Exploiting Discriminant Information in Non negative Matrix Factorization with application to Frontal Face Verification

Stefanos Zafeiriou[†], Anastasios Tefas[†] *Member, IEEE,* Ioan Buciu[†]

and Ioannis Pitas[†] *Senior Member, IEEE,*

[†]Aristotle University of Thessaloniki

Department of Informatics

Box 451

54124 Thessaloniki, Greece

Address for correspondence :

Professor Ioannis Pitas

Aristotle University of Thessaloniki

54006 Thessaloniki

GREECE

Tel. ++ 30 231 099 63 04

Fax ++ 30 231 099 63 04

email: pitas@zeus.csd.auth.gr

Abstract

In this paper, two supervised methods for enhancing the classification accuracy of the *Non-negative Matrix Factorization* (NMF) algorithm are presented. The idea is to extend the NMF algorithm in order to extract features that enforce not only the spatial locality, but also the separability between classes in a discriminant manner. The first method employs discriminant analysis in the features derived from NMF. In this way, a two phase discriminant feature extraction procedure is implemented, namely NMF plus *Linear Discriminant Analysis* (LDA). The second method incorporates the discriminant constraints inside the NMF decomposition. Thus, a decomposition of a face to its discriminant parts is obtained and new update rules for both the weights and the basis images are derived. The introduced methods have been applied to the problem of frontal face verification using the well known XM2VTS database. Both methods greatly enhance the performance of NMF for frontal face verification.

Index Terms

Subspace techniques, non-negative matrix factorization, linear discriminant analysis, frontal face verification.

I. Introduction

Face recognition/verification has attracted the attention of researchers for more than two decades and is among the most popular research areas in the field of computer vision and pattern recognition.

The two problems of face verification and recognition are conceptually different. A recognition system assists a human expert in determining the identity of a test face. In many cases, only the percentage of correctly identified faces within a number of matches is adequate (recognition rate) for evaluating the performance of a face recognition system [1]. By varying the number of matches, the curve of the cumulative match score versus the number of matches is obtained [2]. For details on some state-of-the-art face recognition systems, the interested reader can refer to [1], [3], [4], [5], [6], [7], [8], [9], [10]. A person verification system should decide whether an identity claim is valid or invalid. The performance of face verification systems is measured in terms of the *False Rejection Rate* (FRR) achieved at a fixed *False Acceptance Rate* (FAR). There is a trade-off between FAR and FRR. That is, it is possible to reduce either of them with the risk of increasing the other one. This trade-off between the FAR and FRR can create a curve, where FRR is plotted as a function of FAR. This curve is called *Receiver Operating Characteristic* (ROC) curve [11], [12]. The performance of a verification system is often quoted by a particular operating point of the ROC curve where FAR=FRR. This operating point is called *Equal Error Rate* (EER). Recently, frontal face verification competitions using the XM2VTS [13]- [15] database have been conducted. The interested reader can refer to [13], [14], and to the references therein for the tested face verification algorithms.

The most popular among the techniques used for frontal face recognition/verification are the subspace methods. The subspace algorithms consider the entire image as a feature vector and aim at finding projections (bases) that optimize a given criterion defined over the feature vectors that correspond to different classes. Then, the original high dimensional image space is projected into a low dimensional one. The classification is usually performed according to a simple distance measure at this low dimensional space.

Various criteria have been employed in order to find the bases of the low dimensional spaces. Some of them have been defined in order to find projections that best express the population without using the information about the way the data are separated to different classes, e.g. *Principal Component Analysis* (PCA) [16], NMF [17]). Another class of criteria is the one that deals directly with the discrimination between classes, e.g. LDA [18], [19]. Finally, statistical independence in the low dimensional feature space can be also used as a criterion in order to find the linear projections e.g. *Independent Component Analysis* (ICA)

[4], [20].

One of the oldest and well studied methods for low dimension face representation using criteria that aim at fair facial image representation is the *Eigenfaces* approach [16]. This representation was used in [21] for face recognition. The idea behind the Eigenfaces representation is to choose a linear transformation for dimensionality reduction that maximizes the scatter of all projected samples.

Another subspace method that aims at finding a face representation by using basis images without using class information is NMF [17]. The NMF approach was motivated by the biological aspect that the firing rates of neurons are non-negative. The NMF algorithm, like PCA, represents a facial image as a linear combination of basis images. The difference with PCA is that it does not allow negative elements either in the basis vectors or in the representation weights used in the linear combination of the basis images. This constraint results to radically different bases than PCA. On one hand, the bases of PCA are the Eigenfaces, some of which resemble distorted versions of the entire face. On the other hand the bases of NMF are localized features that correspond better to the intuitive notion of face parts [17]. NMF variants for object recognition have been proposed in [22], [23]. Various distance metrics suitable for the NMF representation space have been proposed in [24]. Methods for initializing the weights and the bases of the NMF decomposition have been proposed in [25]. Theoretical aspects regarding when NMF gives a unique decomposition of an object into its parts are provided in [26].

In [27], a technique for imposing additional constraints to the NMF minimization algorithm has been proposed. This technique, the so-called *Local Non-negative Matrix Factorization* (LNMF), is an extension of NMF and gives even more localized bases. It has been shown that LNMF leads to better classification performance in comparison to NMF and PCA [27]. In [28] the LNMF decomposition has been proposed for face detection. LNMF has also been found to give higher facial expression recognition rate than NMF [29]. To enhance the sparsity of NMF decomposition another approach has been proposed in [30] that is a combination of sparse coding and NMF.

In this paper, we develop a series of techniques for exploiting discriminant information in NMF. The first class of techniques use the NMF basis images in order to discover a low dimensional space and search for discriminant projections in this space. This is similar to Fisherfaces [18], [19], where an initial PCA based dimensionality reduction step is used, before applying LDA in this new space for finding discriminant projections. Of course the motivations of Fisherfaces and the proposed NMF plus LDA method are different. In Fisherfaces, first PCA is used in order to satisfy the invertibility of the within scatter matrix and afterwards LDA is used in this new space. In the proposed NMF plus LDA method, LDA is used along with NMF in order to investigate whether there is any discriminant information in part-based decompositions, like NMF.

The second class of techniques is motivated by LNMF where additional spatial-locality constraints have been considered in the minimization of the cost function of NMF. Instead of spatial locality constraints, we incorporate discriminant constraints inside the NMF decomposition. Here we propose two such techniques, both motivated by the fact that we want a part based decomposition with enhanced discriminant power. The first method gives basis images that are the same for all the different facial classes, while the latter results to a class specific decomposition that is unique for each facial (person) class. The intuitive motivation behind the class-specific methods is to find for every face a unique decomposition into its own discriminant parts. A similar technique has been used in [31], where discriminant constraints have been incorporated in the LNMF cost function. The approach in [31] has given better recognition accuracy than NMF and LNMF, when applied to facial expression recognition. These approaches are consistent with the image representation paradigms of neuroscience which involve sparseness, non-negative constraints, minimization of redundant information and enhanced discriminant power. All the introduced algorithms are applied to the frontal face verification problem.

The outline of this paper is as follows. The problem of frontal face verification and how subspace methods can be applied to this problem is discussed in Section II. The NMF decomposition is revisited in Section III. The NMF plus LDA method is described in Section IV. Methods for incorporating discriminant constraints inside NMF cost and the corresponding decompositions are introduced in Section V. Experimental results are depicted in Section VI. Finally, conclusions are drawn in Section VII.

II. Frontal Face Verification and Subspace Techniques

In this Section, we will briefly outline the problem of frontal face verification and the framework under which a subspace method can be used in order to solve this problem.

Let $\mathcal{U}$ be a facial image database. Each facial image $\mathbf{x} \in \mathcal{U}$ is supposed to belong to one of the K facial (person) classes $\{\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_K\}$ with $\mathcal{U} = \bigcup_{i=1}^{K} \mathcal{U}_i$. For a face verification system that uses the database $\mathcal{U}$, a genuine (or client) claim is performed when a person t provides its facial image $\mathbf{x}$, claims that $\mathbf{x} \in \mathcal{U}_r$ and $t = r$. When a person t provides its facial image $\mathbf{x}$ and claims that $\mathbf{x} \in \mathcal{U}_r$, with $t \neq r$, an impostor claim occurs. The scope of a face verification system is to handle properly these claims by accepting the genuine claims and rejecting the impostor ones.

Let the facial image database $\mathcal{U}$ be comprised by L facial images $\mathbf{x}_j \in \Re_+^F$, where $\Re_+ = [0, +\infty)$ and let the cardinality of each facial class $\mathcal{U}_r$ to be N_r. A linear subspace transformation of the original F-dimensional space onto a M-dimensional subspace (usually $M \ll F$) is a matrix $\mathbf{W} \in \Re^{M \times F}$ estimated using the database $\mathcal{U}$. The new feature vector

$\acute{\mathbf{x}} \in \Re^M$ is given by:

$$\acute{\mathbf{x}} = \mathbf{W}\mathbf{x}. \tag{1}$$

The rows of the matrix $\mathbf{W}$ contain the bases of the lower dimension feature space. The bases matrix $\mathbf{W}$ could be the same for all facial classes of the database or could be unique for each facial class. In case of class-specific image bases, for the reference person r, the set $\mathcal{I}_r = \mathcal{U} - \mathcal{U}_r$, that corresponds to impostor images is used in order to construct the two-class problem (genuine versus impostor class) [12], [32].

After the projection given by (1), a distance metric is chosen in order to measure the similarity of a test facial image to a certain class. This similarity measure can be the L_1 norm, the L_2 norm, the normalized correlation or the Mahalanobis distance [1]. In case of face verification, the algorithm should also learn a threshold on the similarity measure in order to accept or reject a client/impostor claim.

III. NMF Revisited

In this section, we will briefly describe the use of Bregman distances [33]-[35] and how NMF decomposition is obtained. Some notes on how NMF is extended in LNMF in order to give even more sparse basis images are also given.

A. Bregman Distance and Kullback-Leibler Divergence

Let $\phi : \mathcal{D} \rightarrow \Re$ be a continuously differentiable and strictly convex function defined on a closed, convex set $\mathcal{D} \subseteq \Re_+^F$. The Bregman distance associated with the function ϕ is defined for $\mathbf{x}, \mathbf{q} \in \mathcal{D}$ [36]:

$$B_\phi(\mathbf{x}||\mathbf{q}) \triangleq \phi(\mathbf{x}) - \phi(\mathbf{q}) - \nabla\phi(\mathbf{x})(\mathbf{x} - \mathbf{q}) \tag{2}$$

where $\nabla\phi(\mathbf{x})$ is the gradient of ϕ at $\mathbf{x}$. When $\phi(\mathbf{x})$ takes the form of the convex function:

$$\phi(\mathbf{x}) = \sum_i x_i \ln x_i \tag{3}$$

for $\mathbf{x} = [x_1 \dots x_F]^T$, then the Bregman distance is reformulated to Kullback-Leibler (KL) divergence (or relative entropy) between $\mathbf{x}$ and $\mathbf{q}$ [33]-[35] as:

$$KL(\mathbf{x}||\mathbf{q}) \triangleq \sum_i (x_i \ln \frac{x_i}{q_i} + q_i - x_i) \tag{4}$$

where $\mathbf{q} = [q_1 \dots q_F]^T$. It can be shown that, in general, every Bregman distance is non-negative and is equal to zero if and only if its two arguments are equal. More details about optimization algorithms using Bregman distances and KL divergence with linear constraints can be found in [35].

