A Weighted Non-negative Matrix Factorization for Local Representations*

David Guillamet, Marco Bressan and Jordi Vitrià
Centre de Visió per Computador, Dept. Informàtica
Universitat Autònoma de Barcelona,
08193 Bellaterra, Barcelona, Spain.
{davidg, marco, jordi}@cvc.uab.es

Abstract

This paper presents an improvement of the classical Non-negative Matrix Factorization (NMF) approach, for dealing with local representations of image objects. NMF, when applied to global data representations such as faces presents a high ability to represent local features of the original data in an unsupervised way. However, when applied to local representations NMF generates redundant basis. This work implements an improvement on the original NMF approach by incorporating prior knowledge in the form of a weight matrix extracted from the training data. A detailed mathematical description of the inclusion of this weight matrix is provided, and results demonstrating its advantages are included. Furthermore, the original NMF approach lacks a hierarchy of the elements of the estimated basis. A technique to determine an ordered set of discriminant basis is also presented. Finally, the effectiveness of the weighted approach with respect to the classical one is experimentally compared. This is done by implementing a clustering algorithm that automatically extracts object parts from the NMF representation of an image database corresponding to newspapers.

1. Introduction

Visual recognition of objects is one of the most challenging problems in computer vision and artificial intelligence. Approaches to solve this problem have focused on using several methodologies. One of the most used today is appearance based recognition. Under this approach, raw image data is preprocessed in order to find a suitable representation to process the data with pattern recognition techniques. PCA was initially used to describe face patterns in a lower-dimensional space than the image space [10]. Other approaches have also focused on this technique to overcome frequent computer vision problems such as the recognition of objects taken under a wide range of conditions (several viewpoints and illumination conditions) [6], or dealing with partial occlusions by using robust estimation techniques [1, 9]. However, PCA based techniques suffer from several difficulties. Mainly, an image projection to a PCA based space depends on the precise position of relevant objects, on the intensity and shape of background zones, and on intensity and color of illumination.

Recently, a new approach for obtaining a reduced representation of global data has been proposed. This new technique, called Non-negative Matrix Factorization (NMF), was used in the work of Lee and Seung [5] to find parts of objects in an unsupervised way. Non-negative Matrix Factorization differs from other methods by its use of non-negativity constraints. Their work was tested with a set of faces [5] and the obtained NMF basis are localized features that correspond with intuitive notions of the parts of faces.

Some recent approaches focus on the fact that an object can be divided into small windows and that only a subset of them are necessary for identification. Current trends in object recognition focus on the local extraction of information to avoid all the problems generated by global approaches. Is for that reason that in this work, we modify the initial NMF algorithm to obtain reliable results when it is applied to local data representations. This makes possible to adapt previous local approaches to take advantage of this technique. This is an important fact because the original NMF technique works fine with global data representations, but when applied to local representations, it generates redundant basis.

We have modified the global NMF approach by introducing a weight matrix that weighs the training data according to some prior knowledge. Additionally, this work is focused on the representation through color histograms. We show that this weighted NMF approach numerically improves the original version when recovering the original set of color histograms for a particular experiment using color newspa-

*This work is supported by Comissionat per a Universitats i Recerca del Departament de la Presidencia de la Generalitat de Catalunya, Secretaría de Estado de Educacion y Universidades of the Ministerio de Educación, Cultura y Deportes de España, and CICYT grant TIC2000-0399-C02-01.

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

pers. After a set of tests that demonstrate the validity of the weighted NMF, we divide the original color newspapers into parts according to the projected color histograms obtained by both approaches using a previously developed technique [3]. As it turns out, both approaches suffer from not having an ordered list of the basis representing the reduced space. In order to overcome this problem, we present a technique to obtain this ordered list according to the discriminant information that they contain.

2. Non-negative Matrix Factorization (NMF)

The non-negative matrix factorization (NMF) provides a minimum error non-negative representation of the data. The non-negativity constraint leads to a part-based representation because it allows only additive, not subtractive, combinations of the original data [5].

While working with global data, each training vector is a particular object or instance of an object that must be learned. Local data instead, extracts a number of feature vectors from each object. In the training stage, the information within these feature vector can be redundant, and strong similarities can be found among them. These similarities are not taken into account in the classical NMF approach and this results in redundant positive basis. Next section will explain the classical and weighted NMF techniques.

