

LDA-based Non-negative Matrix Factorization for Supervised Face Recognition

Yun Xue^a, Chong Sze Tong^b, Jing Yun Yuan^c

^a School of Physics and Telecommunication Engineering, South China Normal University,
Guangzhou Guangdong 510631, China
Email: xueyun@scnu.edu.cn

^b Department of Mathematics, Hong Kong Baptist University, Hong Kong, China
Email: cstong@math.hkbu.edu.hk

^c Departamento de Matematica - UFPR, Centro Politecnico, CP:19.081, 81531-990, Curitiba, PR, Brazil
Email: jin@mat.ufpr.br

Abstract—In PCA based face recognition, the basis images may contain negative pixels and thus do not facilitate physical interpretation. Recently, the technique of non-negative matrix Factorization (NMF) has been applied to face recognition: the non-negativity constraint of NMF yields a localized parts-based representation which achieves a recognition rate that is on par with the eigenface approach.

In this paper, we propose a new variation of the NMF algorithm that incorporates training information in a supervised learning setting. We integrate an additional term based on Fisher's Linear Discriminant Analysis into the NMF algorithm and prove that our new update rule can maintain the non-negativity constraint under a mild condition and hence preserve the intuitive meaning for the base vectors and weight vectors while facilitating the supervised learning of within-class and between-class information.

We tested our new algorithm on the well-known ORL database, CMU PIE database and FERET database, and the results from experiments are very encouraging compared with traditional techniques including the original NMF, the Eigenface method, the sequential NMF+LDA method and the Fisherface method.

Index Terms—nonnegative matrix factorization, principal component analysis, fisher linear discriminant analysis, eigenface, fisherface

I. INTRODUCTION

As an important problems of image processing, face recognition has received significant attention. It is a typical pattern recognition problem which can be applied to promote the resolution of many other classification problems, and also has lots of potential applications.

Generally, automatic face recognition contains two approaches, namely, constituent-based and face-based methods [1], [2]. The first approach performs recognition based on the spatial relationship between different facial features such as eyes, mouth and nose [3], [4], therefore its performance is highly sensitive to the accuracy of facial feature detection algorithms. However, since extracting facial components accurately is difficult, a small error at this stage may cause a large classification error.

In contrast, the face-based approach defines the face as a whole [5], [6], i.e. the corresponding image is

considered as a two-dimensional pattern and classified based on its underlying statistics.

As an effective face-based approach, Principal Component Analysis (PCA) [5] uses the orthogonal basis images which have a statistical interpretation as the directions of largest variance. Based on this idea, Turk and Pentland [6] built a face recognition system in which the significant features (eigenvectors associated with large eigenvalues) are called eigenfaces. This approach represents a face image using a weighted sum of eigenfaces and the classification is performed by comparing the weight vectors of the test images with those of the reference face images.

However, the traditional Eigenface-based methods suffer from some limitations. Firstly, it is well known that PCA gives a very good representation of the images. Given two images of the same person, the similarity measured under PCA representation is very high. However, for the two images of different persons, the similarity measured is still high. This suggests this method may not enjoy a very high discriminatory ability.

The second problem is that many of the basis images do not have an obvious visual interpretation. This is because PCA allows the coefficients of base vectors and weight vectors to be of arbitrary sign. Since the basis images are used in linear combinations that generally involve complex cancellations between positive and negative numbers, many of them do not have intuitive meaning.

Finally, the PCA approach is based on extracting global face features, so the case of occlusions is difficult to handle.

Recently, a new technique to obtain a linear representation of data has been proposed. This new method, called non-negative matrix factorization (NMF), was first used in the work of Lee and Seung [7] to find parts of objects. NMF differs from other methods by the usage of non-negativity constraints.

Recently NMF has been widely applied in many fields, and it has been shown to outperform PCA in image recognition for certain face databases [8]–[10]. NMF factorizes the image database into two matrix factors, whose entries are all non-negative, and produces a parts-

based representation of images because it allows only additive, not subtractive, combinations of basis images. Therefore the resultant NMF bases are localized features which correspond to the intuitive idea of combining parts to form a whole. In real face databases, the images generally contain natural occlusions such as sunglasses and scarfs. If some such images are contained in the training set, then the basis images would be significantly affected in global methods such as PCA while the local basis images of NMF remain relatively stable and thus obtain better recognition performance (see eg [8]).

However, since the traditional NMF method is an unsupervised learning technique, it's difficult to take advantage of the discrimination information in the training set to boost the classification capability. In this paper we introduce an LDA-based Non-negative Matrix Factorization algorithm which is a new variation to NMF. To take advantage of more information in the training images, we add the Fisher Linear Discriminant into the objective function in NMF algorithm, which will lead to more discriminatory base vectors and weight vectors. Since this algorithm encodes discrimination information for face recognition, it can improve the result for classification.

