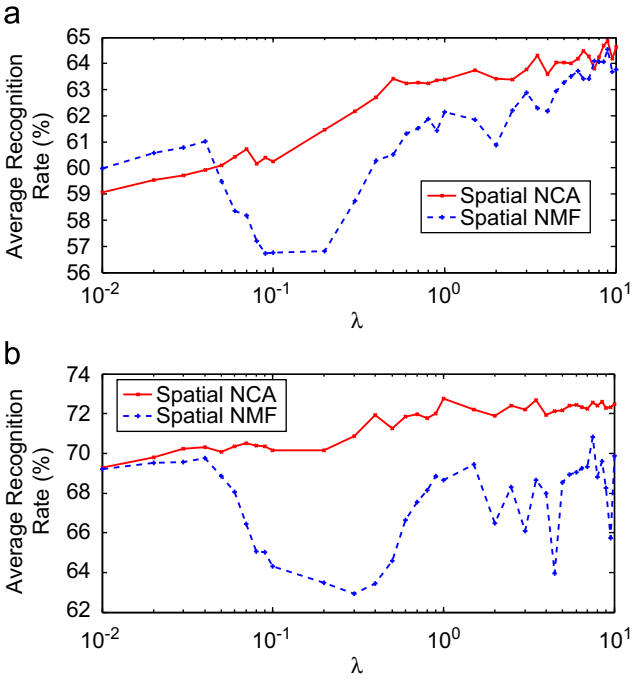

Fig. 10. Illustration of the effect of PDP for recognition on AR. (a) $L=100$ and (b) $L=200$.

Table 8
Spatial NCA and Spatial NMF: data explanation for different numbers of bases (L).

Database	Criterion	Spatial NCA			Spatial NMF		
		$L=36$	$L=64$	$L=100$	$L=36$	$L=64$	$L=100$
CBCL	SD (Eq. (4))	0.8043	0.8448	0.8726	0.8088	0.8561	0.8892
	AOD (Eq. (28))	0.0314	0.0269	0.0227	0.0291	0.0217	0.0163
	MSE	1.3672	0.6955	0.3742	1.4059	0.7121	0.3952

5.5.2. *The number of basis images for data explanation*

We now further discuss the effect of the number of basis images learned by Spatial NCA and Spatial NMF for explaining the given data. Tables 8 and 9 and Figs. 11 and 12 are presented for this purpose on the CBCL and GB2312 datasets respectively. As shown, with more number of basis images, both methods can extract much sparser localized features, i.e. larger SD values and smaller AOD values. Indeed the reconstruction error becomes (much) smaller, which is obvious on GB2312. Even though less basis images are learned, the extracted spatially localized features are still meaningful as image parts shown in Figs. 11 and 12, and it is probably because of the explicit spatial constraint between pixels modeled by the proposed pixel dispersion penalty. However, from another point of view, as shown by the visual results, the spatial localized features become more compact (namely the support area is smaller) which may not more explicitly and obviously explain the parts of image. Hence there should be a balance between achieving high sparsity and intuitively explaining parts in image.

Table 9
Spatial NCA and Spatial NMF: data explanation for different number of bases (L).

Database	Criterion	Spatial NCA			Spatial NMF		
		$L=64$	$L=100$	$L=121$	$L=64$	$L=100$	$L=121$
GB2312	SD (Eq. (4))	0.91	0.9355	0.9457	0.9265	0.9573	0.9703
	AOD (Eq. (28))	0.0197	0.0131	0.011	0.0121	0.0056	0.0032
	MSE	18.2958	12.8823	10.1167	20.3581	15.3095	12.315

