International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 19, No. 4 (2005) 495–511

World Scientific
www.worldscientific.com

NON-NEGATIVE MATRIX FACTORIZATION FRAMEWORK FOR FACE RECOGNITION*

YUAN WANG† and YUNDE JIA‡

Computer Science Department, Beijing Institute of Technology
Beijing, 100081, P. R. China
†*ywang@email.arizona.edu*
‡*yjiar@bit.edu.cn*

CHANGBO HU§ and MATTHEW TURK¶

Computer Science Department, University of California
Santa Barbara, CA 93106, USA
§*cbhu@cs.ucsb.edu*
¶*mturk@cs.ucsb.edu*

Non-negative Matrix Factorization (NMF) is a part-based image representation method which adds a non-negativity constraint to matrix factorization. NMF is compatible with the intuitive notion of combining parts to form a whole face. In this paper, we propose a framework of face recognition by adding NMF constraint and classifier constraints to matrix factorization to get both intuitive features and good recognition results. Based on the framework, we present two novel subspace methods: Fisher Non-negative Matrix Factorization (FNMF) and PCA Non-negative Matrix Factorization (PNMF). FNMF adds both the non-negative constraint and the Fisher constraint to matrix factorization. The Fisher constraint maximizes the between-class scatter and minimizes the within-class scatter of face samples. Subsequently, FNMF improves the capability of face recognition. PNMF adds the non-negative constraint and characteristics of PCA, such as maximizing the variance of output coordinates, orthogonal bases, etc. to matrix factorization. Therefore, we can get intuitive features and desirable PCA characteristics. Our experiments show that FNMF and PNMF achieve better face recognition performance than NMF and Local NMF.

Keywords: Face recognition; non-negative factorization; Fisher discriminant analysis; local facial feature.

1. Introduction

Face recognition is very challenging due to the wide variety of illumination, facial expression and pose variations. It has received extensive attention during the past 20 years, not only because it has several potential applications in areas such as

*This work was partially supported by Grant No. 60473049 from the Chinese National Science Foundation.

Human Computer Interaction (HCI), biometrics and security, but also because it is a prototypical pattern recognition problem whose solution would help in many other classification problems.

Subspace methods have demonstrated their success in numerous visual recognition tasks such as face recognition, face detection and tracking. These methods, such as Principle Component Analysis (PCA)[1,6,15] Fisher Linear Discriminant Analysis (FLDA),[2] Independent Component Analysis (ICA)[1,4,3] and Non-negative Matrix Factorization (NMF),[8,9] learn to represent a face as a linear combination of basis images, but in different ways. The basis images of PCA are orthogonal and have a statistical interpretation as the directions of largest variance. FLDA seeks to find a linear transformation that can maximize the between-class scatter and minimize the within-class scatter. ICA is a linear nonorthogonal transform that yields a representation in which unknown linear mixtures of multidimensional random variables are made as statistically independent as possible. NMF factorizes the image database into two matrix factors whose entries are all non-negative and produces a part-based representation of images because it allows only additive, not subtractive, combinations of basis components. For this reason, the non-negativity constraints are compatible with the intuitive notion of combining parts to form a whole. Because a part-based representation can naturally deal with partial occlusion and some illumination problems, it has received much attention recently.

Li and Feng proposed Local Non-negative Matrix Factorization (LNMF)[5,10] to achieve a more localized NMF algorithm with the aim of computing spatially localized bases from a face database by adding three constraints that modify the objective function in the NMF algorithm.

In this paper, we propose a framework of face recognition by adding NMF constraint and classifier constraints to matrix factorization to get both intuitive features and good recognition results. We present Fisher Non-negative Matrix Factorization (FNMF) by adding the Fisher constraint and NMF to matrix factorization. We also describe PCA Non-negative Matrix Factorization (PNMF) which adds characteristics of PCA and NMF to matrix factorization in order to get intuitive features and characteristics of PCA, such as maximizing the variance of output coordinates, orthogonal bases, etc. Our experiments show that they perform better than NMF and LNMF for face recognition.

2. Previous Work

There is psychological[12] and physiological[11,16] evidence for part-based representations in the brain. Lee and Seung proposed NMF[8,9] for learning parts of faces, and the non-negative constraint added to the matrix factorization is compatible with the intuitive notion of combining parts to form a whole face. However, the NMF algorithm produces global, not spatially localized, parts from the training set. To improve the NMF algorithm, Local NMF (LNMF)[10] was proposed for learning

spatially localized, part-based representations of visual patterns. The remainder of this section will introduce NMF and LNMF.

A database of m face images, each of which contains n non-negative pixel values, is represented by an $n \times m$ matrix V, where each column denotes one of the m facial images. Basis images computed from the database are denoted by an $n \times r$ matrix W, where r is the number of basis images. To reduce the dimensionality of V, r should be less than m. Hence the factorization is

$$V \approx WH \tag{1}$$

where H consists of the coefficients by which a face is represented with a linear combination of basis images.

Many matrix factorizations allow the entries of W and H to be of arbitrary sign. Therefore, the basis images of this kind do not have an obvious visual interpretation because there are complex cancellations between positive and negative numbers when the basis images are used in linear combinations. However, a matrix factorization with a non-negative constraint can produce basis images that have an intuitive meaning, since the entries of W and H are all non-negative.

