Initialization enhancer for non-negative matrix factorization

Zhonglong Zheng\textsuperscript{a,b,}\textsuperscript{*}, Jie Yang\textsuperscript{b}, Yitan Zhu\textsuperscript{c}

\textsuperscript{a}Institute of Information Science and Engineering, P.O. box 143, Zhejiang Normal University, Jinhua 321004, Zhejiang, People’s Republic of China
\textsuperscript{b}Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200030, People’s Republic of China
\textsuperscript{c}Alexandria Research Institute, Virginia Polytechnic Institute and State University, 206 N. Washington Street, Alexandria, VA 22314, USA

Received 25 March 2005; received in revised form 9 September 2005; accepted 12 March 2006
Available online 27 April 2006

Abstract

Non-negative matrix factorization (NMF), proposed recently by Lee and Seung, has been applied to many areas such as dimensionality reduction, image classification image compression, and so on. Based on traditional NMF, researchers have put forward several new algorithms to improve its performance. However, particular emphasis has to be placed on the initialization of NMF because of its local convergence, although it is usually ignored in many documents. In this paper, we explore three initialization methods based on principal component analysis (PCA), fuzzy clustering and Gabor wavelets either for the consideration of computational complexity or the preservation of structure. In addition, the three methods develop an efficient way of selecting the rank of the NMF in low-dimensional space.

Keywords: Non-negative matrix factorization; Principal component analysis; Fuzzy clustering; Gabor wavelet; Dimensionality reduction; Image classification

1. Introduction

In pattern analysis and computer vision, visual recognition of objects is one of the most challenging problems. Approaches to overcoming such problems have focused on several methodologies. Appearance-based representation and recognition is one of the most successfully methods still used today. It involves preprocessing of multidimensional signals, such as face images (Turk and Pentland, 1991), character images (Jiangying Zhou and Lopresti, 1997), speech spectrograms (Bell and Sejnowski, 1995) and so on. In fact, the essence of the preprocessing is the so-called dimensionality reduction. Principal component analysis (PCA), based on second-order statistics, is one of the most popular linear dimensionality reduction methods. It is a well-known fact that PCA is optimal in terms of the reconstruction error (MSE) but not for the separation and classification of classes. In addition, the eigen-face derived from PCA is holistic or can be called dense (Buciu and Pitas, 2004).

Recently, an unsupervised approach called non-negative matrix factorization (NMF) has been proposed by Lee and Seung (Lee and Seung, 1999, 2001). NMF is different by adding its non-negative constraints in contrast to PCA and independent component analysis (ICA) (Barlett et al., 1998). When applied to image analysis and representation, the obtained NMF bases are localized features that correspond with intuitive notions of the parts of the images. It is supported by psychological and physiological evidence that perception of the whole is based on parts-based representations (Mel, 1999). Researchers have proposed several different algorithms based on traditional NMF. Guillamet et al. presented a study of the weighted version of the original NMF (WNMF) (Guillamet et al., 2003). Li et al. developed a variant of NMF, named local non-negative matrix factorization (LNMF), by imposing additional constraints (Li et al., 2001). Buciu et al. proposed a supervised technique called discriminant non-negative matrix factorization (DNMF) by taking into account class information (Buciu and Pitas, 2004). Also,
NMF technique has been applied to many areas (Pauca et al., 2004; Guillamet and Vitrià, 2002; Lee and Lee, 2001).

As we know, however, there are few documents paying much attention to the problem of NMF initialization. In theory, NMF unavoidably converges to local minima. So, the NMF bases will be different given different initialization. The experiments cannot be repeated by others if they do not know the initialization conditions. It is our desire to give proper initialization for a given task in order to improve the performance of NMF either for consideration of computational complexity or for the preservation of data structure. To this end, we explore three techniques, which are PCA, fuzzy clustering and Gabor wavelet, to realize NMF initialization. Furthermore, the three initialization methods provide helpful indications in determining the rank of NMF in low-dimensional space.