2.1. Global NMF

Our object database is represented by a $n \times m$ matrix V, where each column is an n-dimensional non-negative local vector belonging to the original database (m local vectors). Using two new matrices (W and H) we can obtain an approximation of the whole object database (V) as $V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$. The dimensions of the matrix factors W and H are $n \times r$ and $r \times m$, respectively. Usually r is chosen to be smaller than m and n. Each column of matrix W contains a basis vector while each column of H contains the weights needed to approximate the corresponding column in V using the basis from W. In the PCA context, each column of matrix W represents an eigenimage and the matrix factors of H represent the eigenprojections. In contrast to PCA, NMF does not allow negative entries in the matrix factors W and H permitting the combination of multiple basis images to represent an object.

In order to estimate the factorization matrices, an objective function has to be defined. A possible objective function is given by $F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} [V_{i\mu} \log (WH)_{i\mu} - (WH)_{i\mu}]$. This objective function can be related to the likelihood of generating the images in V from the basis W and encondings H. An iterative approach for the obtention of a local maximum of this objective function is given by the following rule [5]: $W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$, $W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$, $H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$. Initialization is performed using positive random initial conditions for matrices W and H. The convergence of the process is also ensured. See [4, 5] for more information.

2.2. Weighted NMF

When using a local representation, similarity between the training vectors can introduce redundancy in the basis W. This redundancy is manifested by the presence of repeated basis. A possible solution to this problem is to introduce a weight on each the training vectors, giving more weight to those vectors with low probability of appearing in the training set. This weighted model can be seen as the result of multiplying both sides of the factorization with a m by m diagonal weight matrix Q and to estimate the basis and encodings for the new factorization model, $VQ \approx WHQ$. Where the diagonal element q_μ corresponds to the weight of training vector μ, with $1 \leq \mu \leq m$. It is also assumed that all the weights sum to unity. The modified objective function in this case takes the form $F_Q = \sum_{\mu=1}^{m} q_\mu \sum_{i=1}^{n} [V_{i\mu} \log((WH)_{i\mu} q_\mu) - (WH)_{i\mu}]$. Now, the iterative update rules to obtain the new matrices subject to this new objective function are defined by: $W_{ia} \leftarrow \frac{W_{ia}}{\sum_\mu q_\mu H_{a\mu}} \sum_\mu \frac{q_\mu V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$, $W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$, $H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$.

As global NMF finds out some redundant basis corresponding to the most frequent training vectors when applied to local data, a good choice to define the weighted matrix Q is giving more weight to the less frequent training vectors. By obtaining the probability of each training vector with respect to the training database and assuming that this probability will hold on the test stage, we invert these probabilities and we take them as the q_μ coefficients. In this way, we equalize the importance of the training vector classes. Thus, the obtained basis will contain a wide variety of behaviours: from the most to the least frequent ones, improving the global NMF capacity of representation that only takes into account the most frequent ones.

3. Object Clustering

Assuming that we have a database of different images that contain similar structures (i.e. newspapers), our goal is to obtain the regions of all the images where the objects have a certain behaviour. Each object instance has a set of representative local vectors $V = \{v_i\}$, that in our particular case is a set of projected local color histograms. For our object clustering problem, we must define a measure of similarity between two projected local color histograms (v_i and v_j) as $s(v_i, v_j) = \sqrt{\sum_{k=1}^{r} (v_i^k - v_j^k)^2}$.

Our object segmentation algorithm is based on finding regions that have a similar behaviours with some other regions in the database. This goal can be achieved by defining

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

a similarity measure M_i^{kl} which reflects the highest similarity between local vector i from object k and all the local vectors from object l.

$$M_i^{kl} = argmin\{s(v_i^k, v_j^l)\} \quad \forall v_j^l \in O_l \tag{1}$$

Thus, an exhaustive search for the most similar local vector in object O_l has to be carried out in order to obtain this similarity measure. This expression was also used by Shams [7] to partition an object in local parts according to its similarity to an object database of similar objects. Denoting as $\mathbf{M_i^k}$ the vector that contains all the l similarity measures M_i^{kl}, we obtain a vector that reflects how a local vector of an object i can be found in the rest of the database. It is expected that, if the database is composed of objects that contain similar regions, the similarity vector of these regions will be have a different behaviour from the other zones of the image. On this expectation, we calculate the correlations:

$$R_{ij}^k = \frac{\mathbf{M_i^{k\,T}} \mathbf{M_j^k}}{\| \mathbf{M_i^k} \| \cdot \| \mathbf{M_j^k} \|} \tag{2}$$

This matrix reflects how the local responses of an object k are correlated between them. If we assume that our object database is composed of objects containing similar regions, we clearly assume that this matrix will contain two different clusters: the similar regions of an object and the other regions. Given this correlation matrix, we use the algorithm developed by Shi and Malik [8] to obtain two or more clusters.

