Fisher Non-negative Matrix Factorization with Pairwise Weighting

Xi Li , Kazuhiro Fukui
Graduate School of Systems and Information Engineering,
University of Tsukuba, JAPAN
{xili@viplab.is,kfukui@cs}.tsukuba.ac.jp

Abstract

Non-negative matrix factorization (NMF) is a powerful feature extraction method for finding parts-based, linear representations of non-negative data . Inherently, it is unsupervised learning algorithm. That is to say, the classical NMF algorithm does not respect the class-specific information. This paper presents an improvement of the classical NMF approach by imposing Fisher constraints. This results in a two-step factorization procedure for discriminative feature extraction. Furthermore, weighting factors for each pairwise scatter is introduced to include the confusability information into the between class covariance matrix. The proposed method has been applied to the problem of face and handwritten digit recognition and the experiments give better performance than previous methods.

1. Introduction

Visual recognition of objects is one of the most studied problems within computer vision and artificial intelligence community. For appearance based recognition, usually the original image matrix is first transformed into the corresponding vector form, which is high dimensional. Subspace methods are used in high dimensional data analysis for dimension reduction. For example, principal component analysis (PCA) is used to describe the input patterns in a lower-dimensional space than the image space and has been successfully applied to several tasks such as face recognition [1]. The purpose of PCA is to find projections that best express the population of the data samples. It does not make use of the information about the way the data are separated to different classes. Fisher linear discriminative analysis (FLDA) is a supervised algorithm that seeks to find a linear transform from the original high dimensional image space into a lower one that can maximize the between-class scatter and minimize the within-class scatters[2].

Recently, a new dimension reduction technique called non-negative matrix factorization (NMF) has been proposed for obtaining a linear representation of data[3]. It differs from other methods by its use of non-negativity constraints. The NMF algorithm tries to approximate non-negative data in a part-based context. More specifically, NMF approximate a given image by a linear combination of basis images which contain no negative elements. Since NMF allows only additive, not subtractive, combinations of basis images, it could produce a sparse part-based representation. The difference between NMF and PCA is that NMF does not allow negative elements either in the basis vectors or in the representations weighting coefficients used in the linear combination of basis images. This constraint is motivated by the biological aspect that the igniting rates of brain neurons are always nonnegative.

The classical NMF method does not efficiently exploit the discrimination information in the training set since inherently the training procedure is implemented in an unsupervised way. Some efforts have been done to improve the performance of NMF by combining NMF with LDA. For example, the within-class scatter and between class scatter of the weight matrix was included in the objective cost function, and new updating rules was derived using Expectation-Maximization[4].. Experiments showed that NMF with LDA achieves better performance compared with classical NMF. But they treat all class pairs equally. But in fact, some class pairs might be much more closer that others, and the discriminative information between them is much more likely to be ignored.

This paper presents an improvement of the classical NMF approach by imposing Fisher constraints. This results in a two-step factorization procedure for discriminative feature extraction. Furthermore, by rewriting the between class scatter in a more explicit way, the weighting factor for each pairwise scatter is introduced to include the confusability information into the between class covariance matrix. The proposed method has been applied to the problem of face and handwritten digit recognition and the experiments give better performance than previous methods.

The following of this paper is organized as follows: In section 2 we recall some basic theories of dimension reduction such as PCA, LDA and NMF. The detail of the proposed pairwise weighting fisher non-negative matrix factorization is described in section 3. Some experimental results based on the proposed method are discussed in section 4. Section 5 draws the conclusion.

2. Review of dimension reduction

This section provides the background theory of dimension reduction and takes the PCA, LDA and NMF as examples.

Suppose $X = \{x_1, \cdots, x_N\}$ is the given image dataset with N samples of dimension D. The images are already vectorized and D equals to the number of pixels in the image. Each data points belongs to exactly one of

the C object classes $\{L_1,\cdots,L_C\}$. The number of images belonging to the i_th class is N_i, thus $\sum_{i=1}^{C} N_i = N$.

PCA tries to find a linear dimensionality reduction transformation that maximizes the scatter of all projected images and decompose the images in terms of basis images. Firstly, all samples are centered by subtracting the mean value of all training sample. PCA can be implemented by solving an eigenvalues decomposition of the covariance matrix $E[XX^T]=USU^T$, where the columns of U are the eigenvectors of the covariance matrix and S is the corresponding diagonal matrix. The dimension reduction can be implemented by projecting the data onto a subspace spanned by the most important eigenvectors according to the lager eigenvalues. The projections of data X can be described as follows: $Y=U_{[1,\cdots,d]}^T X$ where the $D\times d$ matrix $U_{[1,\cdots,d]}$ contains the d eigenvectors corresponding to the d largest eigenvalues.