B. The NMF Algorithm

The basic idea behind NMF is to approximate the image $\mathbf{x}$ by a linear combination of the elements of $\mathbf{h} \in \Re_+^M$ such that $\mathbf{x} \approx \mathbf{Zh}$, where $\mathbf{Z} \in \Re_+^{F\times M}$ is a nonnegative matrix, whose columns sum to one. In order to measure the error of the approximation $\mathbf{x} \approx \mathbf{Zh}$ the $KL(\mathbf{x}||\mathbf{Zh})$ divergence can been used [34].

In order to apply NMF in the database $\mathcal{U}$, the matrix $\mathbf{X} \in \Re_+^{F\times L} = [x_{i,j}]$ should be constructed, where $x_{i,j}$ is the i-th element of the j-th image. In other words the j-th column of $\mathbf{X}$ is the $\mathbf{x}_j$ facial image. NMF aims at finding two matrices $\mathbf{Z} \in \Re_+^{F\times M} = [z_{i,k}]$ and $\mathbf{H} \in \Re_+^{M\times L} = [h_{k,j}]$ such that :

$$\mathbf{X} \approx \mathbf{ZH}. \tag{5}$$

The facial image $\mathbf{x}_j$ after the NMF decomposition can be written as $\mathbf{x}_j \approx \mathbf{Zh}_j$, where $\mathbf{h}_j$ is the j-th column of $\mathbf{H}$. Thus, the columns of the matrix $\mathbf{Z}$ can be considered as basis images and the vector $\mathbf{h}_j$ as the corresponding weight vector. The $\mathbf{h}_j$ vectors can also be considered as the projected vectors of a lower dimensional feature space for the original facial vector $\mathbf{x}_j$.

The defined cost for the decomposition (5) is the sum of all KL divergences for all images in the database. This way the following metric can be formed :

$$D_N(\mathbf{X}||\mathbf{ZH}) = \sum_j KL(\mathbf{x}_j||\mathbf{Zh}_j) = \sum_{i,j}(x_{i,j}\ln(\frac{x_{i,j}}{\sum_k z_{i,k}h_{k,j}}) + \sum_k z_{i,k}h_{k,j} - x_{i,j}) \tag{6}$$

as the measure of the cost for factoring $\mathbf{X}$ into $\mathbf{ZH}$ [34].

The NMF factorization is the outcome of the following optimization problem :

$$\min_{\mathbf{Z},\mathbf{H}} D_N(\mathbf{X}||\mathbf{ZH}) \text{ subject to} \tag{7}$$

$$z_{i,k} \geq 0,\ h_{k,j} \geq 0,\ \sum_i z_{i,j} = 1,\ \forall j.$$

NMF has non-negative constraints on both the elements of $\mathbf{Z}$ and of $\mathbf{H}$; these nonnegativity constraints permit the combination of multiple basis images in order to represent a face using only additions between the different bases. In contrast to PCA [16], [21], no subtractions can occur. For these reasons, the nonnegativity constraints correspond better to the intuitive notion of combining facial parts in order to create a complete face. Additional intuitive explanations why NMF is indeed a sparse part-based decomposition along with experimental verifications of this fact are given in [17], [22], [23], [25], [27], [29]

Recently some theoritical work has been done in order to show when the NMF does give a correct decomposition into parts [26]. Let some object that is comprised of A parts and each part can be in P different positions (in [26] the different positions are viewed as part's articulations). Then, if the images obey the following rules it can be proven that when NMF is applied to this database it can give a correct decomposition into parts[26]:

- Each image $\mathbf{x}$ in the database can be represented as a linear combination of the different parts in the different positions. Both parts and weights of the linear combination obey the nonnegativity constraint.
- The different bases are linear independent.
- The database contains all combinations of parts in the different positions. This constraint require from the database to have a total of A^P images.

Of course, these set of requirements is quite restrictive and cannot be satisfied in the case of facial image databases since it is not feasible to have all the possible images with combinations of different eyes, noses, mouths in different positions. Thus, when NMF is applied to a facial database it can only give an approximation of the decomposition into parts. In all cases NMF is indeed a sparse part-based decomposition [17], [22], [23], [25], [26], [27], [29].

By using an auxiliary function and the Expectation Maximization (EM) algorithm [34], the following update rules for $h_{k,j}$ and $z_{i,k}$ guarantee a non increasing behavior of (6). The update rule for the t-th iteration for $h_{k,j}$ is given by:

$$h_{k,j}^{(t)} = h_{k,j}^{(t-1)} \frac{\sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}{\sum_i z_{i,k}^{(t-1)}} \tag{8}$$

whereas, for the $z_{i,k}$, the update rules are given by:

$$\acute{z}_{i,k}^{(t)} = z_{i,k}^{(t-1)} \frac{\sum_j h_{k,j}^{(t)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t)}}}{\sum_j h_{k,j}^{(t)}} \tag{9}$$

and

$$z_{i,k}^{(t)} = \frac{\acute{z}_{i,k}^{(t)}}{\sum_l \acute{z}_{l,k}^{(t)}}. \tag{10}$$

Since $\mathbf{x}_j \approx \mathbf{Z}\mathbf{h}_j$, a natural way to compute the projection of $\mathbf{x}_j$ to a lower dimensional feature space using NMF is $\acute{\mathbf{x}}_j = \mathbf{Z}^\dagger \mathbf{x}_j$. The pseudo-inverse $\mathbf{Z}^\dagger = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ can be calculated using singular value decomposition methods [37]. In order to proceed to the dimensionality reduction, it has been also claimed that $\mathbf{Z}^T$ can be used as an alternative [31], due to the fact that the calculation of $\mathbf{Z}^\dagger$ may suffer from numerical instability. For a true non-negative dimensionality reduction the matrix $\mathbf{Z}^T$ should be used for feature extraction in the test images.

C. The LNMF algorithm

The idea of NMF decomposition was further extended to the LNMF [27] where additional constraints concerning the spatial locality of the bases were employed in the optimization problem defined in (7).

Let $\mathbf{U} = [u_{i,j}] = \mathbf{Z}^T\mathbf{Z}$, $\mathbf{V} = [v_{i,j}] = \mathbf{H}\mathbf{H}^T$, both being $M \times M$, LNMF aims at learning local features by imposing the following three additional locality constraints on the NMF. The first constraint is to create bases that cannot be further decomposed into more components [27]. Let $\mathbf{z}_j = [z_{1,j} \dots z_{n,j}]^T$ be the jth basis vector. Given the existing constraint that $\sum_i z_{i,j} = 1 \; \forall i$, we want that $\sum_i z_{i,j}^2$ to be as small as possible so that $\mathbf{z}_j$ contains as many zero elements as possible (make the bases as sparse as possible). This is accomplished by imposing $\sum_i u_{i,i}$ to be minimal [27].

Another constraint is to make the bases to be as orthogonal as possible, so as to minimize the redundancy between different bases. This can be imposed by requiring $\sum_{i \neq j} u_{i,j}$ to be minimal [27]. In other words we want the elements of matrix $\mathbf{U}$ that are not in the main diagonal to be as close to zero as possible. The elements of matrix that are not in the main diagonal correspond to the dot product between the different basis vectors and the closest the dot product is to zero the more orthogonal the basis vectors can be considered. The final constraint requires that only the components giving the most important information should be retained. This constraint, requires that $\sum_i v_{i,i}$ is maximized [27]. For additional details the interested reader may refer to [27].

Of course, these constraints do not guarantee that the decomposition will be either orthogonal or the most sparse that can be derived from the training facial database. The only thing that is guaranteed by imposing these heuristic constraints is that the derived decomposition will be more sparse and more orthogonal than the one obtained through NMF. When the above constraints are incorporated in (6), a new cost function is created as:

$$D_L(\mathbf{X}||\mathbf{ZH}) = D_N(\mathbf{X}||\mathbf{ZH}) + \alpha_1 \sum_i u_{i,i} + \alpha_2 \sum_{i \neq j} u_{i,j} - \beta \sum_i v_{i,i} \tag{11}$$

where $\alpha_1, \alpha_2, \beta > 0$ are constants. For simplicity in [27] it was set $\alpha_1 = \alpha_2 = \alpha$. A solution for the minimization of the cost given in (11) subject to the constraints imposed in NMF (7) (non-negative constraints for $z_{i,k}$ and $h_{k,j}$ and the constraint that the columns of the matrix $\mathbf{Z}$ should sum to one), can be found in [27]. In order to ensure that the cost function (11) is nonincreasing, while using a series of approximations in order to eliminate the constants α and β, the following update rule for $h_{k,j}$ is employed:

$$h_{k,j}^{(t)} = \sqrt{h_{k,j}^{(t-1)} \sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}. \tag{12}$$

The update rules for the $z_{i,k}$ are the same as the NMF decomposition and are given by (9) and (10).

IV. NMF PLUS LDA

The two previously presented methods do not use the information about how the various facial images are separated into different facial classes. The most straightforward way in order to exploit discriminant information in NMF is to try to discover discriminant projections for the facial image vectors after the projection to the image bases matrix $\mathbf{Z}^{\dagger}$ (or $\mathbf{Z}^T$). Let the matrix $\mathbf{X}$ that contains all the facial images of the database $\mathcal{U}$, be organized as follows. The j-th column of the database $\mathbf{X}$ is the ρ-th image of the r-th class. Thus, $j = \sum_{i=1}^{r-1} N_i + \rho$.

