A comparative study of NMF, DNMF, and LNMF algorithms applied for face recognition

Ioan Buciu, Nikos Nikolaidis and Ioannis Pitas

Abstract—Three techniques called non-negative matrix factorization (NMF), local non-negative matrix factorization (LNMF), and discriminant non-negative matrix factorization (DNMF), have been recently developed for decomposing a data matrix into non-negative factors named basis images and decomposition coefficients. Although these techniques are closely related to each other since they impose certain common non-negative constraints, the decomposition process of each algorithm involves a different objective function. While NMF approximates in the best possible way the data matrix by the product of its decomposition factors imposing only non-negative constraints, LNMF adds more constraints on the basis images to reduce the redundant information between them and to enlarge their sparseness degree. DNMF imposes more constraints on the coefficients in order to take into account class information. In this paper these methods are used in the context of face recognition to extract features from two image databases (YALE and ORL). Extracted features are further classified by two metric-based classifiers, namely Maximum Correlation Classifier (MCC) and Cosine Similarity Measure (CSM). Besides, Support Vector Machines (SVMs) are also used for classification. Experiments show that when these algorithms are applied along with the aforementioned classifiers, to face recognition task they lead to quite different results, their performance being data dependent.

I. INTRODUCTION

Due to its wide range of applications that includes biometrics, information security, law enforcement and surveillance, etc., face recognition is a research topic heavily investigated by many researchers working on the field of computer vision. A survey on face recognition can be found in [1]. Although a variety of techniques have been proposed to cope with this task, their performance differs from one test image database to another. Face recognition is not an easy task since facial images can be recorded under various conditions such as non-ideal illumination conditions, the presence of occlusions, or different facial expressions. In this paper we compared three subspace methods named non-negative matrix factorization (NMF) [2], local non-negative matrix factorization (LNMF) [3], and discriminant non-negative matrix factorization (DNMF) [4] on the task of feature extraction from two image databases (YALE [5], and ORL [6]). The resulting features are further used to recognize faces from the aforementioned databases by employing two metrics-based classifiers: Maximum Correlation Classifier (MCC), Cosine Similarity Measure (CSM). Support Vector Machines (SVMs) are also used for classification. NMF has been tested in the context of face recognition in [7] along with the MCC. In this paper NMF is employed with CSM and SVM. The face recognition performance of LNMF and NMF has been compared in the ORL database in [3], and LNMF showed superior recognition performance over NMF. In this paper we investigate the behavior of these algorithms in the case of another database namely the YALE face database. The recently introduced DNMF algorithm that was successfully applied to recognize facial expression [4] is also investigated for its potential to recognize faces.

II. FEATURE EXTRACTION

Suppose that each image from a database is stored in a \( m \)-dimensional column vector whose elements are the pixel values obtained by lexicographically scanning an image, i.e. \( x = [x_1, x_2, \ldots, x_m]^T \), where \( m \) represents the total number of image pixels. Given \( n \) images \( x_1, x_2, \ldots, x_n \), we can store them in a matrix \( X \) of size \( m \times n \). Since pixel values correspond to grayscale levels the matrix \( X \) is non-negative. Non-negative matrix factorization (NMF) evaluates two non-negative factors, namely \( W \) (basis images) and \( H \) (coefficients) such that each image \( x_j \), for \( j = 1, 2, \ldots, n \) is a linear combination of the basis images, i.e. \( x_j \approx WH \). Here \( W \) is a matrix of size \( m \times p \) (whose columns store the basis images), and \( H \) is a \( p \times 1 \) vector comprising the linear decomposition coefficients. Due to the non-negativity constraint imposed in the decomposition process the reconstructed image is formed additive composition of basis images. The quality of approximation depends on the cost function associated to the decomposition. Two cost function have been proposed in [2]: Kullback-Leibler divergence \( KL(x|WH)_{NMF} = \sum_i x_i \log \frac{\sum_k w_{ik}h_{kj}}{\sum_j w_{ik}h_{kj}} - x_i + \sum_k W_{ik}h_{kj} \) and squared Euclidean distance \( D(x|WH)_{NMF} = \sum_i \| x_i - \sum_k W_{ik}h_{kj} \|^2 \) between \( x \) and its decomposition \( WH \). When the first cost function is chosen to be minimized, the following multiplicative updating rules, found by applying a Expectation - Maximization (EM) approach, guarantee a non-increasing behavior of the cost function [2]:

\[
h_{kj} = h_{kj} \frac{\sum_i w_{ik}x_{ij}}{\sum_i w_{ik}}.
\]

