Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
J. Vis. Commun. Image R. xxx (2006) xxx–xxx
JOURNAL OF VISUAL Communication & IMAGE Representation
www.elsevier.com/locate/jvci

NMF, LNMF, and DNMF modeling of neural receptive fields involved in human facial expression perception

I. Buciu *,[1], I. Pitas

Aristotle University of Thessaloniki, Department of Informatics, GR-541 24, Box 451, Greece

Received 9 November 2004; accepted 22 June 2006

Abstract

Recently, three learning algorithms, namely non-negative matrix factorization (NMF), local non-negative matrix factorization (LNMF), and discriminant non-negative matrix factorization (DNMF) have been proposed to produce sparse image representations. However, when their input is a database of human facial images, they decompose the images into sparse representations with quite different degree of sparseness. Within a continuum of sparseness ranging from holistic to local image representation, the first algorithm rather tends towards the first extreme, while the second algorithm produces a local representation. The third algorithm provides an image representation that is in between these two extremes. These algorithms decompose the facial images in the database into basis images and their corresponding coefficients. The basis images are learned by the algorithm when human face images are given as input. By analogy to neurophysiology, the basis images could be associated with the receptive fields of neuronal cells involved in encoding human faces. Taken from this point of view, the paper presents an analysis of these three representations in connection to the receptive field parameters such as spatial frequency, frequency orientation, position, length, width, aspect ratio, etc. By analyzing the tiling properties of these bases we can have an insight of how suitable these algorithms are to resemble biological visual perception systems.

Keywords: Image representation; Receptive fields; Facial expressions

1. Introduction

Understanding how the image is processed at each level of the human visual system in order to be transformed into this signal and the type of signal encoding at the receptive fields (RFs) of the neural cells is one of the primary concerns of the neuropsychologists and neurophysiologists. Nowadays, the theoretical and experimental evidence suggests that the human visual system performs object (including face) recognition processing in a structured and hierarchical approach in which neurons become selective to process progressively more complex features of the image structure [1]. Whereas neurons from visual area 1 (V1) are responsible for pro-

* Corresponding author.
E-mail addresses: nelu@zeus.csd.auth.gr (I. Buciu), pitas@zeus.csd.auth.gr (I. Pitas).
[1] On leave from Applied Electronics Department, University of Oradea, 3700 Armatei Romane, No. 5, Romania.

1047-3203/$ - see front matter
doi:10.1016/j.jvcir.2006.06.001

cessing simple visual forms, such as edges and corners (leading to a very sparse image representation), neurons from the visual area 2 (V2) process a larger visual area representing feature groups. As we further proceed to the visual area 4 (V4) and the inferotemporal cortex (IT), we meet neurons having large receptive fields that respond to high- level object descriptions such as ones describing faces or objects. This is equivalent with a decrease in image representation sparseness. While holistic representation treats an image as a whole (global feature) where each pixel has major contribution to representation, sparse representation is characterized by a highly kurtotic distribution, where a large number of pixels have zero value, and small number of pixels have positive or negative values (local features). In its extreme, sparse representation provides a local image representation having only just a very small amount of contributing pixels. Finally, the IT area of the temporal lobe contains neurons whose receptive fields cover the entire visual space. It also contains specialized neurons (face cells) that are selectively tuned for faces. There is now good evidence that there are dedicated areas in temporal cortical lobe that are responsible to process information about faces [2–4]. Moreover, it was found that there are neurons (located in TE areas) with responses related to facial identity recognition, while other neurons (located in the superior temporal sulcus) are specialized only to respond to facial expressions [5].

Models of receptive fields of neuronal cells have been proposed by numerous researchers. There are two types of neural cells: simple and complex ones. It has been shown by Olsahusen and Field [6] that in V1 area the simple cells produces a sparse coding of natural images. Their receptive fields respond differently to visual stimuli having different spatial frequencies, orientations, and directions. Marcelja [7] and Daugman [8] have noticed that the receptive fields of simple cells can be well described by 2D Gabor functions. The main drawback of Gabor function models is that they have many free parameters to be tuned "by hand" in order to tile the joint space or spatial frequency domain to form a complete basis for image representation. Other attempts to model the structure of V1 receptive fields were based on Principal Component Analysis (PCA), which leads to holistic image representation [9,10] and independent component analysis (ICA) [11].

Although the receptive fields of V1 seem to be well described by the models proposed above, there is no conclusive model for cells of the higher cortical levels, especially for face cells. Here, we make an analysis of three recent image representation algorithms namely, the non-negative matrix factorization (NMF), the local non-negative matrix factorization (LNMF), and the discriminant non-negative matrix factorization (DNMF). They all decompose an image database into basis images and the corresponding coefficients. The model, as it is described here, associates the basis images with the receptive fields of neural cells and the coefficients with their firing rates. In particular, we are interested in the representation of facial expression images. We propose a biological plausible model for the facial neurons responsible for biological facial expression recognition. From the computer vision point of view, in this paper we analyze the parameters of the resulting basis images, such as spatial frequency, frequency orientation, position, length, width, aspect ratio, etc., in analogy to the parameters of the spatial neural receptive fields. The analysis of the basis images characteristics to be presented in this paper is motivated by the performance of DNMF algorithm in classifying facial expressions [12]. However, since some constraints are common for these three algorithms, NMF and LNMF are analyzed as well. The results can show us how suitable are these algorithms for modeling biological facial perception systems.

The remaining of the paper is organized as follows. In Section 2, the facial image model is described along with the ways of representing the human face. A brief description of the algorithms investigated in this paper is presented in Section 3. The parameters of the receptive fields obtained through the learned basis images with NMF, LNMF, and DNMF are analyzed in Section 4. The paper ends with some comments and conclusions drawn in Section 5.

2. Human face representation

Face analysis has captured an increased attention from psychologists, anthropologists, and computer scientists due to its special applications in biometrics or human–computer interaction. Research have been done to find the way a human face is represented in the visual system. Both sparse and dense human face representations have been found by neuropsychologists and neurophysiologists to mode aspects of the human visual system. However, as the final goal is either to recognize a face or a particular facial expression, these two representations have their own contribution. While face recognition appears to rely more on a dense image representation (hence producing a holistic appearance of the faces), the information for facial expression

seems to be captured by a more sparse (or even a local) face representation. This difference has been noticed in several research works. Psychologically, the theory that biologically face recognition is a holistic process was explored by Tanaka and Farah [13]. Accordingly, biological face recognition does not simply use parts of the face, but rather the face is perceived as a whole. Their theory is supported by the work of Dailey and Cottrell [14]. Atick and Redlich have demonstrated that receptive fields of retinal ganglion cells can be viewed as local "whitening" filters that remove second-order statistics between pixels in images in a way similar to that of PCA [15]. The use of principal components is consistent with psychological evidence that PCA accounts for some aspects of human memory performance, as shown in the work of Valentine [16].