In the following sections, we will give a brief introduction to the NMF algorithm. Then, we will present the three initialization methods more formally, together with some illustrative simulations on the face image dataset. Finally, we will discuss how to determine the rank of NMF in low-dimensional space through the three initialization techniques.

2. Non-negative matrix factorization algorithm

Let a set of $N$ training images be given as an $m \times N$ matrix $X$, with each column consisting of the $m$ non-negative pixel values of an image. NMF seeks to find non-negative factors $W$ and $H$ such that

$$X \approx \bar{X} = WH,$$  \hspace{1cm} (1)

where $W \in \mathbb{R}^{m \times r}$, $H \in \mathbb{R}^{r \times N}$. The $r$ columns of $W$ are thought of as basis images derived from the NMF algorithm (Lee and Seung, 1999). $H$ is the encoding variable matrix. Dimensionality reduction is achieved when $r < N$.

NMF imposes the non-negative constraints on both $W$ and $H$. As a consequence, the entries of $W$ and $H$ are all non-negative and hence only non-subtractive basis combinations are allowed. It is believed to be compatible to the intuitive notion of combining parts to form a whole. However, PCA requires columns of $W$ to be orthonormal and the rows of $H$ to be mutually orthogonal. It lacks intuitive meaning in that the linear combination of the bases in $W$ usually involves complex cancellations between positive and negative numbers because of the arbitrary sign of the entries in $W$ and $H$.

To find an approximate factorization of NMF, the cost function that quantifies the quality of the approximation has to be defined. Such a cost function can be constructed using some measure of distance between $W$ and $H$. One natural way is simply to use Euclidean distance between two matrices $X$ and $\bar{X}$ to evaluate the approximation:

$$D_E(X, \bar{X}) = \|X - \bar{X}\|^2 = \sum_{ij} (X_{ij} - \bar{X}_{ij})^2.$$  \hspace{1cm} (2)

This is lower bounded by 0, and equals 0 if and only if $X = \bar{X}$. To minimize $\|X - \bar{X}\|^2$ with respect to $W$ and $H$, subject to the constraints $W, H \geq 0$, the multiplicative update rule is described below:

$$H_{zu} \leftarrow H_{zu} \frac{(W^TX)_{zu}}{(W^TWH)_{zu}}, \quad W_{iz} \leftarrow W_{iz} \frac{(XH^T)_{iz}}{(WHH^T)_{iz}}.$$  \hspace{1cm} (3)

The Euclidean distance is invariant under these updates if and only if $W$ and $H$ are at a stationary point of the distance.

Another useful measure is the divergence of $X$ and $\bar{X}$. It is defined as

$$D_D(X||\bar{X}) = \sum_{ij} \left( X_{ij} \log \frac{X_{ij}}{\bar{X}_{ij}} - X_{ij} + \bar{X}_{ij} \right).$$  \hspace{1cm} (4)

It is also lower bounded 0, and vanishes if and only if $X = \bar{X}$. We call it “divergence” instead of “distance” because it is not symmetric in $X$ and $\bar{X}$. To minimize $D_D(X||\bar{X})$ with respect to $W$ and $H$, subject to the constraints $W, H \geq 0$, the corresponding multiplicative update rule is described as follows:

$$H_{zu} \leftarrow \frac{\sum_i W_{iz}X_{iz}(WH)_{zu}}{\sum_k W_{kz}}, \quad W_{iz} \leftarrow \frac{\sum_u H_{zu}X_{zu}(WH)_{zu}}{\sum_v H_{vu}}.$$  \hspace{1cm} (5)

The divergence is also invariant under these updates if and only if $W$ and $H$ are at a stationary point of the divergence (Lee and Seung, 2001).

