Local Non-Negative Matrix Factorization as a Visual Representation

Tao Feng, Stan Z. Li, Heung-Yeung Shum, HongJiang Zhang
Microsoft Research Asia, Beijing Sigma Center, Beijing 100080, China
Contact: tfeng@microsoft.com

Abstract

In this paper, we propose a novel method, called local non-negative matrix factorization (LNMF), for learning spatially localized, parts-based subspace representation of visual patterns. An objective function is defined to impose localization constraint, in addition to the non-negativity constraint in the standard NMF [1]. This gives a set of bases which not only allows a non-subtractive (part-based) representation of images but also manifests localized features. An algorithm is presented for the learning of such basis components. Experimental results are presented to compare LNMF with the NMF and PCA methods for face representation and recognition, which demonstrates advantages of LNMF.

Based on our LNMF approach, a set of orthogonal, binary, localized basis components are learned from a well aligned face image database. It leads to a Walsh function based representation of the face images. These properties can be used to resolve occlusion problem, improve the computing efficiency, and compress the storage requirement of face detection and recognition system.

1 Introduction

Subspace analysis helps to reveal low dimensional structures of patterns observed in high dimensional spaces. A specific pattern of interest can reside in a low dimensional sub-manifold in the original input data space of possibly an unnecessarily high dimensionality. Consider the case of $N \times M$ image pixels, each taking a value in $\{0, 1, \ldots, 255\}$; there is a huge number of possible configurations: $256^{N \times M}$. This space is capable of describing a wide variety of visual object classes or patterns. However, for a specific pattern, such as the human face, the number of admissible configurations is a only tiny fraction of that huge number. In other words, the intrinsic dimensionality is much lower than $N \times M$.

An observation can be considered as a consequence of linear or nonlinear fusion of a small number of intrinsic or latent variables. Subspace analysis is aimed to derive a representation for such a fusion. It is closely related to feature extraction in pattern analysis aimed at discovering and computing intrinsic low dimensions of the pattern from the observation.

For these reasons, subspace analysis has been a major research issue in learning based image analysis, such as object detection and recognition [2, 3, 4, 5, 6, 7]. The significance is twofold: (1) effective characterization of a pattern of interest, or effective classification of different patterns; and (2) dimension reduction.

These are major goals of feature extraction common in both traditional computer vision and current learning based paradigms. However, traditional vision methods prespecifies features of interest, for example, corners, line segments and surface patches. Visual recognition based on such intuitive features has not been very successful in the past. This is perhaps because less intuitive but more crucial information may have been lost in the course of abstracting the image into these features. In contrast, in the current learning paradigm, features are not pre-specified, although may be constrained in much looser way, but learned possibly from a given set of training examples.

Algorithms from both paradigms in effect construct a mapping from the high dimensional input (*e.g.* image) space to a low dimensional feature space. A mapping constructed by visual feature extraction algorithms (*e.g.* corner detection) in traditional computer vision is understandably highly nonlinear and discontinuous. Although a mapping derived by a learning algorithm may also be nonlinear, it is mostly continuous.

Here in this paper, we are interested the linear type of mappings: Dimension reduction from a high dimensional input $\mathbf{x}$ to a low dimensional feature vector $\mathbf{h}$ can be expressed as a linear projection operation as $\mathbf{h} = \mathbf{P}\mathbf{x}$. Reconstruction from is done via a set of basis as $\mathbf{x} = \mathbf{B}\mathbf{h}$. Different learning algorithms derive different basis matrix $\mathbf{B}$ and project matrix $\mathbf{P}$.

The eigen-image method [2, 3, 4] uses principal component analysis (PCA) [8] performed on a set of representative training data to decorrelate second order moments corresponding to low frequency properties. Any image is represented as a linear combination of most significant orthonormal basis components, while least significant components corresponding to lowest eigen-values are discarded to

achieve dimension reduction. Due to the holistic nature of the method, the resulting components are global interpretations, and thus PCA is unable to extract basis components manifesting building parts consisting of localized features.

However, in many applications, a part-based representation, in which object parts are composed of more localized features, offers advantages in object recognition, including stability to local deformations, lighting variations, and partial occlusion. Several methods have been proposed recently for spatially localized, parts-based (non-subtractive or additive) feature extraction.

Local feature analysis (LFA) [9], also based on second order statistics, is a method for extracting, from the holistic PCA basis, local topographic representation in terms of local features. Independent component analysis [10, 11] is a linear non-orthogonal transform leading to a representation in which unknown linear mixtures of multi-dimensional random variables are made as statistically independent as possible. ICA not only decorrelates the second order statistics but also reduces higher-order statistical dependencies. It is found that independent component of natural scenes are localized edge-like filters [12].