Contrary to face representation for biological face recognition, the work of Ellison and Massaro [17] has revealed that the facial expressions are better represented by parts of the face, thus suggesting a non-holistic representation. This is consistent with research results showing that humans respond to information around the eyes independently from motion in the mouth area and that they are able to recognize and distinguish isolated parts of faces. The dissociation between face and facial expression recognition is also noted in the paper of Cotrell et al. [18] who found that PCA (that produce eigenfaces) performs well for face recognition but eigeneyes and eigenmouth (eigenfeatures that are not holistic) perform better in recognizing expressions than eigenfaces, suggesting that non-holistic eigenfeatures might be used to recognize expressions. One of the approaches that has been successfully applied to classify facial actions was ICA. When applied to natural scenes, this approach looks for image components that are as independent as possible from the rest and have similar properties to V1 neural receptive fields, such as orientation selectivity, bandpass, and scaling properties [11]. In a direct comparison between PCA and ICA, Draper et al. [19] found that an holistic approach (PCA) performs the best for face recognition while an approach based on more localized features (ICA) performs better for facial action recognition.

Three sparse image representation methods namely NMF, LNMF, and DNMF have been recently investigated with respect to their performance in facial expression classification [12]. Although they produce sparse representation, their degree of sparseness is quite different. NMF representations are rather holistic (compared to the other two ones), as proven by their small kurtosis value and by visual inspection and has the worst performance in classifying facial expressions [12]. The basis images learned by that algorithm produced localized, oriented, and bandpass Gabor-like features. LNMF produces a rather local image representation while DNMF is situated between NMF and LNMF. As far as the facial expression recognition performance is concerned, DNMF outperforms LNMF approach [12]. This fact might indicate that those features that are important in recognizing facial expression are lost in the case of LNMF in its attempt to obtain a local image representation. Our basic statement supported in the paper is that human facial expression recognition is best modeled by DNMF representation that provides sparse and intuitively meaningful facial image representations. The spatial and frequency characteristics of this representation is elaborated in this work.

3. NMF, LNMF, and DNMF image model representations

Let us suppose that an image is represented by a vector $\mathbf{x}$ having m pixels, $\mathbf{x} = [x_1, \ldots, x_m]^{\mathrm{T}}$. Then, $\mathbf{x}$ can be decomposed in the product of a $m \times p$ matrix $\mathbf{Z}$, whose columns comprised basis functions $\mathbf{z}$, and the coefficients vector, $\mathbf{h} = [h_1, \ldots, h_p]^{\mathrm{T}}$ [6]

$$\mathbf{x} = \mathbf{Zh}. \tag{1}$$

Here, p is the number of images in the database. The choice of basis functions determines the image representation. To perform this decomposition we need to determine the basis functions and their coefficients. From (1) we have

$$\mathbf{h} = \mathbf{Wx}, \tag{2}$$

where $\mathbf{W}$ is a matrix whose rows are the inverse filters. Generally, from (1) and (2) we have that $\mathbf{W} = \mathbf{Z}^{-1}$. When $\mathbf{Z}$ forms orthonormal basis we have $\mathbf{W} = \mathbf{Z}^{\mathrm{T}}$, where T denotes the transpose operator. In biological terms, this decomposition model can be interpreted as follows. The neural cells perform a fully distributed or a sparse coding of the stimulus (image) presented at input in such a way that their neural receptive fields are modeled by the inverse p filters $\mathbf{w}$ of the model and their firing rate is represented by the model coefficients

h. Here, we have limited ourselves to the case where the stimulus is a human face. Non-negative matrix factorization (NMF), as it was proposed by Lee and Seung, is a method that decomposes a given $m \times n$ non-negative matrix **X** into non-negative factors **Z** and **H** such as $\mathbf{X} \approx \mathbf{ZH}$, where **Z** and **H** are matrices of size $m \times p$ and $p \times n$, respectively [20]. Suppose that $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, p$. Then, each element x_{ij} of the matrix **X** can be written as $x_{ij} \approx \sum_k z_{ik} h_{kj}$. The quality of approximation depends on the objective function used. One of the objective functions that can be used is represented by the Kullback–Leibler divergence between **X** and **ZH** [21]

$$D_{\mathrm{NMF}}(\mathbf{X}||\mathbf{ZH}) \triangleq \sum_{i,j} \left(x_{ij} \ln \frac{x_{ij}}{\sum_k z_{ik} h_{kj}} + \sum_k z_{ik} h_{kj} - x_{ij} \right). \tag{3}$$

This expression can be minimized by applying multiplicative update rules subject to $\mathbf{Z}, \mathbf{H} \geqslant 0$. This constraint is natural in many real image processing applications. For example, the grayscale image pixels have non-negative intensities. From biological perspective, its proposers imposed non-negative constraints, partly motivated by the biological aspect that the firing rates of neurons are non-negative. It has been shown that, if the matrix **X** contains images from an image database (one in each matrix column), then the method decomposes them into basis images (columns of **Z**) and the corresponding coefficients (rows of **H**) [20]. The resulting basis images contain parts of the original images, parts that are learned thorough the iterative process in the attempt of approximating **X** by the product **ZH**. In this context, m represents the number of image pixels, n is the total number of images, and p is the number of basis images. The following updating rules for finding the factors h_{kj} and z_{ik} are applied alternatively at each iteration t in an expectation-maximization (EM) manner [21]:

$$h_{kj}^{(t)} = h_{kj}^{(t-1)} \frac{\sum_i z_{ki}^{(t)} \frac{x_{ij}}{\sum_k z_{ik}^{(t)} h_{kj}^{(t-1)}}}{\sum_i z_{ik}^{(t)}}, \tag{4}$$

$$z_{ik}^{(t)} = z_{ik}^{(t-1)} \frac{\sum_j \frac{x_{ij}}{\sum_k z_{ik}^{(t-1)} h_{kj}^{(t)}} h_{jk}^{(t)}}{\sum_j h_{kj}^{(t)}}. \tag{5}$$

They guarantee a nonincreasing behavior of the KL divergence.