2.1. *NMF*

NMF[8] enforces the non-negative constraints on W and H. Thus the basis images can be combined to form a whole face in an intuitive, additive fashion. NMF uses the divergence of V from its approximation $Y = WH$ as the measure of cost for factorizing V into WH. The divergence function, used as an objective function in NMF, is defined as:

$$D(V\|Y) = \sum_{i,j} \left(v_{ij} \log \frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij} \right). \tag{2}$$

NMF factorization is a solution to the following optimization problem

$$\begin{array}{ll} \min\limits_{B,H} & D(V \| WH) \\ \text{s.t.} & W, H \geq 0, \quad \sum_i b_{ij} = 1 \quad \forall j \end{array}$$

where $W, H \geq 0$ indicates that all elements of W and H are to be non-negative; bj are the basis images. This where optimization can be done by using the following multiplicative update rules:

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu} \tag{3a}$$

$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}} \tag{3b}$$

$$H_{a\mu} \leftarrow H_{au} \sum_{i} W_{ia} \frac{V_{i\mu}}{W_{i\mu}}. \tag{3c}$$

2.2. *LNMF*

Local NMF[10] aims to improve the locality of the learned features by imposing additional constraints. It incorporates the following three constraints into the original NMF formulation.

- LNMF attempts to minimize the number of basis components required to represent V. This implies that a basis component should not be further decomposed into more components.
- To minimize redundancy between different bases, LNMF attempts to make different bases as orthogonal as possible.
- Only bases containing the most important information should be retained. LNMF attempts to maximize the total "activity" on each component, i.e. the total squared projection coefficients summed over all training images.

LNMF incorporates the above constraints into the original NMF formulation and defines the following constrained divergence as the objective function:

$$D(V\|Y) = \sum_{i,j}\left(v_{ij}\log\frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}\right) + \alpha\sum_{ij}u_{ij} - \beta\sum_{i}q_{ii} \tag{4}$$

where $Y = WH$, $\alpha, \beta > 0$ are constants, $(W^TW) = U = [u_{ij}]$, and $(HH^T = Q = [q_{ij}])$. This optimization can be done by using a multiplicative update rules which was presented by Li.[10]

3. Fisher NMF

To achieve good recognition results and also get intuitive bases, we propose a novel subspace method using Fisher Linear Discriminant Analysis (FLDA), called Fisher Non-negative Matrix Factorization (FNMF).

FLDA has been successfully applied to the problem of face recognition. The main idea of FNMF is to add the Fisher constraint to the original NMF formulation. Because the columns of the encoding matrix H have a one-to-one correspondence with the columns of the original matrix V, we seek to maximize the between-class scatter and minimize the within-class scatter of H.

We define the following constrained divergence as the new objective function for FNMF:

$$D(V\|Y) = \sum_{i,j}\left(v_{ij}\log\frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij}\right) + \alpha S_W - \alpha S_B \tag{5}$$

where $\alpha > 0$ is a constant, S_W is the within-class scatter of the encoding matrix H, and S_B is the between-class scatter of H. Let n_i denote the number of vectors in the ith class and C the number of classes. We define S_W and S_B as

follows:

$$S_W = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{n_i}\sum_{j=1}^{n_i}(h_j - u_i)(h_j - u_i)^T \tag{6}$$

$$S_B = \frac{1}{C(C-1)}\sum_{i=1}^{C}\sum_{j=1}^{C}(u_i - u_j)(u_i - u_j)^T \tag{7}$$

where $u_i = \frac{1}{n_i}\sum_{j=1}^{n_i} h_j$ denotes the mean values of class i in H.

The following update rules implement a local solution to the above constrained minimization. The convergence proof of FNMF is shown in the appendix.

$$h_{kl} \leftarrow -b + \sqrt{b^2 + 4\left(\sum_i v_{il}\frac{w_{ik}h'_{kl}}{\sum_k w_{ik}h'_{kl}}\right)\left(\frac{2}{n_i C} - \frac{4}{n_i^2(C-1)}\right)} \tag{8}$$

$$w_{kl} \leftarrow \frac{w_{kl}\sum_j v_{kj}\frac{h_{lj}}{\sum_k w_{kl}h_{lj}}}{\sum_j h_{lj}} \tag{9}$$

$$w_{kl} \leftarrow \frac{w_{kl}}{\sum_k w_{kl}} \tag{10}$$

where

$$b = \frac{4}{n_i C(C-1)}\sum_j\left(u_{kj} - \left(u_{ki} - \frac{h'_{kl}}{n_i}\right)\right) - \frac{2}{n_i C}u_{ki} + 1.$$

4. PNMF

4.1. *PCA: Statistical perspective*

4.1.1. *Maximizing the variance of output coordinates*

The property PCA seeks to maximize is the spread of the projection of the sample data on the new axes.

