FACIAL EXPRESSION RECOGNITION USING CLUSTERING DISCRIMINANT NON-NEGATIVE MATRIX FACTORIZATION

Symeon Nikitidis[†*], *Anastasios Tefas*[†], *Nikos Nikolaidis*[†*] *and Ioannis Pitas*[†*]

[*]Informatics and Telematics Institute, Center for Research and Technology Hellas, Greece
[†]Department of Informatics, Aristotle University of Thessaloniki, Greece
{nikitidis,tefas,nikolaid,pitas}@aiia.csd.auth.gr

ABSTRACT

Non-negative Matrix Factorization (NMF) is among the most popular subspace methods widely used in a variety of image processing problems. Recently, a discriminant NMF method that incorporates Linear Discriminant Analysis criteria and achieves an efficient decomposition of the provided data to its discriminant parts has been proposed. However, this approach poses several limitations since it assumes that the underline data distribution forms compact sets which is often unrealistic. To remedy this limitation we regard that data inside each class form various number of clusters and apply a Clustering based Discriminant Analysis. The proposed method combines appropriate discriminant constraints in the NMF decomposition cost function in order to address the problem of finding discriminant projections that enhance class separability in the reduced dimensional projection space. Experimental results performed on the Cohn-Kanade database verified the effectiveness of the proposed method in the facial expression recognition task.

Index Terms— Non-negative matrix factorization, subspace methods, clustering discriminant analysis, facial expression recognition

1. INTRODUCTION

NMF [1] is a matrix decomposition algorithm that requires both the data matrix being decomposed and the yielding factors to contain non negative elements. The non negativity constraint imposed has been exploited by a variety of applications since many types of data in practical problems are non negative. For instance, numerous NMF-based methods operating on data derived from text documents [2, 3] or images, have been developed based on NMF in image processing and pattern recognition and proved efficient compared with other traditional dimensionality reduction algorithms.

Recently numerous practical applications have been proposed, creating specialized NMF based algorithms applied in various problems in diverse fields. A supervised NMF learning method that aims to extract discriminant facial parts, is the Discriminant NMF (DNMF) algorithm proposed in [4]. DNMF combines Fisher's criterion in the NMF decomposition and achieves a more efficient decomposition of the provided data to its discriminant parts, thus enhancing separability between classes compared with conventional NMF. However, the incorporation of Linear Discriminant Analysis (LDA) [5] inside DNMF poses certain deficiencies. More precisely, there are two main disadvantages in this approach. Firstly, LDA assumes that the sample vectors of the classes are generated from underlying multivariate Normal distributions of common covariance matrix but with different means. Secondly, since LDA assumes that each class is represented by a single cluster, the problem of nonlinearly separable classes can not be solved. Unfortunately, in real world applications, data distribution usually do not correspond to compact sets.

To remedy the aforementioned limitations we relax the assumption that each class is expected to consist of a single compact cluster and regard that data inside each class form various clusters, where each one is approximated by a Gaussian distribution. Consequently, we approximate the underlying data samples distribution of each class as a mixture of Gaussians and use the corresponding criteria from the Clustering based Discriminant Analysis (CDA) introduced in [6]. By incorporating these discriminant constraints in the NMF decomposition we derive the proposed method called Subclass Discriminant NMF (SDNMF). The SDNMF algorithm addresses the general problem of finding discriminant projections that enhance class separability in the reduced dimensionality projection space.

2. NMF BASICS

Focusing on the application of the NMF algorithm on facial image data, NMF aims to approximate a facial image by a linear combination of elements the so called basis images, that correspond to facial parts. Thus, the non negativity constraints imply that the combinations of the multiple basis im-

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 211471 (i3DPost).

ages are practically additions of ideally non-overlapping facial parts that attempt to reconstruct accurately the image being decomposed. Let $\mathcal{I}$ be a facial image database comprised of L images belonging to n different classes and $\mathbf{X} \in R_+^{F\times L}$ is the data matrix whose columns are F-dimensional feature vectors obtained by scanning row-wise each facial image in the database. Thus $x_{i,j}$ is the i-th element of the j-th column vector $\mathbf{x}_j$. NMF considers factorizations of the form:

$$\mathbf{X} \approx \mathbf{ZH} \tag{1}$$

where $\mathbf{Z} \in R_+^{F\times M}$ is a matrix containing the basis images, while matrix $\mathbf{H} \in R_+^{M\times L}$ contains the coefficients of the linear combinations of the basis images required to reconstruct each original facial image in the database. Obviously, useful factorizations for real world applications appear when the linear subspace transformation projects data from the original F-dimensional space to a M-dimensional subspace with $M \ll F$.