The projection coefficients for the linear combinations in the above methods can be either positive or negative, and such linear combinations generally involve complex cancellations between positive and negative numbers. Therefore, these representations lack the intuitive meaning of adding parts to form a whole.

Non-negative matrix factorization (NMF) [1] imposes the non-negativity constraints in learning basis images. The pixel values of resulting basis images, as well as coefficients for reconstruction, are all non-negative. This way, only non-subtractive combinations are allowed. This ensures that the components are combined to form a whole in the non-subtractive way. For this reason, NMF is considered as a procedure for learning a parts-based representation [1]. However, the additive parts learned by NMF are not necessarily localized, and moreover, we found that the original NMF representation yields low recognition accuracy, as will be shown.

In this paper, we propose a novel subspace method, called local non-negative matrix factorization (LNMF), for learning spatially localized, parts-based representation of visual patterns. Inspired by the original NMF [1], the aim of this work a NMF representation that truly manifests part-based representation for tasks where feature localization is important. The constraints of sparsity is imposed on coordinates ($\mathbf{h}$) in the low dimensional feature space and locality of features on the basis components ($\mathbf{B}$), in addition to the non-negativity constraint of [1]. A procedure is presented to perform the constrained optimization to learn truly localized, parts-based components. A proof of the convergence of the algorithm has been provided in [13].

In our experiments, with the locality constraint, the basis components ($\mathbf{B}$) of the face images tend to be binary-like. For a congregated train process, we can get a set of binary, orthogonal basis components for the train face images. The binary basis components has advantages on dimension reduction (decreasing the storage needed for face recognition and detection algorithms) and increasing the computing efficiency of algorithms(Binary basis can be computed very fast). This Walsh function like property provides us an approach to resolve the occlusion problem in face detection and recognition problem.

The rest of the paper is organized as follows: Section 2 introduces NMF in contrast to PCA. This is followed by the formulation of LNMF. A LNMF learning procedure is presented and its convergence proved. Section 3 presents experimental results illustrating properties of LNMF and its performance in face recognition as compared to PCA and NMF.

2 Constrained Non-Negative Matrix Factorization

Let a set of N_T training images be given as an $n \times N_T$ matrix $\mathbf{X} = [x_{ij}] = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_T}]$ where a column vector $\mathbf{x}_j$ consists of the n non-negative pixel values of a training image. Denote a set of $m \leq n$ basis vectors by an $n \times m$ matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m]$. (Dimension reduction is achieved when $m < n$). A training image can be represented as a linear combination of the basis vectors $\mathbf{x}_j \approx \mathbf{B}\mathbf{h}_j$ where $\mathbf{h}_j = [h_{1j}, \ldots, h_{mj}]^T$ is an m-element column vector consisting of projected coordinates in the m dimensional feature space, and hence the training image matrix can be approximately factorized as

$$\mathbf{X} \approx \mathbf{B}\mathbf{H} \tag{1}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{N_T}]$ is $m \times N_T$. While Eq.(1) represent the reconstruction, the reverse process, *i.e.* projection, can be done as $\mathbf{h}_j = \mathbf{P}\mathbf{x}_j$ where the projection matrix $\mathbf{P}$ is the (generalized) inverse of $\mathbf{B}$.

The PCA factorization imposes no other constraints than the orthogonality, and hence allows the entries of $\mathbf{B}$ and $\mathbf{H}$ to be of arbitrary sign. Many basis images, *i.e.* eigenfaces in the case of face recognition, lack intuitive meaning; and a linear combination of them generally involves complex cancellations between positive and negative numbers. The NMF and LNMF representations allow only positive coefficients and thus non-subtractive combinations.

2.1 NMF

NMF imposes the non-negativity constraints

$$\mathbf{B}, \mathbf{H} \geq \mathbf{0} \tag{2}$$

such that all entries of $\mathbf{B}$ and $\mathbf{H}$ are non-negative, and hence only non-subtractive combinations are allowed [1]. This is

believed to be compatible to the intuitive notion of combining parts to form a whole, and also consistent with the physiological fact that the firing rate are non-negative.