Assume that the input vector v is reduced to a single output component $h = w^T v$. PCA is looking for a w whose direction maximizes the variance of the output component h, i.e. PCA is looking for a unit vector w which maximize $\sum_i \left(w^T v_i\right)^2$. The projected points onto the axis represented by the vector w are as spread as possible (in a least squares sense). The optimization problem takes the following form:

$$\max \frac{1}{2}\|w^T V\|^2.$$

4.1.2. *Decorrelation: Diagonalization of the covariance matrix*

The existence of correlations among the components (features) of the input signal is a sign of redundancy, therefore from the point of view of transforming the input

representation into one which is less redundant, PCA would like to find a transformation $h = w^T v$ with an output representation h with a diagonal covariance matrix $\sum_h$, i.e. the components of h are uncorrelated.

4.2. *Algorithm*

Of the three constraints that LNMF imposes on NMF, one is similar to PCA in that it constrains the bases to be orthogonal to each other. But orthogonal constraint is only one of the characteristics of PCA.

PCA has been successfully applied to the problem of face recognition. To achieve true characteristics of PCA and also get intuitive bases, we propose another novel subspace method using Principle Component Analysis (PCA), called PCA Non-negative Matrix Factorization (PNMF).

The main idea of PNMF is to add the PCA constraint to the original NMF formulation. Here we add four constraints to NMF:

- Maximizing the Variance of Output Coordinates by $\max \frac{1}{2}\|w^T V\|^2$ and $\max \sum_i (H^T H)_{ii}$.
- Diagonalize the Covariance Matrix of the projected vector by $\min \sum_{i \neq j} ((W^T V)(V^T W))_{ij}$.
- Make different bases as orthogonal as possible by $\min \sum_{i \neq j} w_i^T w_j$ to minimize the redundancy between different bases.

We define the following constrained divergence as the new objective function for PNMF:

$$D(V\|WH) = \sum_{i,j} \left(v_{ij} \log \frac{v_{ij}}{y_{ij}} - v_{ij} + y_{ij} \right) - \alpha \frac{1}{2} \|w^T V\|^2 - \beta \sum_i (HH^T)_{ii} + \gamma \sum_{i \neq j} w_i^T w_j + \lambda \sum_{i \neq j} (W^T V V^T W)_{ij} \tag{11}$$

where $\alpha, \beta, \gamma, \lambda > 0$ are constants.

The following update rules implement a local solution to the above constrained minimization.

$$h_{kl} \leftarrow \sqrt{\sum_i v_{il} \frac{w_{ik} h'_{kl}}{\sum_k b_{ik} h'_{kl}}} \tag{12}$$

$$w_{kl} \leftarrow \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{13}$$

$$w_{kl} \leftarrow \frac{w_{kl}}{\sum_k w_{kl}} \tag{14}$$

where

$$
\begin{aligned}
a &= \lambda \sum_{i \neq j} v_{ki} (v^T)_{jk} - \alpha \sum_i (v)_{ki}^2 \\
b &= \sum_j h_{lj} + 2\gamma \left(\sum_j w_{kj} - w'_{kl} \right) - \alpha \left(\sum_i ((v^T)_{ik} \left(\sum_j (v^T)_{ij} \times w_{jl} \right) \right) \\
&\quad - \sum_i (v^T)_{ik}^2 w'_{kl} \Bigg) \Bigg) + \lambda \left(\sum_j (v_{kj} \left(\sum_z \sum_i (w^T)_{li} v_{iz} \right) \right) \\
&\quad - \sum_i \left(v_{ki} \left(\sum_j w_{lj} v_{ji} \right) \right) - w_{kl} \sum_{i \neq j} v_{ki} (v^T)_{jk} \Bigg) \\
c &= -w'_{kl} \sum_j v_{kj} \frac{h_{lj}}{\sum_k w'_{kl} h_{lj}}.
\end{aligned}
$$

The convergence proof is similar to that of FNMF.

5. Experiments

Our experiments were performed on two benchmarks: the ORL database and a dataset from the FERET database. The Nearest Neighbor (NN) classifier was used for all face recognition experiments. On each benchmark, we reduced the face images from 112×92 to 28×23 for efficiency; this had little effect on the accuracy of recognition.

5.1. *Cambridge ORL database*

We used the ORL face database composed of 400 images: 40 persons, with 10 images of each person. The images were taken at different times, lighting and facial expressions. The faces are in an upright position in frontal view, with a slight left-right rotation. Figure 1 shows some samples images from the database.

Each set of 10 images for a person was randomly partitioned into a training set of five images and a test set of the other five images. The training set was then used to train the PCA, FNMF, PNMF, LNMF and NMF algorithms, and the test set was used to evaluate face recognition. Both methods used the same training and test data.

5.1.1. *Learning basis components*

We used NMF, LNMF and FNMF to learn the basis images of the training set from the ORL database by the update rules described by Eqs. (8)–(10).

Figures 2–5 show, respectively, bases of NMF, LNMF, FNMF and PNMF, which were all learned from the training faces. For comparison of different update rules,

Fig. 1. ORL face samples.

Fig. 2. 25 and 36 basis images of NMF.

the figures show the results of computing both 25 and 36 basis images using the four update rules. The images show that the bases trained by NMF were additive, but not spatially localized, for representation of faces. At the same time, the bases trained from FNMF, PNMF and LNMF are all additive and spatially localized for representing faces.