Discriminant subspace analysis has been extensively studied in pattern recognition and computer vision community. It has been widely used for feature extraction and dimension reduction in object recognition and document classification. One popular method is the Linear Discriminant Analysis (LDA), also known as Fisher Linear Discriminant. Unlike PCA, LDA tries to find an optimal subspace such that the separability of two classes is maximized. The between class scatter matrix S_b and the within-class scatter matrix S_w can be defined as follows:

$$S_b = \sum_{i=1}^{C} N_i (m_i - m)(m_i - m)^T \quad (1)$$

$$S_w = \sum_{i=1}^{C} \sum_{x\in L_i} (x - m_i)(x - m_i)^T \quad (2)$$

Here m_i represents the class mean and m is the global mean of all samples. The objective function to be maximized is:

$$J(\Phi) = \frac{\left|\Phi^T S_b \Phi\right|}{\left|\Phi^T S_w \Phi\right|} \quad (3)$$

and there exists a closed form solution to this optimization problem, i.e. the columns of the optimal Φ are the generalized eigenvectors corresponding to the first d maximal magnitude eigenvalues of the equation

$$S_b \Phi = \lambda S_w \Phi \quad (4)$$

Suppose the given data matrix X is non-negative, which is natural for grey images in object recognition scenario, NMF finds an approximate factorization

$$X = WH \quad (5)$$

into non-negative factors W and H, where W is an $D\times d$ matrix and H an $d\times N$ matrix. Unlike other linear representations such as PCA, the non-negative constraints of NMF makes the representation purely additive. Here the encoding matrix H can be considered as new learned feature vectors based on the basis matrix W of the original data matrix X. The optimal choice of matrices W and H are defined to be those non-negative matrices that minimize the reconstruction error between X and WH. The most common used objective error function to be minimized is the squared error function as follows:

$$E(W,H) = \sum_{i,j} \left(X_{i,j} \log \frac{X_{i,j}}{(WH)_{i,j}} - X_{i,j} + (WH)_{i,j}\right)^2 \quad (6)$$

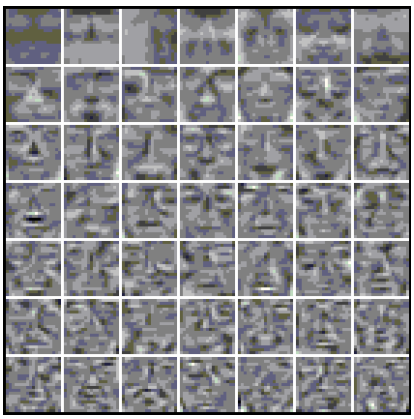
This error function can be interpreted as the likelihood of generating the images in X from the basis matrix W and the encoding matrix H. An iterative multiplicative algorithm was devised for the optimization. It is simple to implement and has good convergent property. The detail procedure is as follows:

$$W_{ia} \leftarrow W_{ia} \sum_j \frac{X_{ij}}{(WH)_{ij}} H_{aj} \quad (7)$$

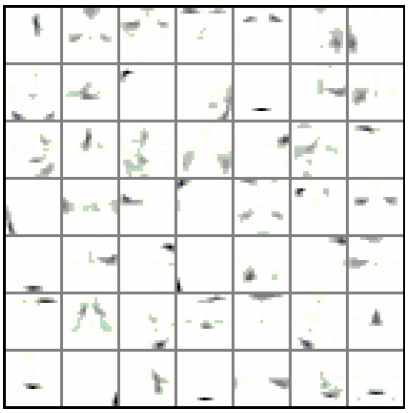
$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}} \quad (8)$$

$$H_{aj} \leftarrow H_{aj} \sum_i W_{ia} \frac{X_{ij}}{(WH)_{ij}} \quad (9)$$

The matrices W and H can be initialized using positive random matrix. Due to the non-negative constraints, the resulting bases provide a sparse and part-based representation of input images. More details can be found in literature[3]. Using face image dataset as example, Figure 1 shows the set of 49 learned basis images for PCA and NMF respectively.



(a)



(b)

Figure 1: The basis images learned via PCA(a) and NMF(b)

3. The proposed method

3.1 Fisher non-negative factorization

The classical NMF method is inherently an unsupervised algorithm and does not use the information of the image labels. Instead of rewriting the objective function to be minimized as in [4], we propose a direct two-steps projection procedure aims to make the NMF more discriminative. Our objective is to find discriminant projections for the image vectors which are already projected to the image basis matrix W. The discriminant information in NMF can be fully exploited.

