Statistics of Feature Extraction by Topographic Independent Component Analysis from Natural Images

RADU MUTIHAC, MARC M. VAN HULLE
Laboratorium voor Neuro- en Psychofysiologie
Katholieke Universiteit Leuven
Campus Gasthuisberg, Herestraat 49, B-3000 Leuven
BELGIUM
{radu.mutihac, marc.vanhulle}@med.kuleuven.ac.be

Abstract: - Our contribution highlights the statistical properties and biological interpretation of the basis vectors (filters) that result from applying topographic independent component analysis (ICA) to feature extraction from patches of natural images. The consistency of the feature sets obtained from various collections of natural image data sets applying topographical ICA (TICA) supports the opinion that the statistical properties of the environmental stimuli enforce a process according to some optimization criterion, which provides a good computational model for the response properties of sensory neurons. However, the basis vector set differs statistically meaningful from one image collection to the other, making the ICA decomposition of natural images unsuitable for a novel approach to image compression.

Key-Words: - Independent component analysis, Topography, Natural images, higher-order statistics

1 Introduction

It has widely been accepted that the sensory neurons have adapted by means of both evolutionary and developmental processes to the statistical properties of the stimuli that have mostly often been encountered in the environment. The importance of determining precise quantitative relationships between environmental statistics and neural processing is manifold. Better understanding of functional properties of neurons and neural systems, and the design of new computational models based on environmental statistics are two main issues. Secondly, finely tuned experimental designs and protocols for probing biological systems can be conceived and, last but not least, improved interfaces between human beings and artificial systems can be designed, with major benefits in our interaction with the environment.

In a neurobiological context, Barlow [1] suggested that a main role of early sensor neurons is to remove statistical redundancy in the sensory input by performing an "efficient coding" of stimuli. Such a task requires the specification of the environment, which amounts to a probability distribution over the space of input signals. One straightforward approach is studying the statistical properties of neural responses to natural stimulations conditions. An alternative way is to conceive statistical generative models of the input data. If the parameters of the model are estimated from natural input data, they are likely to provide deeper insight on the computational properties of the sensory neurons. There is some evidence that out of all visual images possible we see only a very small fraction [2]. If decomposing an image into independent components is one of the principal tasks of simple cells in the primary visual cortex, it would entail that the distribution of their properties to be determined by the statistics of the visual environment. Olshausen and Field [3] modeling visual data with a simple linear generative model showed that the principle of maximizing sparseness (or supergaussianity) of the underlying image components explain the emergence of Gabor-like filters that resemble the receptive fields of simple cells. Running ICA on a large set of calibrated images, and comparing a series of properties of the resulting receptive fields with those of receptive fields measured in simple cells, Hateren and van der Schaaf reported a good similarity [4].

In our simulations, we adopted a similar methodology and employed the TICA model introduced by Hyvärinen *et al*. [5]. TICA relaxes the independence assumption by replacing the conventional topologic ordering based on Euclidian distances of the basis vectors with a new topographic organization based on the dependence in higher-order statistics. The higher-order dependencies, which linear ICA does not remove, are used to define a topographic order such that nearby cells tend to be active (or inactive) at the same time. In this contribution we report our results in applying TICA to feature extraction from natural images. We analyzed statistically the properties of the TICA basis vectors that resulted from different

sets of natural images by performing TICA decomposition. By analogy, it was conjectured to a certain extent that the topographic neighborhoods exhibit properties specific to the complex cells in the mammalian primary visual cortex (V1). Apart from emulating the features of simple cells, like the distributions for spatial frequency bandwidth, orientation tuning bandwidth, aspect ratio, and receptive field length, the topological organization allows the emergence of phase and (partial) shift invariance that characterize complex cells.