5.1.2. *Face recognition on the ORL database*

In this experiment, FNMF, PNMF, LNMF, NMF and PCA were compared for face recognition on the ORL database. Figure 6 shows the recognition results for the five methods. The horizontal axis represents the square root of the number of bases.

This experiment shows that FNMF and PNMF give higher recognition rates than NMF, LNMF and PCA on the ORL database. The recognition rate of NMF was the lowest.

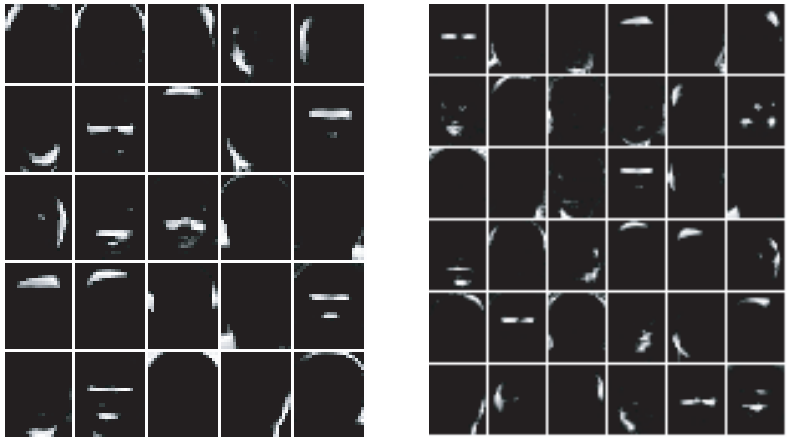


Fig. 3. 25 and 36 basis images of LNMF.

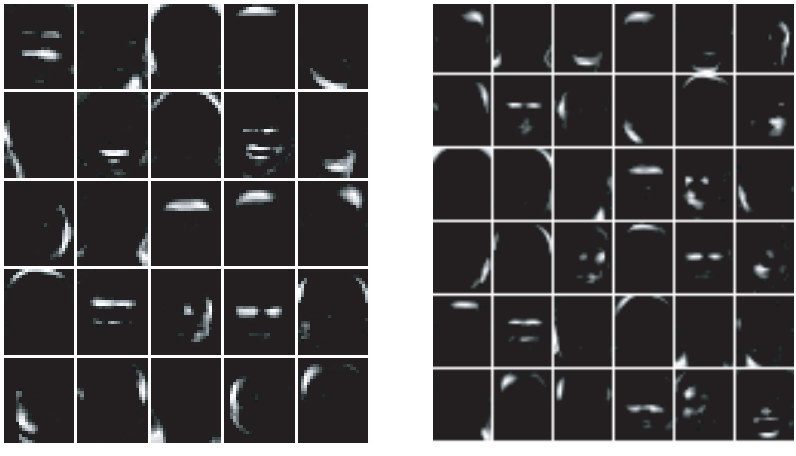


Fig. 4. 25 and 36 basis images of FNMF.

5.1.3. *Occluded face recognition on the ORL database*

In many applications, localized features[13] offer advantages in object recognition, including stability to local deformations, lighting variations, and partial occlusion. NMF does not perform very well in recognizing occluded faces.[5,10] In this experiment, FNMF, PNMF, LNMF and PCA were compared for occluded face recognition on the ORL database. Figure 7 shows the occluded face samples.

Figure 8 shows the recognition results for the four methods. The horizontal axis represents the square root of the number of bases. The subfigures (from left to right, from top to down) respectively show the recognition results of the four algorithms

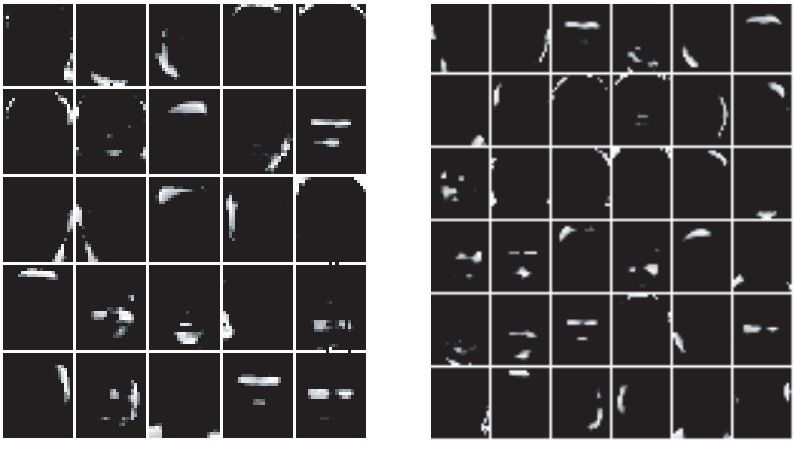


Fig. 5. 25 and 36 basis images of PNMF.

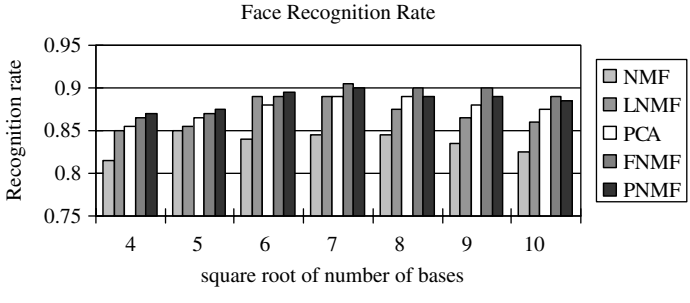


Fig. 6. Face recognition on the ORL database.

