Unsupervised learning of a steerable basis for invariant image representations

Matthias Bethge, Sebastian Gerwinn, and Jakob H. Macke

Max Planck Institute for Biological Cybernetics
Spemannstr. 38
72076 Tübingen, Germany

ABSTRACT

There are two aspects to unsupervised learning of invariant representations of images: First, we can reduce the dimensionality of the representation by finding an optimal trade-off between temporal stability and informativeness. We show that the answer to this optimization problem is generally not unique so that there is still considerable freedom in choosing a suitable basis. Which of the many optimal representations should be selected? Here, we focus on this second aspect, and seek to find representations that are invariant under geometrical transformations occuring in sequences of natural images. We utilize ideas of 'steerability' and Lie groups, which have been developed in the context of filter design. In particular, we show how an anti-symmetric version of canonical correlation analysis can be used to learn a full-rank image basis which is steerable with respect to rotations. We provide a geometric interpretation of this algorithm by showing that it finds the two-dimensional eigensubspaces of the average bivector. For data which exhibits a variety of transformations, we develop a bivector clustering algorithm, which we use to learn a basis of generalized quadrature pairs (i.e. 'complex cells') from sequences of natural images.

1. INTRODUCTION

The representation of images as vectors of pixel intensities is highly redundant[1] and largely obscures the characteristic structure of objects in natural scenes. In particular, a large fraction of variance can be classified as 'appearance' changes which leave the content of the image invariant. For illustration, we can have many different images of the same object due to changes in pose or lighting. Inverting this one-to-many mapping in a way that still allows one to discriminate between different objects is known as the 'invariance problem'.

Invariant representation learning can be viewed as a preprocessing step which removes distracting variance from a data set, so that downstream classifiers or regression estimators perform better. Clearly, it is an inherent part of training a classifier to make its response invariant against all within-class variations. Rather than learning these invariances for each object class individually, however, we observe that there are transformations such as translation, rotation and scaling, which apply to any object independent of its specific shape. This suggests that signatures of such transformations exist in the spatio-temporal statistics of natural images which allow one to learn invariant representations in an unsupervised way.

In a generative model, it is rather straightforward to apply transformations to objects even if they have complicated 3D shape. Estimating the generative coefficients of such models from a given image, however, is generally hard. Therefore, rather than building invariant representations for full-blown objects, the goal of unsupervised methods is to extract low-level image features which behave in a *desirable* way under *common* transformations. What is meant to be '*common*' is determined by the spatio-temporal image statistics. The assessment of whether an image feature behaves in a '*desirable*' way or not can be made precise in terms of an objective function. This should not belie the fact, however, that there are many different objective functions of similar plausibility.

An important goal of this paper is to tie together two different lines of work on invariant filters that have developed rather independently over the last fifteen to twenty years: (a) the work on unsupervised learning, most of which seeks to derive receptive field properties of visual neurons,[2] and (b) the concept of 'steerability',[3] which

M.B.: mbethge@tuebingen.mpg.de S.G.: sgerwinn@tuebingen.mpg.de J.H.M.: jakob@tuebingen.mpg.de

is more common in the computational vision literature and often dedicated to analytic derivations of (steerable) filter kernels. While the filters in (b) are designed to model transformational changes with a clear geometrical meaning, the filters derived by unsupervised learning in (a) are usually determined by a trade-off between signal power and insensitivity to certain appearance changes. Here we seek to clarify the differences and commonalities between the two approaches and also to develop algorithms which combine the advantages of both.

The paper is organized as follows: First, in section 2, we review previous work on unsupervised invariant representation learning which builds upon simultaneous diagonalization of certain covariance matrices. We first explain the diagonalization and then discuss different interpretations that have been used in the context of representation learning. Next, in section 3, we consider the ideal case of perfect predictability to show important differences between the eigenfunctions of symmetric and anti-symmetric covariance matrices. In section 4 we present the theoretical foundations of how steerable filters can be learned from a spectral decomposition of the anti-symmetric part of the time-lagged covariance matrix. Next, we use this method to learn a complete steerable basis for the problem of rotation invariance. Finally, in section 6, we provide a geometric interpretation of the anti-symmetric part of the time-lagged covariance matrix as the *average* bivector. This interpretation allows us to extend the algorithm in order to cope with multiple clusters of bivectors. In particular, we use a bivector clustering algorithm to learn a basis of generalized quadrature pairs from sequences of natural images.