Local non-negative matrix factorization (LNMF) has been developed by Li et al. [22]. This technique is a version of NMF which imposes more constraints on the cost function (3) to increase the degree of image representation sparseness. Therefore, the character of the learned basis images is improved. If we use the notations $[\mathbf{u}_{ij}] = \mathbf{U} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ and $[\mathbf{v}_{ij}] = \mathbf{V} = \mathbf{HH}^{\mathrm{T}}$, the following three additional constraints can be imposed on the NMF basis images and decomposition coefficients:

(1) $\sum_i u_{ii} \to \min$ (maximum sparsity in **H**). This guarantees the generation of more localized features in the basis images **Z**, than those resulting from NMF, since, we impose the constraint that basis image elements are as small as possible.
(2) $\sum_{i \neq j} u_{ij} \to \min$ (maximum orthogonality in **B**). This enforces basis orthogonality, in order to minimize the redundancy between image bases.
(3) $\sum_i v_{ii} \to \max$ (maximum expressiveness in **B**). By means of this constraint, the total energy of the projection coefficients (total squared projection coefficients summed over all training images) is maximized. The new objective function takes the following form:

$$D_{\mathrm{LNMF}}(\mathbf{X}||\mathbf{ZH}) \triangleq D_{\mathrm{NMF}}(\mathbf{X}||\mathbf{ZH}) + \alpha \sum_{ij} u_{ij} - \beta \sum_i v_{ii}, \tag{6}$$

where $\alpha, \beta > 0$ are constants. A solution for the minimization of relation (6) can be found in [22]. Accordingly, if we use the following update rules for image basis and coefficients:

$$h_{kj}^{(t)} = \sqrt{h_{kj}^{(t-1)} \sum_i z_{ki}^{(t)} \frac{x_{ij}}{\sum_k z_{ik}^{(t)} h_{kj}^{(t-1)}}}, \tag{7}$$

$$z_{ik}^{(t)} = \frac{z_{ik}^{(t-1)} \sum_j \frac{x_{ij}}{\sum_k z_{ik}^{(t-1)} h_{kj}^{(t)}} h_{jk}^{(t)}}{\sum_j h_{kj}^{(t)}}, \tag{8}$$

$$z_{ik}^{(t)} = \frac{z_{ik}^{(t)}}{\sum_i z_{ik}^{(t)}}, \quad \text{for all } k \tag{9}$$

the KL divergence is nonincreasing.

NMF and LNMF algorithms do not take into account image class information and treat all images the same way. By modifying the coefficients $\mathbf{H}$ in a such a way that the basis images incorporate class characteristics, we obtain a class-dependent image representation. This is the discriminant non-negative matrix factorization (DNMF) approach [12]. Let us suppose we have $\mathcal{Q}$ distinctive image classes and let n_c be the number of training samples in class $\mathcal{Q}$, $c = 1, \ldots, \mathcal{Q}$. DNMF preserves the LNMF constraints on the basis images and introduces two more constraints on the coefficients $\mathbf{h}_{cl}$, where $c = 1, \ldots, \mathcal{Q}$ and $l = 1, \ldots, n_c$. These are: (1) $\mathbf{S}_w = \sum_{c=1}^{\mathcal{Q}} \sum_{l=1}^{n_c} (\mathbf{h}_{cl} - \boldsymbol{\mu}_c)(\mathbf{h}_{cl} - \boldsymbol{\mu}_c)^{\mathrm{T}} \rightarrow \min$, where $\mathbf{S}_w$ is the within-class scatter matrix and defines the scatter of the projection coefficients of each class around their mean. This dispersion should be as small as possible. (2) $\mathbf{S}_b = \sum_{c=1}^{\mathcal{Q}} (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}} \rightarrow \max$, $\mathbf{S}_b$ denotes the between-class scatter matrix of the projection coefficients and defines the scatter of their class mean around their global mean $\boldsymbol{\mu}$. Each cluster formed by the projection coefficients that belong to the same class must be as far as possible from the other clusters. Here, $\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{l=1}^{n_c} \mathbf{h}_{cl}$ represents the mean vector of class c, $\boldsymbol{\mu} = \frac{1}{n} \sum_{c=1}^{\mathcal{Q}} \sum_{l=1}^{n_c} \mathbf{h}_{cl}$ is the global mean vector. The new objective function is expressed as:

$$D_{\mathrm{DNMF}}(\mathbf{X}||\mathbf{ZH}) \triangleq D_{\mathrm{LNMF}}(\mathbf{X}||\mathbf{ZH}) + \gamma \sum_{c=1}^{\mathcal{Q}} \sum_{l=1}^{n_c} (\mathbf{h}_{cl} - \boldsymbol{\mu}_c)(\mathbf{h}_{cl} - \boldsymbol{\mu}_c)^{\mathrm{T}} - \delta \sum_{c=1}^{\mathcal{Q}} (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^{\mathrm{T}}, \tag{10}$$

where γ and δ are constants. Following the same EM approach used by NMF and LNMF techniques, each element h_{kj} of the coefficients matrix $\mathbf{H}$ is updated as [12]

$$h_{kl(c)}^{(t)} = \frac{2\boldsymbol{\mu}_c - 1 + \sqrt{(1 - 2\boldsymbol{\mu}_c)^2 + 8\xi h_{kl(c)}^{(t-1)} \sum_i z_{ki}^{(t)} \frac{x_{ij}}{\sum_k z_{ik}^{(t)} h_{kl(c)}^{(t-1)}}}}{4\xi}. \tag{11}$$

The elements h_{kl} are then concatenated for all $\mathcal{Q}$ classes as

$$h_{kj}^{(t)} = [h_{kl(1)}^{(t)} | h_{kl(2)}^{(t)} | \ldots | h_{kl(\mathcal{Q})}^{(t)}], \tag{12}$$

where "|" denotes concatenation and $\xi = \gamma - \beta$. The expression for updating the basis image remains unchanged from LNMF. Class-dependent image representation obtained by DNMF is very useful when it comes to classification. Basically, the images are projected into the basis images and the new features are further classified by a classifier [12]. For visualization purpose, Fig. 1 displays the projection of images (which belong to the Cohn–Kanade facial database) coming from three expression classes (anger, disgust, surprise) on the first two basis images shown in Fig. 3. Let us denote by $M1$, $M2$, and $M3$ the mean of the three clusters formed by these projections and the distance between the means by d_{12}, d_{13}, and d_{23}, respectively. Then, for this metric space we have $d_{12} = 4.3$, $d_{13} = 6.7$, and $d_{23} = 7.9$ in the case of NMF, $d_{12} = 11.2$, $d_{13} = 8.2$, and $d_{23} = 18.2$ for LNMF and $d_{12} = 35.8$, $d_{13} = 52.4$, and $d_{23} = 66.9$ for DNMF approaches respectively. The between—classes similarity is larger for DNMF than for the other two approaches. For simplicity in Fig. 1 is shown only M2 and M3 and the distance between them is drawn by a line. It can be noticed that the classes do not overlap in the case of DNMF as much as they do in the case of NMF and LNMF methods.