Li et al. proposed a refinement of NMF by a slight variation of the divergence algorithm, called LNMF (Li et al., 2001). The objective function of LNMF is

$$D_{LD} = \sum_{i=1}^{m} \sum_{j=1}^{N} \left( X_{ij} \log \frac{X_{ij}}{\bar{X}_{ij}} - X_{ij} + \bar{X}_{ij} + aU_{ij} \right) - b \sum_i V_{ii},$$  \hspace{1cm} (6)

where $a$, $b$ are positive constants and $U = W^TW$, $V = HH^T$. The update rule of the LNMF for $W$ is nearly identical to that in Eq. (5). For $H$, the update rule uses the square root to satisfy the additional constraints:

$$H_{zu} \leftarrow \sqrt{\frac{\sum_i W_{iz}X_{iz}}{\sum_z W_{iz}H_{zu}}}.$$  \hspace{1cm} (7)

Having introduced the framework of NMF, we test NMF on our private SJTU-face-database to extract the basis images to give an illustrative impression of NMF. We select 400 images and each of them is cropped to $64 \times 64$ size. The results are shown in Fig. 1. The NMF basis images look a little like those eigenfaces derived from PCA, to some extent. However, the basis faces in Fig. 1(b) could...
approximate the entire collection of faces using only positive combinations instead of arbitrary combination in PCA. The basis faces shown in Fig. 1(c) are derived from LNMF and show more partial patches to represent images. But we have to note that low convergence is one of the weaknesses of LNMF.

3. Initializing NMF with different techniques

As pointed out in Section 2, NMF seeks to find basis images \( W \) and encoding matrix \( H \) with constraints of non-negativity. In theory, the solution to Eq. (1) is not unique under such constraints. Lee and Seung also provided proofs that both multiplicative update rules converge to local minima of the respective objective functions (Lee and Seung, 2001). As a result, it may lead to different outcomes given different initialization conditions. Generally speaking, there are distinct requirements in data analysis or representation for different applications such as visual coding, clustering and so on. It is our desire to design specific initializations for NMF in order to improve its performance either for the consideration of computational complexity or for the preservation of data structure.

3.1. PCA-based initialization

PCA is one of the best known unsupervised feature extraction methods because of its conceptual simplicity and the existence of efficient algorithms that implement it (Mark and Erkki, 2004). Particularly in the face representation task, faces can be economically represented along the eigenface coordinate space, and approximately reconstructed using just a small collection of eigenfaces and their corresponding projections (‘coefficients’). It is optimal representation in the sense of mean-square error.

Given the \( m \times N \) matrix \( X \) as that of in NMF, the average vector \( \psi \) can be computed as:

\[
\psi = \frac{1}{N} \sum_{i=1}^{N} X_i. \tag{8}
\]

Subtract the mean vector, and then the centered matrix is:

\[
\bar{X} = (X_1 - \psi, \ldots, X_N - \psi). \tag{9}
\]

Using SVD to compute the eigenvectors of the \( X^T X \) instead of \( XX^T \) (because usually \( m \gg N \)), up to \( N \) eigenvalues and eigenvectors are obtained. The eigenvector matrix \( W \) is constructed by keeping only \( r \) eigenvectors (corresponding to the \( r \) largest eigenvalues \( \lambda_i \)) as column vectors. And \( H \) is an \( r \times N \) matrix containing the encoding coefficients. The reconstruction error is

\[
E_{\text{mse}} = E[|X - WH|^2]. \tag{10}
\]

The criterion of selecting \( r \) is usually as follows:

\[
\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \geq \alpha. \tag{11}
\]

Comparing Eq. (10) with Eq. (2), we find they are very similar to each other. If \( \alpha \) is large enough (of course not more than 1), PCA can find the direction in input space where most of the energy of the input lies. Motivated by this starting point, we take the PCA-based method into account to initialize NMF. However, the sign of variables in \( W \) and \( H \) of Eq. (10) is arbitrary. In the NMF procedure, all variables in \( W \) and \( H \) are non-negative. In order to utilize \( W \) and \( H \) obtained from PCA, we introduce such a nonlinear operator \( p \) that

\[
p(W) = p(W_+) = \max(0, W_+),
\]

\[
p(H) = p(H_+) = \max(0, H_+). \tag{12}
\]