NMF uses the divergence of $\mathbf{X}$ from $\mathbf{Y}$, defined as

$$D(\mathbf{X}||\mathbf{Y}) = \sum_{i,j} \left(x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij} \right) \tag{3}$$

as the measure of cost for factorizing $\mathbf{X}$ into $\mathbf{BH} \triangleq \mathbf{Y} = [y_{ij}]$. $D(\mathbf{X}||\mathbf{Y})$ reduces to Kullback-Leibler divergence when $\sum_{i,j} x_{ij} = \sum_{i,j} y_{ij} = 1$. An NMF factorization is defined by a solution to the following constrained problem

$$\begin{aligned} \min_{\mathbf{B},\mathbf{H}} \quad & D(\mathbf{X}||\mathbf{BH}) \\ \text{s.t} \quad & \mathbf{B}, \mathbf{H} \geq \mathbf{0}, \ \sum_i b_{ij} = 1 \ \forall j \end{aligned} \tag{4}$$

where $\sum_i b_{ij} = 1$ is for stablizating the computation (see `http://journalclub.mit.edu`). The above optimization can be done by using multiplicative update rules [14], for which a matlab program is available at `http://journalclub.mit.edu` under the "Computational Neuroscience" discussion category. A set of NMF components (columns of $\mathbf{B}$) obtained by using the above learning algorithm on a set of face training data are shown in [1]. There, most of the components present localized and part-based features.

The constrained minimization of Eq.(3) leads to additive decomposition of the data, but not necessarily to basis components consisting of local parts such as eyes and mouth of the face. To evaluate, we applied the algorithm on another data set, which is the ORL face database of AT&T Laboratories Cambridge, and obtained a result shown on the left of Fig.1. On the right of the figure is the result from another face database in which we have aligned the faces in a better quality than the ORL database. These results differ from the above referred one in several aspects: It is holistic rather than localized, and hence not really part-based. We believe that the differences are caused by the different quality of alignment of faces in these databases. Whereas faces in the ORL database are not well aligned, more careful alignment may have been done in Lee and Seung's data. We believe that the desired properties of the NMF results presented in [1], *i.e.* features of localized parts, is ascribed to the good alignment done by the pre-processing of the data, rather than by an inherent ability of the algorithm to learn local parts.

Another reason for this present work is concerning the suitability of NMF basis for object recognition. PCA has been extensively used for face recognition. While NMF is an alternative factorization method, it is natural to evaluate how it compares to PCA in face recognition. Our test with the ORL database concludes that the NMF-based recognition rate is lower than PCA-based, as will be shown in experiments. These motivated us to do an investigation into a rectified NMF model in this paper.

Figure 1: NMF basis components learned from ORL database (left) and a better aligned face database (right). They appear holistic and do not manifest meaningful facial features.

2.2 LNMF

LNMF is aimed at learning localized, part-based features in $\mathbf{B}$ for a factorization $\mathbf{X} \approx \mathbf{BH}$. Denoting $\mathbf{U} = [u_{ij}] = \mathbf{B}^T\mathbf{B}$, $\mathbf{V} = [v_{ij}] = \mathbf{H}\mathbf{H}^T$, both being $m \times m$, the following three additional constraints are imposed on the NMF basis.
(1) Maximum Sparsity in $\mathbf{H}$. $\mathbf{H}$ should contain as many zero components as possible. This requires that a basis component should not be further decomposed into more components so that the number of basis components required to represent $\mathbf{X}$ is minimized. Given the existing constraints $\|b_j\|_1 = \sum_{i=1}^n b_{ij} = 1$ for all j, we wish to minimize $\sum_{i=1}^n b_{ij}^2 = u_{jj}$ so that each $\mathbf{b}_i$ contains as many non-zero elements as possible to be as expressive as possible. The maximum sparsity in $\mathbf{H}$ in imposed as $\sum_{i=1}^n b_{ij}^2 = u_{jj} = \min$.
(2) Maximum Expressiveness of $\mathbf{B}$. As we have seen from the above, sparsity in $\mathbf{H}$ and expressiveness of $\mathbf{B}$ are closely related. The constraint presented here further enhances the maximum sparsity in (1). The idea is that only those components which carry much information about the training examples should be retained. The amount of information about example $\mathbf{x}_j$ carried by component $\mathbf{b}_i$ is measured by the "activity" of the example on the component defined as h_{ij}^2. The total activity of all examples on the component $\mathbf{b}_i$ is $\sum_{j=1}^{N_T} h_{ij}^2$. The total activity on all the learned components is $\sum_{i=1}^{m} \sum_{j=1}^{N_T} h_{ij}^2 = \sum_i v_{ii}$. The maximum expressiveness of $\mathbf{B}$ is imposed as $\sum_i v_{ii} = \max$.
(3) Maximum Orthogonality of $\mathbf{B}$. Different bases should be as orthogonal as possible, so as to minimize redundancy between different bases. This can be imposed by $\sum_{i \neq j} u_{ij} = \min$. Combining this with (1), we require $\sum_{\forall i,j} u_{ij} = \min$.