As our new approach is based on a subspace definition, we have used the Principal Component Analysis (PCA), the original NMF algorithm and the Fisherface method for direct comparison. Moreover, we also use NMF as a method of reducing dimensions and then subject the projected vectors to LDA to extract the feature vectors for face recognition. Since this sequential NMF+LDA process would preserve the unsupervised nature of NMF as initially formulated by Lee and Seung, a direct comparison between this procedure and our algorithm is important. After implementing all these methods on the face image databases, the results from experiments support the conclusion that our new algorithm can achieve a better performance in face recognition.

This paper is organized as follows. Section 2 reviews the background of PCA, NMF and Fisherface. The details of our LDA-based Non-negative Matrix Factorization algorithm are described in section 3. All the testing databases used in this paper are described in section 4. Results from experiments on a face recognition system based on the proposed method are discussed in section 5. Conclusions & future work are presented in section 6.

An earlier version of this paper was presented at the 18th International Conference on Pattern Recognition, 2006 [11]. In this paper, a detailed analysis and proof about our model is provided, and we use more relevant algorithms for the performance comparison based on some well known face databases.

II. REVIEW OF PCA, NMF AND FISHERFACE FOR FACE RECOGNITION

This section provides the background theory of PCA, NMF, and Fisherface for face recognition.

A. Principal component analysis

PCA is originally used to find a low dimensional representation of data [12]. Some details are described as follows.

Let $X = \{X_n \in R^d | n = 1, \dots, N\}$ be a set of vectors. In image processing, they are formed by row or column concatenation of the original image matrix, with d being the product of the width and the height of an image. Let

$$E[X] = \frac{1}{N} \sum_{n=1}^N X_n$$

be the mean vector in the image set. After subtracting it from each vector of X , we get

$$\hat{X} = \{\hat{X}_n, n = 1, \dots, N\} \text{ with } \hat{X}_n = X_n - E[X]$$

Then we define the auto-covariance matrix M for X as

$$M = cov(X) = E(\hat{X} \otimes \hat{X})$$

where $M \in R^{d \times d}$, with elements

$$M(i, j) = \frac{1}{N-1} \sum_{n=1}^N (\hat{X}_n(i) \hat{X}_n(j)), 1 \leq i, j \leq d$$

From matrix theory, it's well known that the auto-covariance matrix for the eigenvectors is diagonal, it follows that the coordinates of the vectors in X with respect to the eigenvectors are un-correlated random variables. Then the PCA of a vector y can be calculated by projecting it onto the subspace which is spanned by d' eigenvectors corresponding to the top d' eigenvalues of the autocorrelation matrix M sorted in descending order. Furthermore, Eigenfaces are the eigenvectors associated with the largest eigenvalues from the PCA method. After representing a face image using a weighted sum of eigenfaces, face recognition is performed by comparing the corresponding weight vectors between the test image and reference faces.

B. NMF method

Non-negative Matrix Factorization (NMF) [7] is a method to obtain a representation of data under non-negativity constraints. These constraints produce a parts-based representation because they allow only additive, not subtractive, combinations, so the corresponding basis images can be understood as localized features that correspond better with intuitive notions of the parts of face images.

If an initial image database is represented as a $n \times m$ matrix V , where each column is a non-negative vector corresponding to a face image, we can find two new non-negative matrices (W and H) to approximate the original data matrix

$$V_{ij} \simeq (WH)_{ij} = \sum_{a=1}^r W_{ia} H_{aj}, W \in R^{n \times r}, H \in R^{r \times m} \quad (1)$$

Each column of matrix W represents a basis vector while each column of H represents the weights used to approximate the corresponding column in V using the bases from W .

For the NMF method, in contrast to PCA, no subtractions can occur, so the non-negativity constraints are compatible with the intuitive idea of combining parts to form a whole face, which is how NMF learns a parts-based representation.

The dependencies between image pixels and the encoding variables are depicted in Fig.1. The top nodes represent an encoding h_1, \dots, h_r (column of H), and the bottom nodes represent an image v_1, \dots, v_n (column of V). The nonnegative value W_{ia} characterizes the extent of influence that the a 'th encoding variable h_a has on the i 'th image pixel v_i . Because of the non-negativity of W_{ia} , the image pixels in the same part of face image will be coactivated when the part is present, and NMF learns by adapting W_{ia} to generate the optimal approximation.