The vector $\mathbf{h}_j$ that correspond to the jth column of the matrix $\mathbf{H}$, is the coefficient vector for the ρth facial image of the rth class and will be denoted as $\boldsymbol{\eta}_\rho^{(r)} = [\eta_{\rho,1}^{(r)} \dots \eta_{\rho,M}^{(r)}]^T$. The mean vector of the vectors $\boldsymbol{\eta}_\rho^{(r)}$ for the class r is denoted as $\boldsymbol{\mu}^{(r)} = [\mu_1^{(r)} \dots \mu_M^{(r)}]^T$ and the mean of all classes as $\boldsymbol{\mu} = [\mu_1 \dots \mu_M]^T$. Then, the within scatter for the coefficient vectors $\mathbf{h}_j$ is defined as:

$$\mathbf{S}_w = \sum_{r=1}^{K} \sum_{\rho=1}^{N_r} (\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})(\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})^T \tag{13}$$

whereas the between scatter matrix is defined as:

$$\mathbf{S}_b = \sum_{r=1}^{K} N_r (\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})^T. \tag{14}$$

The matrix $\mathbf{S}_w$ defines the scatter of sample vector coefficients around their class mean. The dispersion of samples that belong to the same class around their corresponding mean should be as small as possible. A convenient metric for the dispersion of the samples is the trace of $\mathbf{S}_w$. The matrix, $\mathbf{S}_b$ denotes the between-class scatter matrix and defines the scatter of the mean vectors of all classes around the global mean $\boldsymbol{\mu}$. Each class formed by the samples that belong to the same class must be as far as possible from the other classes. Therefore, the trace of $\mathbf{S}_b$ should be as large as possible. By taking into consideration the previous remarks, the well known Fisher discriminant criterion is constructed as:

$$J(\mathbf{\Psi}) = \frac{\text{tr}[\mathbf{\Psi}^T \mathbf{S}_b \mathbf{\Psi}]}{\text{tr}[\mathbf{\Psi}^T \mathbf{S}_w \mathbf{\Psi}]} \tag{15}$$

where tr$[\mathbf{R}]$ is the trace of the matrix $\mathbf{R}$. The maximization of J yields a set of discriminant projections that is given by the columns of the matrix $\mathbf{\Psi}_{opt}$. If $\mathbf{S}_w$ is invertible then the projection matrix $\mathbf{\Psi}_{opt}$ is given by the generalized eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

There is not upper limit for how many bases someone can construct using NMF decomposition in (9) and unless we create a limited number of bases by NMF the matrix $\mathbf{S}_w$ is singular. That is, there always exist vectors $\boldsymbol{\phi}_i$ that satisfy $\boldsymbol{\phi}_i^T \mathbf{S}_w \boldsymbol{\phi}_i = 0$. These vectors turn out to be very effective if they satisfy $\boldsymbol{\phi}_i^T \mathbf{S}_b \boldsymbol{\phi}_i > 0$ at the same time [3], [6], [38]. In that case the Fisher discriminant criterion degenerates into the following between-class scatter criterion:

$$J_b(\mathbf{\Phi}) = \text{tr}[\mathbf{\Phi}^T \mathbf{S}_b \mathbf{\Phi}] \ (\mathbf{\Phi} = [\ldots \boldsymbol{\phi}_i \ldots], \ ||\boldsymbol{\phi}_i|| = 1). \tag{16}$$

We will use the main results of [6] in order to extract discriminant features using an arbitrary number of NMF bases. The discriminant features are then extracted by the minimization of the criterions (15) and (16). The discriminant projections that are derived from the (15) will be called *regular discriminant projections* (or *regular NMFfaces*) while the ones created by (16) will be called *irregular discriminant projections* (or *irregular NMFfaces*).

Let the total scatter matrix of the feature vectors $\mathbf{h}_j$ be defined as:

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b \tag{17}$$

it is easy to prove that the matrix $\mathbf{S}_t$ is a compact and self-adjoint operator in $\Re^M$ [6]. Thus, its eigenvector system forms an orthonormal bases for $\Re^M$ [6].

Let $\mathcal{O}$ and $\mathcal{O}^\perp$ be the two complementary spaces spanned by the orthonormal eigenvectors that correspond to no-zero and to zero eigenvalues of $\mathbf{S}_t$, respectively. It is easy to prove, using the theory developed in [6], that $\mathcal{O}^\perp$ does not contain any discriminant information in respect to the criterion (15) and (16). The isomorphic mapping in order to move from the feature space of the vectors $\mathbf{h}_j$ to $\mathcal{O}$ is the matrix $\mathbf{\Pi}$ whose columns are the orthonormal eigenvectors of $\mathbf{S}_t$ that correspond to its non-zero eigenvalues. In order to find the non-zero eigenvectors of $\mathbf{S}_t$ efficiently, we can use algorithms like [21].

Let $\check{\mathbf{S}}_w$ and $\check{\mathbf{S}}_b$ be the within scatter and the between scatter matrices in the space $\mathcal{O}$. These matrices are given by $\check{\mathbf{S}}_w = \mathbf{\Pi}^T\mathbf{S}_w\mathbf{\Pi}$ and by $\check{\mathbf{S}}_b = \mathbf{\Pi}^T\mathbf{S}_b\mathbf{\Pi}$. In the space $\mathcal{O}$ the matrix $\check{\mathbf{S}}_w$ is still singular. Let $\mathbf{\Xi}_1$ and $\mathbf{\Xi}_2$ be the orthonormal eigenvectors that correspond to non-zero and to zero eigenvectors of the matrix $\check{\mathbf{S}}_w$, respectively.

In the space spanned by the vectors contained in $\mathbf{\Xi}_1$ the discriminant projections are given by the columns of the matrix $\mathbf{\Theta}_1$ that are the eigenvectors of $\tilde{\mathbf{S}}_w^{-1}\tilde{\mathbf{S}}_b$, where $\tilde{\mathbf{S}}_w = \mathbf{\Xi}_1^T\check{\mathbf{S}}_w\mathbf{\Xi}_1$ and $\tilde{\mathbf{S}}_b = \mathbf{\Xi}_1^T\check{\mathbf{S}}_b\mathbf{\Xi}_1$. In the space that is spanned by the columns of $\mathbf{\Xi}_2$ it can be easily proven that $\hat{\mathbf{S}}_b = \mathbf{\Xi}_2^T\check{\mathbf{S}}_b\mathbf{\Xi}_2$ is not singular [6]. Thus, the discriminant projections in this space are given by the matrix $\mathbf{\Theta}_2$ that has as columns the orthonormal eigenvectors of $\hat{\mathbf{S}}_b$.

The linear transform that extracts the regular discriminant features (will be called *regular NMFfaces* in the rest of the paper) using NMF is:

$$\mathbf{\Phi}_1 = \mathbf{\Theta}_1^T\mathbf{\Xi}_1^T\mathbf{\Pi}^T\mathbf{Z}^\dagger, \tag{18}$$

whereas, the linear transform that extracts the irregular discriminant features (will be called *irregular NMFfaces* in the rest of the paper) using NMF is:

$$\mathbf{\Phi}_2 = \mathbf{\Theta}_2^T \mathbf{\Xi}_2^T \mathbf{\Pi}^T \mathbf{Z}^{\dagger} \tag{19}$$

where $\mathbf{Z}$ is the decomposition of NMF given by (9). The total number of discriminant projections derived from this procedure is $2(K-1)$.

V. LDA INCORPORATED INSIDE NMF

In this subsection, we introduce alternatives to NMF plus LDA by incorporating discriminant constraints inside the cost function to be minimized for obtaining the new decompositions. Two different discriminant decompositions are proposed. These decompositions are motivated by the need of finding basis images that correspond to discriminant parts of faces. The first is the same for all facial classes in the database. The second one uses alternative discriminant constraints and gives a decomposition that is different for every facial class. This class-specific decomposition is intuitively motivated by the theory that humans memorize different discriminant features (e.g. noses, eyes) for different faces and use these features for recognizing them or verifying the identity of a face [39], [40]. The interested reader may refer to [39], [40] and to references within for different theories and technologies for human and machine recognition of faces.

A. The DNMF Algorithm

In order to incorporate discriminant constraints into the NMF decomposition we substitute the locality constraints of LNMF with discriminant constraints. This way, a modified divergence can be constructed that is derived from the minimization of the Fisher criterion. This is done by requiring tr$[\mathbf{S}_w]$ to be as small as possible while tr$[\mathbf{S}_b]$ is required to be as large as possible. The new cost function is given by:

$$D_d(\mathbf{X}||\mathbf{Z}_D\mathbf{H}) = D_N(\mathbf{X}||\mathbf{Z}_D\mathbf{H}) + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b]. \tag{20}$$

where γ and δ are constants. Following the same EM approach used by NMF [34] and LNMF [27] techniques, we come up with the following update rules for the weight coefficients $h_{k,j}$

that belongs to the r-th facial class:

$$h_{k,j}^{(t)} = \frac{T_1 + \sqrt{T_1^2 + 4(2\gamma - (2\gamma + 2\delta)\frac{1}{N_r})h_{k,j}^{(t-1)} \sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}}{2(2\gamma - (2\gamma + 2\delta)\frac{1}{N_r})}. \quad (21)$$

The detailed derivation of (21) along with the definition of T_1 are given in Appendix I. The update rules for the bases $\mathbf{Z}_D$ are the same as in NMF and can be given by (9) and (10). The above decomposition is a supervised non-negative matrix factorization method that decomposes the facial images into parts while, enhancing the class separability. This method will be called *Discriminant Non-negative Matrix Factorization* (DNMF) in the rest of the paper. The matrix $\mathbf{Z}_D^{\dagger} = (\mathbf{Z}_D^T \mathbf{Z}_D)^{-1}\mathbf{Z}_D^T$, which is the pseudo-inverse of $\mathbf{Z}_D$, is then used for extracting the discriminant features as $\acute{\mathbf{x}} = \mathbf{Z}_D^{\dagger}\mathbf{x}$. The $\mathbf{Z}_D^T$ can be used instead of $\mathbf{Z}_D^{\dagger}$ for a true non-negative dimensionality reduction. It is interesting to notice here that there is no restriction on how many dimensions we may keep for $\acute{\mathbf{x}}$ and that the bases of the DNMF are common for all the different facial classes in the database.

B. The CSDNMF Algorithm

In this subsection alternative discriminant constraints are integrated inside the cost function (6). The minimization procedure of the new cost function yields a *Class-Specific Discriminant Non-negative Matrix Factorization* (CSDNMF) method. In order to formulate the CSDNMF decomposition, the facial image vectors of the genuine claims to the reference person r are in the first $N_G = N_r$ columns of the matrix $\mathbf{X}$. Then, the columns from $N_r + 1$ to L correspond to impostor claims. The total number of impostor claims is $N_I = L - N_r$. The coefficient vector $\mathbf{h}_j$ of the image $\mathbf{x}_j$ that corresponds to the ρth image of the genuine class will be denoted as $\boldsymbol{\eta}_{\rho}^{(G)}$. If the facial vector $\mathbf{x}_j$ is the ρth image of the impostor class then the corresponding coefficient vector $\mathbf{h}_j$ will be denoted as $\boldsymbol{\eta}_{\rho}^{(I)}$.

In the previous section, we have seen that the tr$[\mathbf{S}_w]$ should be small whereas tr$[\mathbf{S}_b]$ should be large for the vectors $\mathbf{h}_j$.

In this section we replace these constraints with others that suite better with the face verification problem. Let a distance metric (e.g. the L_2 norm) be used in order to quantify the similarity of a test facial image vector $\mathbf{x}_j$ to a given facial class. It sounds reasonable to require that the feature vectors corresponding to the genuine class, should have great similarity with the mean image of the genuine class (small distance metric value with the mean image of the genuine facial class), while the feature vectors of the impostor class should have small similarity with the mean image of the reference facial class (large distance metric value with the mean image of the genuine facial class).

