HOLISTIC AND LOCAL IMAGE REPRESENTATIONS FOR HUMAN FACE ANALYSIS - I

Ioan Buciu[1], Ioannis Pitas[2], and Ioan Nafornita[3]

[1]Dept. of Electronics, University of Oradea
Universitatii 1, 410087, Oradea, Romania
phone: +40-259-408195, email: ibuciu@uoradea.ro
web: http://webhost.uoradea.ro/ibuciu/

[2]Dept. of Informatics, Aristotle University of Thessaloniki
GR – 541 24,Box 451, Thessaloniki, Greece
phone: +30-231-099-6361, email: pitas@aiia.csd.auth.gr
web: http://poseidon.csd.auth.gr/EN/

[3]Electronics and Communications Faculty, "Politechnica" University of Timisoara
Bd. Vasile Parvan nr.2 , 300223, Timisoara, Romania
phone: +40-256-40-3302, fax : +40-256-40-3301, email: ioan.nafornita@etc.upt.ro
web: http://hermes.etc.upt.ro/personal/nafornitaIoan.html

ABSTRACT

The mechanism of image processing at each level of the human visual system (HVS), mechanism which includes signal encoding at various HVS receptive fields (RFs) of the neural cells is one of the primary concerns of the neuropsychologists, neurophysiologists and even computer vision scientists. Two main theories exist with respect to the face analysis, encoding and representation in the HVS. The first one claims that the face is represented globally (known also as holistic) where the features have "holon"-like appearance. The second theory suggests that a more appropriate human face representation would be given by a sparse representation, where only a few neural cells are triggered. Although impressive work has been reported in the literature concerning both theories, no common agreement was established yet. However, a link exists, as, nowadays, the theoretical and experimental evidence brought evidence that the HVS performs face analysis (encoding, storing, face recognition, facial expression recognition) in a structured and hierarchical way, where both representations have their own contribution and goal. Basically, according to neuropsychological experiments, it is believed that, for face recognition, the encoding appears to be more likely global, while the sparse (or local) image representation is for the facial expression analysis and classification tasks. However, face and facial expression analysis is not only a concern from the neuropsychology field. Applications where the human face plays a central role are provided by facial biometrics and facial expression analysis. In the light of the computer vision perspective, various techniques developed by the computer scientists dealing with face and facial expression recognition fall in the same two image representation approaches. Accordingly, the findings from neuroscience are well correlated with the nature of image representation provided by the mathematical models of these techniques, i.e. the techniques which were found to perform better for face recognition yield a holistic image representation, contrary to those techniques which are more suitable for facial expression recognition and lead to a sparse or local image representation. This first part of the paper describes the most representative techniques for image representation concerning the holistic approaches in conjunction with face and facial recognition tasks, including authors' personal contribution to these areas.

1. INTRODUCTION

The bio-inspired mathematical models of image formation and encoding try to simulate the *efficient storing*, *organization* and *coding* of data in the human cortex. This is equivalent with embedding constraints in the model design regarding the dimensionality reduction, redundant information minimization, mutual information minimization, non-negativity constraints, class information, etc. The visual pathway is depicted in Figure 1. It starts from the retina and ends at the two regions of inferotemporal cortex – IT (PIT and AIT). Multiple representations of the retinal space are mapped onto the cortex in a manner that preserves the visual topology. These representations define the visual modules: V1, V2, V4, IT.

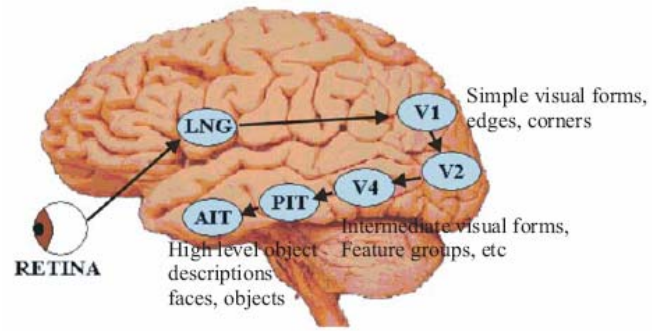

Figure 1. Visual pathway structure in HVS. Information passes from the retina to the lateral geniculate nucleus (LGN) before arriving in cortical area V1. Further processing occurs in areas V2 and V4 and the posterior and anterior inferotemporal (IT) cortex (PIT and AIT).