2 The ICA Model for Image Data

The basic stationary noiseless linear ICA model assumes that $\mathbf{s}(t) \in \tilde{N}^M$ and $\mathbf{x}(t) \in \tilde{N}^N$ are two random (column) vectors with zero mean and finite covariance, with the components of $\mathbf{s}(t)$ being statistically independent and at most one gaussian, $\mathbf{A}$ is a rectangular constant full column rank $N \times M$ matrix with at least as many rows as columns ($N \geq M$), and t is the sample index (e.g. time or point) assumed to take discrete values $t = 1,2,...,T$:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{M} \mathbf{a}_i s_i(t), \quad t = 1,2,...,T \tag{1}$$

The columns $\{\mathbf{a}_i\}$, $i = 1,2,...,M$ of $\mathbf{A}$ are the ICA basis vectors. Then the ICA problem can be formulated as follows: given T realizations of $\mathbf{x}(t)$, estimate both the matrix $\mathbf{A}$ and the corresponding realizations of $\mathbf{s}(t)$. In an alternative context, the ICA decomposition (1) is equivalent with sparse coding [6]. Most ICA algorithms are searching for a separation matrix $\mathbf{W}$ to demix data on the basis of various estimation principles of independence.

There is clear evidence that the distribution of natural images is nongaussian [7]. Hence it seems reasonable to consider a static monochrome image $I(x,y)$ as a linear superposition of some *features* or basis functions $\{a_i(x,y)\}$, $i = 1,2,...,M$

$$I(x,y) = \sum_{i=1}^{M} a_i(x,y) s_i \tag{2}$$

where each image $I(x,y)$ has different stochastic coefficients s_i, $i = 1,2,...,M$. In order to comply with the underlying ICA assumptions, the coefficients $\{s_i\}$ are assumed nongaussian and mutually independent. Estimating the model amounts to determining the values of s_i and $a_i(x,y)$ for all indexes i and points (x,y), given a sufficient number of observations of images, such as image patches $I(x,y)$. If we restrict the study to the case where $a_i(x,y)$ form an invertible linear system, then $s_i = \langle w_i, I \rangle$, where $w_i(x,y)$, $i = 1,2,...,M$ denote the inverse filters and $\langle w_i, I \rangle = \sum_{x,y} w_i(x,y) I(x,y)$ stands for the dot product. The inverse filters $w_i(x,y)$ can be identified as the *receptive fields* of the model simple cells, and the coefficients s_i as their activities when presented with a given image patch $I(x,y)$. When this model is estimated with input data consisting of patches of natural scenes, the obtained filters $w_i(x,y)$ exhibit the three principal properties of simple cells in V1: they are spatially localized, oriented, and band-pass in different spatial frequency bands Quantitative comparison of obtained filters $\{w_i(x,y)\}$ with those measured by single-cell recordings of the macaque cortex showed a close match for most of the parameters [7].

3 The Topographic ICA Model

The model (2) is nevertheless inappropriate to describe the response of complex cells due to their properties of phase and (limited) shift invariance [Pollen and Ronner, 1983]. In classic ICA, the latent variables $\{s_i\}$ have no particular order and no relationship between them is assumed whatsoever. This is in compliance with the assumption of complete statistical independence of the latent variables. However, there are applications in which ICA does not completely remove the dependence between components, which may be quite informative. Since many ICA estimation methods constrain the components to be uncorrelated, it seems reasonable to preserve the uncorrelatedness in any further extension of ICA. Hyvärinen and Hoyer [9] proposed a higher-order correlation based on *energies*, which can be intuitively interpreted as a simultaneous activation of the units

$$cov\left(s_i^2 s_j^2\right) = E\left\{s_i^2 s_j^2\right\} - E\left\{s_i^2\right\} E\left\{s_j^2\right\} \neq 0 \tag{3}$$

if s_i and s_j are close in topography. In the generative model $\mathbf{x} = \mathbf{A}\mathbf{s}$ of TICA, the central issue is to define the joint density of $\mathbf{s}$ based on the topography (e.g. simultaneous activation or inactivation of the nearby cells). The topography is generally defined by specifying a neighborhood function $h(i,j)$ that express the proximity between the components i and j. Its common form is a monotonically decreasing function of some distance

measure, so that $h(i,j)$ comes out as a matrix of hyperparameters of the model. Hereafter $h(i,j)$ is assumed known and fixed. The components $\{s_i\}$ of **s** are defined by means of their variances $\{s_i^2\}$, which are assumed random variables generated according to a model specified on the basis of topography. Then the variables $\{s_i\}$ are generated mutually independent using some conditional distributions. So the dependence among the s_i's is implied by the dependence of their variances

$$s_i = f\left(\sum_{k=1}^{M} h(i,k)\, u_k\right) z_i \tag{4}$$

with $\{u_k\}$ the higher-order independent components used to generate the variances, f a scalar nonlinearity, and z_i is a random variable with the same distribution as s_i, if s_i^2 is fixed to unity. The variables u_i and z_i are mutually independent.