In order to define the similarity of the projection $\mathbf{h}_j$ of the facial image $\mathbf{x}_j$ to a given class r in the feature space of the coefficients, the L_2 norm can be used as:

$$d_r(\mathbf{h}_j) = ||\mathbf{h}_j - \boldsymbol{\mu}^{(G)}||^2 \tag{22}$$

where $\boldsymbol{\mu}^{(G)}$ is the mean vector of the vectors $\boldsymbol{\eta}_\rho^{(G)}$. The use of other similarity measures like L_1 or the normalized correlation has not given a closed form for the update rules. However, the experimental results using these measures were similar. In the reduced feature space of the vectors $\mathbf{h}_j$ we demand that the similarity measures $d_r(\boldsymbol{\eta}_\rho^{(I)})$ (impostor similarity measures) to be maximized while minimizing the similarity measures $d_r(\boldsymbol{\eta}_\rho^{(G)})$ (genuine similarity measures). Then the optimization problem for the class r is the maximization of:

$$\frac{1}{N_I}\sum_{\mathbf{x}_j \in \mathcal{I}_r} d_r(\mathbf{h}_j) = \frac{1}{N_I}\sum_{\rho=1}^{N_I} ||\boldsymbol{\eta}_\rho^{(I)} - \boldsymbol{\mu}^{(G)}||^2 = \mathrm{tr}[\mathbf{W}_r], \tag{23}$$

where $\mathbf{W}_r = \frac{1}{N_I}\sum_{\rho=1}^{N_I}(\boldsymbol{\eta}_\rho^{(I)} - \boldsymbol{\mu}^{(G)})(\boldsymbol{\eta}_\rho^{(I)} - \boldsymbol{\mu}^{(G)})^T$. The second optimization problem is the minimization of:

$$\frac{1}{N_G}\sum_{\mathbf{x}_j \in \mathcal{U}_r} d_r(\mathbf{h}_j) = \frac{1}{N_G}\sum_{\rho=1}^{N_G} ||\boldsymbol{\eta}_\rho^{(G)} - \boldsymbol{\mu}^{(G)}||^2 = \mathrm{tr}[\mathbf{B}_r], \tag{24}$$

where $\mathbf{B}_r = \frac{1}{N_G}\sum_{\rho=1}^{N_G}(\boldsymbol{\eta}_\rho^{(G)} - \boldsymbol{\mu}^{(G)})(\boldsymbol{\eta}_\rho^{(G)} - \boldsymbol{\mu}^{(G)})^T$.

We impose these two additional constraints in the cost function given in (7) as:

$$D_c(\mathbf{X}||\mathbf{Z}_r\mathbf{H}_r) = D_N(\mathbf{X}||\mathbf{Z}_r\mathbf{H}_r) + \zeta\text{tr}[\mathbf{B}_r] - \theta\text{tr}[\mathbf{W}_r]. \tag{25}$$

where $\zeta, \theta > 0$ are constants. The decomposition is person specific (different bases $\mathbf{Z}_r$ for each reference face class r). For $j = 1, \ldots, N_G$ (genuine class), the update rule for the coefficients $h_{k,j}$ of the reference person r is given by:

$$h_{k,j}^{(t)} = \frac{T_2 + \sqrt{T_2^2 + 4\frac{1}{N_G}(2\zeta - (2\zeta + 2\theta)\frac{1}{N_G})h_{k,j}^{(t-1)} \sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}}{2\frac{1}{N_G}(2\zeta - (2\zeta + 2\theta)\frac{1}{N_G})} \tag{26}$$

whereas the update rule for the weight coefficients of the impostor class ($j = N_G + 1, \ldots, L$) is given by:

$$h_{k,j}^{(t)} = \frac{T_3 + \sqrt{T_3^2 - 8N_I\theta h_{k,j}^{(t-1)} \sum_i z_{i,k}^{(t-1)} \frac{x_{i,j}}{\sum_l z_{i,l}^{(t-1)} h_{l,j}^{(t-1)}}}}{4\theta} \tag{27}$$

where T_2 and T_3 are given in Appendix II. The update rules for the bases matrix $\mathbf{Z}_r$ for the reference person r are the same as in NMF decomposition and can be given by (9) and (10). When someone claims that a test image $\mathbf{x}$ corresponds to a reference facial class r, then $\mathbf{x}$ is projected using the pseudo-inverse of $\mathbf{Z}_r$, $\mathbf{Z}_r^{\dagger}$, matrix as $\acute{\mathbf{x}} = \mathbf{Z}_r^{\dagger}\mathbf{x}$. In the same manner as NMF and DNMF the matrix $\mathbf{Z}_r^T$ can be used for a true non-negative dimensionality reduction.

VI. Experimental Results

A. Database Description

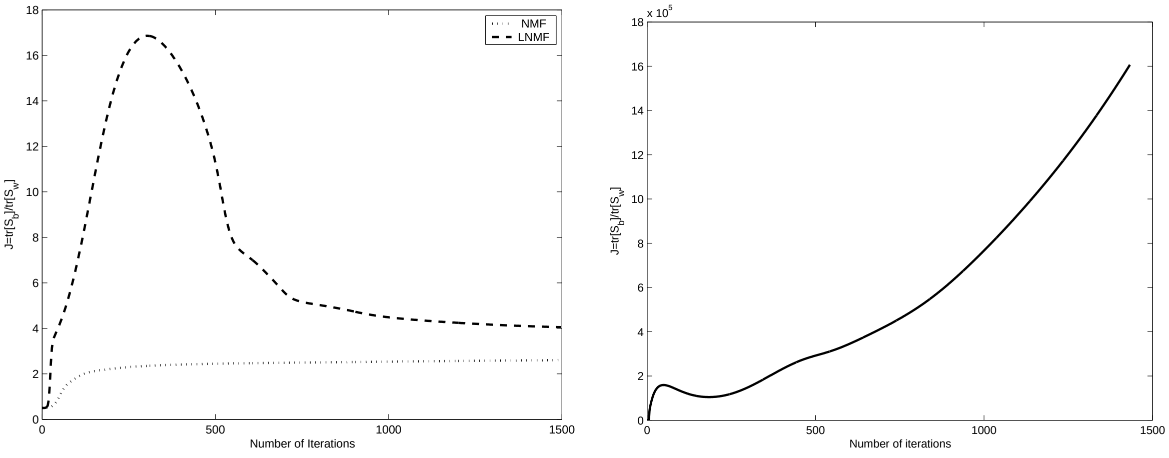
The experiments were conducted in the XM2VTS database using the protocol described in [15]. The images were aligned semi-automatically according to the eyes position of each facial image using the eye coordinates. The facial images were down-scaled to a resolution of 64×64 pixels. Histogram equalization was used for normalizing the facial image luminance.

The XM2VTS database contains 295 subjects, 4 recording sessions and two shots (repetitions) per recording session. The XM2VTS database provides two experimental setups namely, Configuration I and Configuration II [15]. Each configuration is divided into three different sets: the training set, the evaluation set and the test set. The training set is used to create client and impostor models for each person. The evaluation set is used to learn the verification decision thresholds. In case of multimodal systems, the evaluation set is also used to train the fusion manager [15]. For both configurations the training set has 200 clients, 25 evaluation impostors and 70 test impostors. The two configurations differ in the distribution of client training and client evaluation data. For additional details concerning XM2VTS database the interested reader can refer to [15].

B. Training Procedure

In the training phase, the basis images corresponding to the NMF (Section III-B), the LNMF (Section III-C), the proposed DNMF (Section V-A), the proposed CSDNMF (Section V-B), the Eigenfaces, the Fisherfaces and the proposed NMFfaces (regular and irregular discriminant bases of NMF plus LDA method proposed in Section IV) are found. For all the approaches except from CSDNMF the bases are common for all facial classes. In the case of CSDNMF, the training set is used for calculating for each reference person r a different set of bases for feature selection. A convenient way for having an insight of the class separability is to compute the quantity $J = \text{tr}[\mathbf{S}_b]/\text{tr}[\mathbf{S}_w]$ in the training set [41]. In Figure 1, J is plotted versus the number of iterations used in the decomposition. Note that there is a significant scale difference in the y-axis of Figures 1a and 1b. This indicates a much better class separability in case of DNMF compared to the ones obtained either by NMF or by LNMF (class separability is measured in respect to J).

By imposing only non-negativity constraints, the features extracted by NMF have a rather holistic appearance. This can be seen in Figure 2(a). LNMF greatly improves the bases image sparseness and minimizes redundant information by imposing locality constraints. The proposed DNMF and CSDNMF also minimize the redundant information while maximizing



(a) (b)

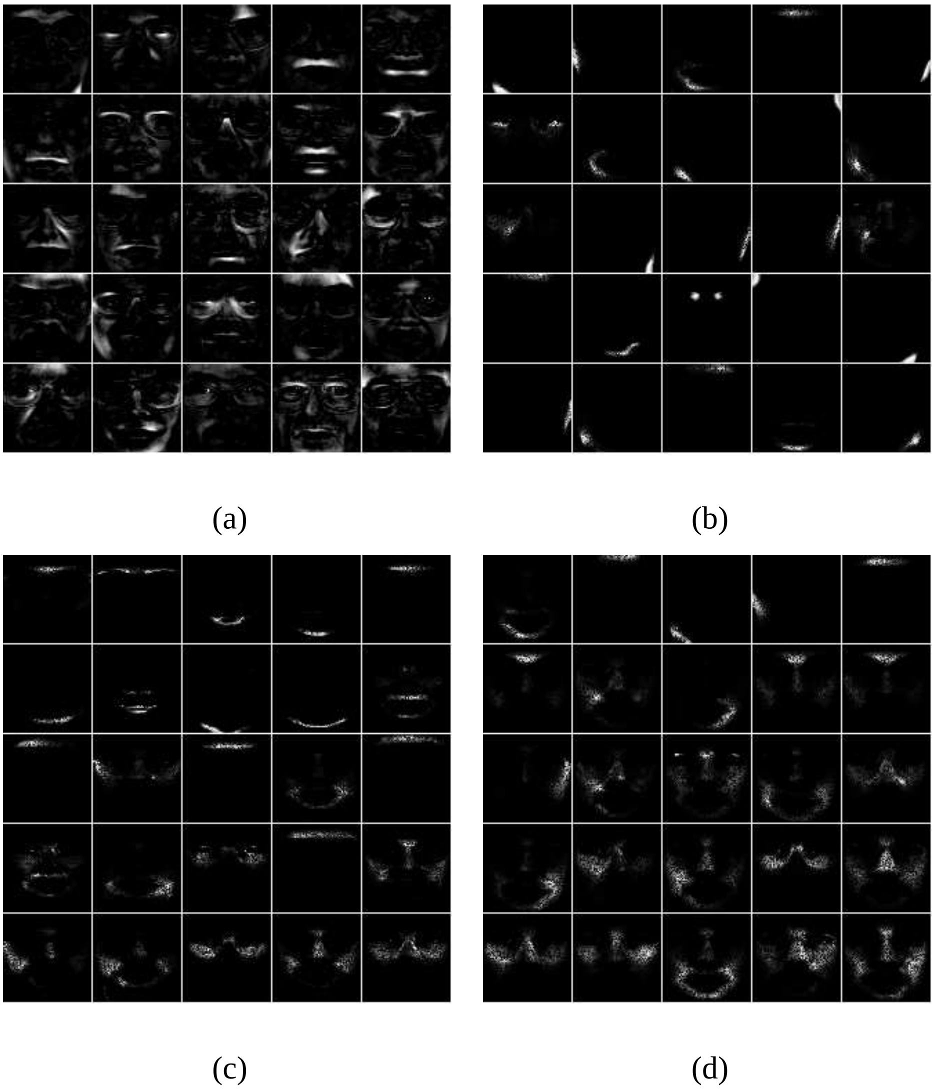
Fig. 1. (a) J plotted versus the number of iterations for the NMF and LNMF (b) J plotted versus the number of iterations for DNMF.

class separability (the class separability is measured in respect to J). To quantify the degree of sparseness of basis images, someone can measure the normalized kurtosis of a base image $\mathbf{z}$ defined as [20]:

$$\kappa(\mathbf{z}) = \frac{\sum_i (z_i - \bar{z})^4}{(\sum_i (z_i - \bar{z})^2)^2} - 3. \tag{28}$$

where $\mathbf{z} = [z_1 \dots z_F]^T$ and $\bar{z} = \frac{1}{F}\sum_{i=1}^{F} z_i$. The largest the number of kurtosis the sparsest an image is. It was experimentally found that the average kurtosis over the maximum number of 199 basis images are: $\bar{k}_{NMF} = 8.12$, $\bar{k}_{LNMF} = 160.58$, $\bar{k}_{DNMF} = 26.88$ and $\bar{k}_{CSDNMF} = 33.88$.