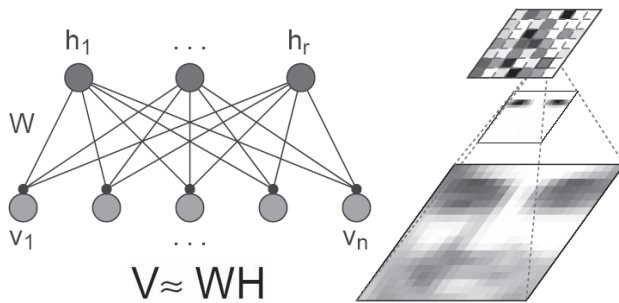


Figure 1. Probabilistic hidden variables model underlying non-negative matrix factorization [7].

The update rule for NMF is derived as below:

First define an objective function to measure the similarity between V and WH :

$$F = \sum_{i=1}^n \sum_{j=1}^m [V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}] \quad (2)$$

Then an iterative algorithm to reach a local maximum of this objective function is given [7]:

$$\widehat{W}_{ia}^{t+1} = W_{ia}^t \sum_j \frac{V_{ij}}{(W^t H^t)_{ij}} H_{aj}^t \quad (3)$$

$$W_{ia}^{t+1} = \frac{\widehat{W}_{ia}^{t+1}}{\sum_j W_{ja}^{t+1}} \quad (4)$$

$$H_{aj}^{t+1} = H_{aj}^t \sum_i W_{ia}^{t+1} \frac{V_{ij}}{(W^{t+1} H^t)_{ij}} \quad (5)$$

The convergence of the process is proved in [13].

The flow chart of the NMF algorithm is as below.

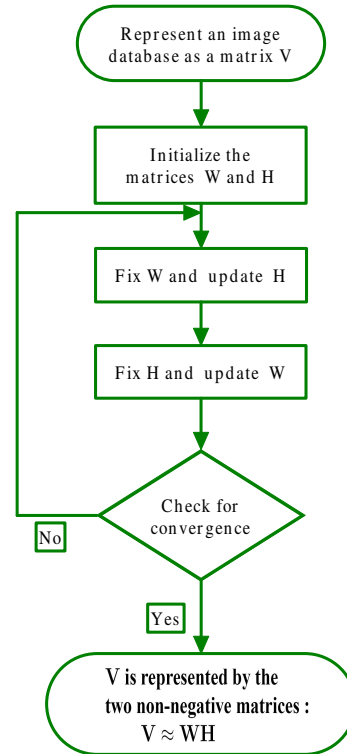


Figure 2. Flowchart of the NMF algorithm (after [14]).

In face recognition, NMF is applied as follows. Firstly, in the feature extraction stage, all the training images form the original data matrix V , then the bases W_1 can be obtained using the above update rules (3)-(5). Next, let $W^+ = (W_1^T W_1)^{-1} W_1^T$, then each training face image V_t is projected onto the linear space as a feature vector $H_t = W^+ V_t$ which is then used as a prototype feature point. A probe image V_p to be classified is represented as $H_p = W^+ V_p$. Finally, we classify the probe images using the nearest neighbour classification scheme. At this stage, some suitable distance between the weight vectors of the probe image and training image, $\text{dist}(H_p, H_t)$, is calculated, then the probe image is classified to the class with the minimum distance.

C. Fisherface and sequential NMF+LDA

As a classical technique in multivariate statistics, LDA has also been successfully applied in pattern recognition [15]. It utilizes class-specific information within the training images and finds a feature space in which the ratio of the between-class scatter and the within-class scatter is maximized.

However, for the face recognition problem, the within-class scatter matrix is often singular. In order to overcome this difficulty, Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman [16] proposed the Fisherface method, which combines PCA with the standard LDA method to reduce the dimensionality of feature vectors. Nowadays, it is one of the most popular feature extraction and dimension reduction techniques used in the face-based approach.

In the Fisherface procedure, none of the class information is incorporated in the PCA stage, while all the PCA-based feature vectors are subsequently used to extract the LDA-based feature vectors so that the classification is more robust. Similarly, we could also first use NMF without any class information to reduce the dimensionality of feature vectors, and then use the LDA method to take advantage of the class information in the same way. Since this sequential NMF+LDA process is a simple procedure to preserve the nature of NMF as initially formulated by Lee and Seung [7], we will use it and Fisherface as benchmarks for comparison with our LDA-based NMF method, which also exploits the same class information within the training images.