The type of image encoding is related to the number of neurons that are active (respond) to a certain piece of information represented by a specific sensory stimulus caused by the image. A ***dense*** image representation emerges if a *large* cell population with overlapping sensory input is activated and contributes to the image encoding. On the other hand, a system based on a holistic encoding suffers from slow training, requires heavy training and is likely to produce redundant image representations. Its main advantage is given by the large capacity of making new associations. Dense encoding is closely related to the holistic. The term holistic refers to an image representation which stores a face as a perceptual whole, without explicitly specifying its parts (components). The term component describes the separated parts of the face (e.g. eyes, nose, mouth, and chin) that are perceived independently as distinct parts of the whole.

The HVS often serves as an informal standard for evaluating systems. Therefore, not surprisingly, most face analysis approaches rely on bio-inspired models. To be plausible, these computer vision models have to share some characteristics and constraints with their organic models. A common characteristic of the proposed HVS models is the dimensionality reduction principle of image space. Dimensionality reduction operates by decomposing the high dimensional data into a lower dimensional subspace (yielding the so-called basis images) where each original image can be reconstructed by linearly (or nonlinearly) combination of the resulting basis images using the encoding coefficients. It is commonly accepted that the intrinsic dimensionality of the space of possible faces is much lower than that of the original image space. Basically, the latent variables incorporated there are discovered by decomposing (projecting) the image onto a linear (nonlinear) low dimensional image subspace. By reference to neuroscience, the receptive fields can be modeled by the basis images of the image subspace and their firing rates can be represented by the decomposition coefficients. Mathematically, the decomposition is described by $\mathbf{X} = \mathbf{ZH}$, where $\mathbf{X}$ is an m x n matrix comprising n images in its columns (original m-dimensional image space), $\mathbf{Z}$ is an m x p corresponds to the basis images (image subspace) and $\mathbf{H}$ denotes the encoding coefficients for the lower p - dimensional subspace (i.e. $p < m$).

2. HOLISTIC APPROACHES

One of the most popular techniques for dimensionality reduction is PCA [1], which represents faces by their projection onto a set of orthogonal axes (also known as principal components, eigenvectors, eigenfaces, or basis images) pointing into the directions of maximal covariance in the facial image data. By defining the covariance matrix with $\mathbf{C}_x = E\left\{(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T\right\}$ where $\boldsymbol{\mu}_x$ denotes the mean image. PCA solution is found by solving the equations system $\mathbf{C}_x\mathbf{Z}_{PCA} = \boldsymbol{\lambda}\mathbf{Z}_{PCA}$, with $\boldsymbol{\lambda}$ as eigenvectors.

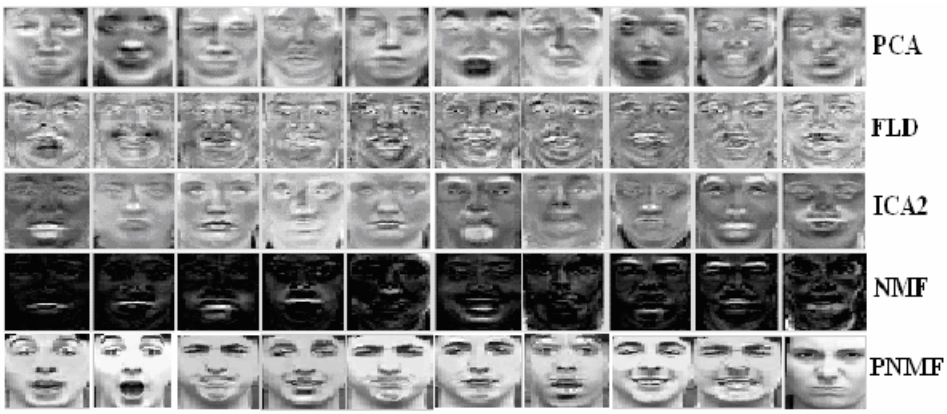

Figure 2. Holistic subspace image representation. From top to bottom, each row depicts 10 basis images (**Z**) corresponding to PCA, FLD, ICA2, NMF, and PNMF (degree 7). For PCA the basis images are ordered by decreasing variance, for ICA2 and NMF by decreasing kurtosis. Cohn-Kanade AU-coded facial expression database has been used as image samples.