The incorporation of the above constraints leads the following constrained divergence as the objective function for LNMF:

$$D(\mathbf{X}||\mathbf{BH}) = \sum_{i,j}\left(x_{ij}\log\frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}\right) + \alpha\sum_{i,j}u_{ij} - \beta\sum_{i}v_{ii} \quad (5)$$

where $\alpha, \beta > 0$ are some constants (these constants will be eliminated in the derivation of a minimization algorithm). A local solution to the above constrained minimization, as an LNMF factorization, can be found by using the following three step update rules:

$$h_{kl} = \sqrt{h_{kl}\sum_i x_{il}\frac{b_{ik}}{\sum_k b_{ik}h_{kl}}} \quad (6)$$

$$b_{kl} = \frac{b_{kl}\sum_j x_{kj}\frac{h_{lj}}{\sum_k b_{kl}h_{lj}}}{\sum_j h_{lj}} \quad (7)$$

$$b_{kl} = \frac{b_{kl}}{\sum_k b_{kl}} \quad (8)$$

The derivation and a proof of the convergence is provided in [13].

2.3 Face Recognition in LNMF Subspace

Face recognition in the PCA, NMF or LNMF linear subspace is performed as follows where $\mathbf{B}^+ = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$:

1. Feature extraction. Each training face image $\mathbf{x}_i$ is projected into the linear space as a feature vector $\mathbf{h}_i = \mathbf{B}^+\mathbf{x}_i$ which is then used as a prototype feature point. A query face image $\mathbf{x}_q$ to be classified is represented as $\mathbf{h}_q = \mathbf{B}^+\mathbf{x}_q$.
2. Nearest neighbor classification. Some suitable distance between the query and each prototype, $d(\mathbf{h}_q, \mathbf{h}_i)$, is calculated. The query is classified to the class to which the closest prototype belongs.

3 Experiments

3.1 Data Preparation

The Cambridge ORL face database is used for deriving PCA, NMF and LNMF bases. There are 400 images (112×92) of 40 persons, 10 images per person (Fig.2 shows the 10 images of one person). The images are taken at different times, varying lighting slightly, facial expressions (open/closed eyes, smiling/non-smiling) and facial details (glasses/no-glasses). All the images are taken against a dark homogeneous background. The faces are in up-right position of frontal view, with slight left-right out-of-plane rotation. Each image is linearly stretched to the full range of pixel values of [0,255].

Figure 2: Face examples from ORL database.

The set of the 10 images for each person is randomly partitioned into a training subset of 5 images and a test set of the other 5. The training set is then used to learn basis components, and the test set for evaluate. All the compared methods take the same training and test data.

3.2 Learning Basis Components

LNMF, NMF and PCA representations with $25, 36, 49, 64, 81, 100, 121$ basis components are computed from the training set. The matlab package from `http://journalclub.mit.edu` is used for NMF. NMF converges about 5-times faster than LNMF. Fig.3 shows the resulting LNMF and NMF components for subspaces of dimensions 25 and 81. Higher pixel values are in in darker color; the components in each LNMF basis set have been ordered (left-right then top-down) according to the significance value v_{ii}. The NMF bases are as holistic as the PCA basis (eigenfaces) for the training set. We notice the result presented in [1] does not appear so, perhaps because the faces used for producing that result are well aligned. The LNMF procedure learns basis components which not only lead to non-subtractive representations, but also manifest localized features and thus truly parts-based representations. Also, we see that as the dimension (number of components) increases, the features formed in the LNMF components become more localized.

By applying the localization and orthogonal criteria for the basis component, we get a set binary bases. Figure 4 shows the histograms of the basis components of the LNMF (left) and NMF (right). It can be shown that the LNMF components are binary images. This leading to a Walsh function based represention of face images. Two important characteristics of the Walsh functions are their compactness (representing the lower order functions requires fewer samples), and the simplicity and quickness of their compution.