III. LDA-BASED NMF

A. Our model

In the original NMF model and Eigenface model, the training face images are used collectively without reference to the class membership of the training faces. To improve the performance for face recognition, we propose to add the Fisher discriminant to the original NMF algorithm formulation. Because the columns of the weight matrix H have a one-to-one correspondence with the columns of the image matrix V , we naturally hope to maximize the between-class scatter and simultaneously minimize the within-class scatter of H . Based on this idea, we define the new objective function:

$$F_1 = \sum_{i=1}^n \sum_{j=1}^m [V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}] + \alpha S_W - \alpha S_B, \quad (6)$$

where $\alpha > 0$ is a regularization parameter, S_W is the within-class scatter of the weight matrix H , and S_B is the between-class scatter of H .

Let M_{tr} denote the number of vectors in each class and C denote the number of classes. We define S_W and S_B as follows:

$$S_B = \frac{1}{C(C-1)} \sum_{i=1}^C \sum_{j=1}^C (\mu_i - \mu_j)^T (\mu_i - \mu_j) \quad (7)$$

$$S_W = \frac{1}{CM_{tr}} \sum_{i=1}^C \sum_{k \in C_i} (H(:,k) - \mu_i)^T (H(:,k) - \mu_i) \quad (8)$$

Here C_i means the i 'th person's images, $\mu_i = \frac{1}{M_{tr}} \sum_{k \in C_i} H(:,k)$ denotes the mean vector of the i 'th class in H .

In the next section, we shall derive the update rules for our model. Note that a similar idea has been discussed and the relevant model developed in [17]. But their final update rule is entirely different from ours. In [18], the authors' idea is also similar to our work, however, our model uses just one Lagrangian multiplier α in the new objective function which is much more convenient for parameter selection in practical experiments. Moreover,

we used a different definition for S_W and S_B based on unbiased estimation, and the final update rules are also different from theirs. Furthermore, we shall also prove that using our update rules, the non-negativity constraints for all the coefficients in W and H can be satisfied automatically under a mild condition, and this issue of maintaining non-negativity was not addressed in [18].

B. Modified update rules

To derive our modified update rules, we will use a technique which minimizes an objective function by using an auxiliary function similar to that used in [13].

Definition 1: $G(H, H')$ is an auxiliary function for the function $F(H)$ if

$$G(H, H') \geq F(H), G(H, H) = F(H)$$

Then we can use the following lemma:

Lemma 1: If G is an auxiliary function, then F is nonincreasing by using this update rule:

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

where H^t and H^{t+1} denote the obtained weight matrices after t iterations and $t+1$ iterations respectively.

Proof 1: Since $F(H^{t+1}) \leq G(H^{t+1}, H^t)$, $G(H^{t+1}, H^t) \leq G(H^t, H^t)$ and $G(H^t, H^t) = F(H^t)$ we can conclude $F(H^{t+1}) \leq F(H^t)$.

Using this lemma, we can obtain the corresponding update rule for the objective function F_1 . Noting that the new terms in function F_1 are just related to the weight matrix H , we can directly adopt the update rule for W in the original NMF algorithm, then just deduce the iterative algorithm for H^{t+1} when fixing W and given H^t .

First, we can design the following auxiliary function for F_1 :

Lemma 2: Define

$$G_1(H, H^t) = \sum_{i,j} (V_{ij} \log V_{ij} - V_{ij}) + \sum_{i,j} (WH)_{ij} - \sum_{i,j,k} V_{ij} \frac{W_{ik} H_{kj}^t}{\sum_n W_{in} H_{nj}^t} (\log(W_{ik} H_{kj})) - \log \frac{W_{ik} H_{kj}^t}{\sum_n W_{in} H_{nj}^t} + \alpha S_W - \alpha S_B$$

It's an auxiliary function for $F_1(H)$.

Proof 2: Proof $G_1(H, H^t) = F_1(H)$ is straightforward to verify. To show that $G_1(H, H^t) \geq F_1(H)$, we use the convexity of the log function to get the following inequality:

$$-\log \left(\sum_k W_{ik} H_{kj} \right) \leq - \sum_k \alpha_{ijk} \log \frac{W_{ik} H_{kj}}{\alpha_{ijk}},$$

if $\alpha_{ijk} \geq 0$ and $\sum_k \alpha_{ijk} = 1$

Setting $\alpha_{ijk} = \frac{W_{ik} H_{kj}^t}{\sum_k W_{ik} H_{kj}^t}$, based on the above inequality we can conclude

$$-\log(WH)_{ij} \leq - \sum_k \frac{W_{ik} H_{kj}^t}{\sum_n W_{in} H_{nj}^t} (\log W_{ik} H_{kj})$$

$$-\log \frac{W_{ik}H_{kj}^t}{\sum_n W_{in}H_{nj}^t}), \quad \text{i.e. } G_1(H, H^t) \geq F_1(H)$$