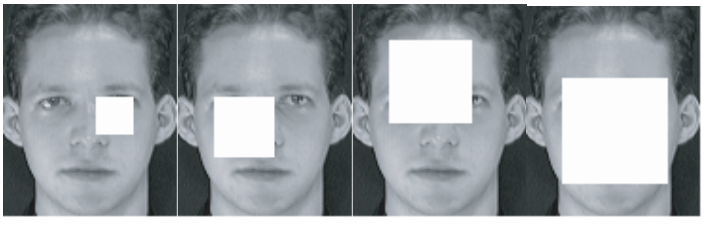


Fig. 7. Occluded face samples, occluding patch sizes of (from left to right) 20×20, 32×32, 44×44, 56×56 pixels.

versus the size of $S \times S$ occluding patch for $S \in \{20, 32, 44, 56\}$. From Fig. 8, we can see that although PCA gets good results when the number of bases is low or when the occluding patches is small, FNMF and PNMF performs better with the increase of the number of bases and the size of the occluding patch.

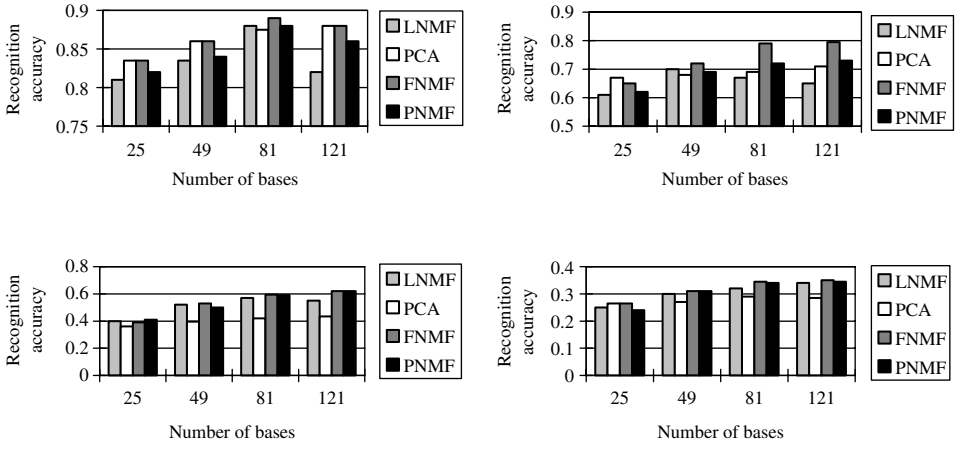


Fig. 8. Recognition accuracies versus the number $(25, 49, 81, 121)$ of basis components, with different occluding patch sizes of (from left to right, then top to down) 20×20, 32×32, 44×44, 56×56 pixels on ORL database.

5.1.4. *Face sample data projection energy on the new bases*

For the algorithm to have the statistical perspective of PCA, we are looking for a unit vector $\mathbf{w}$ which maximizes $\sum_i (W^T v_i)^2$. In other words, the projected points onto the axis represented by the vector $\mathbf{w}$ are as spread as possible (in a least squares sense). The following experiment shows the projection energy of face sample data on the bases of PNMF, LNMF and NMF. We use $\sqrt{\sum_i (W^T v_i)^2}$ to stand for projection energy. In the updating rules of PNMF, LNMF and NMF, we fix $\sum_i w_{ij} = 1$.

The abscissa is the iteration number of the three updating rules, from 1 to 3,000. The vertical axis is $\sqrt{\sum_i (W^T v_i)^2}$. Figure 9 shows that we can get larger projection energy when using PNMF than LNMF and NMF.

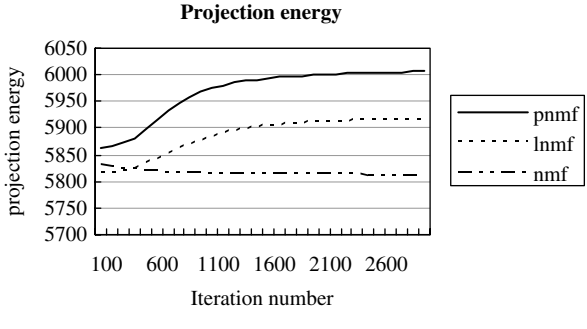


Fig. 9. Data projection energy on PNMF, LNMF and NMF.

5.1.5. *Get orthogonal bases*

Different bases should be as orthogonal as possible, so as to minimize redundancy between different bases. At the same time, PCA enforces the orthogonal constraint on the bases; therefore, we hope to get bases that are more orthogonal.

The following experiment compares the extent of orthogonality between PNMF and LNMF. Given that $W^T W = U$, $\frac{U_{ii}}{\sum_j U_{ji}}$ will reflect the extent of orthogonality of the bases.

When respectively using PNMF and LNMF to get sixteen bases (the number of basis components is 16) and fixing the iteration number of both updating rules 3,000, we get the results shown in Fig. 10. The abscissa of the figure is the sixteen bases. The vertical axis is $\frac{U_{ii}}{\sum_j U_{ji}}$. We can see that the sixteen bases of PNMF are more orthogonal than LNMF.