In this case, the reconstruction error is

\[
E_{\text{mse+}} = E[|X - p(W)p(H)|^2]. \tag{13}
\]

Taking the mean matrix \( \psi \) composed of mean vector \( \psi \) into consideration, we rewrite Eq. (13) as

\[
E_{\text{mse+}} = E[|X - (p(W)p(H) + \psi)|^2], \tag{14}
\]

where \( X \) is the original data matrix. Obviously, Eq. (14) is equivalent to Eq. (2). If we initialize both factors \( W \) and \( H \) of NMF with \( p(W) \) and \( p(H) \), then it is equal to minimize Eq. (13) or minimize Eq. (2).

The reason why we adopt the PCA-based method to initialize \( W \) and \( H \) lies in two aspects. One is that there is much information contained in PCA subspace such that...
they give a proper approximation to the original data. However, the approximation cannot be controlled by random initialization. The other is that the operator $p$ enhances the sparseness of $W$: one possible goal of NMF. On the contrary, random initialization is somewhat counterintuitive to this. The visual images of $random(W)$ and $p(W)$ are shown in Fig. 2, respectively. Comparing the images shown in Fig. 2(b) and those in Fig. 1(d), we can conclude that operator $p$ makes the basis images sparser. The sparsity of basis images will slightly reduce the computational complexity. Once an element is zero, it remains zero under the multiplicative update rules. So it does not need to compute an element if it is initialized to be zero.

To initialize NMF using the PCA-based method, we have to choose the rank of the low-dimensional space. As stated in Eq. (11), we choose $r = 25$ when setting $z = 0.9$. The ordering of columns in $W$ is unimportant only if the ordering of rows in $H$ is kept consistent. Fig. 3 illustrates the improvement of the basis images after different iterations using Euclidean distance NMF. For visualization, each basis image is normalized to have elements with a maximum value 1.

We find from Fig. 3 that NMF starts to model faces from the beginning of the random initialization. As the algorithm progresses, the random initialization begins to lose its original noise and more closely resemble the faces expected. However, the PCA-based initialization already works in the “face space” with predefined parts. After 300 iterations, the basis images obtained by random initialization are distinctly different from those initial images shown in Fig. 2(a). However, with PCA-based initialization, the algorithm has a better start at the beginning to emphasize facial components such as eyes, nose and eyebrows. Comparing the basis images after 1000 iterations with those initial images shown in Fig. 2(b), the reader will find they are very similar to each other.

We further analyze the basis images $W$ based on the results shown in Fig. 3. One natural analysis, illustrated in Fig. 6, is the minimization error of resulting approximation. Here, the criterion of error analysis is measured by the Frobenius norm:

$$
\|X - WH\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{N} c_{ij}^2}.
$$

Fig. 6 shows that the approximation error from both PCA-based and random initialization decreases fast before 200 iterations. After 300 iterations, the error rate decreases very slowly. Before 300 iterations, the PCA-based method almost achieves superior approximation than the random one because of its better initialization. In the long term, random initialization will achieve a lower error than the PCA-based method. This is because PCA-based initialization enforces certain restrictions on $W$ and $H$, and the NMF algorithm using Euclidean distance measure cannot pull it out of a local minima. However, it is doubtful if this long-term error is meaningful because of its very low decreasing rate.

The other straightforward analysis is the sparsity and orthogonality of the normalized basis images $W$. For the sparsity of $w_i$, we define the measure as

$$
\|w\|_p = \left(\sum_{i=1}^{m} |w_i^p|\right)^{1/p}, \quad 0 < p \leq 1.
$$

In fact, this is a quasi-norm for $0 < p < 1$ since it does not satisfy the triangle inequality, but only the weaker conditions: $\|x + y\|_p \leq 2^{-1/p} (\|x\|_p + \|y\|_p)$ where $p' = p/(p-1)$ is the conjugate exponent of $p$; and $\|x + y\|_p \leq \|x\|_p + \|y\|_p$ (Donoho, 2001).