Assuming h_i is the i_th column of encoding matrix H with corresponding labels L_i, then the between class scatter Ξ_w and the within class scatter Ξ_b can be defined as,

$$\Xi_b = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T \quad (10)$$

$$\Xi_w = \sum_{i=1}^{C} \sum_{h \in L_i} (h - \mu_i)(h - \mu_i)^T \quad (11)$$

where

$$\mu_i = \frac{1}{N_i} \sum_{h \in L_i} h \quad \text{and} \quad \mu = \frac{1}{N} \sum_i h_i \quad (12)$$

Again, the objective function to be maximized can be constructed based on Fisher discriminant criterion:

$$J(\Psi) = \frac{\left|\Psi^T \Xi_b \Psi\right|}{\left|\Psi^T \Xi_w \Psi\right|} \quad (13)$$

Like LDA, the optimal choice of transformation Ψ is the solution of the generalized eigenvectors problem:

$$\Xi_b \Psi = \lambda \Xi_w \Psi \quad (14)$$

And the final linear transform T that extracts the regular discriminant features using NMF is

$$T = \Psi^T W^{\dagger} \quad (15)$$

$W^{\dagger}$ denotes the pseudoinverse of W and can be calculated efficiently as $(W^T W)^{-1} W^T$.

3.2 Pairwise weighting

We further extend our fisher non-negative factorization algorithm by incorporating into the discriminant criterion a pairwise weighting function. Like [5], we rewrite the between class scatter matrix for the NMF transformed image data, i.e., the data points for the encoding matrix H into the following form:

$$\Xi_b = \frac{1}{N^2} \sum_{i=1}^{C-1} \sum_{j=i+1}^{C} N_i N_j w_{ij} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \quad (16)$$

where $\{w_{ij}\}$ is a set of non-negative weights assigned to class pair (i, j). The weighted between class scatter matrix is a generalization of the original one, that is to say, if all w_{ij} equal to 1, the new weighted between class scatter matrix is exactly the same as conventional between class scatter matrix. We put more weight on those class pairs which are closer. In other words, those class pairs which are more confusable to be classified are weighted more and those class pairs which are less confusable to be classified are weighted less. In the experiments of this work, the weighting function is chosen as the square of the inverse of the Euclidean distance between class centers, i.e.:

$$w_{ij} = \frac{1}{\left\|\mu_i - \mu_j\right\|^2} \quad (17)$$

Experiments show that the introducing of the weighting function can significantly improve the recognition rate, which will be demonstrated in next section.

4. Experiment results

The well-known ORL face database[6] and the USPS digit database[7] are used to test the proposed algorithm.

4.1 Experiments with the ORL database

In ORL database, there are 40 persons and each person consists of 10 different images taken at different times, varying lighting conditions, facial expressions and other facial details(glasses / no glasses). The size of each image is 92×112 pixels with 256 grey levels per pixel. The original images are resized to a resolution of 24×32 and the facial image luminance is normalized via histogram equalization. Each set of the 10 images for a specific person is randomly partitioned into a training set of 5 images and a testing set of the other 5 images. Figure 2 shows the similarity matrix between the 10 subjects.

We compare the recognition performance of principal component analysis (PCA) , linear discriminant analysis (LDA) , classical non-negative factorization (NMF) , the proposed Fisher non-negative factorization (F-NMF) and Fisher non-negative factorization with pairwise weighting (WF-NMF). The feature dimension for the PCA, NMF and the first step of factorization of F-NMF and WF-NMF is chosen to be 20,40,60,…,140. For LDA and the second factorization step of F-NMF and W F-NMF, the reduced dimension is bounded by C-1, where C is the number of classes. We chosen it to be C-1 since it achieves best results. Here the fisher discriminant analysis is implemented in the space of encoding matrix, which is low dimensional, the small sample size problem does not exist. The weighting function is chosen as described in sec3.2. Figure 3 illustrates the plot of recognition rate versus the dimension of the reduced space. It can be clearly seen that the proposed Fisher non-negative matrix factorization outperforms the classical PCA,LDA and the original NMF algorithm. This is expected since the class-specific discriminant information has been fully exploited in the F-NMF. The WF-NMF can furthermore improve the recognition rate and achieves the best performance.

4.2 Experiments with the USPS database

The USPS digit database contains a training set with 7291 images and a test set with 2007 images including

handwritten digits of 0-9. All images are 16×16 grey images and are normalized to the range of [0 1]. Feature dimensionality d = 50,100,…,250 are tested. Figure 4 shows the similarity matrix between the 10 digits. It can be seen that some subject pairs are much similar, which renders them easily to be confused each other, such as 3 and 8. The comparison of recognition rates is plotted in Figure 5. Again, the WF-NMF achieves the best result.