5.2. *FERET subset*

There are 70 persons in the FERET[14] subset. Each person has six different frontal-view images. There are three different illuminations and two different facial expressions for each illumination. Figure 11 shows some samples.

Each set of six images for a person was randomly partitioned into a training set of three images and a test set of the other three images. The training set was then used to train PCA, FNMF, PNMF, LNMF and NMF, and the test set was used to evaluate face recognition. Both methods used the same training and test data.

Figure 12 compares the results of PCA, NMF, LNMF, FNMF and PNMF on the FERET database. The horizontal axis represents the square root of the number of bases.

The experiment shows that FNMF and PNMF performed better than LNMF, PCA and NMF for this data set.

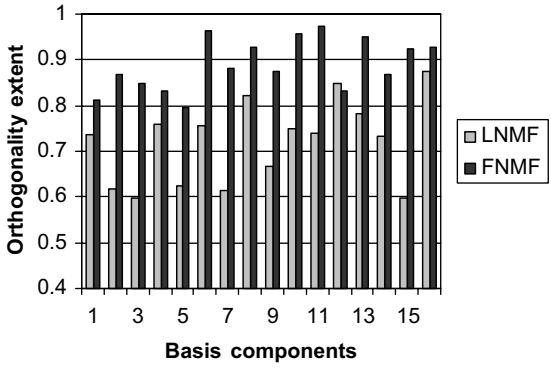


Fig. 10. The extent of orthogonality of bases.

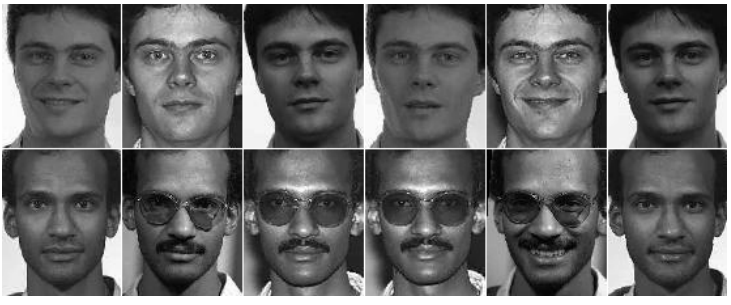


Fig. 11. FERET dataset face samples.

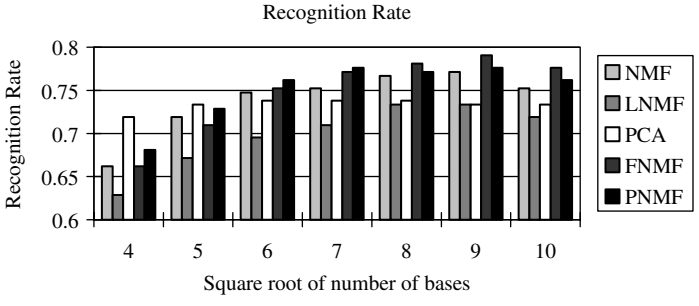


Fig. 12. Face recognition on FERET database.

5.3. *Discussion*

From the experimental results, we find that the proposed FNMF and PNMF techniques can obtain additive and spatially localized bases from a training set and achieve a higher recognition rate than LNMF and NMF. We also found that FNMF, PNMF and LNMF perform well on the ORL database, which has little illumination variation, but LNMF is not as good as FNMF, PNMF and NMF on the FERET dataset.

Why does LNMF perform better on ORL database but worse on FERET dataset than NMF? We have the following explanations. The FERET dataset we used are frontal view faces with less rotation and variance of expressions than ORL. The main differences among the faces in FERET are glasses or no glasses, long and short hair, and illumination. Though NMF and LNMF are both parts based methods, NMF's parts are global. By themselves, the global features are more informative than local features when the face does not change significantly. In that case, NMF can perform better than LNMF. However, when there is occlusion, change of facial expression, harsher illumination change, etc. this performance edge goes away. FNMF and PNMF can give good results on FERET because they well represent the local discriminant characteristics.

Other constraints may be imposed on NMF, and in future work we will explore promising ones. For example, constraints that are part of ICA computation may be implemented to form ICA-NMF. We will experiment with this and other modifications to the basic NMF approach.

6. Conclusion

In this paper, we presented two new constrained non-negative matrix factorization algorithms, called Fisher non-negative matrix factorization (FNMF) and PCA non-negative matrix factorization (PNMF), for face recognition and learning part-based subspace representations. The main idea of FNMF and PNMF is to start with NMF as a framework and then add useful constraints to the matrix factorization: the Fisher constraint and characteristics of PCA. We showed that using FNMF and PNMF results in intuitive basis images, and performs better than LNMF and NMF on face recognition on the ORL database and FERET dataset.