4. Receptive fields modeled by NMF, LNMF, and DNMF

We trained NMF, LNMF, and DNMF on a database consisting of facial expressions derived from Cohn–Kanade AU-coded facial expression database [23]. The facial action (action units) that are described in the image annotations have been converted into emotion class labels according to [24]. Fig. 2 depicts the results,

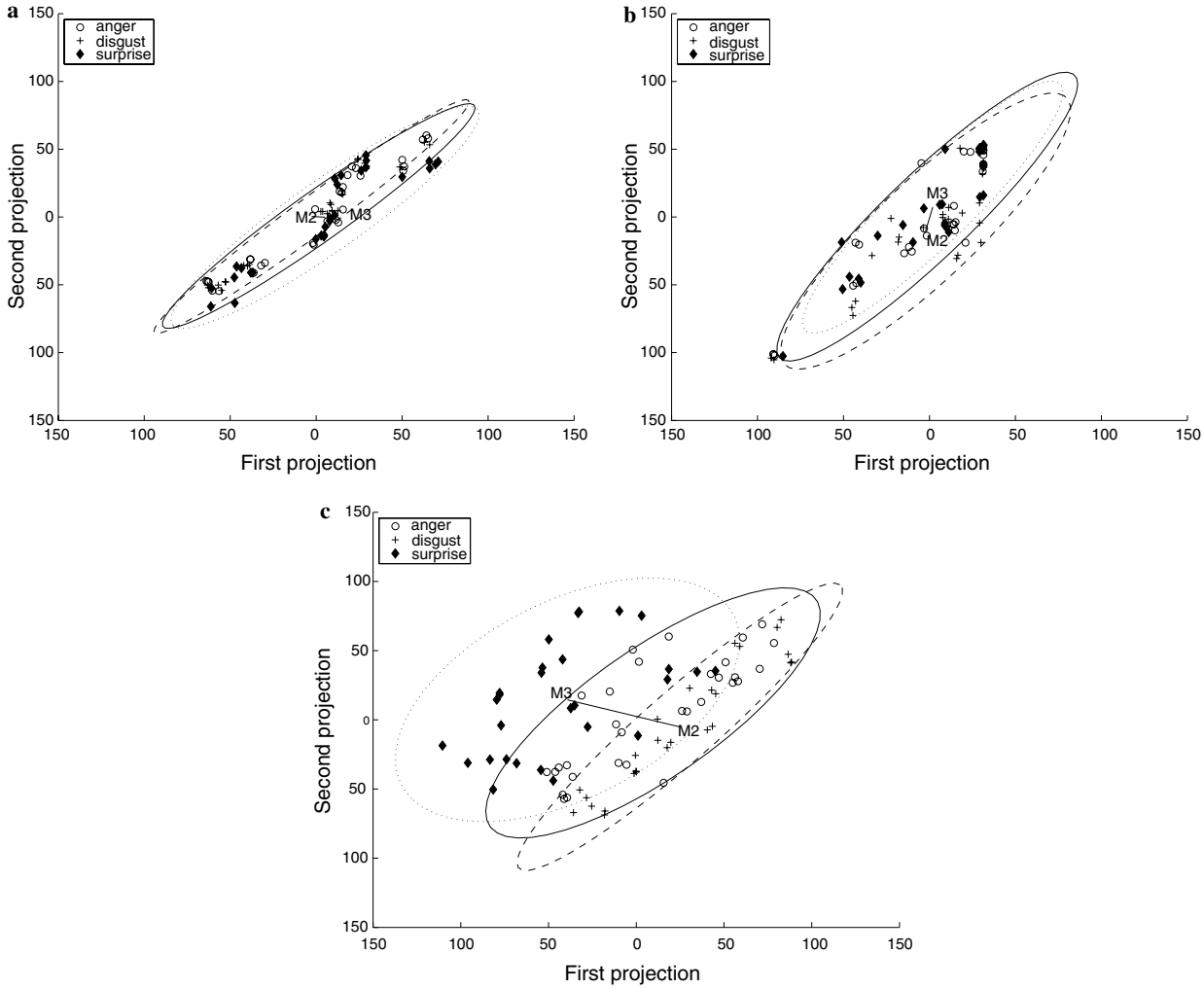

Fig. 1. Scatter plot of the clusters formed by the projection of three expression classes (anger, disgust, surprise) on the first two basis images shown in Fig. 3 for (a) NMF, (b) LNMF, and (c) DNMF. $M2$ and $M3$ represents the mean of the clusters corresponding to "disgust" and "surprise" classes and the distance between these means is drawn by a line. The ellipse encompasses the distribution with a confidence factor of 90%.

as they are reported in [12], obtained on the facial expression recognition task corresponding to the above mentioned database for the all three algorithms.

We worked on a subspace of 144 basis images ($p = 144$). Once the basis images are calculated we compute the 144 inverse filters $\mathbf{W} = \mathbf{Z}^{-1}$ (to be called receptive field (RF) masks) corresponding to the basis images for all three algorithms. Twenty five receptive field masks for NMF, LNMF, and DNMF are shown in Fig. 3. As can be seen from the Fig. 3a, NMF produces neither oriented nor localized masks. The features discovered by NMF have a larger space coverage than those obtained by LNMF or DNMF, thus capturing redundant information. On the contrary, the LNMF receptive field masks are oriented and localized. Mask domain denotes the mask region where mask coefficients are large (above a certain threshold). Some of them have domain of almost a single pixel. Neurophysiologically, one single pixel representation is similar of having a grandmother cell where a specific image is represented by one neuron (with a very small receptive field size). Furthermore, the features discovered by LNMF have rather random position in the image domain. Receptive field masks produced by DNMF are sparse but contain less localized and oriented domain than LNMF. In addition it contains non-oriented features. Probably the most important issue related to the DNMF RFs masks is the