For comparison a number of 25 images for the NMF, the LNMF, the proposed DNMF and the CSDNMF are given in Figure 2. In Figures 2a and 2b the images are ordered row-wise according to their descending degree of sparseness, calculated according to (28). Obviously DNMF and CSDNMF is a compromise between NMF and LNMF in terms of sparseness. Probably, the most important issue concerning the DNMF and the CSDNFM algorithm, that has been experimental verified, is the fact that almost all features found by its basis images are represented by the salient face features, such as eyes, eyebrows or mouth. As can be seen the features retrieved by LNMF have random positions that can not be directly attributed to facial features.



(a) (b)

(c) (d)

Fig. 2. A set of 25 basis images for (a) NMF, (b) LNMF (c) DNMF (d) CSDNMF.

By a visual inspection of the images of Figure 3, it can be seen that Eigenfaces, Fisherfaces and regular NMFfaces (it also holds for the irregular) resemble degraded versions of faces. The basis images in Figure 3a-3c are sorted in descending order of their corresponding eigenvalue.

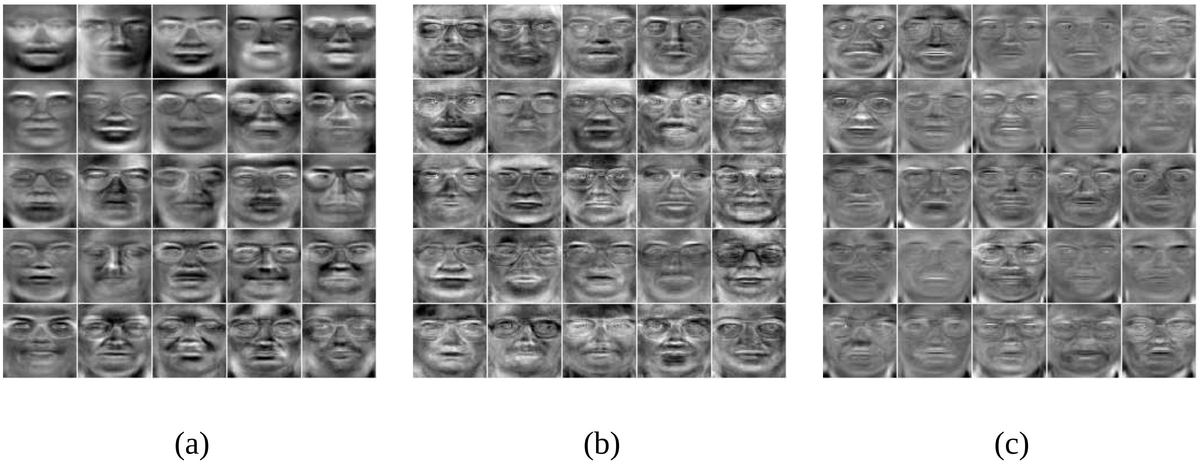


Fig. 3. A set of 25 basis images for (a) EigenFaces, (b) FisherFaces (c) the regular NMFfaces.

The parameters γ and δ in the DNFM cost (20) and the parameters ζ and η in (25) should be carefully selected. Due to the fact that the cost function defined by the proposed DNMF and CSDNMF is formed by several terms that are simultaneously optimized (minimized of maximized), its global optimization suffers. Although the cost functions (20) and (25) are globally minimized, each term has each own rate of convergence. The parameters γ and δ govern the convergence speed for minimizing tr$[\mathbf{S}_w]$ and maximizing tr$[\mathbf{S}_b]$, while the parameters ζ and θ govern the convergence speed for tr$[\mathbf{W}_r]$ and tr$[\mathbf{B}_r]$. An automated way of choosing the parameters γ and δ for the proposed DNMF and ζ and η for the proposed CSDNMF is to use an adaptive formulation for them rather than a fixed one. Starting with small parameter values, the algorithm proceeds while, at each iteration step, the degree of sparseness is checked using the kurtosis and the algorithm restarts with new parameter values. This is repeated till the kurtosis exceeds a certain threshold (we have chosen as a threshold the average kurtosis to be greater than 20).

In our experiments we have tested values for γ and δ in the range $[0, 1]$ (this also holds for the case ζ and η). We have seen that very small values of these constants speed up the decrease of tr$[\mathbf{S}_w]$, the increase of tr$[\mathbf{S}_b]$ and the minimization of $D_d(\mathbf{X}||\mathbf{ZH})$. However, the algorithm may stop too early and the number of iterations might not be sufficient to reach a local minimum for $D_d(\mathbf{X}||\mathbf{ZH})$. A premature stop can affect the process of correctly learning the basis images that might not be sparse anymore. The best results have been obtained when choosing values in the range $[0.1, 0.5]$.

C. Experimental Results in Configuration I

The training set of the Configuration I contains 200 persons with 3 images per person. The evaluation set contains 3 images per client for genuine claims and 25 evaluation impostors with 8 images per impostor. Thus, evaluation set gives a total of $3 \times 200 = 600$ client claims and $25 \times 8 \times 200 = 40.000$ impostor claims. The test set has 2 images per client and 70 impostors with 8 images per impostor and gives $2 \times 200 = 400$ client claims and $70 \times 8 \times 200 = 112.000$ impostor claims. The maximum number of Eigenfaces [21] given by the training set is 599. The number of classes is 200 and, thus, the number of Fisherfaces [19] is 199. For NMF plus LDA, 1000 basis images have been created initially using NMF and after the regular and irregular discriminant information has been found according to (18) and (19) that gives a total of 398 projections (199 regular NMFfaces and 199 irregular NMFfaces). For NMF, LNMF, DNMF and CSDNMF, 199 bases have been also considered for comparison.

The facial images have been then projected using these bases into a low dimensional feature space and the normalized correlation was used in order to define the similarity measure between two faces as:

$$D(\mathbf{x}_r, \mathbf{x}_t) = \frac{\acute{\mathbf{x}}_r^T \acute{\mathbf{x}}_t}{||\acute{\mathbf{x}}_r||||\acute{\mathbf{x}}_t||} \tag{29}$$

where $\mathbf{x}_r$ and $\mathbf{x}_t$ are the reference and the test facial image, respectively while $\acute{\mathbf{x}}_r$ and $\acute{\mathbf{x}}_t$ are their projections to one of the subspace. Of course other similarity metrics are suitable like L_1,L_2 or the Mahalanobis distance [1] but in the specific database the normalized correlation or (the cosine distance) has given the best results for all the tested methods. For completeness experiments using the L_2 norm are presented for the CSDNMF method since the L_2 norm has been used for formulating the CSDNMF decomposition (25).

In case of NMF plus LDA two different discriminant projection are found by (18) and (19). Thus, two different similarity values are created by $D_g(\mathbf{x}_r, \mathbf{x}_t) = \frac{(\mathbf{\Phi}_1 \mathbf{x}_r)^T (\mathbf{\Phi}_1 \mathbf{x}_t)}{||\mathbf{\Phi}_1 \mathbf{x}_r|| ||\mathbf{\Phi}_1 \mathbf{x}_t||}$ and by $D_u(\mathbf{x}_r, \mathbf{x}_t) = \frac{(\mathbf{\Phi}_2 \mathbf{x}_r)^T (\mathbf{\Phi}_2 \mathbf{x}_t)}{||\mathbf{\Phi}_2 \mathbf{x}_r|| ||\mathbf{\Phi}_2 \mathbf{x}_t||}$ for the regular and the irregular discriminant information, respectively. In [6] it has been proposed to use a simple fusion technique by weighting the irregular score with some empirical coefficient. Instead of using the empirical parameter we used the evaluation set of the Configuration I in order to learn a discriminant weighting vector $\mathbf{w}$ using also LDA. The final similarity measure between the facial image vectors $\mathbf{x}_r$ and $\mathbf{x}_t$ is given by:

$$D_t(\mathbf{x}_r, \mathbf{x}_r) = \mathbf{w}^T [D_g(\mathbf{x}_r, \mathbf{x}_t) \ D_u(\mathbf{x}_r, \mathbf{x}_t)]^T. \tag{30}$$

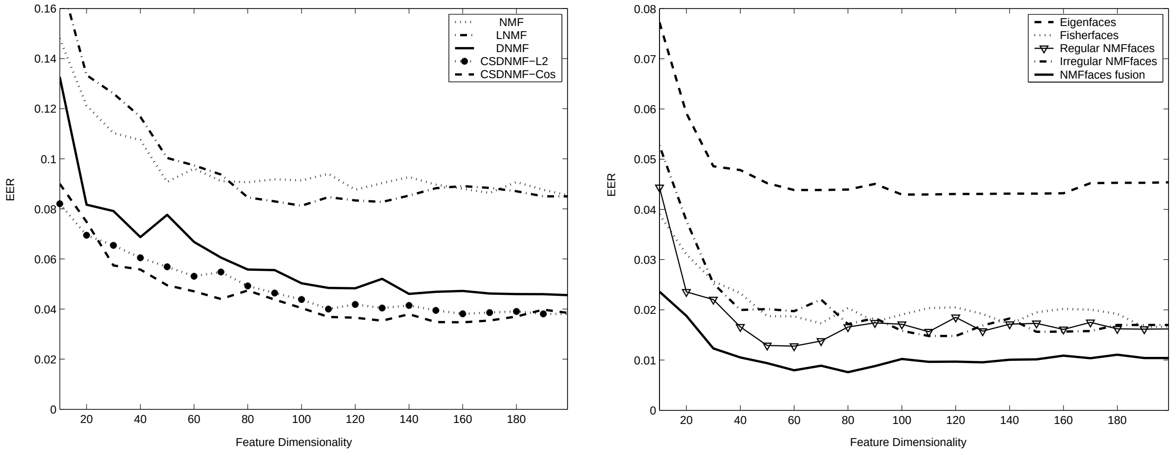
The similarity measures for each person, calculated in both evaluation and training set form the distance vector $\mathbf{d}(r)$. The elements of the vector $\mathbf{d}(r)$ are sorted in descending order and are used for the person specific thresholds on the distance measure. Let $T_Q(r)$ denote the Q-th order statistic of the vector of distances, $\mathbf{d}(r)$ (the Q-th smallest distance in the vector). The threshold of the person r is chosen to be equal to $T_Q(r)$. Let $\mathbf{x}_r^1$, $\mathbf{x}_r^2$ and $\mathbf{x}_r^3$ be the 3 instances of the person r in the training set. A claim of a person (with a facial image $\mathbf{x}_t$) to the identity r is considered valid if $\max_j \{D(\mathbf{x}_r^j, \mathbf{x}_t)\} < T_Q(r)$. Obviously when varying Q, different pairs of FAR and FRR can be created and that way a ROC curve is produced and the EER can be measured [15].