Assuming that we have divided an object into N different parts using the segmentation algorithm of Shi and Malik [8], we can learn the discrimination of each object part with respect to the whole object database. Thus, we can extract a measure of discrimination of each object part by considering the energy of its vectors M as:

$$E_{\text{part}}^k = \frac{1}{P} \sum_{i=1}^{P} \| M_i^k \| \quad \forall M_i^k \in \text{Part}(O_k) \tag{3}$$

and P indicates the number of local vectors in object part Part(O_k). By sorting the local regions according to their energy (E_{part}^k), we can obtain the most discriminant local part of an object with respect to the whole object database. To obtain a more detailed information about this clustering technique, refer to [3].

4. Discriminant Histogram Basis

Using the PCA technique we obtain an ordered list of the eigenvectors and we can use a certain number of them according to the task we are working on. Nevertheless, using the NMF technique we only obtain an unordered set of basis.

Given that the NMF technique finds two matrices W and H that maximize a likelihood expression, if a certain color dominates over the others, it would be expected to find more basis histograms representing this color because NMF tries to maximise an objective function. This behaviour is corrected when we use the Weighted NMF. By averaging the set of local projected histograms per object (matrix H), we know how all the histogram basis (matrix W) are used in the whole the database. More formally, we have n objects (O_n), each of them containing a set of local histograms $V_{O_i} = \{v_{O_i}\}$ that are projected in the NMF space from where we obtain a set of coefficients $H_{O_i,j}$,

$$\begin{aligned}
O_1 \rightarrow V_{O_1} = \{v_{O_1}\} \quad &\Rightarrow \quad v'_{O_1,i} = W H_{O_1,i} \quad \forall v_i \in V_{O_1} \\
O_2 \rightarrow V_{O_2} = \{v_{O_2}\} \quad &\Rightarrow \quad v'_{O_2,i} = W H_{O_2,i} \quad \forall v_i \in V_{O_2} \\
&\vdots \\
O_n \rightarrow V_{O_n} = \{v_{O_n}\} \quad &\Rightarrow \quad v'_{O_n,i} = W H_{O_n,i} \quad \forall v_i \in V_{O_n}
\end{aligned} \tag{4}$$

where $H_{O_j,i}$ are the NMF projected coefficients of vector v_i from object O_j and $v'_{O_j,i}$ is the reconstructed vector of the original $v_{O_j,i}$. Now, averaging all the NMF projected coefficients per object, we can obtain a set of representative vectors per each histogram basis as

$$F_j = \left[\frac{1}{|V_{O_1}|} \sum_{i=1}^{|V_{O_1}|} H_{O_1,i}^j, \cdots, \frac{1}{|V_{O_n}|} \sum_{i=1}^{|V_{O_n}|} H_{O_n,i}^j \right] \tag{5}$$

where F_j is an n-dimensional vector that encodes how the j-th histogram basis is used in the whole database. Considering the whole set of histogram basis, we can cluster these vectors according to their similarities of usage using the work of Shi and Malik [8]. Once we obtain a set of clusters dividing the histogram basis according to their level of usage, we can sort them from the most unused to the most used. Thus, we obtain an ordered list of the histogram basis where the most unused set would represent the most discriminant histogram basis. To obtain more detailed information about how to find a discriminant basis using this technique, refer to [2].

5. Experimental Results

Three different experiments were conducted to demonstrate the feasibility of the Weighted NMF approach with respect to the classical one. The first experiment numerically proves that the Weighted NMF approach outperforms the classical NMF in terms of the reconstruction error of color images (newspapers). In this experiment, we also observe that with the same initial conditions and with the same training and testing data, Weighted NMF is able to obtain a better representation. To complement the numerical experiment and in order to observe visually the differences in performance, we present a visual comparison between NMF

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

and WNMF in the context of extracting object parts. Finally, because NMF and WNMF lack of a hierarchy on the elements of the estimated basis, we also present a technique to cluster this set of basis according to the level of discriminant information that they contain.