To measure the cost of the decomposition in (1), one popular approach is to use the Kullback-Leibler (KL) divergence metric [7, 8]. Thus the cost of the decomposition in (1) can be measured as the sum of all KL divergences between all images in the database and their respective reconstructed versions, obtained from the factorization. Consequently, the cost for factorizing $\mathbf{X}$ into $\mathbf{ZH}$ is evaluated as:

$$D_{NMF}(\mathbf{X}||\mathbf{ZH}) = \sum_{j=1}^{L} KL(\mathbf{x}_j||\mathbf{Zh}_j) =$$

$$= \sum_{j=1}^{L}\sum_{i=1}^{F}\left(x_{i,j}\ln(\frac{x_{i,j}}{\sum_k z_{i,k}h_{k,j}}) + \sum_k z_{i,k}h_{k,j} - x_{i,j}\right). \tag{2}$$

Using the Expectation Maximization (EM) algorithm and an appropriately designed auxiliary function, it has been shown in [9] that the following multiplicative update rules update $h_{k,j}$ and $z_{i,k}$, yielding the desired factors, while guarantee a non increasing behavior of the cost function in (2). The update rule for the t-th iteration for $h_{k,j}^{(t)}$ is given by:

$$h_{k,j}^{(t)} = h_{k,j}^{(t-1)} \frac{\sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}{\sum_i z_{i,k}^{(t-1)}}, \tag{3}$$

while for $z_{i,k}^{(t)}$ the update rule is given by:

$$\acute{z}_{i,k}^{(t)} = z_{i,k}^{(t-1)} \frac{\sum_j h_{k,j}^{(t)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t)}}}{\sum_j h_{k,j}^{(t)}}. \tag{4}$$

Finally, the basis images matrix $\mathbf{Z}$ is normalized so that its column vectors elements sum up to one:

$$z_{i,k}^{(t)} = \frac{\acute{z}_{i,k}^{(t)}}{\sum_l \acute{z}_{l,k}^{(t)}}. \tag{5}$$

3. PROPOSED METHOD

In this section we present the performed clustering based discriminant analysis and demonstrate how the derived discriminant constraints are incorporated in the NMF decomposition cost function creating the proposed SDNMF optimization problem. Next, we derive the proposed multiplicative update rule that optimize SDNMF.

3.1. Clustering based Discriminant Analysis

To facilitate CDA in the n-class facial image database $\mathcal{I}$, let us denote the number of clusters composing the r-th class by C_r, the total number of formed clusters in the database by C, where $C = \sum_i^n C_i$, and the number of facial images belonging to the θ-th cluster of the r-th class by $N_{(r)(\theta)}$. Let us also define the mean vector for the θ-th cluster of the r-th class by $\boldsymbol{\mu}^{(r)(\theta)} = [\mu_1^{(r)(\theta)} \dots \mu_M^{(r)(\theta)}]^T$ which is evaluated over the $N_{(r)(\theta)}$ facial images, while vector $\boldsymbol{\eta}_\rho^{(r)(\theta)} = [\eta_{\rho,1}^{(r)(\theta)} \dots \eta_{\rho,M}^{(r)(\theta)}]^T$ corresponds to the feature vector of the ρ-th image of the θ-th cluster of the r-th class. Using the above notations we can define the within cluster scatter matrix $\mathbf{S}_w$ as:

$$\mathbf{S}_w = \sum_{r=1}^{n}\sum_{\theta=1}^{C_r}\sum_{\rho=1}^{N_{(r)(\theta)}} \left(\boldsymbol{\eta}_\rho^{(r)(\theta)} - \boldsymbol{\mu}^{(r)(\theta)}\right)\left(\boldsymbol{\eta}_\rho^{(r)(\theta)} - \boldsymbol{\mu}^{(r)(\theta)}\right)^T \tag{6}$$

and the between cluster scatter matrix $\mathbf{S}_b$ as:

$$\mathbf{S}_b = \sum_{i=1}^{n}\sum_{r,r\neq i}^{n}\sum_{j=1}^{C_i}\sum_{\theta=1}^{C_r} \left(\boldsymbol{\mu}^{(i)(j)} - \boldsymbol{\mu}^{(r)(\theta)}\right)\left(\boldsymbol{\mu}^{(i)(j)} - \boldsymbol{\mu}^{(r)(\theta)}\right)^T. \tag{7}$$