The properties of the TICA model are discussed in detail by Hyvärinen, Hoyer, and Inki in [4]. The model is a missing variables model in which its likelihood cannot be obtained in a closed form. An approximation for the likelihood of the model results if following the derivation as in ICA [10]. An analytical approximation was derived by assuming further simplifications, namely constant marginal densities for $\{u_i\}$ and constant conditional densities for $\{s_i\}$ as a gaussian, and the nonlinearity is defined as $f\left(\sum_{k=1}^{M} h(i,k)\, u_k\right) = \left(\sum_{k=1}^{M} h(i,k)\, u_k\right)^{-\frac{1}{2}}$, hence the log likelihood (actually an approximation of its lower bound) becomes [4]

$$log\ L(\mathbf{W}) \cong \sum_{t=1}^{T}\sum_{j=1}^{M} G\left(\sum_{i=1}^{M} h(i,j)\left(\mathbf{w}_i^T \mathbf{x}(t)\right)^2\right) + T\ log\left|det\ \mathbf{W}\right| \tag{5}$$

where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_N)^T = \mathbf{A}^{-1}$, and $\mathbf{x}(t)$, $t = 1,2,...,T$ are the observations of **x**. Practically, if data are sparse, convergence is achieved for almost any function G that is convex for some nonnegative argument.

The model can be solved by maximizing $log\ L(\mathbf{W})$. The data, which are assumed zero-mean, are first whitened $\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s}$. If $\mathbf{V}\mathbf{A}$ is invertible, the new separating matrix becomes $\mathbf{W} = (\mathbf{V}\mathbf{A})^{-1}$ and we can constrain its rows $\{\mathbf{w}_i^T\}$ to form an orthonormal system [5], [11]. The orthonormal basis in the whitened space amounts to decorrelating the estimated components, so that their dependency in higher-order statistics remains. A simple gradient algorithm can be derived for updating the (weight) vectors $\{\mathbf{w}_i\}$ [4]

$$D\mathbf{w}_i \propto E\left\{\mathbf{z}\left(\mathbf{w}_i^T \mathbf{z}\right) r_i\right\} \tag{6}$$

where $r_i = \sum_{k=1}^{M} h(i,k)\, g\left(\sum_{j=1}^{M} h(k,j)\left(\mathbf{w}_j^T \mathbf{z}\right)^2\right)$ is a modulation factor and the function g is the derivative of a convex function such as $G(y) = -a\sqrt{y} + b$. The scaling constant a and the normalizing constant b are determined so as to produce a probability density in compliance with the constraints imposed on $\{\mathbf{w}_i\}$. Actually, the vectors $\{\mathbf{w}_i\}$ must be normalized to unit variance and orthogonalized after every step in (6). If we denote the matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M)^T$, the method involving matrix square roots can be used following $\mathbf{W} \leftarrow \left(\mathbf{W}\mathbf{W}^T\right)^{-\frac{1}{2}}\mathbf{W}$. The original mixing matrix of the unwhitened data can be computed after learning $\{\mathbf{w}_i\}$ such as $\mathbf{A} = (\mathbf{W}\mathbf{V})^{-1}$. The rows of $\mathbf{A}^{-1}$ provide the filters (weight vectors) in the original, not whitened space.

4 Experimental

Previous experiments with natural images were missing basically three pictorial elements that are routinely present in common environments at various scales: (i) square and oblique corners of buildings, rooms, miscellaneous technical equipment, etc.; (ii) humans and human faces, (iii) text (letters and figures). Our main interest was to extract the underlying features of surrounding images and not to simulate the long-lasting natural conditions responsible for the functioning of visual simple and complex nervous cells.