5.1. Weighted NMF and NMF

This experiment compares the capacity of representation of both algorithms, the NMF and WNMF. We have selected 5 different color newspapers of 360×549 (see fig. (1)) each of them containing particular local characteristics. Having 10 instances of each newspaper, we have divided all of them using a predefined grid obtaining a large amount of local windows (200 per image). Each local window contains approximately 1000 pixels and is represented using a 8-bin color histogram ($512D$) and all of them are used as input in the learning process of the NMF matrices (W and H). Because of the initial random conditions of the NMF approaches, we have initialized each of them (NMF and Weighted NMF) with the same random matrices, for accurate comparison. The selection of the weights (matrix Q) in the WNMF approach is performed using a leave one out technique that tests the probability of finding the training histogram in the whole training database. Once we have obtained matrix W for both approaches, other local color histograms were projected. These additional histograms were extracted from the same objects but using a different grid. Finally, the reconstruction error of both approaches was compared ($E = \sqrt{\sum (v'_{O_j,i} - v_{O_j,i})^2}$). This experiment was performed with a different number of basis $r = 30, 35, 40, 45, 50, 55$. Results are shown in fig. (2) where we can see the evolution of the reconstruction error of the testing set through the number of iterations. We have to note that for the training stage, we learned our models using 400 iterations of the algorithm because they tend to be stable. For the testing stage we only made 50 iterations. Fig. (2) shows how the reconstruction error is stabilized within only a few iterations. From figure (2) we can conclude that using a small number of basis $r = 30, 35, 40$ (from (a) to (c)), classical NMF outperforms the Weighted NMF. When the number of desired basis is increased ($r = 45, 50, 55$), Weighted NMF outperforms the classical approach.

Figure 1. Five different color newspapers (360×549) containing different local regions.

In fig. (2), we obtain that the weighted NMF can not outperform the original one given that it does not have enough

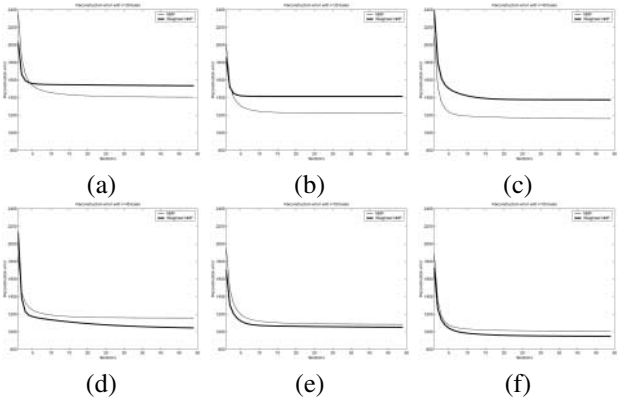


Figure 2. NMF (light line) and WNMF (solid line) reconstruction error (y axis) through 50 iterations (x axis) of the testing data.

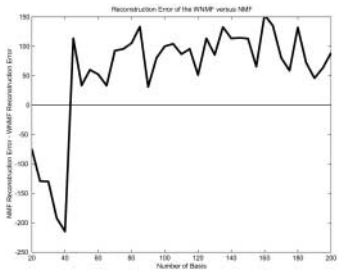


Figure 3. Reconstruction error of WNMF versus NMF using from $r = 20$ to $r = 200$ basis.

basis to represent specific tonalities when using less than 40 basis. But, considering more than $r = 40$ basis we obtain that the weighted NMF always outperforms the classical one. To show the behaviour on the performance of the Weighted NMF with respect to the classical one, we have trained the model using different basis (from $r = 20$ to $r = 200$) in steps of 5 and the results are shown in fig. (3). In this figure, a positive value means that WNMF is working better than NMF in terms of reconstruction error, and a negative value, the opposite.

From fig. (3) we can really check that with a few basis, the classical NMF is better than the weighted one. This behaviour is justified by the fact that, having a large amount of training vectors, the classical NMF tries to find the best basis that represent all this space. With only a few set of basis, NMF captures the most used color regions but without generating redundant basis. The weighted NMF tries to give more weight to specific color regions which might not be relevant in terms of reconstruction error, but that would surely avoid redundancy with an increased number of basis. For this reason, when we expect to find more than 40 basis, the Weighted NMF can improve the classical NMF.