Matrix $\mathbf{S}_w$ represents the scatter of sample vector coefficients around their cluster mean. It is rationale to desire the dispersion of those samples that belong to the same cluster of a certain class to be as small as possible, since this would denote a high concentration of these samples around their cluster mean and consequently, more compact clusters formation. In order to measure the samples dispersion inside clusters we compute the trace of the within cluster scatter matrix $\mathbf{S}_w$. Furthermore, matrix $\mathbf{S}_b$ defines the scatter of the mean vectors between all clusters that belong to different classes. To separate clusters belonging to different classes we desire to maximize the mean difference between every cluster of a certain class to every cluster of each other class. Therefore, the trace of $\mathbf{S}_b$ is desired to be as large as possible.

3.2. Subclass Discriminant Non-negative Matrix Factorization (SDNMF)

In order to incorporate clustering based discriminant constraints derived from the performed CDA in the NMF decomposition, we reformulate the NMF cost function adding

appropriate penalty terms. Since we desire the trace of matrix $\mathbf{S}_w$ to be as small as possible and at the same time the trace of $\mathbf{S}_b$ to be as large as possible, the new cost function is formulated as:

$$D_{SDNMF}(\mathbf{X}||\mathbf{ZH}) = D_{NMF}(\mathbf{X}||\mathbf{ZH}) + \frac{\alpha}{2}\text{tr}[\mathbf{S}_w] - \frac{\beta}{2}\text{tr}[\mathbf{S}_b] \tag{8}$$

where α and β are positive constants, tr[.] is the trace operator, while $\frac{1}{2}$ is used to simplify subsequent derivations. Consequently, the new minimization problem is formulated as:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{H}} D_{SDNMF}(\mathbf{X}||\mathbf{ZH}) \\ &\text{subject to:} \quad z_{i,k} \geq 0 \quad , h_{k,j} \geq 0, \quad \sum_i z_{i,k} = 1, \forall i,j,k. \end{aligned} \tag{9}$$

which requires the minimization of (8) subject to the non-negativity constraints applied on the elements of both the weights matrix $\mathbf{H}$ and the basis images matrix $\mathbf{Z}$.

The constrained optimization problem in (9) is solved by introducing Lagrangian multipliers $\boldsymbol{\phi} = [\phi_{i,k}] \in R^{F\times M}$ and $\boldsymbol{\psi} = [\psi_{j,k}] \in R^{M\times L}$ each one associated with constraints $z_{i,k} \geq 0$ and $h_{k,j} \geq 0$, respectively. Thus the Lagrangian function $\mathcal{L}$ is formulated as:

$$\begin{aligned} \mathcal{L} &= D_{NMF}(\mathbf{X}||\mathbf{ZH})) + \frac{\alpha}{2}\text{tr}[\mathbf{S}_w] - \frac{\beta}{2}\text{tr}[\mathbf{S}_b] + \\ &+ \sum_i\sum_k \phi_{i,k}z_{i,k} + \sum_j\sum_k \psi_{j,k}h_{j,k} \Leftrightarrow \\ \mathcal{L} &= D_{NMF}(\mathbf{X}||\mathbf{ZH})) + \frac{\alpha}{2}\text{tr}[\mathbf{S}_w] - \frac{\beta}{2}\text{tr}[\mathbf{S}_b] + \\ &+ \text{tr}[\boldsymbol{\phi}\mathbf{Z}^T] + \text{tr}[\boldsymbol{\psi}\mathbf{H}^T]. \end{aligned} \tag{10}$$