For the orthogonality of $w_i$, we define the measure as the sum of the $(N-1)$ inner products:

$$
O(w_i) = \sum_{j=1, j \neq i}^{r} w_i^T w_j, \quad i = 1, \ldots, r.
$$

If $W$ is an orthogonal matrix, $O(w_i)$ equals 0. Otherwise, $O(w_i)$ will increase with the orthogonality decreasing. The sparsity and orthogonality results of random and PCA-based initialization after 1000 iterations are shown in Fig. 4, respectively. Here we see that the basis images in $W$ of PCA-based initialization after 1000 iterations are more...
sparse and orthogonal than those obtained by the random method.

3.2. Clustering-based initialization

In conventional clustering, each sample in the original dataset $X$ is either assigned to or not assigned to a group. The fuzzy clustering method, which applies the concept of fuzzy sets to cluster analysis, gives membership to groups at each point of the dataset by a membership function (Mihaela et al., 2002). It has been proved to be very suited to deal with the imprecise data. The advantage of fuzzy clustering is its adaptation to noisy data and classes that are not well separated. The fuzzy c-means (FCM) clustering algorithm used in this paper is described as follows:

(a) Fix the number of cluster $c$, fuzzier $q$, termination tolerance $\varepsilon$ and any inner product norm metric for $\mathbb{R}^n$.  

Fig. 3. (a)–(j) Improvement of basis images for random and PCA-based methods at different iterations. The top two rows are the results of random initialization; the other two are the results of PCA-based initialization.
(b) Initialize cluster center matrix $W$.
(c) Compute dissimilarity matrix $D_{i,j}$ which is the dissimilarity between $w_i$ and $x_j$.
(d) Compute the membership matrix
\[ u_{i,j} = \frac{1}{\sum_{j=1}^{N} (D_{i,j}/D_{i,j})^{1/(q-1)}}. \]
(e) Update cluster center $W$
\[ w_i = \frac{\sum_{j=1}^{N} u_{i,j}^q x_j}{\sum_{j=1}^{N} u_{i,j}^q}. \]
(f) Calculate objective function
\[ J = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{i,j}^q D_{i,j} \]
and if the tolerance $\varepsilon$ is satisfied, the algorithm is terminated; otherwise, go to (c).

Before we carried out the FCM, the column vectors $x_i$ of the original data matrix $X$ are normalized to have elements with a maximum value 1. Using the FCM algorithm, we can obtain the centroid matrix $W$ and the membership $u$ of
each sample in the dataset. The centroid $W$ is shown in Fig. 5(a) given $c = 25$. To some extent, each basis image contained in $W$ is a combination of one or more faces in the original dataset. In Fig. 5(a), some images look more blurry than others because the corresponding cluster contains more faces than other clusters. Motivated by this, let us think about the following least-square problem:

$$X \approx WL, \quad L = (W^TW)^{-1}W^TX,$$

where the existence of $(W^TW)^{-1}$ depends on the linear independence of the columns of $W$ (Dhillon and Modha, 2001). Unfortunately, we have to note that the elements in the least-squares solution $L$ are unconstrained in sign. Therefore, we have to seek another kind of matrix $H$ whose elements are non-negative to approximate $X$ together with $W$ in order to initialize NMF. Assume that each cluster derived from FCM is compact; a natural way is to replace every $x_i$ in $X$ with its corresponding cluster centroid $w_i$. This leads to the following equation:

$$H_{ij} = \begin{cases} 1, & i := \arg \max_i (U_{ij}), \quad j = 1, \ldots, N, \\ 0, & \text{otherwise} \end{cases}$$

In other words, the values of elements in $H$ are automatically determined by the membership matrix. The maximum value in columns $U_{ij}$ corresponds to 1 in $H_{ij}$. Then the approximation is $X \approx WH$. Let us compare the objective function Eq. (2) with the cost function of the FCM algorithm under this constraint. In this case, we rewrite Eq. (2) as

$$D_E(X, \tilde{X}) = \sum_{j=1}^{N} \|X_j - W_{cj}\|^2,$$

where $W_{cj}$ denotes the centroid that the sample $X_j$ belongs to.