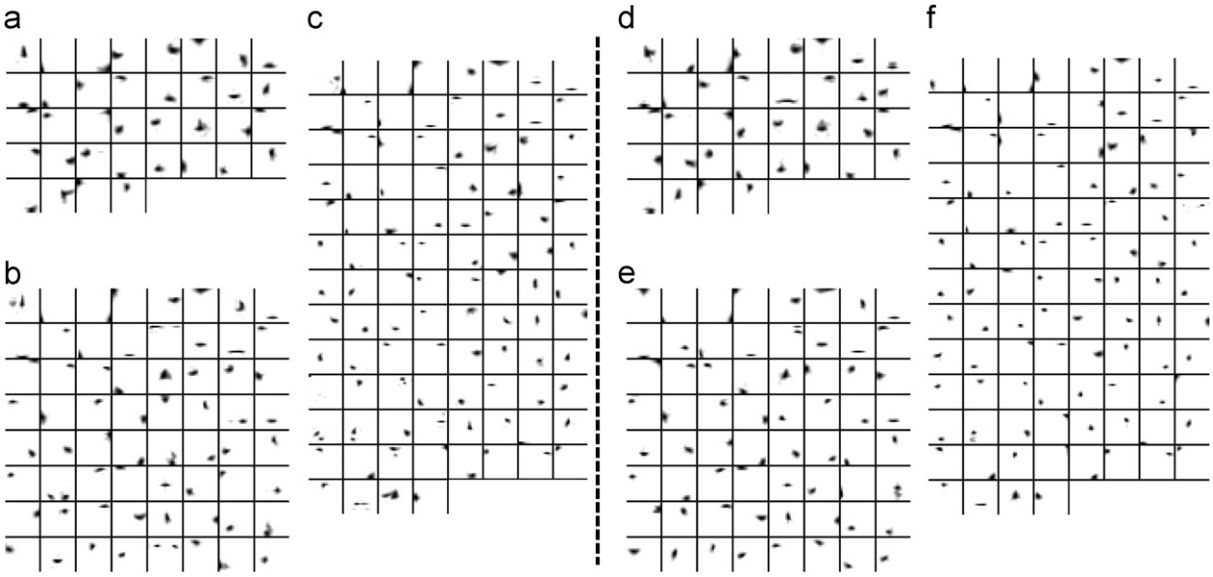

Fig. 11. Spatial NCA and Spatial NMF: basis images for different numbers of bases (L) on CBCL. (a) Spatial NCA: $L=36$, (b) Spatial NCA: $L=64$, (c) Spatial NCA: $L=100$, (d) Spatial NMF: $L=36$, (e) Spatial NMF: $L=64$, and (f) Spatial NMF: $L=100$.

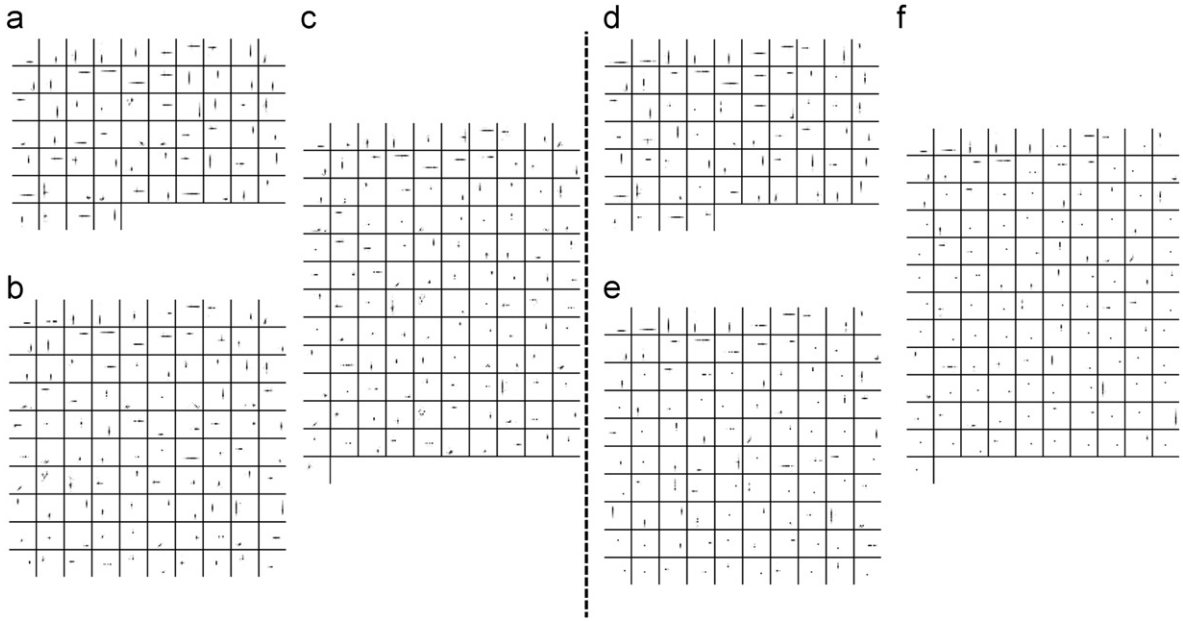

Fig. 12. Spatial NCA and Spatial NMF: basis images for different numbers of bases (L) on GB2312. (a) Spatial NCA: $L=64$, (b) Spatial NCA: $L=100$, (c) Spatial NCA: $L=121$, (d) Spatial NMF: $L=64$, (e) Spatial NMF: $L=100$, and (f) Spatial NMF: $L=121$.