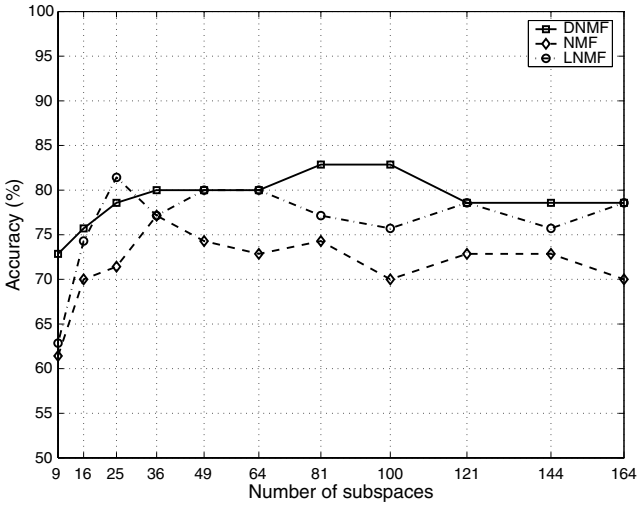

Fig. 2. Accuracy (correct classification in percentage) achieved on the facial expression recognition task corresponding to the Cohn–Kanade database for the NMF, LNMF, and DNMF algorithm, respectively, and for different number of subspaces (p). Details of the experiment can be found in [12].

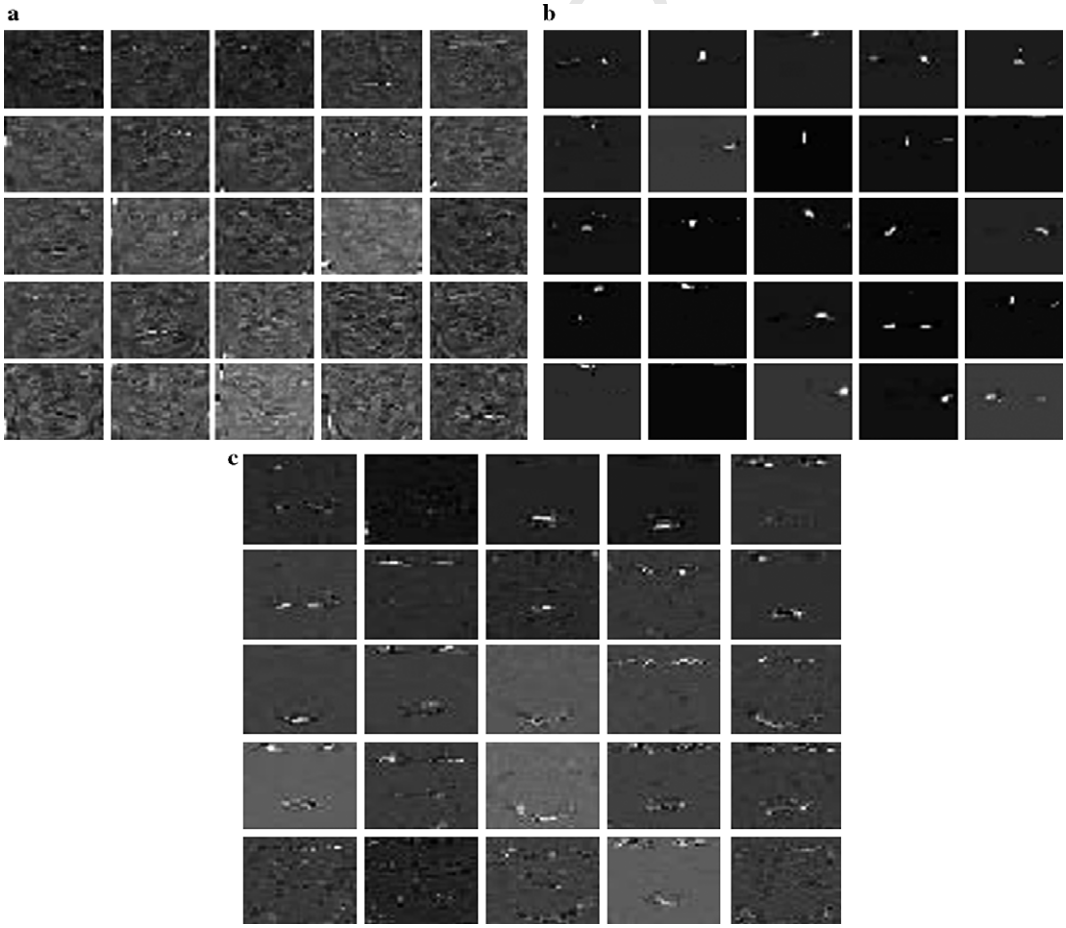

Fig. 3. Sample receptive field masks corresponding to basis images learned by (a) NMF, (b) LNMF, and (c) DNMF. They were ordered according to a decreasing degree of sparseness.

fact that almost all their domain correspond to salient face features such as eyes, eyebrows, or mouth that are of great relevance to facial expressions. While discarding less important information (e.g., nose and cheeks, which is not the case for NMF), DNMF preserves local spatial information of salient facial features (that are almost absent in the case of LNMF). The preservation of the spatial facial topology correlates well with the findings of Tanaka et al. [25] who argued that some face cells require the correct spatial facial feature configuration in order to be activated for facial expression recognition. We have noticed in our experiments that the degree of sparseness corresponding to basis images extracted by DNMF did not increase after a number of iterations. We believe this is caused by those patterns in the basis images that encode meaningful class information (such as those corresponding to salient facial features) and they cannot be disregarded as the iterations proceed further. The degree of RF masks sparseness can be quantified by measuring the normalized kurtosis of a base image $\mathbf{z}$ (one column of $\mathbf{Z}$) defined as $k(\mathbf{z}) = \frac{\sum_i (z_i - \bar{z})^4}{(\sum_i (z_i - \bar{z})^2)^2} - 3$, where z_i are the mask pixels and $\bar{z}$ denotes the sample mean of $\mathbf{z}$. The average kurtosis for the three representations over 144 basis images are: $\bar{k}_{\mathrm{NMF}} = 7.51$, $\bar{k}_{\mathrm{LNMF}} = 152.89$, and $\bar{k}_{\mathrm{DNMF}} = 22.57$.