The performance of the methods that project to face-part like bases as NMF, LNMF, the proposed DNMF and CSDNMF algorithms for various feature dimensions is illustrated in Figure 4a. The best EER achieved for the proposed CSDNMF is 3.4% and 3.7% when the normalized correlation (cosine) and the L_2 norm has been used, respectively, while keeping more than 110 dimensions. The best performance of the proposed DNMF is 4.61%. The best EER for NMF and LNMF is more than 8%. That is, a decrease of more than 4% in terms of EER has been achieved by incorporating the proposed discriminant constraints in the cost of NMF. Even though NMF, LNMF, DNMF and CSDNMF are optimization methods that depend on the initialization of the bases and may get trapped to local minima we have not verified large deviations in verification performance when starting with different initial values (the standard deviation for the best performance after 10 restarts was about 0.2% in terms of EER). An alternative to random initialization is a structured initialization that has proposed in [25].

The performance of the methods that project to face bases like Eigenfaces, Fisherfaces and NMFfaces (regular and irregular) for various feature dimensions is illustrated in Figure 4b. The best EER achieved was 0.8% when 80 regular and 80 irregular projections have been kept. The best EER for Fisherfaces has been 1.6% and for Eigenfaces 4.3%. Unfortunately, the EER of the tested methods does not decrease monotonically with the number of image bases kept. This fact has been verified in other face recognition subspace methods like [1], [3], [4], [6] where the performance does not always increase with the number of the kept dimensions.

Therefore, the proposed NMFfaces scheme has the best verification performance. Unfortunately the decompositions like the proposed DNMF and the proposed CSDNMF have worst performance in comparison to the proposed NMFfaces and Fisherfaces. We have experimentally found that the training set contains limited discriminant information for the DNFM and CSDNFM methods (only 3 images per facial class) to be trained properly. We have also found that when adding in the training set images extracted from video of the same session (about 60-100 images per person) of the training set of the Configuration I (which is different from the session the test images have been extracted) a decrease of about

$2 - 2.5\%$ in the terms of EER has been verified. We have also used these images for training NMFfaces and Fisherfaces and no significant improvement in performance has been verified (about $0.2 - 0.3\%$ in terms of EER).



(a)

(b)

Fig. 4. EER for Configuration I: a) EER plotted versus feature dimensionality for part-based decompositions as LNMF, NMF, the proposed CSDNMF (using cosine and L_2 metrics) and the proposed DNMF; b) EER plotted versus feature dimensionality for Eigenfaces, FisherFaces and the proposed NMFfaces (regular, irregular and fusion).

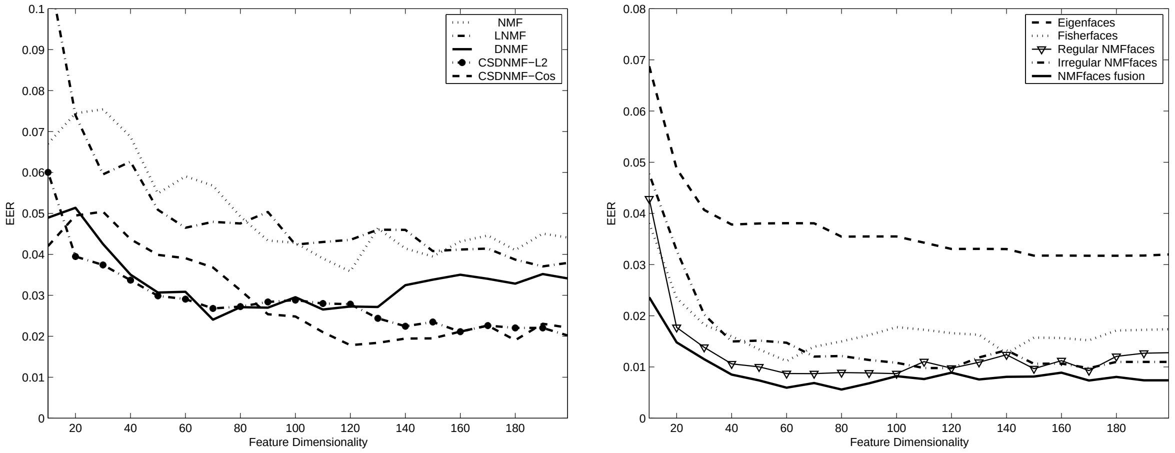
D. Experimental Results in Configuration II

The Configuration II differs from the Configuration I in the distribution of client training and client evaluation data. The training set of the Configuration I contains 200 persons with 4 images per person. The evaluation set contains 2 images per client for genuine claims. Thus, the evaluation set gives a total of $2 \times 200 = 400$ genuine claims. The training set contains 4 references images for each client. The same approach as in Configuration I has been used for accepting a claim as valid and for threshold calculation.

Figure 5a depicts the plot of the EER versus the dimensionality of the feature vectors for face-part like bases. As can be seen, CSDNMF have the best performance in comparison to the

NMF, LNMF and DNMF. The minimum EER achieved when projecting to CSDNMF bases has been equal to 1.8% and 2.2% when the cosine and L_2 norm has been used, respectively. For DNFM the minimum EER has been measured about 2.6%, while for NMF and LNMF the EER has been found equal to 3.7%.

Figure 5b depicts the plot of the EER versus the dimensionality of the feature vectors. As can be seen, the fusion of the two different NMFfaces (regular and irregular) have the best performance and the minimum achieved EER has been 0.6$\%$ when keeping 80 dimensions. For the Fisherfaces the best EER has been 1.2%, while for the Eigenfaces has been 3.1%.



(a) (b)

Fig. 5. EER for Configuration II a) EER plotted versus feature dimensionality for part-based decompositions as LNMF, NMF, CSDNMF (using cosine and L_2 metrics) and the DNMF; b) EER plotted versus feature dimensionality for Eigenfaces, FisherFaces and the proposed NMFfaces (regular, irregular and fussion).

VII. Conclusions

A series of novel techniques for supervised facial feature extraction has been developed. The new techniques are based on the NMF decomposition that find basis images which are intuitively related to face parts. The first discriminant technique gives basis images that are holistic and is comprised of two different phases, namely NMF and LDA thus producing the so-called NMFfaces. The other class of techniques aim at finding face decompositions in discriminant parts by integrating discriminant constraints inside the cost of NMF. The new subspace techniques have been applied to frontal face verification. A significant improvement of the performance of NMF has been verified in the frontal verification problem when the proposed constraints are incorporated. The proposed NMFfaces though outperform the well-known Fisherfaces and Eigenfaces in face verification.

Acknowledgment

This work is funded by the integrated project BioSec IST-2002-001766 (Biometric Security, http://www.biosec.org), under Information Society Technologies (IST) priority of the 6th Framework Programme of the European Community.

Appendix I

Derivation of the DNMF decomposition

In order to derive the coefficients of DNMF we have used an auxiliary function similar to those in the EM algorithm in [34]. Let G be an auxiliary function for $Y(\mathbf{F})$ if $G(\mathbf{F},\mathbf{F}^{(t-1)}) \geq Y(\mathbf{F})$ and $G(\mathbf{F},\mathbf{F}) = Y(\mathbf{F})$. If G is an auxiliary function of Y, then Y is nonincreasing under the update $\mathbf{F}^t = \arg\min_{\mathbf{F}} G(\mathbf{F},\mathbf{F}^{(t-1)})$[34]. With the help of the auxiliary function the update rules for the coefficients $\mathbf{H}$ and for the bases $\mathbf{Z}_D$ of DNMF can be derived. By fixing the matrix $\mathbf{Z}_D$, the matrix $\mathbf{H}$ is updated by minimizing $\mathbf{Y}_d(\mathbf{H}) = D_d(\mathbf{X}||\mathbf{Z}_D\mathbf{H})$ defined in (20). Let the function $\mathbf{G}_d$ be defined as:

$$\begin{aligned}\mathbf{G}_d(\mathbf{H},\mathbf{H}^{(t-1)}) &= \textstyle\sum_i \sum_j (x_{i,j} \ln x_{i,j} - x_{i,j}) + \\ &\textstyle\sum_i \sum_j \sum_k \frac{z_{i,k} h^{(t-1)}_{k,j}}{\sum_l z_{i,l} h^{(t-1)}_{l,j}} (\ln(z_{i,k} h_{k,j}) - \ln \frac{z_{i,k} h^{(t-1)}_{k,j}}{\sum_l z_{i,l} h^{(t-1)}_{l,j}}) + \\ &\textstyle\sum_i \sum_j \sum_k z_{i,k} h_{k,j} + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b].\end{aligned} \tag{31}$$

This function $\mathbf{G}_d(\mathbf{H},\mathbf{H}^{(t-1)})$ is an auxiliary function for $Y_d(\mathbf{H})$. It is straightforward to show that $\mathbf{G}_d(\mathbf{H},\mathbf{H}) = Y_d(\mathbf{H})$. In order to prove that $\mathbf{G}_d(\mathbf{H},\mathbf{H}^{(t-1)}) \geq Y_d(\mathbf{H})$ since,

$\ln(\sum_k z_{i,k}h_{k,j})$ is convex, the following inequality holds:

$$-\ln(\sum_k z_{i,k}h_{k,j}) \leq -\sum_k a_k \ln\frac{z_{i,k}h_{k,j}}{a_k} \tag{32}$$

for all non-negative a_k that satisfy $\sum_k a_k = 1$. By letting $a_k = \frac{z_{i,k}h_{k,j}^{(t-1)}}{\sum_l z_{i,l}h_{l,j}^{(t-1)}}$ we obtain:

$$-\ln(\sum_k z_{i,k}h_{k,j}) \leq \sum_k \frac{z_{i,k}h_{k,j}^{(t-1)}}{\sum_l z_{i,l}h_{l,j}^{(t-1)}}(\ln(z_{i,k}h_{k,j}) - \ln\frac{z_{i,k}h_{k,j}^{(t-1)}}{\sum_l z_{i,l}h_{l,j}^{(t-1)}}). \tag{33}$$

From (33) it is straightforward to show that $\mathbf{G}_d(\mathbf{H},\mathbf{H}^{(t-1)}) \geq Y_d(\mathbf{H})$. Thus $\mathbf{G}_d(\mathbf{H},\mathbf{H}^{(t-1)})$ is an auxiliary function of $Y_d(\mathbf{H})$.