5.5.3. The coefficient part in spatial NCA

Finally, we have an in-depth discussion on the coefficient part in Spatial NCA. Different from Spatial NMF, Spatial NCA releases the non-negativity constraint on coefficients. That is, there may be negative values appear in the coefficient matrix **H**. We now have a discussion on these data values. In Fig. 13, we present the coefficient vectors for an image in CBCL across three values of λ, where the horizontal axis is the entry index of a coefficient vector

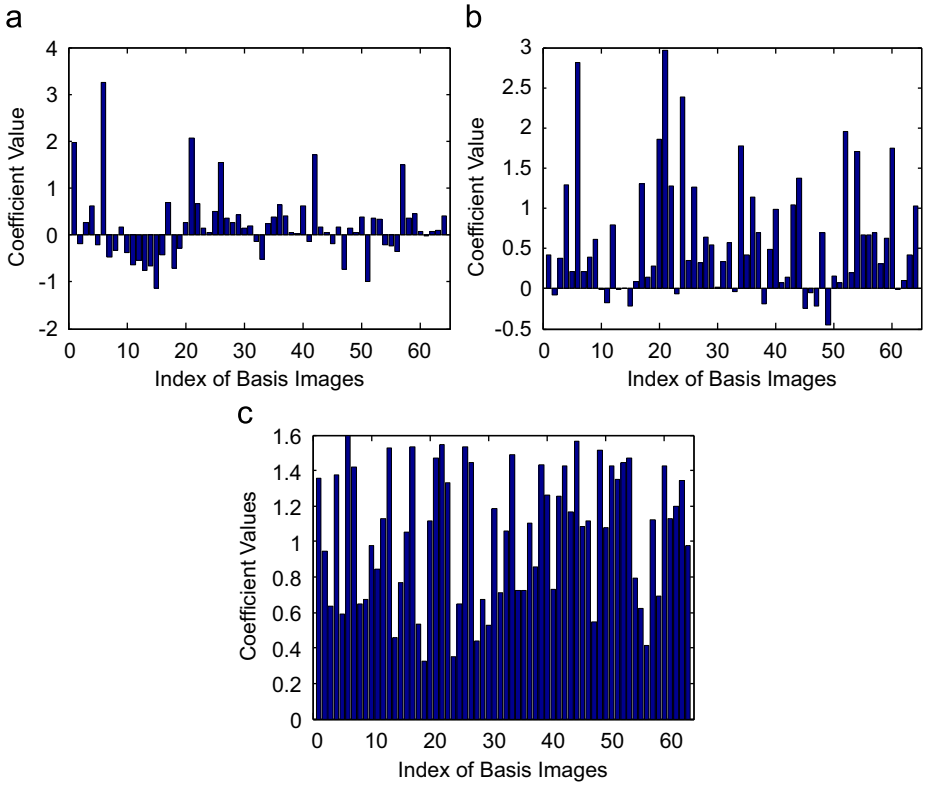

Fig. 13. Example of the distribution of coefficients learned by Spatial NCA for an image sample in CBCL: *x*-axis is the index of basis image, *y*-axis is the corresponding coefficient value. (a) $\lambda=0$, (b) $\lambda=0.1$ and (c) $\lambda=1$.