The basis images corresponding to PCA are typically ordered according to the decreasing amount of variance they represent, i.e., the respective eigenvalues. Here $\mathbf{Z}_{PCA}$ comprises the eigenimages. PCA-based representations of human faces provides us a dense encoding and the post-processed images have holistic ("ghostlike") appearances, as drawn in the first row of figure 2. The principal components produce an image representation with minimal quadratic error. One of the proposed general organizational principles of the HVS refers to redundancy reduction. In PCA, this is guaranteed by imposing orthogonality among the basis images, thus redundancy is minimized. The nature of information encoded in the basis images was analyzed by O'Toole et al. [2]. They found that the first basis images (containing low spatial frequency information) were most discriminative for classifying gender and race, while the basis images with small eigenvalues (corresponding to a middle range of spatial frequencies) contain valuable in-formation for face recognition.

PCA has been successfully applied to face recognition [3], and facial expression recognition, respectively [4], [5]. One statistical limitation of PCA is that it only decorrelates the input data (second-order statistics) without addressing higher-order statistics between image pixels. It is well known and accepted that, at least for natural stimuli, important in-formation (e.g. lines, edges) is encoded in the higher-order statistics. Another limitation is related to the poor face recognition results for PCA when the faces are recorded under strong illumination variations. Another holistic subspace image representation is obtained by a class-specific linear projection method based on Fisher's linear discriminant (FLD) [6]. This technique projects the images onto a subspace where the classes are maximally separated by maximizing the between-classes scatter matrix and minimizing the within-class scatter matrix at the same time. If we denote the set of all N = | x | data divided into *c* classes with $X \equiv \{x_1, x_2, ..., x_c\}$, then the inter-class scatter matrix $\mathbf{S}_w$ is defined as $\mathbf{S}_w = \sum_{i=1}^{c} \sum_{x_k \in x_i} (x_k - \mu_i)(x_k - \mu_i)^T$ while the between-class scatter matrix $\mathbf{S}_b$ is defined as $\mathbf{S}_b = \sum_{i=1}^{c} | x_i | (\mu_i - \mu)(\mu_i - \mu)^T$, where μ_i is the mean image of class x_i and μ is the mean of all data. Here, $\mathbf{Z}_{FLD}$ satisfies

$$\mathbf{Z}_{FLD} = \arg\max_{Z} \frac{\left|\mathbf{Z}^T \mathbf{S}_B \mathbf{Z}\right|}{\left|\mathbf{Z}^T \mathbf{S}_w \mathbf{Z}\right|}.$$

The solution for finding Z_{FLD} is to solve the generalized eigeinvalues problems: $\mathbf{S}_B \mathbf{Z}_{LDA} = \lambda \mathbf{S}_w \mathbf{Z}_{LDA}$. The basis images obtained through FLD are depicted in the second row of figure 2. This approach has been shown to be efficient in recognizing faces, outperforming PCA. Although this method seems to be more robust than PCA when small variation in illumination conditions appears, it fails in case of strong illumination changes. This is due to the assumption of linear separability of the classes. This assumption is violated, when strong changes in illumination occur. Another drawback of this method is that it needs a large number of training image samples for reasonable performance. Furthermore, the projection onto too few subspace dimensions does not guarantee the linear class separability, hence the method will yield poor performance. Along with redundancy reduction, another principle of HVS image coding mechanism is given by phase information encoding. It was shown that methods relying only on second order statistics capture the amplitude spectrum of images but not the phase.