The update rules are derived from setting $\frac{\partial G_d(\mathbf{H},\mathbf{H}^{(t-1)})}{\partial h_{k,l}}$ to zero for all the $h_{k,l}$. Let $h_{k,l}$ be the l-th element of the ρ-th image for the r-th class, thus, $h_{k,l} = \eta_{\rho,k}^{(r)}$. We need to calculate the partial derivatives $\frac{\partial \text{tr}[\mathbf{S}_w]}{\partial h_{k,l}}$ and $\frac{\partial \text{tr}[\mathbf{S}_b]}{\partial h_{k,l}}$. The partial derivative of the $\frac{\partial \text{tr}[\mathbf{S}_w]}{\partial h_{k,l}}$ is given by:

$$\begin{aligned}\frac{\partial \text{tr}[\mathbf{S}_w]}{\partial \eta_{\rho,k}^{(r)}} &= \frac{\partial \sum_i \sum_{c=1}^{K}\sum_{m=1}^{N_c}(\eta_{m,i}^{(c)} - \mu_i^{(c)})^2}{\partial \eta_{\rho,k}^{(r)}} = \sum_i\sum_{c=1}^{K}\sum_{m=1}^{N_c}\frac{\partial(\eta_{m,i}^{(c)} - \mu_i^{(c)})^2}{\partial \eta_{\rho,k}^{(r)}} \\ &= \sum_{m=1,m\neq\rho}^{N_c}\frac{\partial(\eta_{m,k}^{(r)} - \mu_k^{(r)})^2}{\partial \eta_{\rho,k}^{(r)}} + \frac{\partial(\eta_{\rho,k}^{(r)} - \mu_k^{(r)})^2}{\partial \eta_{\rho,k}^{(r)}} \\ &= -\sum_{m=1,m\neq\rho}^{N_c} 2(\eta_{m,k}^{(r)} - \mu_k^{(r)})\frac{1}{N_r} + 2(\eta_{\rho,k}^{(r)} - \mu_k^{(r)})(1-\frac{1}{N_r}) = 2(\eta_{\rho,k}^{(r)} - \mu_k^{(r)}).\end{aligned} \tag{34}$$

For the partial derivative $\frac{\partial \text{tr}[\mathbf{S}_b]}{\partial \eta_{\rho,k}^{(r)}}$ we have:

$$\begin{aligned}\frac{\partial \text{tr}[\mathbf{S}_b]}{\partial \eta_{\rho,k}^{(r)}} &= \frac{\partial \sum_i \sum_{c=1}^{K} N_c(\mu_i^{(c)} - \mu_i)^2}{\partial \eta_{\rho,k}^{(r)}} = \sum_i\sum_{c=1}^{K} N_c\frac{\partial(\mu_i^{(c)} - \mu_i)^2}{\partial \eta_{\rho,k}^{(r)}} = \sum_{c=1}^{K} N_c\frac{\partial(\mu_k^{(c)} - \mu_k)^2}{\partial \eta_{\rho,k}^{(r)}} \\ &= \sum_{c,c\neq r}^{K} N_c\frac{\partial(\mu_k^{(c)} - \mu_k)^2}{\partial \eta_{\rho,k}^{(r)}} + N_r\frac{\partial(\mu_k^{(r)} - \mu_k)^2}{\partial \eta_{\rho,k}^{(r)}} \end{aligned} \tag{35}$$

$$= -\frac{1}{L}\sum_{c,c\neq r}^{K} 2N_c(\mu_k^{(c)} - \mu_k) + 2N_r(\mu_k^{(r)} - \mu_k)(\frac{1}{N_r} - \frac{1}{L}) = 2(\mu_k^{(r)} - \mu_k). \tag{36}$$

Using (34) and (35) we have:

$$\frac{\partial G_d(\mathbf{H},\mathbf{H}^{(t-1)})}{\partial h_{k,l}} = -\sum_i x_{i,l}\frac{z_{i,k}h_{k,l}^{(t-1)}}{\sum_n z_{i,n}h_{n,l}^{(t-1)}}\frac{1}{h_{k,l}} + \sum_i z_{i,k} + 2\gamma(h_{k,l} - \mu_k^{(r)}) - 2\delta(\mu_k^{(r)} - \mu_k) = 0 \tag{37}$$

The quadratic equation (37) can be expanded as:

$$\begin{array}{l} -\sum_i x_{i,l}\frac{z_{i,k}h_{k,l}^{(t-1)}}{\sum_n z_{i,n}h_{n,l}^{(t-1)}} + h_{k,l} + 2\gamma h_{k,l}^2 - (2\gamma+2\delta)(\frac{1}{N_r}\sum_\lambda h_{k,\lambda})h_{k,l} + 2\delta\mu_k h_{k,l} = 0 \Leftrightarrow \\ -\sum_i x_{i,l}\frac{z_{i,k}h_{k,l}^{(t-1)}}{\sum_n z_{i,n}h_{n,l}^{(t-1)}} + (1 - (2\gamma+2\delta)(\frac{1}{N_r}\sum_{\lambda,\lambda\neq l} h_{k,\lambda}) + 2\delta\mu_k)h_{k,l} + \\ +(2\gamma - (2\gamma+2\delta)\frac{1}{N_r})h_{k,l}^2 = 0. \end{array} \tag{38}$$

By solving the quadratic equation (38) the update rules can be derived as:

$$h_{k,l} = \frac{T_1 + \sqrt{T_1^2 + 4(2\gamma - (2\gamma+2\delta)\frac{1}{N_r})h_{k,l}^{(t-1)}\sum_i z_{i,k}^{(t-1)}\frac{x_{i,j}}{\sum_n z_{i,n}^{(t-1)}h_{n,l}^{(t-1)}}}}{2(2\gamma - (2\gamma+2\delta)\frac{1}{N_r})} \tag{39}$$

where T_1 is given by:

$$T_1 = (2\gamma+2\delta)(\frac{1}{N_r}\sum_{\lambda,\lambda\neq l} h_{k,\lambda}) - 2\delta\mu_k - 1. \tag{40}$$

Appendix II

Derivation of the CSDNMF decomposition

The derivation of CSDNMF decomposition results in the same way as the decomposition of DNMF. Let r be the reference facial class. In a similar manner to Appendix I we can prove that $\mathbf{G}_c(\mathbf{H}_r, \mathbf{H}_r^{(t-1)})$ is an auxiliary function of $Y_c(\mathbf{H}_r) = D_c(\mathbf{X}||\mathbf{Z}_r\mathbf{H}_r)$ defined in (25), where $\mathbf{G}_c(\mathbf{H}_r, \mathbf{H}_r^{(t-1)})$ is given by:

$$\begin{array}{rl} \mathbf{G}_c(\mathbf{H}_r, \mathbf{H}_r^{(t-1)}) = & \sum_i\sum_j(x_{i,j}\ln x_{i,j} - x_{i,j}) + \\ & \sum_i\sum_j\sum_k \frac{z_{i,k}h_{k,j}^{(t-1)}}{\sum_l z_{i,l}h_{l,j}^{(t-1)}}(\ln(z_{i,k}h_{k,j}) - \ln\frac{z_{i,k}h_{k,j}^{(t-1)}}{\sum_l z_{i,l}h_{l,j}^{(t-1)}}) + \\ & \sum_i\sum_j\sum_k z_{i,k}h_{k,j} + \zeta\mathrm{tr}[\mathbf{B}_r] - \theta\mathrm{tr}[\mathbf{W}_r]. \end{array} \tag{41}$$

In this decomposition we have two different update rules. One for the genuine class and one for the impostor class. For $l = 1, \ldots, N_G$ (genuine class) the update rules for the coefficients $h_{k,l}$ for the reference person r are given by letting $\frac{\partial G_c(\mathbf{H}_r, \mathbf{H}_r^{(t-1)})}{\partial h_{k,l}} = 0$. Then,

$$\frac{\partial G_c(\mathbf{H}_r, \mathbf{H}_r^{(t-1)})}{\partial h_{k,l}} = -\sum_i x_{i,l}\frac{z_{i,k}h_{k,l}^{(t-1)}}{\sum_n z_{i,n}h_{n,l}^{(t-1)}}\frac{1}{h_{k,l}} + \sum_i z_{i,k} + 2\zeta(h_{k,l} - \mu_k^{(G)})\frac{1}{N_G} - 2\theta(\mu_k^{(G)} - \mu_k^{(I)})\frac{1}{N_G} = 0. \tag{42}$$

The quadratic equation (42) is expanded as:

$$-\sum_i x_{i,l}\frac{z_{i,k}h_{k,l}^{(t-1)}}{\sum_n z_{i,n}h_{n,l}^{(t-1)}} + h_{k,l} + 2\zeta\frac{1}{N_G}h_{k,l}^2 - (2\zeta+2\theta)\frac{1}{N_G}(\frac{1}{N_G}\sum_\lambda h_{k,\lambda})h_{k,l} + 2\theta\frac{1}{N_G}\mu_k^I h_{k,l} = 0 \Leftrightarrow$$
$$-\sum_i x_{i,l}\frac{z_{i,k}h_{k,l}^{(t-1)}}{\sum_n z_{i,n}h_{n,l}^{(t-1)}} + (1-(2\zeta+2\theta)\frac{1}{N_G}(\frac{1}{N_G}\sum_{\lambda,\lambda\neq l} h_{k,\lambda}) + 2\theta\frac{1}{N_G}\mu_k^{(I)})h_{k,l}+$$
$$+\frac{1}{N_G}(2\zeta-(2\zeta+2\theta)\frac{1}{N_G})h_{k,l}^2 = 0. \tag{43}$$

By solving the quadratic equation (43) the update rules for the $h_{k,l}$ of the genuine class are:

$$h_{k,l} = \frac{T_2 + \sqrt{T_2^2 + 4\frac{1}{N_G}(2\zeta-(2\zeta+2\theta)\frac{1}{N_G})h_{k,l}^{(t-1)}\sum_i z_{i,k}^{(t-1)}\frac{x_{i,j}}{\sum_n z_{i,n}^{(t-1)}h_{n,l}^{(t-1)}}}}{2\frac{1}{N_G}(2\zeta-(2\zeta+2\theta)\frac{1}{N_G})} \tag{44}$$

where T_2 is given by:

$$T_2 = (2\zeta+2\theta)\frac{1}{N_G}(\frac{1}{N_G}\sum_{\lambda,\lambda\neq l} h_{k,\lambda}) - 2\theta\frac{1}{N_G}\mu_k^{(I)} - 1. \tag{45}$$

The update rules for the coefficients $h_{k,l}$ for the impostor class of the reference person r are given by letting $\frac{\partial G_c(\mathbf{H}_r,\mathbf{H}_r^{(t-1)})}{\partial h_{k,l}} = 0$:

$$\frac{\partial G_c(\mathbf{H}_r,\mathbf{H}_r^{(t-1)})}{\partial h_{k,l}} = -\sum_i x_{i,l}\frac{z_{i,k}h_{k,l}^{(t-1)}}{\sum_n z_{i,n}h_{n,l}^{(t-1)}}\frac{1}{h_{k,l}} + \sum_i z_{i,k} - 2\frac{1}{N^I}\theta(h_{k,l} - \mu_k^{(G)}) = 0 \tag{46}$$

where $j = N_G + 1, \ldots, L$. By solving the quadratic equation (46) the update rules for the $h_{k,l}$ are given by

$$h_{k,l} = \frac{T_3 + \sqrt{T_3^2 - 8N_I\theta h_{k,l}^{(t-1)}\sum_i z_{i,k}^{(t-1)}\frac{x_{i,j}}{\sum_n z_{i,n}^{(t-1)}h_{n,l}^{(t-1)}}}}{4\theta} \tag{47}$$

where T_3 is given by:

$$T_3 = 2\theta\mu_k^{(G)} + N_I. \tag{48}$$

REFERENCES

[1] L. Chengjun, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 572 – 581, May 2004.