We have described the spatial distribution of the receptive field masks in terms of 4 spatial parameters: average domain location (x, y), domain orientation (0°, 90°, 45°, and 135°, respectively) directions, and aspect ratio. The aspect ratio is defined as l/w, where l and w are the length and width of the receptive fields calculated as follows [26]:

$$\begin{aligned} l_k &\equiv \sqrt{\sum_{x,y} (x\sin(\theta) + y\cos(\theta))^2 \bar{\mathbf{z}}_k^2}, \\ w_k &\equiv \sqrt{\sum_{x,y} (x\cos(\theta) - y\sin(\theta))^2 \bar{\mathbf{z}}_k^2} \end{aligned} \tag{13}$$

over (x, y) image space. These RF masks domain parameters calculated over the facial image database are represented in Fig. 4. We can notice in Fig. 4a that the RF masks do not cover the entire space. For NMF and DNMF they are centrally distributed and cover the image center which is in par with a similar characteristic of V4 receptive fields. LNMF features are rather distributed marginally as shown in Fig. 4a. Unlike NMF, where domain orientation is at oblique angles (45° and 135°), LNMF emphasizes more horizontal and vertical features. DNMF puts approximately the same emphasis on horizontal and oblique features and slightly less stress on vertical ones. The oblique features are represented due to the chin contour (as it can be seen from Fig. 3c) where DNMF acts like a local edge detector.

The aspect ratio of NMF ranges from 0.6 to 1.6 with a mean at 1.09 and a standard deviation of 0.19. LNMF aspect ratios range from 2 to 11 with a mean at 1.65 and standard deviation 2.04. DNMF aspect ratios range from 0.5 to 2.2 with mean 1.03 and standard deviation 0.26. The higher average aspect ratio of LNMF indicates that its receptive fields are more elongated horizontally then those of NMF or DNMF.

To characterize the frequency distribution of RF masks we have computed their spatial frequency and orientations from their Discrete Fourier Transform: $F_k(u, v) = \frac{1}{NM}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\mathbf{z}(x, y)\exp[-j2\pi(ux/N + vy/M)]$, where $u = 1, \ldots, N-1$ and $v = 1, \ldots, M-1$ are spatial frequency coordinates in the horizontal and vertical directions, respectively, expressed in cycles/image and N and M are the number of rows and columns in the basis image, respectively. The two-dimensional spatial frequency are represented in polar coordinates (r, φ), where r denotes the absolute spatial frequency, φ orientation, $u = r\cos(\varphi)$ and $v = r\sin(\varphi)$. Thus, the optimal spatial frequency (orientation) is defined as the spatial frequency (orientation) of the peak in the amplitude (phase) spectrum.

Fig. 5 presents the optimal spatial frequency and optimal orientation for NMF, LNMF, and DNMF receptive field masks found by taking the peak of the spectrum. Figs. 5a–c indicate that the features are evenly spread in all orientation in the frequency domain for all three representation studied. Regarding radial spectrum distribution, NMF shows peak at a high spatial frequency bands (approximately between 0.7 and 0.9 cycles/image) as shown in Fig. 5d. LNMF features are distributed within a lower frequency band (of 0.25–0.45 cycles/image) as shown in Fig. 5e. A bandpass spectrum shape is shown by DNMF in Fig. 5f. The RFs power spectrum covers a larger spatial frequency band at [0.45, 0.8] cycles/image, capturing a larger radial spectrum.

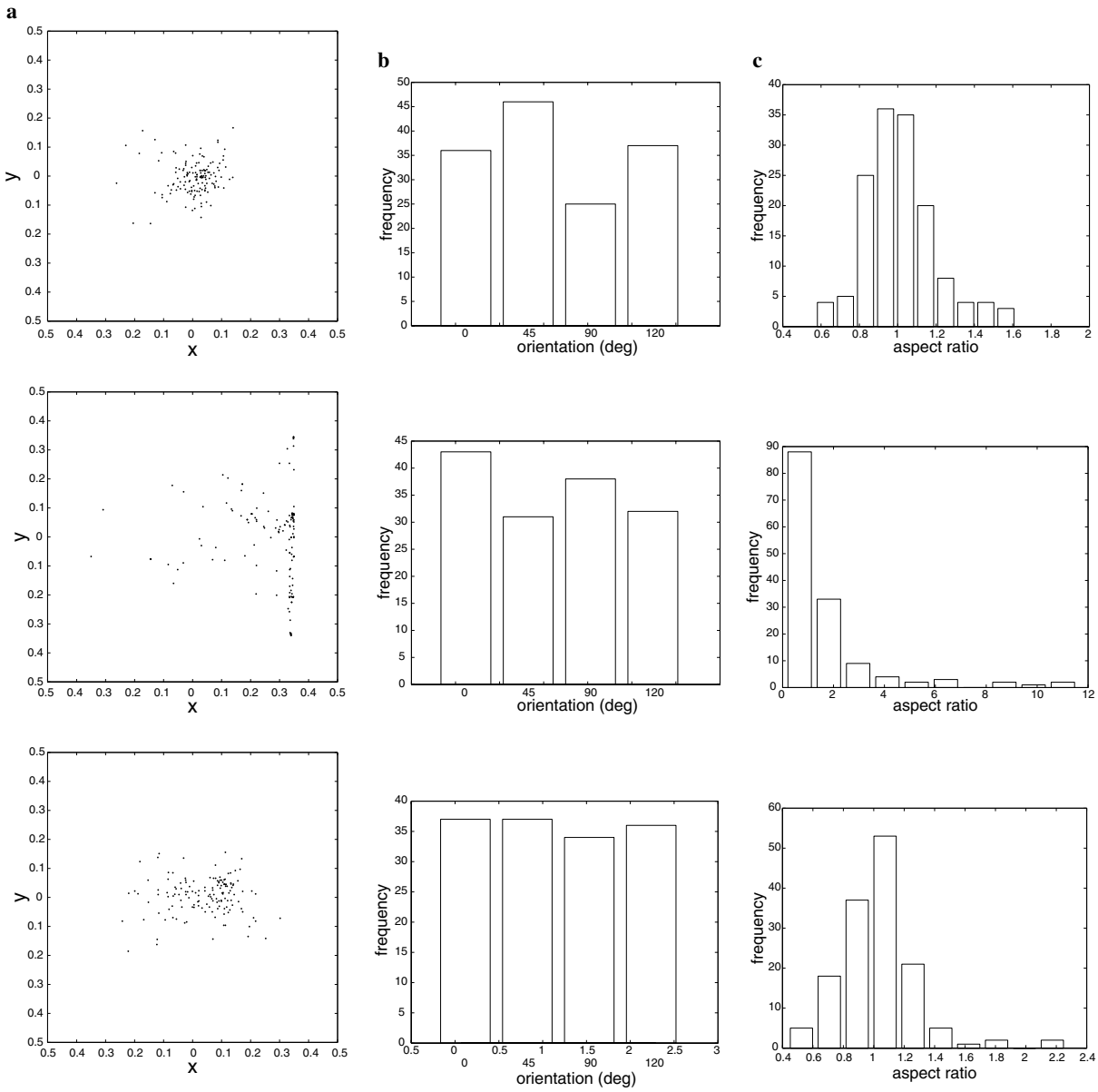

Fig. 4. Spatial characteristics or FS masks domain for NMF (top), LNMF (middle), and DNMF (bottom) receptive fields (RFs): (a) average location of RF domain; (b) histogram of RF domain orientations in degrees (0°, 45°, 90°, and 135°) and (c) length-to-width aspect ratio of RF spatial domain.