Appendix

Convergence Proof of FNMF

Our update rules are based on a technique which minimizes an objective function $L(X)$ by using an auxiliary function. $G(X, X')$ is defines as an auxiliary function for $L(X)$ if $G(X, X') \geq L(X)$ and $G(X, X) = L(X)$ are satisfied. If G is an auxiliary function, then $L(X)$ is nonincreasing when X is updated by

$$X^{(t+1)} = \arg\min_{x} G(X, X^{(t)}) \tag{A.1}$$

because

$$L(X^{(t+1)}) \leq G(X^{(t+1)}, X^{(t)}) \leq G(X^{(t)}, X^{(t)}) = L(X^{(t)}).$$

H is updated by minimizing $L(H) = D(V \| WH)$ with W fixed.

We construct an auxiliary function for $L(H)$ as

$$G(H, H') = \sum_{i,j} v_{ij} \log v_{ij} - \sum_{i,j,k} v_{ij} \frac{w_{ik} h'_{kj}}{\sum_k w_{ik} h'_{kj}} \left(\log(w_{ik} h_{kj}) - \log \frac{w_{ik} h'_{kj}}{\sum_k w_{ik} h'_{kj}} \right) + \sum_{i,j} y_{ij} - \sum_{i,j} v_{ij} + \alpha S_W - \alpha S_B. \tag{A.2}$$

$G(H, H') = L(H)$ is easily verified, so we will just prove $G(H, H') \geq L(H)$ as follows. Because $\log(\sum_k W_{ik} h_{kj})$ is a convex function, the following holds for all i, j and $\sum_k \sigma_{ijk} = 1$:

$$-\log \left(\sum_k w_{ik} h_{kj} \right) \leq - \sum_k \sigma_{ijk} \log \frac{w_{ik} h_{kj}}{\sigma_{ijk}}$$

where $\sigma_{ijk} = \frac{w_{ik}h'_{kj}}{\sum_k w_{ik}h'_{kj}}$. So

$$-\log\left(\sum_k w_{ik}h_{kj}\right) \leq -\sum_k \frac{w_{ik}h'_{kj}}{\sum_k w_{ik}h'_{kj}}\left(\log w_{ik}h_{kj} - \log\frac{w_{ik}h'_{kj}}{\sum_k w_{ik}h'_{kj}}\right)$$

which is $G(H, H') \geq L(H)$.
In order to minimize $L(H)$ w.r.t. H, we can update H using

$$H^{(t+1)} = \arg\min_H G(H, H^{(t)}).$$

H can be found by letting $\frac{\partial G(H,H')}{\partial h_{kl}} = 0$ for all k, l, since

$$\begin{aligned}\frac{\partial G(H,H')}{\partial h_{kl}} = &\sum_i v_{il}\frac{w_{ik}h'_{kl}}{\sum_k w_{ik}h'_{kl}}\frac{1}{h_{kl}} + \sum_i w_{ik} + \alpha\frac{2}{n_i C}(h_{kl} - u_{ki}) \\ &- \alpha\frac{4}{n_i C(C-1)}\sum_j\left(u_{kj} - \left(\frac{h_{kl}}{n_i} + u_{ki} - \frac{h'_{kl}}{n_i}\right)\right). \qquad \text{(A.3)}\end{aligned}$$

In fact, α is a constant, so we can try some values for α to get a relatively optimal one; here we define $\alpha = 1$. n_i corresponds to the number of face vectors in the class to which h_{kl} belongs, and C is the number of face classes. We find that

$$h_{kl} = \frac{-b + \sqrt{b^2 + 4\left(\sum_i v_{il}\frac{w_{ik}h'_{kl}}{\sum_k w_{ik}h'_{kl}}\right)\left(\frac{2}{n_i C} - \frac{4}{n_i^2(C-1)}\right)}}{2\left(\frac{2}{n_i C} - \frac{4}{n_i^2(C-1)}\right)} \qquad \text{(A.4)}$$

where

$$b = \frac{4}{n_i C(C-1)}\sum_j\left(u_{kj} - \left(u_{ki} - \frac{h'_{kl}}{n_i}\right)\right) - \frac{2}{n_i C}u_{ki} + 1.$$

$\left(\frac{2}{n_i C} - \frac{4}{n_i^2(C-1)}\right)$ is just a positive constant when $n_i \geq 2$ (we can easily ensure $n_i \geq 2$) and has little effect on the update rules, so we can replace (4) by (8).

Just like updating H, we can update W by minimizing $L(W) = D(V\|WH)$ with H fixed. The auxiliary function for $L(W)$ is

$$\begin{aligned}G(W, W') = &\sum_{i,j} v_{ij}\log v_{ij} - \sum_{i,j,k} v_{ij}\frac{w'_{ik}h_{kj}}{\sum_k w'_{ik}h_{kj}}\left(\log(w_{ik}h_{kj}) - \log\frac{w'_{ik}h_{kj}}{\sum_k w'_{ik}h_{kj}}\right) \\ &+ \sum_{i,j} y_{ij} - \sum_{i,j} v_{ij} + \alpha S_W - \alpha S_B. \qquad \text{(A.5)}\end{aligned}$$

We can prove $G(W, W) = L(W)$ and $G(W, W') \geq L(W)$ in the same way as proving $G(H, H') = L(H)$ and $G(H, H') \geq L(H)$. By letting $\frac{\partial G(W,W')}{\partial w_{kl}} = 0$, we find

$$w_{kl} = \frac{w'_{kl}\sum_j v_{kj}\frac{h_{lj}}{\sum_k w'_{kl}h_{lj}}}{\sum_j h_{lj}}.$$

According to the above analysis, we conclude that the three step update rules lead to a sequence of nonincreasing values of $D(V\|WH)$, and hence a local minimum.