We built up, accordingly, *5* distinct sets of *24* digital images each as different as possible in terms of subject, texture, and scales, selected from FreeFoto.com [12], which is a large database featuring *50* main sections with over *1000* sub-headings of free photographs on the Internet. The selected images were first auto equalized, then converted to monochrome uncompressed TIFF format with *8*-bit pixel depth (i.e., *256* gray levels), and subsequently cropped down to size 256×256 pixels, with a resolution of *150* pixel/inch. Original

images were randomly rotated in order to avoid any bias caused by camera orientation and/or light source position, and then square 16×16 pixel image patches were randomly cropped. We selected images with both unimodal and multimodal histograms, out of which *48,000* image patches were randomly extracted from each set and stored as columns of a matrix **X** of size 256×48000. The mean gray scale value of the patches was substracted and the matrix **X** was normalized to unit variance. Then, the dimensionality was reduced (i.e., low-pass filtering) by running principal component analysis (PCA) and retaining the principal *192* components in decreasing order of mean projected variance in the data space. Subsequently to PCA, the data were sphered by zero-phase whitening filter, which equated as a multiplication of the data by the inverse of its squared covariance matrix. Now data were contained in a *192*-dimensional subspace spanned by the *192* most energetic basis vectors that formed an orthonormal system, but not for the original data space. For visualization purposes and in order to avoid border effects, the topography was chosen as *2D* torus lattice as suggested first by Kohonen [13]. We used only one neighborhood size S_m of a 3×3 square around each unit. It meant that the neighborhood function could be expressed as $h(i, j) = 1$ if $\exists m$: $i,j \in S_m$ and zero otherwise. The form of function G was the simplest possible amended with a small constant e for numerical stability $G(y) = -a\sqrt{e + y} + b$ [4]. Then the gradient method was used to maximize the approximation of the likelihood (5) over the *48,000* image patches under the constraint of orthonormality of the *192* filters in the whitened space. Each running of the algorithm on a PC with Pentium 4 processor at *1.5* GHz took around *40* hours until reaching a similar value of the objective function in all cases.

5 Results and Discussion

We used *5* sets of *24* different images each containing as general as possible natural images (i.e. landscapes, animals, mountains, etc), as well as human made objects that are frequently encountered in common living environments (Fig. 1). The point was to generate a faithful representation of image statistics that is continuously influencing our visual system in daily life (though not realistic at the human evolution scale).

Fig. 1 Samples from a natural image set. *Left*: common landscape. *Right* – group of buildings.

A typical topographic ICA vector basis (i.e., a subset of *192* image features) and the corresponding spatial ICA filters are presented in Fig. 2 for one of the experimental image set.

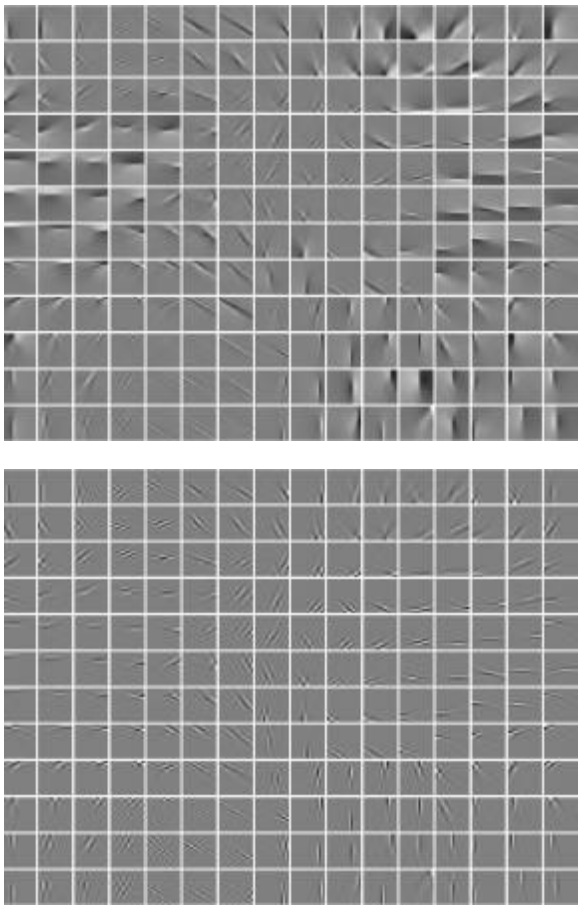

Fig. 2 *Top:* typical feature decomposition of natural images in topographic order (the columns of the mixing matrix). *Bottom*: the corresponding spatial filters (the rows of the separation matrix). The image features (i.e., the basis vectors) are the patterns that optimally stimulate their corresponding ICA filters, while minimally stimulating any other ICA filter.