2. UNSUPERVISED REPRESENTATION LEARNING FROM SPATIO-TEMPORAL CORRELATION ANALYSIS

A large body of work on invariant representation learning[4–11] is ultimately based on linear correlation analysis[12] oftentimes endowed with a nonlinear feature space.[11,13–15] A crucial commonality in all of this work is the use of spectral decompositions of covariance matrices or Gram matrices as it is well known from principal component analysis (PCA).

In this paper, we focus particularly on the following generalized spectral decomposition of the covariance function $C_x(\tau) = \langle \mathbf{x}(t+\tau)\,\mathbf{x}^T(t)\rangle$ of a wide-sense stationary vector-valued process $\mathbf{x}(t) \in R^d$, which is at the core of many studies on representation learning from temporal sequences:

$$C_x(0)^{-1/2} C_x(\tau) C_x(0)^{-1/2} = U D V^T \tag{1}$$

The r.h. side is the singular value decomposition (SVD) with U, V being orthogonal (i.e. $UU^T = U^T U = VV^T = V^T V = I_d$) and $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is diagonal with non-negative entries placed in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. If a singular value λ_k is well separated from all other singular values (i.e. $|\log(\lambda_k) - \log(\lambda_j)| \geq \epsilon > 0$) then the k-th column of the left eigenbasis U and the k-th column of the right eigenbasis V are uniquely determined. Furthermore, if $C_x(\tau) = C_x(\tau)^T = C_x(-\tau)$ is symmetric, then $U = V$, and representing the process $\mathbf{x}(t)$ by using the new representation $\mathbf{y}(t) := U^T C_x(0)^{-1/2} \mathbf{x}(t)$ yields the covariance matrices

$$C_y(0) = (U^T C_x(0)^{-1/2}) C_x(0) (U^T C_x(0)^{-1/2})^T = U^T \underbrace{C_x(0)^{-1/2} C_x(0) C_x(0)^{-1/2}}_{I_d} U = I_d \tag{2}$$

$$C_y(\tau) = (U^T C_x(0)^{-1/2}) C_x(\tau) (U^T C_x(0)^{-1/2})^T = U^T \underbrace{C_y(0)^{-1/2} C_x(\tau) C_x(0)^{-1/2}}_{UDU^T} U = D\,, \tag{3}$$

which are both diagonal. Again, if the singular values are well separated, then the new representation $\mathbf{y}(t) := U^T C_x(0)^{-1/2} \mathbf{x}(t)$ is unique in decorrelating the different components $\langle y_j(t) y_k(t)\rangle$ and $\langle y_j(t) y_k(t+\tau)\rangle$ for all $j \neq k$ at the same time.

There are multiple interesting interpretations to the very same decomposition in Eq. 1, and thus it has been motivated from various perspectives. In the next section, we present four of these interpretations including *oriented PCA, 2nd-order independent component analysis, slow feature analysis, and canonical correlation analysis/predictive coding*. In particular, we explain their differences and meanings specific to the problem of invariant representation learning.

2.1. Dimensionality reduction vs predictive coding in invariant representation learning

If a representation is invariant against certain changes in the image, then the information about these changes is lost. Therefore, it is important to tell apart two different sorts of problems that are confounded in the task of invariant representation learning: (a) Which image informations are irrelevant and hence can be discarded? (b) Which representation should be used to represent the remaining relevant information?