(1)

\[
w_{ik} = w_{ik} \frac{\sum_j x_{ij}h_{kj}}{\sum_j h_{kj}}.
\]

(2)

Local non-negative matrix factorization (LNMF) has been developed by Li et al [3] in order to increase the basis images sparseness. Also, by appropriate modification of the cost function by imposing basis orthogonality the redundant...
information captured by the basis images is minimized. The new cost function becomes:

\[ D(X|WH)_{LNMF} = KL(X|WH)_{NMF} + \alpha \sum_{i,k} u_{ik} - \beta \sum_{k} v_{kk} + \gamma S_w(h) - \delta S_b(h), \]

subject to \( W, H \geq 0 \). Here \( \gamma \) and \( \delta \) are constants. Both \( S_w(h) \) and \( S_b(h) \) are associated to the coefficient matrix. Notice that by eliminating the last two terms we obtain LNMF cost function. Obviously, by minimizing within-class scatter matrix we want the dispersion of samples that belong to the same class around their corresponding mean to be as small as possible, while by maximizing the between-class scatter matrix each cluster formed by the samples that belong to the same class is moved as far as possible from the other clusters.

In the light of the same EM strategy the DNMF algorithm updates the coefficients as follows [4]:

\[ h_{kli}(c) = \frac{2\mu_c - 1 + \sqrt{(1 - 2\mu_c)^2 + 8\xi h_{kli}(c)} \sum_i w_{ki} \sum_k w_{ik} h_{kli}(c)}{4\xi} \]

where \( \xi \) is a constant. The elements \( h_{kli} \) are then concatenated for all \( Q \) classes as:

\[ h^{(t)}_{kli} = [h_{kli(1)} | h_{kli(2)} | \ldots | h_{kli(Q)}] \]

where \( ^{t} \) denotes concatenation. Since there is no change in the cost function related to the basis images with respect to LNMF their update follows the same rules (5) and (6).

### III. Classification Procedure

Let us now split the \( n \) face images into a training set \( n_{tr} \) and a disjoint test set \( n_{te} \) with the corresponding matrices \( X_{tr} \) and \( X_{te} \), respectively. The training images \( X_{tr} \) are used in the expression for evaluating the decomposition factors. In the classical classification problem, we construct a classifier where the output (predicted value) of the classifier for a test image \( x_{te} \) is \( \hat{l} \). Since \( X_{tr} = WH \), the feature vectors used for classification are formed as \( h_{te} = W^{-1}x_{te} \), where \( x_{te} \) is now a zero mean training face. A new test feature vector \( h_{te} \) is then formed as \( h_{te} = W^{-1}x_{te} \), where \( x_{te} \) is a zero mean test face. The recognition error is defined as the percentage of misclassified face images when \( \{ l(h_{te}) \neq \hat{l}(h_{te}) \} \). Once we have formed \( Q \) classes of new feature vectors the following three classifiers are employed to classify a new test image:

1. **Cosine similarity measure (CSM).** This approach is based on the nearest neighbor rule and uses as similarity the angle between a test feature vector and a training one. We choose the feature vector \( h_{te} \) for which \( \hat{l} = \arg\min_{d_c} \{d_c \} \), where \( d_c = \| h_{te} - h_{tr} \| \) and \( d_c \) is the cosine of the angle between a test feature vector \( h_{te} \) and the training one \( h_{tr} \).

2. **Maximum correlation classifier (MCC).** The second classifier is a minimum Euclidean distance classifier. The Euclidean distance from \( h_{te} \) to \( h_{tr} \) is expressed as \( \| h_{te} - h_{tr} \|^2 = -2g_c(h_{te}) + (h_{te})^T h_{tr} \), where \( g_c(h_{te}) = (h_{te})^T h_{tr} - \frac{1}{2} \| h_{te} \|^2 \) is a linear discriminant function of \( h_{te} \). A test image is classified by this classifier by computing \( Q \) linear discriminant functions and choosing the MCC:

\[ l_{MCC} = \arg\max_{c} g_c(h_{te}) \]

3. **Support vector machines (SVMs).** For SVMs the class membership for a new test vector \( h_{te} \) is given by the sign of the following decision function [8]:

\[ f(h_{te}) = \sum_{i=1}^{n_{tr}} \alpha_i l_i K(h_{te}, h_{tr}) + b \]

where \( K(h_1, h_2) \) is a kernel function that defines the dot product between \( \Phi(h_1) \) and \( \Phi(h_2) \) in an higher-dimensional
Hilbert space $\mathcal{H}$, $\Phi$ denoting a nonlinear map $\Phi : \mathbb{R}^m \rightarrow \mathcal{H}$. $\alpha_i$ are nonnegative Lagrange multipliers associated with the quadratic optimization problem:

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2} y^T y + E \sum_{i=1}^{n(\tau)} \rho_i \\
\text{subject to} & \quad l_i (y^T \Phi(h_i^{(\tau)}) + b) \geq 1 - \rho_i, \ i = 1, \ldots, n(\tau) \quad (11)
\end{align*}$$

In (11), $y$ and $b$ are the parameters of the optimal hyperplane in $\mathcal{H}$ that attempts to separate the classes. That is, $y$ is the normal vector to the hyperplane, $|b|/\|y\|$ is the perpendicular distance from the hyperplane to the origin, with $\|y\|$ denoting the Euclidian norm of vector $y$. $E$ is a parameter which penalizes the errors and $\rho_i$ are positive slack variables. Frequently used kernel functions are the polynomial kernel, $K(h_i, h_j) = (h_i^T h_j + s)^q$ and the Exponential Radial Basis Function (ERBF) kernel, $K(h_i, h_j) = \exp\{-\gamma |h_i - h_j|\}$. We used $q = 1$ (equivalent to a linear classifier), $q = 2, 3, 4$ and $\gamma = 0.005$ in our experiments. To handle multi-class classification we chose the Decision Directed Acyclic Graph (DDAG) learning architecture proposed by Platt et al. [9].

**IV. DATABASES DESCRIPTION**

Two different public available databases have been chosen to work with. The Yale face database contains 165 grayscale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. For computational reasons the image size was reduced to $42 \times 31$ pixels. The second database used was the ORL face database that contains ten different images for forty distinct subjects. All images have been shot against a dark homogeneous background with the subjects in an upright, frontal position with tolerance for some side movement. For the experiments the images were downsampled to $42 \times 31$ pixels.

**V. EXPERIMENTAL RESULTS**

For the Yale database, the first six image samples of each subject were used to form the training data set while the remaining five samples were used as test images. In the case of the ORL database, the images have been randomly split in 200 samples for training data set and the rest 200 of them as test data set. Figures 1 and 2 depict the recognition error for the CSM and the MCC classifier, respectively, on the Yale database versus the number of basis images. The smallest errors is attributed to NMF algorithm in conjunction with the CMS metric. As far as SVMs are concerned, the best results were obtained by the linear kernel. Thus, we do not report
results obtained using polynomial and ERBF kernels. Figure 3 shows that NMF still outperforms both LNMF and DNMF, DNMF being the second best. Performing face recognition on ORL database lead to somehow opposite results. For this database, LNMF provided the lowest recognition error regardless of the classifier involved as can be seen from Figures 4, 5 and 6. NMF behaved the worst especially when it was associated with MCC classifier which is consistent with the results reported in [3]. DNMF algorithm is situated in the middle regarding its performance.

VI. DISCUSSIONS AND CONCLUSION

The potential of NMF, LNMF and DNMF algorithms in the face recognition task has been investigated in this paper. Features extracted by these techniques from two different databases were classified by using two metric-based classifiers (CSM, MCC) and SVMs. The experiments showed that CSM and SVM yield better recognition performance than MCC. Overall, as we expected, SVMs performed the best, followed by CSM and MCC. As far as the feature extraction approach is concerned, based on the results obtained by the algorithms involved, it seems that NMF is more robust to illumination changes than LNMF and DNMF since the variation of lighting condition for the faces pertaining to Yale database is much more intense than for images from the ORL database. Contrary to the results obtained for ORL, where LNMF gave the highest recognition rate, when face recognition is performed on the YALE database the best results are obtained by the NMF algorithm. Although for the ORL database, generally, the faces are in frontal position, this database also contains poses where the face is slightly rotated. This can contribute to the performance of LNMF since this algorithm is rotation invariant (up to some degree) since it generates local features in contrast to NMF which yields more distributed features. Despite the fact that DNMF approaches was successfully applied to recognize facial expression, in the case of YALE database we found its performance to be inferior to NMF and slightly inferior when compared to LNMF applied for ORL database.

ACKNOWLEDGMENT

This work has been conducted in conjunction with the "SIMILAR" European Network of Excellence on Multimodal Interfaces of the IST Programme of the European Union (www.similar.cc).

REFERENCES