References

1. K. Back, B. A. Draper, J. R. Beveridge and K. She, PCA versus ICA: a comparison on the FERET data set, *Int. Conf. Computer Vision, Pattern Recognition and Image Processing* (2002), pp. 824–827.
2. P. N. Belhumeur, J. P. Hespanha and D. Kriegman, Eigenfaces versus Fisherfaces: recognition using class specific linear projection, *IEEE Trans. PAMI* **19**(7) (1997) 711–720.
3. A. J. Bell and T. J. Sejnowski, The "independent components" of natural scenes are edge filters, *Vis. Res.* **37** (1997) 3327–3338.
4. P. Common, Independent component analysis — A new concept? *Sign. Process.* **36** (1994) 287–314.
5. T. Feng, S. Z. Li, H. Shum and H. J. Zhang, Local non-negative matrix factorization as a visual representation, in *Proc. 2nd Int. Conf. Development and Learning.* Washington DC (June, 2002), pp. 178–183.
6. I. T. Jolliffe, *Principal Component Analysis* (Springer-Verlag, New York, 1986).
7. D. D. Lee and H. S. Seung, Unsupervised learning by convex and conic coding, *Adv. Neural Inform. Proc. Syst.* **9** (1997) 515–521.
8. D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* **401** (1999) 788–791.
9. D. D. Lee and H. S. Seung, Algorithms for non-negative matrix factorization, in *Proc. Neural Information Processing Systems* (2000), pp. 556–562.
10. S. Z. Li, X. W. Hou and H. J. Zhang, Learning spatially localized, part-based representation, *IEEE Conf. Computer Vision and Pattern Recognition* (2001), pp. I-207–I-212.
11. N. K. Logothetis and D. L. Sheinberg, Visual object recognition, *Ann. Rev. Neurosci.* **19** (1996) 577–621.
12. S. E. Palmer, Hierachical structure in perceptual representation, *Cogn. Psychol.* **9** (1997) 441–474.
13. P. Penev and J. Atick, Local feature analysis: a general statistical theory for object representation, *Neural Syst.* **7**(3) (1996) 477–500.
14. P. J. Phillips, H. Moon, S. Rizvi and P. Rauss, The FERET evaluation methodology for face recognition algorithms, *IEEE Trans. PAMI* **22**(10) (2000) 1090–1104.
15. M. Turk and A. Pentland, Eigenfaces for recognition, *J. Cogn. Neurosci.* **3** (1991) 71–86.
16. E. Wachsmuth, M. W. Oram and D. I. Perrett, Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque, *Cereb. Cortex* **4** (1994) 509–522.
17. M.-H. Yang, Kernel eigenfaces versus Kernel fisherfaces: face recognition using kernel methods, *IEEE Conf. Automatic Face and Gesture Recognition* (2002), pp. 215–220.

Yuan Wang received a B.E. from Harbin Institute of Technology in 2001, an M.E. from Beijing Institute of Technology in 2004. He is pursuing his Ph.D. in the Department of Electrical Computer Engineering at University of Arizona. He has worked as an intern in Microsoft Research Asia, and is currently a research assistant in Optical Science Center at the University of Arizona.

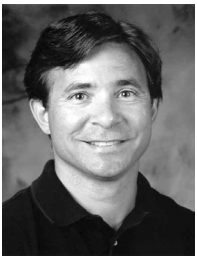
His research interests include eye tracking, face recognition, face tracking, handwriting recognition, human–computer interaction, etc.

Yunde Jia is a professor in the Department of Computer Science and engineering, currently a vice dean of the School of Information Science and Technology at Beijing Institute of Technology. He earned a bachelor's, master's and Ph.D. degrees in mechatronics from Beijing Institute of Technology in 1983, 1986 and 2000, respectively. He found the Computer Vision Lab in 1997 at the university just after working as a visiting researcher in Carnegie Mellon University.

His research interests include computer vision, media computing and intelligent human–computer interaction.

Changbo Hu received a B.E. from Zhejiang University in 1992, an M.E. from Anhui University in 1995, and a Ph.D. from National Lab of Pattern Recognition at the Institute of Automation in 2001. He has worked as an intern in Microsoft Research Asia, a postdoc in Computer Science Department at the University of California, Santa Barbara, and is currently a project scientist in Robotics Institute, Carnegie Mellon University.

His research interests include face tracking, face recognition, facial expression analysis, human–computer interaction, etc.



Matthew Turk received a B.S. from Virginia Tech in 1982, an M.S. from Carnegie Mellon University in 1984, and a Ph.D. from the Massachusetts Institute of Technology in 1991. He has worked for Martin Marietta Denver Aerospace, LIFIA/IMAG (Grenoble, France), Teleos Research, and Microsoft Research, and is currently an associate professor in Computer Science and Media Arts and Technology at the University of California, Santa Barbara. He co-directs the UCSB Four Eyes Lab (ilab.cs.ucsb.edu) which focuses on research in imaging, interaction, and innovative interfaces.