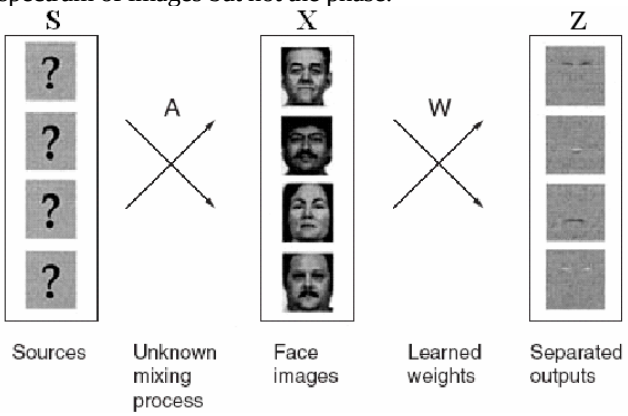

Figure 3. The fiducial features are considered as independent. What we see is the human faces **X** composed of those supposedly independent features (nose, mouth, eyebrows, etc) **S** mixed through **A** to form the whole face. By applying ICA, the goal is to identify the umixing matrix **W** to retrieve the independent features from **Z** [20].

The phase spectrum can be captured by employing higher order statistics as independent image components [7]. There are several optimization principles taken into account when extracting independent components. The one described in [7] is based on the maximal information transfer between neurons and, among all the proposed ICA techniques, it seems to be the most plausible approach from the neuroscientific point of view. Bartlett et al. [8] used two ICA configurations to represent faces for recognition. The general idea is depicted in Figure 3. PCA was carried out prior to ICA for dimensionality reduction. An intermediate step for "whitening" the data has been introduced between PCA and ICA processing. The data were then decomposed into basis images and decomposition coefficients. Their second ICA configuration (ICA2) yields holistic basis images very similar to those produced by PCA. Such basis images are depicted in the third row of figure 2. In that case, ICA is applied to the projection matrix containing the principal components. Under this architecture, the linear decomposition coefficients are as independent as possible.

A recently proposed subspace image decomposition technique is the Non-negative Matrix Factorization (NMF)

[9], which allows the data to be described as a combination of elementary features that involve only additive parts to form the whole. Both basis images and decomposition coefficients are constrained to be non-negative. Allowing only addition for recombining basis images to produce the original data is justified by the intuitive notion of combining parts to form the whole image. Another argument for imposing non-negativity constraints comes from neuroscience and is related to the non-negative firing rate of neurons. Finally, the non-negativity constraint arises in many real image processing applications. For example, the pixels in a grayscale image have non-negative intensities. Euclidean distance and *Kull-back-Leibler* (KL) divergence were originally proposed as objective functions for minimizing the difference between the original image data and their decomposition product expressed by:

$$f_{\mathrm{NMF}}(\mathbf{X} \,\|\, \mathbf{ZH}) = \sum_{i,j}\left(x_{ij} \ln \frac{x_{ij}}{\sum_k z_{ik} h_{kj}} + \sum_k z_{ik} h_{kj} - x_{ij} \right)$$

The factors are updated according to:

$$h_{kj}^{t} = h_{kj}^{t-1} \frac{\sum_i z_{ki} \dfrac{x_{ij}}{\sum_k z_{ik} h_{kj}^{t-1}}}{\sum_i z_{ik}} \quad z_{ik}^{t} = z_{ik}^{t-1} \frac{\sum_j \dfrac{x_{ij}}{\sum_k z_{ik}^{t-1} h_{kj}} h_{jk}}{\sum_j h_{kj}}$$

It has been noticed in several works that, for some databases, the NMF decomposition rather produces a holistic image representation. The representation could be affected by the inaccurate image alignment procedure performed on the original database prior to NMF.