Consequently, the optimization problem in (9) is equivalent to the minimization of the Lagrangian $\arg\min_{\mathbf{Z},\mathbf{H}} \mathcal{L}$. To minimize $\mathcal{L}$, we first obtain its partial derivatives with respect to $z_{i,j}$ and $h_{i,j}$ and set them equal to zero:

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial h_{i,j}} &= -\sum_k \frac{x_{k,j}z_{k,i}}{\sum_l z_{k,l}h_{l,j}} + \sum_l z_{l,i} + \psi_{i,j} + \frac{\alpha}{2}\frac{\partial\text{tr}[\mathbf{S}_w]}{\partial h_{i,j}} \\ &- \frac{\beta}{2}\frac{\partial\text{tr}[\mathbf{S}_b]}{\partial h_{i,j}} = 0 \\ \frac{\partial\mathcal{L}}{\partial z_{i,j}} &= -\sum_l \frac{x_{i,l}h_{j,l}}{\sum_k z_{i,k}h_{k,l}} + \sum_l h_{j,l} + \phi_{i,j} + \frac{\alpha}{2}\frac{\partial\text{tr}[\mathbf{S}_w]}{\partial z_{i,j}} \\ &- \frac{\beta}{2}\frac{\partial\text{tr}[\mathbf{S}_b]}{\partial z_{i,j}} = 0. \end{aligned} \tag{11}$$

According to KKT conditions [10], $\phi_{i,j}z_{i,j} = 0$ and also $\psi_{i,j}h_{i,j} = 0$. Consequently, we obtain the following equalities:

$$\begin{aligned} &\left(\frac{\partial\mathcal{L}}{\partial h_{i,j}}\right) h_{i,j} = 0 \Leftrightarrow -\sum_k \frac{x_{k,j}z_{k,i}}{\sum_l z_{k,l}h_{l,j}}h_{i,j} + \sum_l z_{l,i}h_{i,j} \\ &+ \alpha\left(h_{i,j} - \mu_i^{(r)(\theta)}\right)h_{i,j} - \frac{\beta}{N_{(r)(\theta)}}\mu_i^{(r)(\theta)}(C - C_r)h_{i,j} \\ &+ \frac{\beta}{N_{(r)(\theta)}}\sum_{m,m\neq r}^{n}\sum_{g=1}^{C_m}\mu_i^{(m)(g)}h_{i,j} = 0 \end{aligned} \tag{12}$$

$$\left(\frac{\partial\mathcal{L}}{\partial z_{i,j}}\right) z_{i,j} = 0 \Leftrightarrow -\sum_l \frac{x_{i,l}h_{j,l}}{\sum_k z_{i,k}h_{k,l}}z_{i,j} + \sum_l h_{j,l}z_{i,j} = 0. \tag{13}$$

Solving the resulting from equation (12) quadratic function for $h_{i,j}$, leads to the proposed multiplicative update rule for the weight coefficients which for the t-th iteration is defined as:

$$h_{i,j}^{(t)} = \frac{A + \sqrt{A^2 + T}}{2\left(\alpha - \left[\alpha + \frac{\beta}{N_{(r)(\theta)}}(C - C_r)\right]\frac{1}{N_{(r)(\theta)}}\right)} \tag{14}$$

$$\text{where} \quad T = 4\left(\alpha - \left[\alpha + \frac{\beta}{N_{(r)(\theta)}}(C - C_r)\right]\frac{1}{N_{(r)(\theta)}}\right)\cdot h_{i,j}^{(t-1)}\sum_k z_{k,i}^{(t-1)}\frac{x_{k,j}}{\sum_n z_{k,n}^{(t-1)}h_{n,j}^{(t-1)}},$$

$h_{i,j}$ denotes the j-th feature element of the ρ-th image belonging to the θ-th cluster of the r-th facial class and A is defined as:

$$\begin{aligned} A &= \left(\alpha + \frac{\beta}{N_{(r)(\theta)}}(C - C_r)\right)\frac{1}{N_{(r)(\theta)}}\sum_{\lambda,\lambda\neq j} h_{i,\lambda} \\ &- \frac{\beta}{N_{(r)(\theta)}}\sum_{m,m\neq r}^{n}\sum_{g=1}^{C_m}\mu_i^{(m)(g)} - 1. \end{aligned} \tag{15}$$

The update rule for $z_{i,j}$ is directly derived by solving (13) and is the same as in (4).