The FCM cost function is transformed to

$$J = \sum_{j=1}^{N} D_j.$$ 

It is interesting that they are equivalent to each other. Therefore, it is appropriate using the FCM-based method to initialize NMF. The iteration results are shown in Fig. 5(b)–(f) under this initialization.

The experiment results illustrated in Fig. 5 show that the basis images do not change dramatically even after 500 iterations. They are different from those of random initialization shown in Fig. 3. The reason is that the clusters derived from FCM on this small dataset are compact enough that the approximation error is very small at the head start. So it is easily prone to converge to some certain local minima. And the multiplicative update rule cannot change basis much and pull it out because of its better approximation. The approximation error is shown in Fig. 6.

However, as the basis images in $W$ are the centroids of different clusters, they are combinations of one or more faces. So the bases become denser than a single face image and the orthogonality decreases. The sparsity and orthogonality of $W$ using Eq. (17) is shown in Fig. 7. Compared with those shown in Fig. 4(c) and (d), the values in Fig. 7 are much larger.

### 3.3. Gabor-based initialization

Gabor wavelet is a powerful tool in image feature extraction and Lee described the good performance of Gabor wavelet in rotation, scaling and translation property (Daugman, 1988; Liu and Wechsler, 2001). The Gabor wavelet (Lee, 1996) can be defined as follows

$$\psi_{\mu,\nu}(z) = \frac{|k_{\mu,\nu}|^2}{\sigma^2} \exp\left(-\frac{\|k_{\mu,\nu}\|^2 z^2}{2\sigma^2}\right) \left[\exp(k_{\mu,\nu}z - e^{-\sigma^2/2})\right],$$

where $\mu$ and $\nu$ define the orientation and scale of the Gabor kernels, and the wave vector $k_{\mu,\nu}$ is defined as

$$k_{\mu,\nu} = k_\omega e^{i\phi_{\mu,\nu}}.$$
Once these parameters are given, the Gabor feature representation of an image $I(z)$ is

$$G_{\mu_s}(z) = I(z) \ast \psi_{\mu_s}(z),$$

where $z = (x, y)$ and $\ast$ is the convolution operator.

When convolved with Gabor wavelets, the image is transformed into a set of image features at certain scales and orientations. Conversely, the image can be reconstructed from these image features. Motivated by this standpoint, we try to initialize NMF with the Gabor-based method. In order to utilize Gabor wavelets to get the initial $W$ for NMF, we just compute the mean face of the dataset. That is to say, we compute the center of all the images in the dataset. The mean-face image is shown in Fig. 8(a).

Here, we choose five scales and five orientations for the sake of comparison although four scales and eight orientations are selected for most cases. The initial basis and iterative results are show in Fig. 8.

The approximation error using Gabor-based initialization at different iteration is illustrated in Fig. 6. The error curve is similar to that of random initialization. In this case, all elements in the right factor $H$ are initialized to be 1. Imprecisely speaking, the samples in the original dataset are approximated by the computed mean face under such initialization. It is obvious that such approximation is inferior to both the clustering-based method and the PCA-based method. However, the features contained in the basis images are different from those of clustering-based and PCA-based methods because of the scale and orientation property. Although the basis images after 1000 iterations also look whole-face alike, there may exist some constraints (unexplored) to strengthen the features in terms of scale or orientation according to different applications. The sparsity and orthogonality of the bases are similar Fig. 9.
4. Discussion of choosing rank $r$

We have discussed PCA-based, clustering-based and Gabor-based methods in detail in the previous sections to initialize NMF. For the sake of comparison, all experiments choose rank $r = 25$ for the low-dimensional space.