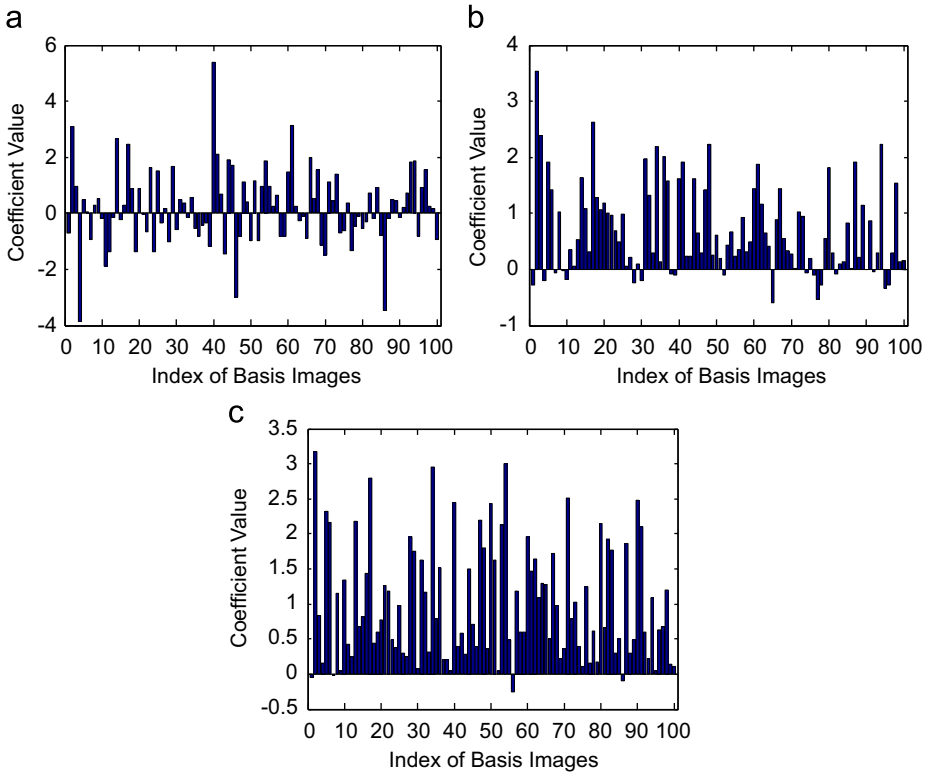

Fig. 14. Example of the distribution of coefficients learned by Spatial NCA for an image sample in AR dataset. (a) $\lambda=0$, (b) $\lambda=0.1$ and (c) $\lambda=1$.

and the vertical axis is the value for each entry. Similar figure is also illustrated for an image in AR dataset in Fig. 14. Note that, since the scale of the basis matrix **W** is different for different λ values, the entry of each coefficient vector is therefore scaled by the scale of the corresponding basis image, so that all coefficient values are investigated at the same magnitude level in the figure. We find that when $\lambda=0$, that is when Spatial NCA becomes NCA, more negative coefficients are observed, and when λ gets much

larger, for example from $\lambda = 0.1$ to $\lambda = 1$, less negative values are observed. The coefficients tend to be non-negative when the proposed pixel dispersion penalty plays a more important role in Spatial NCA. This is not a single observation and similar things can be observed for other images in the same or different datasets when using Spatial NCA as well. Actually, it happens intuitively and naturally, because when more weight is on pixel dispersion penalty, the learned factors are more spatially localized, i.e. the size of the local patch becomes smaller and smaller and the learned basis images are much less overlapped, and therefore all non-negative local parts should be additively combined together in order to reconstruct an image due to the reconstruction error constrained in the criterion. However, even though the negative coefficient becomes small, it still has impact on recognition, e.g. Spatial NCA still achieves better performance than Spatial NMF in general when $\lambda = 1$, as shown in Table 2. This suggests allowing the negative coefficients is a way to balance the reconstruction and recognition performance during extraction of non-negative basis images, and we see from the reported experiments that this can be an important issue for sparse matrix factorization techniques. Such a balance was also investigated and recognized as an important issue for discriminant subspace methods [35]. Even if almost all entries of the learned coefficient matrix by Spatial NCA with a very large λ are ultimately non-negative, Spatial NCA still allows negative values exist in the coefficient matrix **H** in the initial stage of optimization, and our results imply this flexibility may help achieve more robust performance for face recognition under illumination and occlusion. From another point of view, the solution set for Spatial NMF is just a subset of the one for Spatial NCA, and hence it is possible that Spatial NCA performs more flexibly and better.

6. Conclusion

We in this paper explore a novel penalty function called pixel dispersion penalty in order to guide matrix factorization techniques to learn spatially localized non-negative basis images without using any additional pre-specified structured patterns. The pixel dispersion penalty has directly explored the spatial relationship between pixels. Based on the proposed pixel dispersion penalty, we have developed spatial non-negative matrix factorization (Spatial NMF) and spatial non-negative component analysis (Spatial NCA). Extensive experiments have been conducted in order to quantify the performance of the proposed pixel dispersion penalty and the two developed methods against related methods. We find that the pixel dispersion penalty performs more effective for extracting spatially localized basis images and overall better for face recognition under illumination and occlusion. Moreover, by using pixel dispersion penalty, allowing subtractive and additive combinations of non-negative basis images at the same time would yield a more flexible and also effective matrix factorization technique. This leads to (1) good balance between extracting spatially localized basis images and extracting informative basis images and (2) overall better and more robust classification performance. As in [18–24,36], the future developments of this work can be to combine the proposed Spatial NMF and Spatial NCA with the manifold learning smoothed, and (semi-)supervised learning together in order to obtain much better recognition performance. Also, developing an online learning model which can deal with large-scale data processing is also our future work.