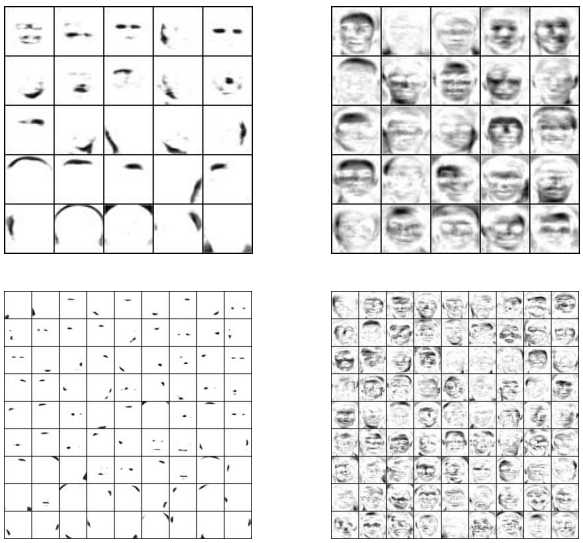

Figure 3: LNMF (left) and NMF (right) bases of dimensions 25 (row 1),81 (row 2). Every basis component is of size 112×92 and the displayed images are re-sized to fit the paper format. The LNMF representation is both parts-based and local, whereas NMF is parts-based but holistic.

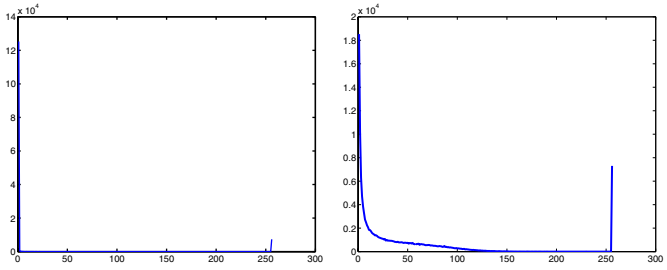

Figure 4: Histogram of the component basis LNMF (left) and NMF (right) bases of dimensions 81. The LNMF representation is binary, whereas NMF is not.

It leads to a significant dimension compression for the face images and improvement of computing efficiency (Binary vector can be computed very quickly on digital computers).

3.3 Reconstruction

Fig.5 shows reconstructions in the LNMF, NMF and PCA subspaces of various dimensions for a face image in the test set which corresponds to the one in the middle of row 1 of Fig.2. As the dimension is increased, more details are recovered. We see that while NMF and PCA reconstructions look similar in terms of the smoothness and texture of the reconstructed images, with PCA presenting better reconstruction quality than NMF. Surprisingly the LNMF representation, which is based on more localized features, provides smoother reconstructions than NMF and PCA.

Figure 5: Reconstructions of the face image in the (left to right) 25, 49, 81 and 121 dimensional (top-down) LNMF, NMF, and PCA subspaces.

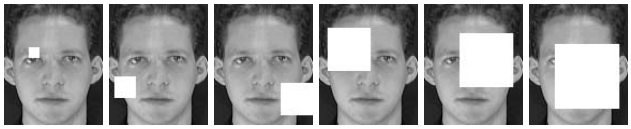

Figure 6: Examples of random occluding patches of sizes (from left to right) 10x10, 20x20, ..., 50x50, 60x60.

3.4 Face Recognition

The LNMF, NMF and PCA representations are comparatively evaluated for face recognition using the images from the test set. The recognition accuracy, defined as the percentage of correctly recognized faces, is used as the performance measure. Tests are done with varying number of basis components, with or without occlusion. The occlusion is simulated in an image by using a white patch of size $s \times s$ with $s \in \{10, 20, \ldots, 60\}$ at a random location; see Fig.6 for examples.

Figs.7 and 8 show recognition accuracy curves under various conditions. Fig.7 compares the three representations in terms of the recognition accuracies versus the number $m \times m$ of basis components for $m \in \{5, 6, \ldots, 10, 11\}$. The LNMF yields the best recognition accuracy, slightly better than PCA whereas the original NMF gives very low accuracy. Fig.8 compares the three representations under varying degrees of occlusion and with varying number of basis components, in terms of the recognition accuracies versus the size $s \times s$ of occluding patch for $s \in \{10, 20, \ldots, 50, 60\}$. As we see, although PCA yields more favorable results than LNMF when the patch size is small, the better stability of the LNMF representation under partial

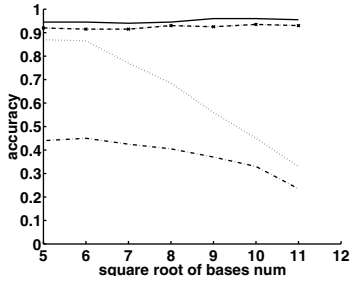

Figure 7: Recognition accuracies as function of the number (in 5x5, 6x6, ..., 11x11) of basis components used, for the LNMF (solid) and NMF (dashed) and PCA (dot-dashed) representations.