Based on this lemma, we minimize $F_1(H)$ by using the following rule

$$H^{t+1} = \arg \min_H G_1(H, H^t)$$

The minimum of $G_1(H, H^t)$ is obtained by setting $\frac{\partial G_1(H, H^t)}{\partial H_{\beta\gamma}} = 0$ for any β, γ . Let $H(:, \gamma)$ be the weight vector of an image belonging to the i_1^{th} class. Since

$$\begin{aligned} \frac{\partial G_1(H, H^t)}{\partial H_{\beta\gamma}} &= - \sum_i V_{i\gamma} \frac{W_{i\beta}H_{\beta\gamma}^t}{\sum_n W_{in}H_{n\gamma}^t} \frac{1}{H_{\beta\gamma}} + \sum_i W_{i\beta} \\ &+ \frac{2\alpha H_{\beta\gamma}}{CM_{tr}} - \frac{2\alpha \sum_{\gamma_1 \in C_{i_1}} H_{\beta\gamma_1}}{CM_{tr}^2} \\ &- \frac{4\alpha}{M_{tr}C(C-1)} \sum_{j \neq i_1} (\mu_{i_1}(\beta) - \mu_j(\beta)), \end{aligned}$$

we obtain a quadratic equation for $H_{\beta\gamma}$:

$$aH_{\beta\gamma}^2 + bH_{\beta\gamma} + d = 0$$

where

$$\begin{aligned} b &= 1 - \frac{6\alpha}{CM_{tr}^2} \sum_{n \in C_{i_1}, n \neq \gamma} H_{\beta n}^t + \frac{4\alpha}{M_{tr}C(C-1)} \sum_{j \neq i_1} \mu_j(\beta) \\ a &= \frac{2\alpha M_{tr} - 6\alpha}{CM_{tr}^2}, \quad d = -H_{\beta\gamma}^t \sum_i V_{i\gamma} \frac{W_{i\beta}^{t+1}}{(W^{t+1}H^t)_{i\gamma}} \end{aligned}$$

Then we derive the formula for the bigger of the two real roots, namely, $H_{\beta\gamma}^{t+1}$:

$$H_{\beta\gamma}^{t+1} = \frac{-b + \sqrt{b^2 - 4ad}}{2a} \quad (9)$$

Using the derivation in [13], we can obtain the update rules for W :

$$\widehat{W}_{\lambda\beta}^{t+1} = W_{\lambda\beta}^t \sum_{\gamma} \frac{V_{\lambda\gamma}}{(W^t H^t)_{\lambda\gamma}} H_{\beta\gamma}^t \quad (10)$$

$$W_{\lambda\beta}^{t+1} = \frac{\widehat{W}_{\lambda\beta}^{t+1}}{\sum_{\gamma} \widehat{W}_{\gamma\beta}^{t+1}} \quad (11)$$

Comparing these final update rules with the original NMF rules (i.e. equations (5) and (9)), we see that the main difference between our new algorithm and the traditional NMF method is the new update rule for H which integrates the class information in the weight vectors of the training set.

C. Proof of Non-negativity

From the previous section, the update rules for W is the same as in the conventional NMF and its non-negativity is ensured [13]. However, the incorporation of class information via LDA has yielded very different update rule for the weight matrix H . In this section, we shall establish a sufficient condition to ensure its non-negativity.

We shall proceed by induction and assume that $H^t > 0$. Then we have

$$H^t, W^{t+1} > 0 \Rightarrow d < 0$$

We also note that

$$M_{tr} > 3 \Leftrightarrow a > 0$$

then since $d < 0$, from (9) we have

$$M_{tr} > 3 \Rightarrow a > 0 \Rightarrow b^2 - 4ad > b^2 \Rightarrow H_{\beta\gamma}^{t+1} > 0$$

And since $H^0 > 0$, by induction we have $H^t > 0, \forall t$. Therefore, we have established a sufficient condition for the non-negativity of H , namely $M_{tr} > 3$, i.e. the number of training images must be 4 or more.

Based on the above analysis, we conclude that our update rules will produce a sequence of nonincreasing values of the new objective function, which is at least convergent to a locally optimal solution and satisfies the non-negativity constraints as long as sufficient number of training images are used.

Lastly, we note that although the condition $M_{tr} > 3$ is not strictly necessary for the non-negativity of H^{t+1} (since $H_{\beta\gamma}^{t+1} > 0$ whenever $b < 0$), in practice the sign of b is data-dependent, thus it's advisable to treat the condition as if it is a necessary condition such that the update rules won't break down in any case and maintains non-negativity. Since most practical face databases generally include many images for each person, it's a mild condition in real applications.