Acknowledgments

This research was supported by the National Natural Science of Foundation of China (nos. 61102111, 61173084, 61103155), the NSFC-GuangDong (No. U0835005), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20110171120051), and the 985 Project at Sun Yat-sen University under Grant no. 35000-3181305.

References

[1] B.W. Mel, Think positive to find parts, Nature 401 (1999) 759–760.
[2] A.R. Webb, Statistical Pattern Recognition, 2nd ed., John Wiley & Sons, Ltd, UK, 2002.
[3] M. KirKirbyby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Transaction Pattern on Analysis and Machine Intelligence 12 (1) (1990) 103–108.
[4] R. Jenatton, G. Obozinski, F. Bach, Structured sparse principal component analysis, in: International Conference on Artificial Intelligence and Statistics, 2010.
[5] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[6] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, 2000.
[7] P.O. Hoyer, Nonnegative sparse coding, in: IEEE Workshop Neural Networks for Signal Processing, 2002.
[8] S.Z. Li, X. Hou, H. Zhang, Q. Cheng, Learning spatially localized parts-based representations, in: IEEE Conference on Computer Vision and Pattern Recognition, 2001.
[9] R. Zass, A. Shashua, Nonnegative sparse PCA, in: Advances in Neural Information Processing Systems, 2006.
[10] W.-S. Zheng, S.Z. Li, J.H. Lai, S. Liao, On constrained sparse matrix factorization, in: IEEE International Conference on Computer Vision, 2007.
[11] P.O. Hoyer, Non-negative matrix factorization with sparseness constraints, Journal of Machine Learning Research 5 (12) (2004) 1457–1469.
[12] J. Huang, T. Zhang, D. Metaxas, Learning with structured sparsity, in: International Conference on Machine Learning, 2009.
[13] L. Jacob, G. Obozinski, J. Vert, Group lasso with overlap and graph lasso, in: Annual International Conference on Machine Learning, ACM, 2009, pp. 433–440.
[14] L. He, L. Carin, Exploiting structure in wavelet-based Bayesian compressive sensing, IEEE Transactions on Signal Processing 57 (9) (2009) 3488–3497.
[15] R. Baraniuk, V. Cevher, M. Duarte, C. Hegde, Model-based compressive sensing, IEEE Transactions on Information Theory 56 (4) (2010) 1982–2001.
[16] P. Zhao, G. Rocha, B. Yu, Grouped and Hierarchical Model Selection Through Composite Absolute Penalties, Technical Report 703, Department of Statistics, UC Berkeley.
[17] R. Jenatton, J. Audibert, F. Bach, Structured variable selection with sparsity-inducing norms, Arxiv preprint arXiv:0904.3523.
[18] D. Cai, X. He, X. Wu, J. Han, Non-negative matrix factorization on manifold, in: IEEE International Conference on Data Mining, IEEE, 2008, pp. 63–72.
[19] T. Zhang, B. Fang, Y. Tang, G. He, J. Wen, Topology preserving non-negative matrix factorization for face recognition, IEEE Transactions on Image Processing 17 (4) (2008) 574–584.
[20] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative non-negative matrix factorization with fast gradient descent, IEEE Transactions on Image Processing 20 (7) (2011).
[21] H. Liu, Z. Wu, Non-negative matrix factorization with constraints, in: Twenty-Fourth AAAI Conference on Artificial Intelligence, 2010.
[22] J. Yang, S. Yang, Y. Fu, X. Li, T. Huang, Non-negative graph embedding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[23] X. Liu, S. Yan, H. Jin, Projective nonnegative graph embedding, IEEE Transactions on Image Processing 19 (5) (2010) 1126–1137.
[24] S. Zafeiriou, A. Tefas, I. Buciu, I. Pitas, Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, IEEE Transactions on Neural Networks 17 (3) (2006) 683–695.
[25] A. Pascual-Montano, J.M. Carazo, K. Kochi, D. Lehmann, R.D. Pascual-Marqui, Nonsmooth non-negative matrix factorization (nsNMF), IEEE Transaction on Pattern Analysis and Machine Intelligence 28 (3) (2006) 403–415.
[26] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, in: Advances in Neural Information Processing Systems, 2007.
[27] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1) (2006) 49–67.
[28] H. Park, H. Kim, One-sided non-negative matrix factorization and non-negative centroid dimension reduction for text classification, in: Proceedings of the 4th Workshop on Text Mining, SDM, 2006.
[29] C.H.Q. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, IEEE Transaction on Pattern Analysis and Machine Intelligence 32 (1) (2010) 45–55.
[30] C.-J. Lin, Projected gradient methods for nonnegative matrix factorization, Neural Computation 19 (10) (2007) 2756–2779.
[31] D. Donoho, V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts?, in: Advances in Neural Information Processing Systems, 2003.
[32] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression database, IEEE Transaction on Pattern Analysis and Machine Intelligence 25 (12) (2003) 1615–1619.