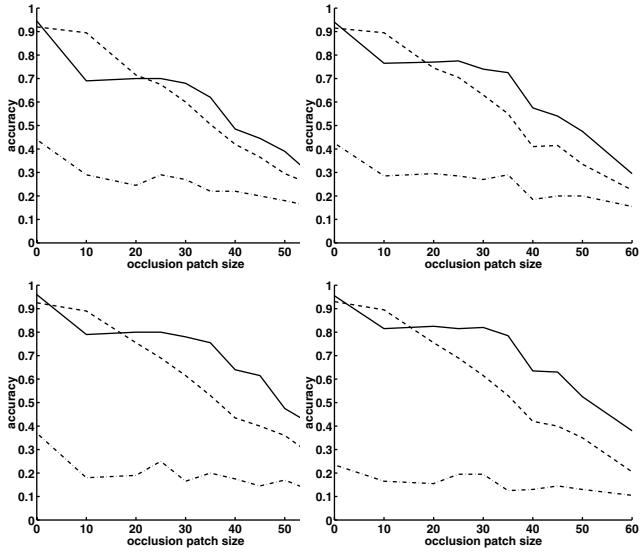

Figure 8: Recognition accuracies versus the size (in 10x10, 20x20, ..., 60x60) of occluding patches, with 25, 49, 81, 121 basis components (left-right, then top-down), for the LNMF (solid) and NMF (dashed) and PCA (dot-dashed) representations.

occlusion becomes clear as the patch size increases.

4 Conclusion

In this paper, we have proposed a new method, local non-negative matrix factorization (LNMF), for learning spatially localized, part-based subspace representation of visual patterns. The work is aimed to learn localized of features in NMF basis components suitable for tasks such as face recognition. An algorithms is presented for the learning and its convergence proved. Experimental results have shown that we have achieved our objectives: LNMF derives bases which are better suited for a localized representation than PCA and NMF, and leads to better recognition results than the existing methods.

The LNMF and NMF learning algorithms are local minimizers. They give different basis components from different initial conditions. We will investigate how this affects the recognition rate. Further future work includes the following topics. The first is to develop algorithms for faster convergence and better solution in terms of minimizing the objective function. The second is to investigate the ability of the model to generalize, *i.e.* how the constraints, the non-negativity and others, are satisfied for data not seen in the training set.The third is to compare with other methods for learning spatially localized features such as LFA [9] and ICA [12].

References

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

[2] L. Sirovich and M. Kirby, ""Low-dimensional procedure for the characterization of human faces"," *Journal of the Optical Society of America A*, vol. 4, no. 3, pp. 519–524, March 1987.

[3] M. Kirby and L. Sirovich, ""Application of the Karhunen-Loeve procedure for the characterization of human faces"," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, January 1990.

[4] Matthew A. Turk and Alex P. Pentland, ""Face recognition using eigenfaces."," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hawaii, June 1991, pp. 586–591.

[5] David Beymer, Amnon Shashua, and Tomaso Poggio, ""Example based image analysis and synthesis"," A. I. Memo 1431, MIT, 1993.

[6] A. P. Pentland, B. Moghaddam, and T. Starner, ""View-based and modular eigenspaces for face recognition"," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1994, pp. 84–91.

[7] H. Murase and S. K. Nayar, ""Visual learning and recognition of 3-D objects from appearance"," *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.

[8] K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, Boston, 2 edition, 1990.

[9] P. Penev and J. Atick, ""Local feature analysis: A general statistical theory for object representation"," *Neural Systems*, vol. 7, no. 3, pp. 477–500, 1996.

[10] C. Jutten and J. Herault, ""Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture"," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[11] P. Comon, ""Independent component analysis - a new concept?"," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[12] A. J. Bell and T. J. Sejnowski, ""The 'independent components' of natural scenes are edge filters"," *Vision Research*, vol. 37, pp. 3327–3338, 1997.

[13] S. Z. Li, X. W. Hou, and H. J. Zhang, ""Learning spatially localized, parts-based representation"," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hawaii, December 11-13 2001, p. ???

[14] D. D. Lee and H. S. Seung, ""Algorithms for non-negative matrix factorization"," in *Proceedings of Neural Information Processing Systems*, 2001, vol. 13, pp. 556–562.