Depending on the variability of the input data there will be some point from where the classical NMF starts to gener-

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

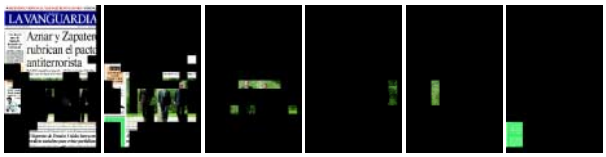
ate redundant basis and the Weighted NMF will start working better. In our specific problem, this point is found when we use $r = 40$ basis and it is useful to know this point if we have to choose the best representation for a given problem.

5.2. Clustering Parts of Objects

In this section we present a graphical application of the Weighted NMF approach that can be used to appreciate how it outperforms the classical NMF. Using the previous explained clustering algorithm (section 3) with the projected histograms of both NMF approaches, we demonstrate that WNMF outperforms NMF because it is able to represent some specific behaviours that are not initially detected. Figs. (4) and (5) show two different newspapers divided into local parts. These figures contain an ordered list of the object parts from the most common object part (usually newspaper title and text) to the most discriminant.

(a) (b) (c)

Figure 4. Different parts obtained by applying the classical NMF based clustering algorithm.



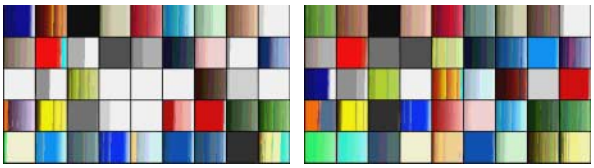
(a) (b) (c) (d) (e) (f)

Figure 5. Different object parts obtained by applying the WNMF based clustering algorithm (compare with those in fig. (4)).

Figs. (4) and (5) show one newspaper that visually reflects an important behaviour of the WNMF: its ability to separate specific tonalities in different object parts. This is shown in fig. (5.c) to (4.f) given that they contain different green tonalities separated in different object parts. This behaviour is not manifested with the classical NMF given that nearly all the green tonalities are compacted in one cluster.

5.3. Discriminant Histogram Basis

In this section, we show that all the unordered histogram basis obtained by both NMF approaches can be sorted according to the level of discriminant information that they contain. For this experiment, we focus on the particular case of using $r = 45$ histogram basis. The histogram basis obtained by both NMF approaches are shown in fig. (6). As it can be seen, the histogram basis obtained using the classical approach of the NMF contains several repeated histograms. Some of these redundant set of obtained basis contain white tonalities because this is the most frequent color in all the newspapers. This problem is minimized using the weighted NMF because all the color histogram basis are obtained by considering some previous knowledge about how the input data is used in the whole database. Thus, all the redundant set of histogram basis obtained by the classical NMF are eliminated and some specific tonalities that are not initially considered by the classical NMF have emerged improving the representation capacity of the algorithm.



(a) (b)

Figure 6. Histogram basis obtained by both NMF approaches. (a) Using the classical NMF approach and (b) using the WNMF.

As seen in fig. (6), the weighted NMF reduces the number of histogram basis concerning white regions giving more emphasis to the specific tonalities. However, these set of basis do not contain any information about which of them are the most important. NMF can be described as a dimensionality reduction algorithm as PCA but there is no way to know which basis are more relevant and contain more information to describe the original data. As noted in the original work that describes the classical NMF [5], NMF generates a sparse basis that, after combining some of them, it is possible to obtain a global object. This sparsity property leads to consider that the histogram basis found in this particular study are only used in some specific regions of the color newspapers. This means that we can use this property to cluster the obtained basis according to how often they are used.

Given the matrix factor H for both approaches of the NMF and as it has been explained in section 4, we have obtained all $F_1, \ldots F_{45}$ (expression (5)). By applying the algorithm developed by Shi and Malik [8], we obtain a set of histogram clusters. When a cluster contains different histogram basis, it means that this set of histograms have the same behaviour and that they appear in the same newspapers. This set of clusters is ordered taking into account how often its histograms are used in the whole database. Figs. (7) and (8) show the different obtained clusters for both NMF approaches.

As visually observed in figs. (7) and (8), we can really appreciate that the Weighted NMF generates a more sparse set of histogram basis clusters. The main difference between these two sets of histogram basis is that the clus-

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE

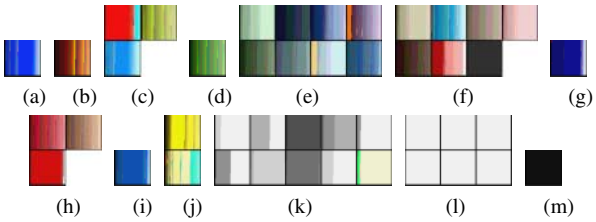


Figure 7. Ordered clusters of histogram basis obtained with NMF. The most discriminant one (a) and the least discriminant (m).