NMF, LNMF, and DNMF receptive fields show a low, high and bandpass frequency spectrum, respectively. Redundancy reduction is also obtained by suppressing the low spatial frequency in order to whiten the power spectrum of images, therefore this is done by highpass filtering [29]. This is consistent with what LNMF performs through $\sum_{i \neq j} u_{ij} \rightarrow \min$, and, thus having receptive fields similar to highpass filters (see Fig. 5a and Fig. 5d). On the other hand, the high frequency components contain only little power from the image source and, therefore, it is not robust to noise. To avoid this, highpass frequency must be eliminated. The combination of noise and redundancy reduction optimizes the information transfer, resulting a bandpass filtering. However, as noticed in [29], the balance between highpass and lowpass filtering depends on the signal to noise ratio of the input signal, which depends on the ambient light level.

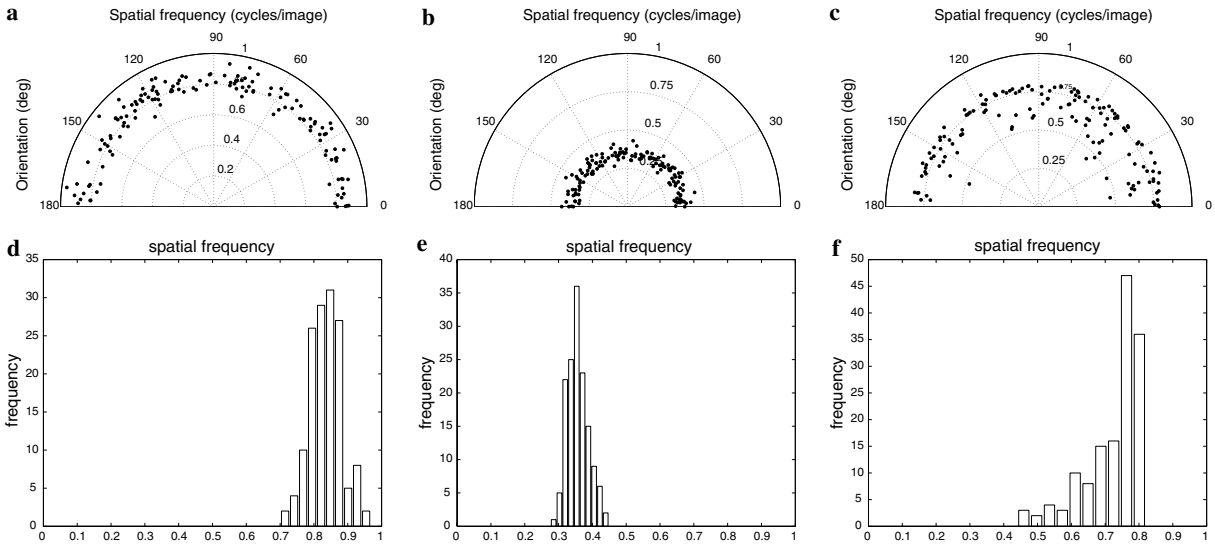

Fig. 5. The optimal orientation and optimal spatial frequency for RF masks corresponding to (a) NMF, (b) LNMF, and (c) DNMF receptive fields. The histogram of the distribution of 144 RFs in the spatial-frequency corresponding to (d) NMF, (e) LNMF, and (f) DNMF approaches.

5. Discussion and conclusion

There are many models proposed for biological facial analysis in the human visual system. On one side, the computer scientists try to find reliable methods that give satisfactory results for face or facial expression recognition. On the other side, psychologists and neurophysiologists try to understand how the human face is perceived by the human visual system, and develop models based on various experiments. Not surprisingly, some models proposed by the computer scientists, such as PCA, ICA, or Gabor image decomposition, have been accepted as biologically plausible, since they share common properties with biological vision models. In this paper, three other models (NMF, LNMF, and DNMF) were investigated. Although the main goal of this paper was to analyze their receptive field masks, it is worthwhile to mention common properties and differences between these three methods in order to draw a general conclusion. Table 1 summarizes several common and specific characteristics of these models.

The basic principle of efficient information transfer (and hence efficient coding) is to reduce the redundancy of the input signal. It is well-known that the natural stimuli (images) contain a large amount of redundant information that loads the dynamic range of the transmission channel without transferring information [15,6]. Generally, the term efficient coding and information redundancy reduction was associated with finding principal or independent components in representing a set of images. One fundamental difference between the methods mentioned in the Introduction and these three algorithms analyzed in this paper is that neither NMF,

Table 1
Characteristics of NMF, LNMF, and DNMF methods

	Decomposition method		
	NMF	LNMF	DNMF
Non-negative constraints	Yes	Yes	Yes
Redundancy reduction	No	Yes	Yes
Sparseness degree	Holistic	Local	Sparse
Class-dependent learning	No	No	Yes
Learning type	Unsupervised	Unsupervised	Supervised
Salient feature extraction	Yes	No	Yes
Spat. freq. bandwidth	Lowpass	Highpass	Bandpass

LNMF, nor DNMF assume features independence. ICA and other methods that rely on this assumption work well when they are applied on natural scenes. Definitely, natural images can contain more independent features than facial images. Here, each image has the same features (eyes, mouth, etc.) spatially located in approximately the same position. This might be a reason why ICA performed worse than NMF, LNMF, and DNMF when it comes to classify facial expressions [12].