[2] P. J. Phillips, "Matching pursuit filters applied to face identification," *IEEE Transactions on Image Processing*, vol. 7, no. 8, pp. 1150–1164, Aug. 1998.

[3] L. Juwei, K.N. Plataniotis, and A.N. Venetsanopoulos, “Face recognition using kernel direct discriminant analysis algorithms,” *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 117–126, 2003.

[4] L. Chengjun and H. Wechsler, “Independent component analysis of Gabor features for face recognition,” *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 919–928, July 2003.

[5] B.-L. Zhang, H. Zhang, and S. S. Ge, “Face recognition by applying wavelet subband representation and kernel associative memory,” *IEEE Transactions on Neural Networks*, vol. 15, no. 1, pp. 166–177, 2004.

[6] J. Yang, A.F. Frangi, J. Yang, D. Zhang, and Z. Jin, “KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 230–244, 2005.

[7] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang, “Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble,” *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 875–886, 2005.

[8] H. Zhang, B. Zhang, W. Huang, and Q. Tian, “Gabor wavelet associative memory for face recognition,” *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 275–278, 2005.

[9] M.J. Er, W. Chen, and S. Wu, “High-speed face recognition based on discrete cosine transform and RBF neural networks,” *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 679 – 691, 2005.

[10] W. Zheng, L. Zhao, and Z. Cairong, “Foley-Sammon optimal discriminant vectors using kernel approach,” *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 1 – 9, 2005.

[11] C. Kotropoulos, A. Tefas, and I. Pitas, “Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions.,” *Pattern Recognition*, vol. 33, no. 12, pp. 31–43, Oct. 2000.

[12] A. Tefas, C. Kotropoulos, and I. Pitas, “Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 735–746, 2001.

[13] J. Matas, M. Hamou, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Bigun, N. Capdevielle, W. Gerstner, S. Ben-Yacouba, Y. Abdelaoued, and E. Mayoraz, “Comparison of face verification results on the XM2VTS database,” in *ICPR*, Barcelona, Spain, 3-8 September 2000, pp. 858–863.

[14] K. Messer, J.V. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, C. Sanderson, J. Czyz, L. Vandendorpe, S. Srisuk, M. Petrou, W. Kurutach, A. Kadyrov, R. Paredes, B. Kepenekci, F.B. Tek, G.B. Akar, F. Deravi, and N. Mavity, “Face verification competition on the XM2VTS database,” in *AVBPA03*, Guildford, United Kingdom, 9-11 June 2003, pp. 964–974.

[15] K. Messer, J. Matas, J.V. Kittler, J. Luettin, and G. Maitre, “XM2VTSDB: The extended M2VTS database,” in *AVBPA’99*, 1999, pp. 72–77.

[16] M. Kirby and L. Sirovich, “Application of the karhunen-loeve procedure for the characterization of human faces.,” *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[17] D.D. Lee and H.S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[18] D. L. Swets and J. Weng, “Using discriminant eigenfeatures for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[19] 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.

[20] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski, “Face recognition by independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, 2002.

[21] M. Turk and A. P. Pentland, “Eigenfaces for recognition.,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[22] D. Guillamet, J. Vitria, and B. Schiele, “Introducing a weighted non-negative matrix factorization for image classification,” *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2447 – 2454, 2003.

[23] L. Weixiang and N. Zheng, “Non-negative matrix factorization based methods for object recognition,” *Pattern Recognition Letters*, vol. 25, no. 9-10, pp. 893–897, 2004.

[24] D. Guillamet and Vitria, “Evaluation of distance metrics for recognition based on non-negative matrix factorization,” *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1599 – 1605, 2003.

[25] S. Wild, J. Curry, and A. Dougherty, “Improving non-negative matrix factorizations through structured initialization,” *Pattern Recognition*, vol. 37, pp. 2217–2232, 2004.

[26] D. Donoho and V. Stodden, “When does non-negative matrix factorization give a correct decomposition into parts ?,” *Advances in Neural Information Processing Systems*, vol. 17, 2004.

[27] S.Z. Li, X.W. Hou, and H.J. Zhang, “Learning spatially localized, parts-based representation,” in *CVPR*, Kauai, HI, USA, December 8-14 2001, pp. 207–212.

[28] X. Chen, L. Gu, S.Z. Li, and H-J. Zhang, “Learning representative local features for face detection,” in *CVPR*, Kauai, HI, USA, December 8-14 2001, pp. 1126–1131.

[29] I. Buciu and I. Pitas, “Application of non-negative and local non negative matrix factorization to facial expression recognition,” in *ICPR*, Cambridge, United Kingdom, 23-26 August 2004, pp. 288–291.

[30] P.O. Hoyer, “Non-negative sparse coding,” in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Martigny, Valais, Switzerland, September 4-6 2002, pp. 557–565.

[31] I. Buciu and I. Pitas, “A new sparse image representation algorithm applied to facial expression recognition,” in

MLSP, Sao Lus, Brazil, Sep. 29 - Oct. 1st 2004.

[32] C. Kotropoulos, A. Tefas, and I. Pitas, “Frontal face authentication using discriminating grids with morphological feature vectors.,” *IEEE Transactions on Multimedia*, vol. 2, no. 1, pp. 14–26, Mar. 2000.

[33] 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.

[34] D.D. Lee and H.S. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2000, pp. 556–562.

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

[36] L.M. Bregman, “The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming,” *U.S.S.R. Computational Mathematics and Mathematical Physics*, vol. 1, pp. 200–217, 1967.

[37] G.H. Golub and C.F. VanLoan, *Matrix Computations*, third ed. John Hopkins Univ. Press, 1996.

[38] L. Juwei, K.N. Plataniotis, and A.N. Venetsanopoulos, “Face recognition using LDA-based algorithms,” *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, 2003.

[39] Rama Chellappa, Charles L. Wilson, and Saad Sirohey, “Human and machine recognition of faces: A survey.,” *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, May 1995.

[40] W. Zhao, R. Chellappa, P.-J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” *ACM Computing Survey*, vol. 35, pp. 399–458, 2003.

[41] T.K. Ho and M. Basu, “Complexity measures of supervised classification problems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 289–300, 2002.

PLACE
PHOTO
HERE

Stefanos Zafeiriou was born in Thessaloniki, Greece in 1981. He received the BS degree from the Department of Informatics of Aristotle University of Thessaloniki with honors in 2003. He is currently a researcher and teaching assistant and he is studying towards a PhD at the Department of Informatics at the University of Thessaloniki. His current research interests lie in the areas of signal and image processing, pattern recognition and computer vision as well as in the area of watermarking for copyright protection and authentication of digital media.

PLACE
PHOTO
HERE

Anastasios Tefas received the B.Sc. in informatics in 1997 and the Ph.D. degree in informatics in 2002, both from the Aristotle University of Thessaloniki, Thessaloniki, Greece. From 1997 to 2002, he was a researcher and teaching assistant in the Department of Informatics, University of Thessaloniki. From 2003 to 2004, he was a temporary lecturer in the Department of Informatics, University of Thessaloniki where he is currently, a senior researcher. He has co-authored over 50 journal and conference papers. His current research interests include computational intelligence, pattern recognition, digital signal and image processing, detection and estimation theory, and computer vision.

PLACE
PHOTO
HERE

Ioan Buciu Ioan Buciu was born in Oradea, Romania, in 1971. He received the Diploma of Electrical Engineering in 1996 and Master of Science in Microwave in 1997 both from the University of Oradea, Romania. From 1997 to 2000 he served as teaching Assistant in the Department of Electrical and Computer Engineering at the University of Oradea. He is currently a researcher and he is studying toward a PhD at the Artificial Intelligence and Information Analysis Lab, Department of Informatics, at Aristotle University of Thessaloniki. His current research interests lie in the areas of signal, image processing, pattern recognition, machine learning and artificial intelligence. Also, his area of expertise includes face analysis, support vector machines and image representation.

PLACE
PHOTO
HERE

Ioannis Pitas received the Diploma of Electrical Engineering in 1980 and the PhD degree in Electrical Engineering in 1985 both from the Aristotle University of Thessaloniki, Greece. Since 1994, he has been a Professor at the Department of Informatics, Aristotle University of Thessaloniki. From 1980 to 1993 he served as Scientific Assistant, Lecturer, Assistant Professor, and Associate Professor in the Department of Electrical and Computer Engineering at the same University. He served as a Visiting Research Associate at the University of Toronto, Canada, University of Erlangen- Nuernberg, Germany, Tampere University of Technology, Finland, as Visiting Assistant Professor at the University of Toronto and as Visiting Professor at the University of British Columbia, Vancouver, Canada. He was lecturer in short courses for continuing education. He has published over 145 journal papers, 380 conference papers and contributed in 18 books in his areas of interest. He is the co-author of the books Nonlinear Digital Filters: Principles and Applications (Kluwer, 1990), 3-D Image Processing Algorithms (J. Wiley, 2000), Nonlinear Model-Based Image/Video Processing and Analysis (J. Wiley, 2001) and author of Digital Image Processing Algorithms and Applications (J. Wiley, 2000). He is the editor of the book Parallel Algorithms and Architectures for Digital Image Processing, Computer Vision and Neural Networks (Wiley, 1993). He has also been an invited speaker and/or member of the program committee of several scientific conferences and workshops. In the past he served as Associate Editor of the IEEE Transactions on Circuits and Systems, IEEE Transactions on Neural Networks, IEEE Transactions on Image Processing, EURASIP Journal on Applied Signal Processing and co-editor of Multidimensional Systems and Signal Processing. He was general chair of the 1995 IEEE Workshop on Nonlinear Signal and Image Processing (NSIP95), technical chair of the 1998 European Signal Processing Conference and general chair of IEEE ICIP 2001. His current interests are in the areas of digital image and video processing and analysis, multidimensional signal processing, watermarking and computer vision.