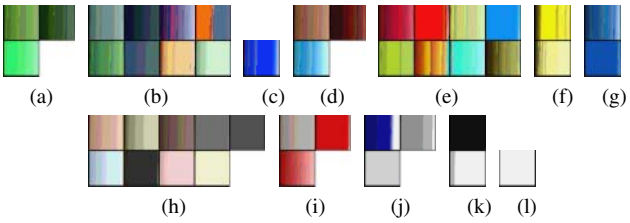


Figure 8. Ordered clusters of histogram basis obtained with WNMF. The most discriminant one (a) and the least discriminant (l).

ter referenced as fig. (7.l) contains 6 histogram basis that are representing the same white color. Using the Weighted NMF, this cluster is represented by fig. (8.l) where only appears one white color histogram. Another important difference is that cluster represented by fig. (8.a) contains 3 different histogram basis representing different green tonalities that are not present with the classical NMF raising to consider this cluster as the most discriminant one. Other differences can also be seen in the most frequently used clusters (figs. (7.k) and (8.j)). One example is the color histograms containing gray values, in the classical NMF we obtain nearly 9 different color histograms containing different levels of gray color tonalities but with the Weighted NMF we only obtain 5 color histograms with this tonality.

6. Conclusions

In this paper we have presented a technique to manage local data representations using the Non-negative Matrix Factorization (NMF) technique. This technique has been used with global representations leading to generate interesting results in the sense that each obtained basis corresponds to different object parts of the input data. This means that it can be seen as a dimensionality reduction algorithm that finds a representation of the internal data in a different way than traditional techniques such as PCA. Working with local representations, the classical NMF fails to obtain a good model of the original data leading to generate some redundant basis while trying to minimize an error function. However, we demonstrate that adding a weight matrix we achieve a high performance in the obtained results by representing local behaviours that are not previously considered. This is because of the weight matrix takes into account how each local vector can be found in the original training database.

We have done different tests to demonstrate the superior performance of Weighted NMF with respect to the classical approach. The numerical one shows the ability of the Weighted NMF to obtain a closer representation of the original data in terms of the reconstruction error. We have found that using only a few components, the classical NMF can find a better representation with respect to the Weighted NMF but when the number of desired basis is increased, the classical NMF finds some redundant basis that are eliminated with the Weighted NMF. The visual experiment shows different object parts obtained from the projected basis using both approaches. Because of the high capacity of representing specific behaviours, the Weighted NMF finds clusters that are not considered with the original one. This fact implies that when the classical NMF finds redundant basis, the Weighted NMF can reduce this redundancy by considering other specific behaviours improving the representation capacity. Finally, we have also developed a method to sort the basis obtained by both approaches according to the level of discriminant information that they contain.

However, further analysis of the Weighted NMF has to be done in the sense that it can be seen as a dimensionality reduction algorithm and it can be compared to the traditional techniques such as PCA.

References

[1] M. Black and A. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation. *IJCV*, Vol. 26, n. 1, pp. 63-84, 1998.

[2] D. Guillamet and J. Vitrià. Discriminant Basis for Object Classification. In *Proc. of Intern. Conf. on Image Analysis and Processing (ICIAP)*, 2001.

[3] D. Guillamet and J. Vitrià. Unsupervised Learning of Part-Based Representations. In *Proc. of CAIP*,LNCS 2124, pages 700-708, 2001.

[4] D. Lee and H. Seung. Algorithms for Non-negative Matrix Factorization. *NIPS*, 2000.

[5] D. Lee and H. Seung,. Learning the parts of objects by non-negative matrix factorization. *Nature*, Vol. 401, pp. 788-791, 1999.

[6] H. Murase and S.K. Nayar. Visual Learning and recognition of 3D objects from appearance. *International Journal of Computer Vision*, Vol. 14, pp. 5-24, 1995.

[7] L. Shams. Development of Visual Shape Primitives. PhD thesis, University of Southern California, 1999.

[8] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Trans. PAMI*, Vol. 22, n. 8, pp. 888-905, 2000.

[9] F. de la Torre and M. Black. Robust Principal Component Analysis for Computer Vision. *in Proc. of ICCV*, 2001.

[10] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, Vol. 3, pp. 71-86, 1991.

ISBN 0-7695-1272-0/01 $10.00 (C) 2001 IEEE