4. EXPERIMENTAL RESULTS

We compared the performance of the proposed SDNMF method with the DNMF and the conventional NMF algorithm on the facial expression recognition problem using the popular Cohn-Kanade [11] database. In order to form the training and test sets, face detection was performed and the resulting Regions Of Interest (ROIs) were manually aligned with respect to the eyes position. Each facial image in the database was isotropically scaled, so as to have fixed size of 30×40 pixels (where 30 and 40 are the columns and rows of the image, respectively) and was converted to grayscale. Each such fixed size facial image was scanned row-wise so as to form a feature vector $\mathbf{x} = [f_1 \ldots f_{1200}]^T$ (f_i being the

luminance of the i-th pixel) which is used to form the training and test sets.

We have performed 5-fold cross-validation on the available data samples where the training set was used in order to learn the basis images for the low dimensional projection space while the test set to report the facial expression recognition accuracy rates in the respective learned projection space. Classification was performed by feeding the projected to the lower dimensional discriminant facial expression representations to a linear SVM classifier.

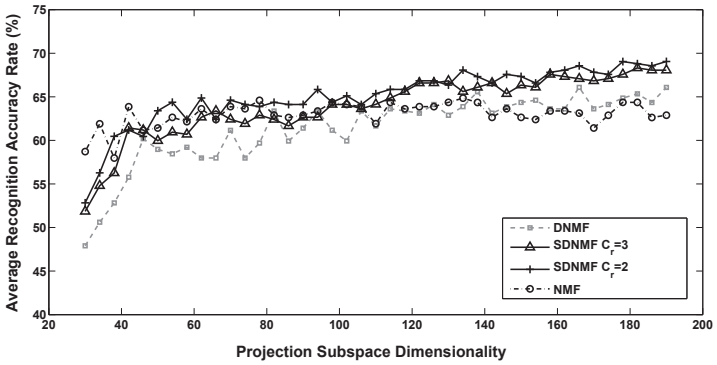

Fig. 1. Average facial expression recognition accuracy rate versus the dimensionality of the projection subspace in the Cohn-Kanade database.

Figure 1 shows the average expression recognition accuracy rates versus the projection subspace dimensionality. The highest measured recognition rates achieved by each examined method, as well as, the respective subspace dimensionality are summarized in Table 1. As it can be seen SDNMF outperforms both NMF and DNMF methods.

Table 1. Best average expression recognition accuracy rates in Cohn-Kanade database

Method	Accuracy Rate	Subspace Dimensionality
SDNMF $C_r = 2$	$\mathbf{69.05}\%$	190
SDNMF $C_r = 3$	68.31%	182
DNMF	66.08%	166
NMF	64.85%	134

5. CONCLUSIONS

We proposed a novel method that addresses the general problem of finding discriminant projections that enhance class separability by incorporating CDA in the NMF decomposition. To solve the SDNMF problem, we develop a multiplicative update rule that considers not only samples class origin but also clusters formation inside each class. We compared the performance of SDNMF algorithm with NMF and DNMF and the experimental results verified the effectiveness of the proposed method in the facial expression recognition task.

6. REFERENCES

[1] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

[3] V.P. Pauca, F. Shahnaz, M.W. Berry, and R.J. Plemmons, "Text mining using nonnegative matrix factorizations," in *IEEE International Conference on Data Mining (ICDM)*, 2004, pp. 452–456.

[4] Stefanos Zafeiriou, Anastasios Tefas, Ioan Buciu, and Ioannis Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 683–695, 2006.

[5] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, July 1997.

[6] X. Chen and T. Huang, "Facial expression recognition: a clustering-based approach," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1295–1302, 2003.

[7] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Statistical learning algorithms based on bregman distances," in *Proceedings of the Canadian Workshop on Information Theory*, Toronto, Canada, 1997.

[8] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, adaboost and bregman distances," *Computational Learing Theory*, pp. 158–169, 2000.

[9] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, 2000, pp. 556–562.

[10] R. Fletcher, *Practical methods of optimization; (2nd ed.)*, Wiley-Interscience, New York, NY, USA, 1987.

[11] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," March 2000, pp. 46–53.