IV. TESTING DATABASES USED IN THIS PAPER

A modern face recognition system [19], [20] has to deal with realistic situations and achieve good performance on challenging databases which should include considerable pose variations and real-life illumination changes etc. Therefore, we use the following well known face databases for the performance evaluation of our method.

A. ORL database

The Olivetti database, as the name suggests, originated at the Olivetti Research Laboratory (ORL) in England. In ORL database, there are 40 persons and each person consists of 10 images with different facial expressions and illumination, small scale and small rotation.

B. CMU PIE database

In CMU Pose, Illumination, and Expression (PIE) database, there are 68 persons and each person has 13 pose variations ranged from right profile image to left profile image and 43 different lighting conditions, 21 flashes with ambient light on or off. In our experiments, for each person, we select 56 images including 13 poses with neutral expression and 43 different lighting conditions in frontal view. For all frontal view images, we apply alignment based on two eye center and nose center points but no alignment is applied on the other images with pose.

C. FERET database

The Facial Recognition Technology (FERET) database was sponsored by the Department of Defenses Counterdrug Technology Development Program [21]. We selected 120 persons, 6 frontal-view images for each individual. Face image variations in these 720 images include illumination, facial expression, partial occlusion and aging [21]. All images are aligned by the centers of eyes and mouth.

To reduce the computational complexity, we normalized all the images to the same resolution, 23×28 , by nearest neighbour interpolation. We also normalized the pixel values of each image in the above databases to $[0, 1]$.

V. EXPERIMENTS

In this section, we build a face recognition system to provide a direct comparison of PCA, NMF, Fisherface, sequential NMF+LDA and our LDA-based NMF using images from databases described in Sect.IV, which include complicated variations in illumination, pose and expression.

In all the experiments, we randomly selected M_{tr} images per person from the databases to form a training set and use the remainder as the test set. Then the system adopts the aforementioned 5 approaches and all of them consist of two stages, namely, training and recognition stages. Training stage computes the representational bases for training images and converts them into training image representations, then all the representations are stored into the library. Recognition stage translates the probe image into probe image representation using the representational bases, and then, matches it with those training images stored in the library to identify the face image.

The detailed procedure for PCA, Fisherface and NMF can be found in references [6], [9], [16], so we just mention the two stages for our LDA-based NMF.

A. Training stage

There are 3 major steps in the training stage:

First, we use an $n \times m$ matrix V_1 to represent all the training images in one database.

In the second step, LDA-based NMF algorithm is applied to V_1 and we can find two new matrices (W_1 and H_1) s.t.

$$(V_1)_{ij} \simeq (W_1 H_1)_{ij} = \sum_{a=1}^r (W_1)_{ia} (H_1)_{aj}$$

as described in the section 3 to obtain the base matrix W_1 .

Finally, let $W^+ = (W_1^T W_1)^{-1} W_1^T$, then each training face image V_t is projected onto the linear space as a feature vector $H_t = W^+ V_t$ which is then used as a prototype feature point.

B. Recognition stage

The recognition stage can be divided into 2 steps.

1) *Feature extraction*: Each probe image V_p to be classified is represented as $H_p = W^+ V_p$. Then, we will get the weight matrix H_2 for the probe set.

2) *Nearest neighbor classification*: In this step, some suitable distance between the probe image and training image, $\text{dist}(H_p, H_t)$, is calculated, then the probe image is classified to the class which the closest training image belongs to.

C. Distance measures and parameter selection

The Mahalanobis distance is one of the most widely-used distance measures in pattern recognition tasks [15], [22], so it's applied as a replacement for the Euclidean distance in this classification task.

The definition of Mahalanobis distance is as below:

$$d(X, Y) = \sqrt{(X - Y)' \sigma^{-1} (X - Y)} \quad (12)$$

Where X, Y are feature vectors of length n obtained by different methods, and σ is the auto-covariance matrix for weight vector of training images. In our experiments, we use the full covariance matrix (as opposed to the simpler diagonal covariance matrix) since it always obtains the better result for our face databases.

The regularization parameter α in the new cost function (6) is important as it determines the amount of weight to be assigned to the discriminant factor. According to our experience, the best results are generally obtained when α is about 0.1.

D. Results from experiments

A set of experiments were conducted on the above system, then we evaluated the classification performance of LDA-based NMF algorithm with the Mahalanobis distance metrics and compared it with the result of PCA, NMF, Fisherface and the sequential NMF+LDA algorithm.