Is there any other selection of the rank $r$? The answer is definitely “yes”. Generally speaking, the rank of low-dimensional space $r$ is difficult to determine in dimensionality reduction task. However, the criterion of how to choose rank $r$ is alternative along with different applications. For example, the online image retrieval demands the rank $r$ be as small as possible to realize real time requirement (Gunther and Beretta, 2001). For image compression, however, the rank $r$ should not be too small in order to achieve better approximation to the originals. In addition, the common users who are not familiar with the specific application or property of the original dataset would feel puzzled to select a particular rank $r$ for the low-dimensional space.

The initialization methods mentioned above give some suggestions in determining the rank $r$ comparing with random initialization, to some extent. For the PCA-based method, it is mainly concerned with energy conservation as stated in Eq. (11). This standpoint is very popular in PCA-related applications. It still gives directions when used to initialize NMF. For the clustering-based method, the clusters are the indication of selected rank $r$ under this type of initialization. It is believed that there are some other clustering algorithms that can be adopted to initialize NMF (Wild, 2003). For the Gabor-based method, it originates from its multi-scale and orientation property. When some other constraints are imposed on scale and orientation, the method may show its potential benefit for some specific applications.

If we compute NMF given different values $r$ in order to, for example, gain better approximation, it could be computationally very expensive. For different datasets, the rank $r$ has to be re-computed without rank selection strategy. It is the reason why we use these methods to initialize NMF. Furthermore, we find that the error decreases dramatically before 200 iterations. Both PCA-based and clustering-based methods achieve better approximation accuracy than random initialization from head start, though they are inferior in the long term. In our view, the approximation accuracy of PCA-based and clustering-based methods could be satisfactory in most cases because the decreasing error rate is small enough to be ignored in the long run. Even for the Gabor-based method, the iteration result is nearly the same as that of random initialization. One can expected to benefit from the proposed methods in choosing rank $r$.

5. Conclusion

It has been proven that non-negative matrix factorization is a useful tool in the analysis of a diverse range of data. One of its most salient properties is that the resulting bases are often intuitive and easy to interpret because they are non-negative and sparse. Researchers often take random initialization into account when utilizing NMF. In fact, random initialization may make the experiments unrepeatable because of its local minima property, although neural networks are not (Baldi and Hornik, 1995). In this paper, we discuss the NMF algorithm, with emphasis on the initialization problem. The proposed methods are based on PCA, clustering and Gabor techniques. Comparing with random initialization, the three methods demonstrate their superiority either in the fast convergence in prophase or in the structure preservation. We believe that the performance of NMF will be enhanced with the proposed initialization methods in data analysis.

Acknowledgement

The authors wish to acknowledge that this work is supported by Fundamental Project of Shanghai under grant number 03DZ14015. The authors would like to thank Dr. Yonggang Wang for useful discussion.

References

Zhonglong Zheng received the B.A. degree in EE from University of Petroleum, China in 1999 and doctoral degree in the institute of image processing & pattern recognition of Shanghai JiaoTong University, China. Now he is a associate professor in the institute of information science and engineering, Zhejiang Normal University. His current research interests include image analysis, video tracking, face detection and recognition, and data mining.

Jie Yang received a Ph.D. degree in computer science from University of Hamburg, Germany in 1994. Dr. Yang is now the Professor and Vice-Director of Institute of Image Processing & Pattern Recognition of Shanghai JiaoTong University, China. His research interests include artificial intelligence, image processing, computer vision, data mining and pattern recognition. He is the author of more than 80 scientific publications and one book.

Yitan Zhu received his B.A. and M.S. degree in pattern recognition and intelligent system from Shanghai Jia Tong University, China. Now he is a Ph.D. student of Alexandria Research Institute, Virginia Polytechnic Institute and State University.