Among the statistical characteristics of natural images, particularly important are their nongaussianity and statistical redundancy [14]. The covariance properties of natural images can be used to derive basis functions that are similar to receptive fields found physiologically in primary visual cortex (i.e., oriented band-pass filters). Alternatively, a quantitative study of the topographic organization can be carried out by computing the correlations of the energies between components of the entire data space. The model predicts a gradually vanishing of the correlations with the distance on the topographic grid. The analysis may fail because of two reasons: (*i*) if the basis vector system is underdetermined, the neighborhoods are two small, and/or there are not enough data, (*ii*) if the image patches do not comply entirely with the model (i.e., they are not a linear superposition of invariant features). We expect that for larger neighborhoods, the preponderance of spatial frequency to decrease in favor of orientation and location.

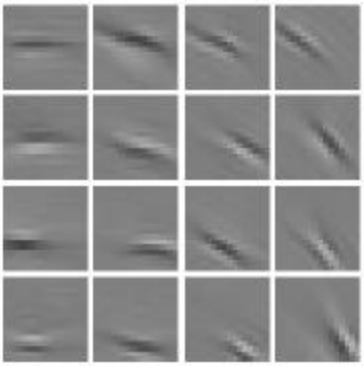

Fig. 3 The characteristics of the basis vectors for natural images as a function of a relative position on the topographic grid.

The similarity between the neighborhoods in TICA and the receptive fields of complex cells in the primary visual cortex of mammalians is supported by the tendency of nearby basis vectors (simple cells) in the topographic map to be of similar orientation and frequency but having very different phases (Fig. 3). Maximizing the independence (e.g. the sparseness) of linear filter outputs, the model provides simple cell properties. Maximizing the independence of the norms of the projections on linear subspaces, the model yields complex cell properties. Both cases stand for the importance of dependence reduction as a strategy for sensory information processing. It was argued that sparse coding simplifies further processes in the visual system since it comes out with a representation of the stimulus that help detection of coincidences [15]. As the cells preferentially respond to oriented edges or lines, they may be interpreted as edge or line detectors. Then the oriented edge features can be interpreted as a sparse representation of the image. It means that over an ensemble of images a particular feature will seldom be significantly active.

We applied nonparametric statistics to assess the statistical significance of the feature sets produced by topographic ICA for 5 distinct collections of images. The ICA basis vectors were sorted in decreasing order of their mean variance. Then the Kruskal-Wallis analysis of ranks was applied to test the statistical significance of the basis vector sets under the null hypothesis that the features are drawn from the same distribution. The results indicated statistically distinct populations from which the features were extracted given the size of the collection, the size of images, the number of image patches considered.

6 Conclusion

All variants of ICA are seeking for some form of statistical independence of the estimates and describe the images in terms of linear superposition. Yet natural images are not formed by sums of independent components since image formation often obeys the rules of occlusion rather than addition of light. Analysis of statistical relationships in images reveals nonlinear dependencies across space as well as across scale and orientation [16]. Still image decomposition by various forms of ICA performs image invariant-feature extraction, which we proved to be statistically relevant.

The main utility of topography is visualization [13], which shows the connections between components and possibly adding some information. Classic ICA applied to natural images yields a linear decomposition into Gabor-like linear features that resemble the receptive fields of simple cells. Topographic ICA organizes image features in compliance with the defined topography. The basic conclusion drawn from running ICA on natural image patches is that ICA filters are localized and oriented, whereas the ICA basis functions are oriented and not clearly localized, which makes difficult to notice any multiscale properties. The ICA filters come out with more sparsely distributed (kurtotic) outputs when applied to an ensemble of natural scenes in comparison with other filters like PCA or zero-phase whitening filters (ZCA) [17]. Moreover, the topographic neighborhoods resemble

complex cells in their response by exhibiting phase and shift invariance [18]. In this way TICA shows simultaneous emergence of complex cell properties and topographic organization following the same principle of defining topography by simultaneous activation (or inactivation) of neighbors.