In all the experiments, M_{tr} images per person were selected from the database to form a training set and the rest are used as the test set. The Nearest Neighbor Classification method is adopted to obtain the corresponding recognition rate. Following [23], to produce reliable results, we take 10 repetitions of the random split for each database and then obtain the results in terms of the average recognition rate (over the 10 different splits) as well as the standard deviation of the recognition rates, which we incorporate in the result plots in Fig.3, Fig.4 and Fig.5 as error bars indicating plus or minus one standard deviation from the mean.

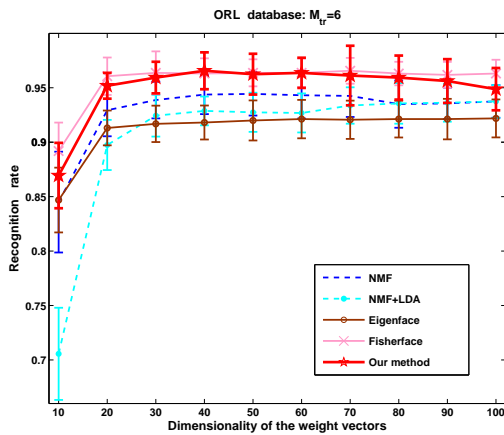


Figure 3. Average recognition rates according to the 5 methods for ORL database when $M_{tr} = 6$.

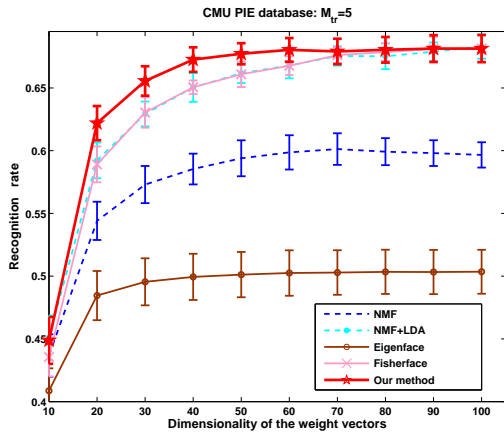


Figure 4. Average recognition rates according to the 5 methods for CMU PIE database when $M_{tr} = 5$.

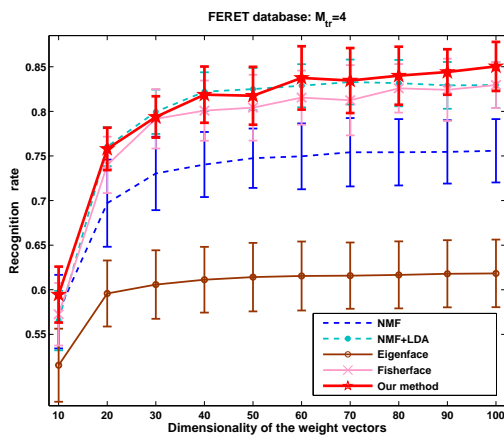


Figure 5. Average recognition rates according to the 5 methods for FERET database when $M_{tr} = 4$.

From these figures, we can see that the NMF class of

methods and the Fisherface performs much better than the PCA approach. We believe that the superior performance may be attributed to either the use of LDA (in whatever form) or the advantage of the parts-based representation of the NMF-based methods.

For the ORL database, our method is superior to the traditional NMF, sequential NMF+LDA and Eigenface, and is comparable to the Fisherface method.

For the CMU PIE database, which contains more complicated pose variations and real-life illumination changes, our method is significantly superior to all the other methods when the dimensionality of feature vectors are not very big. But the differences become negligible as the dimensionality increases and our method is practically the same as the Fisherface and the NMF+LDA, but still much better than the traditional NMF and the Eigenface.

For the FERET database, once again the three LDA-based methods outperform the traditional NMF which in turns outperforms the Eigenface method. The advantage of our method over Fisherface is still evident (albeit to a lesser extent) while the sequential NMF+LDA gives very comparable results to our approach except at large dimensionality when our method is marginally better.

To summarize, our LDA-based NMF algorithm is much better than the traditional NMF and Eigenface, and in specific cases, is at least comparable to the Fisherface method and the sequential NMF+LDA method which both use the same class information in the training set. However, our method is the only method that consistently achieves the best results for all the experimental databases used.

VI. CONCLUSIONS & FUTURE WORK

In this paper, we proposed a new constrained nonnegative matrix factorization algorithm, called LDA-based NMF, for the face recognition problem. Its basic idea is to add the Fisher discriminant to the cost function of the nonnegative matrix factorization model. We showed that it can perform better than PCA and the original NMF for face recognition. Furthermore, we found its performance can be better than Fisherface and the sequential NMF+LDA method which also take advantage of the same class information in the training images. However, the improvement depends on the face database used and the dimensionality of the feature vectors. This suggests that our method can be improved if used in conjunction with a feature selection algorithm.