More recently, a generalized NMF variant was developed in [10]. The approach called Polynomial Non-negative Matrix Factorization (PNMF) relies on the kernel for mapping the original data into a nonlinear feature space followed by a data decomposition process where both factors remain non-negative. High order dependency between the basis images is retrieved while keeping the non-negativity constraints on both basis images and coefficients. If the transformed data is denoted by $\mathbf{F} = [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), ..., \phi(\mathbf{x}_n)]$, with the l - dimensional vector $\phi(\mathbf{x}_j) = [\phi(\mathrm{x})_1, \phi(\mathrm{x})_2, ..., \phi(\mathrm{x})_l]^T \in \mathsf{F}$, a matrix $\mathbf{Y} = [\phi(\mathbf{z}_1), \phi(\mathbf{z}_2), ..., \phi(\mathbf{z}_n)]$ can be found that approximates the transformed set, such that, each vector $\phi(\mathbf{x})$ is describes as a linear combination as $\phi(\mathbf{x}) \approx \mathbf{Yh}$. The basis and the coefficients are then updated according to:

$$\mathbf{H}^{t} = \mathbf{H}^{t-1} .* (\mathbf{K}_{zx}^{t-1} ./ (\mathbf{K}_{zz}^{t-1} * \mathbf{H}^{t-1}))$$

$$\mathbf{B}^{t} = \mathbf{B}^{t-1} .* (\mathbf{X}\mathbf{K}_{xz}^{'t-1} ./ (\mathbf{B}^{t-1} * \mathbf{\Omega} * \mathbf{K}_{zz}^{'t-1})),$$

where $\mathbf{K}_{zx} = \langle \phi(\mathbf{z}_i), \phi(\mathbf{x}_i) \rangle$, $\mathbf{K}_{xz} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{z}_i) \rangle$ and $\mathbf{K}_{zz} = \langle \phi(\mathbf{z}_i), \phi(\mathbf{z}_o) \rangle$ are kernel matrices. Also, $\omega_{rr} = \sum_{j=1}^{n} H_{rj}, \quad r = 1, ..., p$ and $\mathbf{s}_r = \sum_{i=1}^{m} \mathbf{B}_{rj}, \quad r = 1, ..., p$. Here, the kernel function involved is the polynomial kernel, i.e. $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$.

A measure for quantifying the degree of sparseness in image representations is provided by the normalized kurtosis. If the basis images are stored as columns of a matrix $\mathbf{Z}$ the kurtosis of a base image z is defined as $k(\mathbf{z}) = \dfrac{\sum_i (z_i - \bar{z})^4}{\left(\sum_i (z_i - \bar{z})^2\right)^2} - 3$ where z_i are the elements of $\mathbf{z}$ (pixels of base image) and $\bar{z}$ denotes the sample mean of $\mathbf{z}$. The average normalized kurtosis for the 49 basis images are: $k_{PCA} = 1.22$, $k_{FLD} = 1.23$, $k_{ICA2} = 0.93$, $k_{NMF} = 5.93$, $k_{PNMF} = -0.4117$. Thus, by far, NMF is the sparsest representation among ones represented in Figure 2. A negative value for the PNMF's kurtosis, method which provides the most global image representation, indicates a sub-Gaussian distribution for the basis image entries.

REFERENCES

[1] P.J.B.Hancock, R.J.Baddeley, L.S.Smith, "The principal components of natural images", *Network Computation in Neural Systems* 3 (1), pp. 61 –70, 1992.

[2] A. O'Toole, H. Abdi, K. Deffenbacher, and D. Valentin, "Low-dimensional representation of faces in higher dimensions of the face space", *Journal of the Optical Society of America A*, 10(3), pp. 405–411, 1993.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *Cognitive Neuroscience*, 3(1), pp. 71–86, 1991.

[4] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vision Research*, 41, pp. 1179– 1208, 2001.

[5] C. Padgett and G. Cottrell, "Representing face images for emotion classification", *Advances in Neural Information Processing Systems*, vol. 9, pp. 894–900, 1997.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fish-erfaces: Recognition Using Class Specific Linear Projection", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 19(7), pp. 711-720, 1997.

[7] A.J.Bell and T.J.Sejnowski, "The "independent components" of natural scenes are edge filters", *Vision Research* 37, pp. 3327 –3338, 1997.

[8] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. Neural Networks*, 13(6), pp. 1450-1464, 2002.

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

[10] I. Buciu, N. Nikolaidis and I. Pitas, "Non-negative matrix factorization in polynomial feature space", *IEEE Transactions on Neural Networks*, [in Press], 2008.