Sparsity is another important issue that comes from neurophysiological field and has several advantages over holistic or local representations [27]. It is argued that the tuning of the neurons in the temporal cortex that respond preferentially to faces represents a trade-off between fully distributed encoding (holistic or global representation, as NMF result) and a grandmother cell type of encoding (local representation, achieved by LNMF) [28]. This trade-off seems to be provided by DNMF representation.

The next three characteristics, namely class-dependent learning, training type, and salient feature extraction are closely related to each other. NMF and LNMF are unsupervised approaches while DNMF is supervised one. In a feature extraction framework supervised learning is often necessary to guide feature development. Forcing a class-dependent learning by means of new constraints on coefficients expression, combined with the sparsity constraint on basis images (i.e., relation $\sum_i u_{ii} \rightarrow \min$), leads to a DNMF sparse image representation where the salient facial features (emotion-specific patterns that contribute most to expression recognition) are selected from the entire face image while the contribution of irrelevant features is diminished. However, it should be noticed that this class-dependent approach is rather a condition which comes from pattern recognition domain.

As a general conclusion, when comparing these three matrix factorization algorithms with each other, we favor DNMF since it fulfills several requirements: its enhances the class separability (which a pattern recognition issue) compared to the first two approaches, minimizes the redundancy over basis images (similar to efficient coding principle) and leads to a moderate sparse image representation (a neurophysiological issue). We found that, when DNMF is applied to faces, the receptive fields obtained by its basis images are bandpass filters covering the entire frequency orientation domain. Neurophysiology studies must be performed in order to validate the values of the parameters of the DNMF receptive fields.

Acknowledgments

We are grateful to Bruno Olsahusen and Patrik Hoyer for helpful discussions. This work has been conducted in conjunction with the "SIMILAR" European Network of Excellence on Multimodal Interfaces of the IST Programme of the European Union (www.similar.cc).

References

[1] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nature Neuroscience 2 (1999) 1019–1025.
[2] R. Desimone, Face selective cells in the temporal cortex of monkey, Journal of Cognitive Neuroscience 3 (1991) 1–8.
[3] D.I. Perret, E.T. Rolls, W. Caan, Visual neurons responsive to faces in the monkey temporal cortex, Experimental Brain Research 47 (1982) 329–342.
[4] N. Kanwisher, J. McDermott, M.M. Chun, The fusiform face area: a module in human extrastriate cotrex specialized for face perception, Journal of Neuroscience 17 (1997) 4302–4311.
[5] M.E. Hasselmo, E.T. Rolls, G.C. Baylis, V. Nalwa, The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey, Behavioral Brain Research 32 (1989) 203–218.
[6] B.A. Olshausen, D.J. Field, Natural image statistics and efficient coding, Network Computation in Neural Systems 7 (2) (1996) 333–339.
[7] S. Marcelja, Mathematical description of the responses of simple cortical cells, Journal of the Optical Society of America 70A (11) (1980) 1297–1300.
[8] J.G. Daugman, Two-dimensional spectral analysis of cortical receprive field profile, Vision Research 20 (1980) 847–856.
[9] P.J.B. Hancock, R.J. Baddeley, L.S. Smith, The principal components of natural images, Network Computation in Neural Systems 3 (1) (1992) 61–70.
[10] C. Fyfe, R. Baddeley, Finding compact and sparse-distributed representations of visual scenes, Network Computation in Neural Systems 6 (3) (1995) 333–344.
[11] A.J. Bell, T.J. Sejnowski, The 'independent components' of natural scenes are edge filters, Vision Research 37 (1997) 3327–3338.
[12] I. Buciu, I. Pitas, A new sparse image representation algorithm applied to facial expression recognition, IEEE Workshop on Machine Learning for Signal Processing (2004) 539–548.

[13] J.W. Tanaka, M.J. Farah, Parts and wholes in face recognition, Quarterly Journal of Experimental Psychology: Human Experimental Psychology 46A (1993) 225–245.
[14] M.N. Dailey, G.W. Cottrell, Organization of face and object recognition in modular neural network models, Neural Networks 12 (1999) 1053–1073.
[15] J.J. Atick, A.N. Redlich, What does the retina know about the natural scene? Neural Computation 4 (1992) 196–210.
[16] T. Valentine, A unified account of the effects of distinctiveness, inversion, and race in face recognition, Quarterly Journal of Experimental Psychology 43A (1991) 161–204.
[17] J.W. Ellison, D.W. Massaro, Featural evaluation, integration, and judgement of facial affect, Journal of Experimental Psychology: Human Experimental Psychology: Human Perception and Performance 23 (1) (1997) 213–226.
[18] G.W. Cottrell, M.N. Dailey, C. Padgett, R. Adolphs, Is all face processing holistic?, Philosophical Transactions of the Royal Society of London 352 (10) (1997) 1203–1219.
[19] B.A. Draper, K. Baek, M.S. Bartlett, J.R. Beveridge, Recognizing faces with PCA and ICA, in: Computer vision and image understanding, vol. 91: Special issue on Face Recognition, 2003, pp. 115–137.
[20] D.D. Lee, H.S. Seung, Learning the parts of the objects by non-negative matrix factorization, Nature 401 (1999) 788–791.
[21] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Advances Neural Information Processing Systems 13 (2001) 556–562.
[22] S.Z. Li, X.W. Hou, H.J. Zhang, Learning spatially localized, parts-based representation, International Conference on Computer Vision and Pattern Recognition (2001) 207–212.
[23] T. Kanade, J. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proceedings of IEEE International Conference on Face and Gesture Recognition, 2000, pp. 46–53.
[24] M. Pantic, L.J.M. Rothkrantz, Expert system for automatic analysis of facial expressions, Image and Vision Computing 18 (11) (2000) 881–905.
[25] K. Tanaka, C. Saito, Y. Fukada, M. Moriya, Integration of form, texture, and color information in the inferotemporal cortex of the macaque, Vision, Memory and the Temporal Lobe (1990) 101–109.
[26] K.P. Kording, C. Kayser, W. Einhauser, P. Konig, How are complex cell properties adapted to the statistics of natural stimuli?, Journal of Neurophysiology 91 (1) (2004) 206–212.
[27] P. Foldiak, Sparse coding in the primate cortex, The Handbook of Brain Theory and Neural Networks, second ed., MIT Press, Cambridge, MA, 2002, pp. 1064–1068.
[28] E.T. Rolls, A. Treves, The relative advantages of sparse versus distributed encoding for associative neural networks in the brain, Network 1 (1990) 407–421.
[29] J.J. Atick, Could information theory provide an ecological theory of sensory processing, Network 3 (1992) 213–251.