Though informative, the basic ICA model is essentially linear and non-adaptive, ignoring phenomena that commonly occur in human visual system, like contrast adaptation, contrast normalization, nonlinearities involved in orientation tuning, adaptation to various stimulus statistics, to cite the main issues only. In feature extraction from patches of natural images, ICA comes out with feature sets that statistically belong to different distributions distinct if different image collections are subject of analysis. Although our results reported herein are qualitatively rejecting the uniqueness of feature extraction from natural image patches by topographic ICA decomposition, the lack of image calibration and the amount of processed data are the main limiting factors to a full quantitative comparison.

Acknowledgments

R.M. is supported by a postdoc grant from the European Community, FP5 (QLG3-CT-2000-30161). M.M.V.H. is supported by research grants received from the Fund for Scientific Research (G.0185.96N), the National Lottery (Belgium) (9.0185.96), the Flemish Regional Ministry of Education (Belgium) (GOA 95/99-06; 2000/11), the Flemish Ministry for Science and Technology (VIS/98/012), and the European Community, FP5 (QLG3-CT-2000-30161 and IST-2001-32114). R.M. is grateful to A. Hyvärinen for his pertinent comments and suggestions.

References:

[1] H.B. Barlow, Possible principles underlying the transformation of sensory messages, in *Sensory Communication*, W.A. Rosenblith (Ed.), MIT Press, Cambridge MA, 1961, pp. 217-234

[2] D.L. Ruderman, W. Bialek, Statistics of natural images: Scaling in the woods, *Phys. Rev. Lett.* Vol. 73, No. 6, 1994, pp. 814-817

[3] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature* Vol. 381, 1996, pp. 607-609.

[4] J.H. van Hateren, A. van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex, *Proc. Royal. Society ser. B*, Vol. 265, 1998, pp. 359-366.

[5] A. Hyvärinen, P.O. Hoyer, M. Inki, Topographic independent component analysis, *Neural Comput*. Vol. 13, No. 7, 2001, pp. 1527-1558.

[6] P. Comon, Independent component analysis, A new concept? *Signal Processing* Vol. 36, No. 3, 1994, pp. 287-314.

[7] E.P. Simoncelli, B.A. Olshausen, Natural image statistics and neural representation, *Annu. Rev. Neurosci.* Vol. 24, 2001, pp. 193-216.

[8] D. Pollen, S. Ronner, Visual cortical neurons as localized spatial frequency filters, *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 13, 1983, pp. 907-916.

[9] A. Hyvärinen, P.O. Hoyer, Topographic independent component analysis as a model of V1 organization and receptive fields, *Neurocomput*. Vols. 38-40, 2001, pp. 1307-1315.

[10] D.-T. Pham, P. Garrat, C. Jutten, Separation of a mixture of independent sources through a maximum likelihood approach, *Proc. EUSIPCO*, 1992, pp. 771-774.

[11] J.-F. Cardoso, B.H. Laheld, Equivariant Adaptive Source Separation, *IEEE Trans. Signal Proces*. Vol. 44, No. 12, 1996, pp. 3017-3030.

[12] FreeFoto.com image database available at http://www.freefoto.com/

[13] T. Kohonen, Self-Organizing Maps, Spriger-Verlag, Berlin, 1995.

[14] D.J. Field, Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, Vol. 4, No. 12, 1987, pp. 2379-2394.

[15] H.B. Barlow, What is the computational goal of the neocortex? in *Large Scale Neuronal Theories of the Brain*, C. Koch (Ed.), MIT Press, Cambridge MA, 1994, pp. 1-22.

[16] E.P. Simoncelli, O. Schwartz, Image statistics and cortical normalization models, in: Advances in Neural Information Processing, Vol. 11, M.S. Kearns, S.A. Solla, D.A. Kohn, (Eds.), 1999, pp. 153-159.

[17] A.J. Bell, T.J. Sejnowski, The "independent components" of natural scenes are edge filters. *Vision Res.* Vol. 37, 1997, pp. 3327-3338.

[18] A. Hyvärinen, P.O. Hoyer, Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces, *Neural Comput*. Vol. 12, No. 7, 2000, pp. 1705-1720.