[33] I. Buciu, N. Nikolaidis, I. Pitas, Nonnegative matrix factorization in polynomial feature space, IEEE Transactions on Neural Networks 19 (6) (2008) 1090–1100.
[34] A.M. Martinez, A.C. Kak, PCA versus LDA, IEEE Transaction on Pattern Analysis and Machine Intelligence 23 (2) (2001) 228–233.
[35] S. Fidler, D. Skocaj, A. Leonardis, Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (3) (2006) 337–350.
[36] Shengcai Liao, Zhen Lei, Stan Z. Li, Nonnegative matrix factorization with Gibbs random field modeling, in: Proceedings of the 2nd IEEE International Workshop on Subspace, in conjunction with ICCV 2009, Kyoto, Japan, September 27, 2009.

Wei-Shi Zheng received his Ph.D. degree in Applied Mathematics at Sun Yat-sen University, China, 2008. After that, he has been a Postdoctoral Researcher on the European SAMURAI Research Project at the Department of Computer Science, Queen Mary University of London, UK. He has joined Sun Yat-sen University as a faculty under the one-hundred-people program of Sun Yat-sen University in 2011. He has published widely in IEEE TPAMI, IEEE TNN, IEEE TIP, Pattern Recognition, IEEE TSMC-B, IEEE TKDE, ICCV, CVPR and AAAI. His current research interests are in object association and categorization for visual surveillance. He is also interested in discriminant/sparse feature extraction, dimension reduction, kernel methods in machine learning, transfer learning, and face image analysis.

JianHuang Lai was born in 1964. He received the M.Sc. degree in applied mathematics in 1989 and the Ph.D. degree in mathematics in 1999 from Sun Yat-sen University, Guangzhou, China. He joined Sun Yat-sen University in 1989, where currently, he is a Professor with the Department of Electronics and Communication Engineering, School of Information Science and Technology. He has published over 50 papers in the international journals, book chapters, and conferences. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelets and their applications. Dr. Lai had successfully organized the International Conference on Advances in Biometric Personal Authentication' 2004, which was also the Fifth Chinese Conference on Biometric Recognition (Sinobiometrics'04), Guangzhou, in December 2004. He has taken charge of more than four research projects, including NSFC (number 60144001, 60 373 082, 60675016), the Key (Key grant) Project of Chinese Ministry of Education (number 105 134), and NSF of Guangdong, China (number 021 766, 06023194). He has published over 60 papers. He serves as a board member of the Image and Graphics Association of China and also serves as a board member and secretary-general of the Image and Graphics Association of Guangdong.

Shengcai Liao received the B.S. degree in mathematics and applied mathematics from the Sun Yat-sen University, Guangzhou, China, in 2005 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010.

He is currently a Post Doctoral Fellow in the Department of Computer Science and Engineering, Michigan State University. His research interests include computer vision, pattern recognition, and machine learning, with a focus on image and video analysis, particularly face recognition, object detection and recognition, video surveillance, and sparse matrix factorization. He serves as a regular reviewer for several international journals include IJCV, T-PAMI, TIP, and Neurocomputing. He has also served as a program committee member or reviewer for several international conferences, include ICCV, CVPR, ICPR, ICB, BTAS, etc.

Dr. Liao was awarded the Excellence Paper of Motorola Best Student Paper and the 1st Place Best Biometrics Paper in the International Conference on Biometrics on 2006 and 2007, respectively, for his work on face recognition.

Ran He received the BS degree in computer science from the Dalian University of Technology of China and the PhD degree in pattern recognition and intelligent system from the Institute of Automation, Chinese Academy of Sciences, in 2009. He is currently an assistant professor with NLPR (National Laboratory of Pattern Recognition), Institute of Automation, Chinese Academy of Science, Beijing, China. His research interests include information theoretic learning and computer vision. He serves as an associate editor of Neurocomputing(Elsevier). He is a member of the IEEE Computer Society.