5. Conclusion

This paper presents a novel algorithm named pairwise weighting Fisher non-negative matrix factorization (WF-NMF), which combines the merit of the biological inspired part-based representation for NMF and the discriminative power of the LDA. Furthermore, weighting factors for each pairwise scatter is introduced to incorporate the confusability information into the between class covariance matrix. The proposed method has been applied to the problem of recognition using real life dataset and achieves better performance than previous methods.

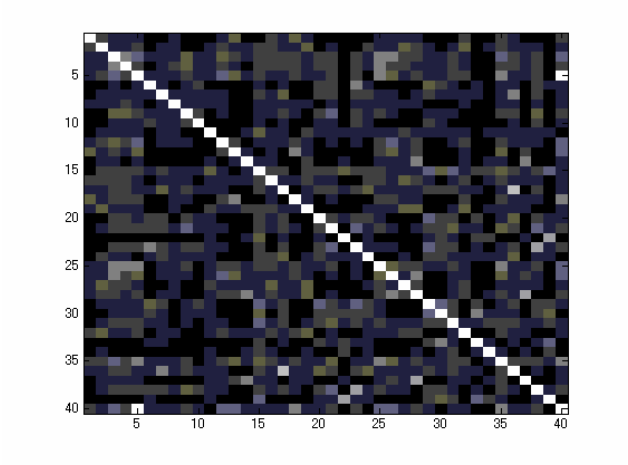


Figure 2: The similarity matrix for ORL database

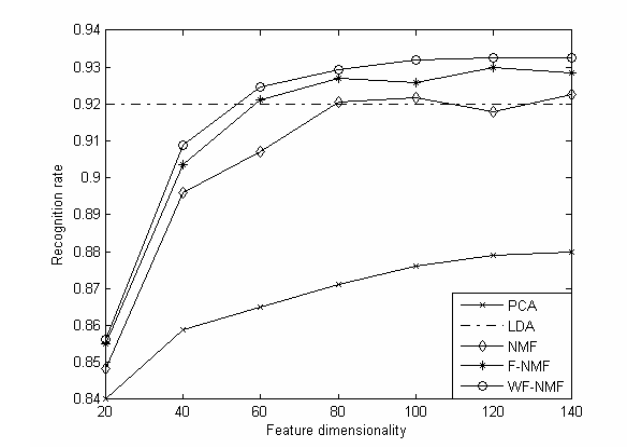


Figure 3: The performance comparison of PCA, LDA, NMF, F-NMF and WF-NMF for ORL database.

References

[1] M.Turk and A. Pentland. Eigenfaces for recognition. J. Cognitive Neuroscience,3(1):71-86,1991

[2] Peter N. Belhumeur, Joao P. Hespanha and David J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 19 (7): 711-720,1997

[3] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 788-791,1999

[4] Wang Y, Jia Y, Hu C and Murk M., Fisher non-negative matrix factorization for learning local features. Asian Conference on Computer Vision,Korea,2004

[5] M. Loog and R.P.W. Duin. Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(6):732–739, 2004

[6] F. Samaria and A. Harter, Parameterization of a stochas-tic model for human face identification, in Second IEEE-Workshop on Applications of Computer Vision , 138-142,1994

[7] http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html

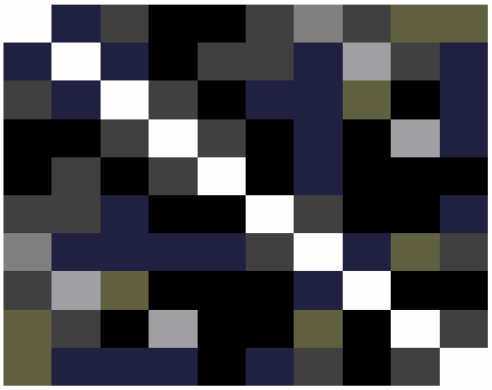


Figure 4: The similarity matrix for USPS database

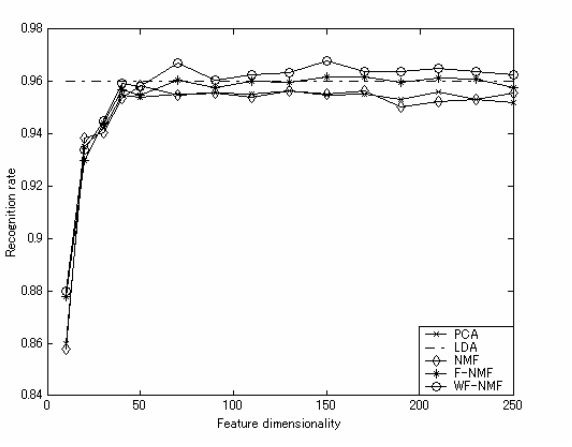


Figure 5: The performance comparison of PCA, LDA, NMF, F-NMF and WF-NMF for USPS database.