Recently, several different NMF algorithms [17], [24] have been proposed, and the advantages and disadvantages of them are compared based on different aspects. But for the task of face recognition, we still need further work to enhance their performance. Based on the findings in this paper, we believe that the introduction of suitable additional constraint can lead to more robust base vectors and improve the final recognition result.

As with the PCA method, NMF also suffers from the problem of large computational load. To resolve this limitation, we are considering the use of image transform

to reduce the computational complexity and give better recognition accuracy. In particular, wavelet transform has been a very popular tool for image analysis, and we are considering applying our modified NMF method to selected subbands in our future work.

ACKNOWLEDGEMENTS

This research was partially supported by the National Natural Science Foundation of China (Grant Nos. 60877068) and Science and Technology Planning Project of Guangdong Province of China under Grant 2009B010800051, 2009B010900056.

REFERENCES

- [1] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, 1995.
- [2] G. Chow and X. Li, "Towards a system for automatic facial feature detection," *Pattern Recognition*, vol. 26, no. 12, pp. 1739–1755, 1993.
- [3] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, and N. Otsu, "Face recognition system using local autocorrelations and multiscale integration," *IEEE Trans. PAMI*, vol. 18, no. 10, pp. 1024–1028, 1996.
- [4] K. M. Lam and H. Yan, "Locating and extracting the eye in human face images," *Pattern Recognition*, vol. 29, no. 5, pp. 771–779, 1996.
- [5] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE Trans. PAMI*, vol. 12, pp. 103–108, 1990.
- [6] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.
- [7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
- [8] D. Guillaumet and J. Vitrià, "Classifying faces with non-negative matrix factorization," in *Proceedings of the 5th Catalan Conference for Artificial Intelligence (CCIA 2002)*. Castello de la Plana, Spain: Springer-Verlag, 2002, pp. 24–31.
- [9] —, "Non-negative matrix factorization for face recognition," in *CCIA '02: Proceedings of the 5th Catalonian Conference on AI*. London, UK: Springer-Verlag, 2002, pp. 336–344.
- [10] Y. Xue, C. S. Tong, Y. Chen, and W. S. Chen, "Clustering-based initialization for non-negative matrix factorization," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 525–536, 2008.
- [11] Y. Xue, C. S. Tong, W. S. Chen, W. P. Zhang, and Z. Y. He, "A modified non-negative matrix factorization algorithm for face recognition," in *ICPR (3)*. IEEE Computer Society, pp. 495–498.
- [12] A. K. Jain, *Fundamentals of digital image processing*. Prentice Hall, 1989, pp. 163–175.
- [13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Adv. Neural Info. Proc. Syst.*, vol. 13, pp. 556–562, 2001.
- [14] Y. Xue and C. S. Tong, "Evaluation of distance measures for nmf-based face recognition," in *Proceedings of the 2006 International Conference on Computational Intelligence and Security—Part I*, Y. ming Cheung, Y. Wang, and H. Liu, Eds., Nov. 2006, pp. 651–656.
- [15] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley, 2001.
- [16] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 711–720, 1997.
- [17] Y. Wang, Y. Jia, C. Hu, and M. Turk, "Fisher non-negative matrix factorization for learning local features," *Asian Conference on Computer Vision, Korea*, January 27–30, 2004.
- [18] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Trans. on Neural Networks*, vol. 17, no. 3, pp. 683–695, 2006.
- [19] Fabien Cardinaux, Conrad Sanderson, and Samy Bengio, "User Authentication via Adapted Statistical Models of Face Images," *IEEE Transaction on Signal Processing*, vol. 54, no. 1, pp. 361–373, 2005.
- [20] C. Sanderson, S. Bengio, and Y. Gao, "On transforming statistical models for non-frontal face verification," *Pattern Recognition*, vol. 39, no. 2, pp. 288–302, 2006.
- [21] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. PAMI*, vol. 22, no. 10, pp. 1090–1104, 2000.
- [22] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, 2007.
- [23] P. S. Aleksic and A. K. Katsaggelos, "Audio-visual biometrics," *IEEE Proceedings*, vol. 94, pp. 2025–2044, November 2006.
- [24] W. Liu and J. Yi, "Existing and new algorithms for non-negative matrix factorization," Tech. Rep.

Yun Xue received his Ph.D. degree in image processing from Hong Kong Baptist University, Hong Kong, in 2007. He is currently an associate professor in South China Normal University and a member of the IEEE and ACM. He has published more than twenty papers. His current research interests include pattern recognition and data mining.