In unsupervised learning, question (a) is usually addressed by resorting to some heuristics. For instance, principal component analysis (PCA) is a popular method frequently used for the purpose of dimensionality reduction which seeks to maximize the signal energy. **Oriented PCA**[5] is a generalization of that method which is also employed in the context of invariant representation learning.[10,11,16] It models the data $\mathbf{x} \in R^d$ as a superposition $\mathbf{x} = \mathbf{s} + \mathbf{n}$ of a relevant signal $\mathbf{s}$ and irrelevant noise $\mathbf{n}$ and seeks to find a subspace spanned by $\mathbf{y} = W\mathbf{x}$, $W \in R^{m \times d}$, $m < d$ which maximizes the signal-to-noise ratio. Analogous to plain PCA, it is instructive to start with the one-dimensional case $m = 1$. In this case, oriented PCA searches for a filter $\mathbf{w}$ which optimizes the generalized Rayleigh quotient

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T C_x \mathbf{w}}{\mathbf{w}^T C_n \mathbf{w}} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T C_n^{-1/2} C_x C_n^{-1/2} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} = \arg\min_{\mathbf{w}} \frac{\mathbf{w}^T C_x^{-1/2} C_n C_x^{-1/2} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \tag{4}$$

where C_x denotes the total covariance of the data and C_n stands for the noise covariance. Note that in case of isotropic noise (i.e. if $C_n \propto I_d$) the optimization reduces to the usual eigenvalue problem of plain PCA. If C_n is not proportional to the identity matrix, one can use the generalized eigenvalue problem spelled out in Eq. 1 to find the optimum with respect to the Rayleigh quotient (Eq. 4). In the case of more than one dimension (i.e. $m > 1$) the signal-to-ratio is maximized for all filter matrices $W \in R^{m \times d}$ which maximize the ratio of the corresponding determinants:

$$W_o = \arg\max_{W} \frac{\det(W^T C_x W)}{\det(W^T C_n W)}. \tag{5}$$

It does not come as much of a surprise that the eigenvectors of the m largest eigenvalues of $C_x^{-1/2} C_n C_x^{-1/2}$ take a maximum for this ratio, but this maximum is not unique: any other $\tilde{W} := AW_o$ with arbitrary $A \in R^{m \times m}$ of full rank (i.e. $\det(A) \neq 0$) equivalently maximizes the signal-to noise ratio.

In order to use oriented PCA for unsupervised representation learning, the crucial question is how to decompose the data covariance C_x into signal and noise covariance. In invariant representation learning, the idea of "slowness" has been employed as heuristic[11,17–19] building on the assumption that the perceived environment changes slowly and the quickly varying components in the sensory input carry less important local information. The optimization for slowness in **slow feature analysis (SFA)**[11] mirrors the reasoning behind oriented PCA. In the case of one dimension, the objective function of SFA is given by

$$\begin{aligned} \mathbf{w} &= \arg\min_{\mathbf{w}} \frac{\mathbf{w}^T \langle (\mathbf{x}(t+\tau) - \mathbf{x}(t))(\mathbf{x}(t+\tau) - \mathbf{x}(t))^T \rangle \mathbf{w}}{\mathbf{w}^T \langle \mathbf{x}(t)\mathbf{x}(t)^T \rangle \mathbf{w}} \\ &= \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T \langle \mathbf{x}(t)\mathbf{x}(t)^T \rangle \mathbf{w}}{\mathbf{w}^T \langle (\mathbf{x}(t+\tau) - \mathbf{x}(t))(\mathbf{x}(t+\tau) - \mathbf{x}(t))^T \rangle \mathbf{w}} \end{aligned} \tag{6}$$

which can be interpreted as a special case of oriented PCA for which the noise covariance is set proportional to the covariance $\langle (\mathbf{x}(t+\tau) - \mathbf{x}(t))(\mathbf{x}(t+\tau) - \mathbf{x}(t))^T \rangle$ of changes at a given time scale τ.

For SFA, it has never been specified what the objective function is in case of many dimensions $m > 1$. The interpretation of Eq. 6 in terms of a signal-to-noise ratio offers a reasonable extension to the multi-dimensional case via Eq. 5. However, it must not be forgotten that the optimal filter matrix $W \in R^{m \times d}$ is always ambiguous for $m > 1$ because left multiplication with an arbitrary $A \in R^{m \times m}$ of full rank leaves the signal-to-noise ratio invariant.

Therefore, only the subspace but not the spanning basis is distinct in terms of the objective function (Eq. 5). Clearly, the generalized eigendecomposition for finding this subspace suggests to use the eigenbasis returned by the standard algorithm, but neither maximization of the signal-to-noise ratio nor maximization of slowness

provides a theoretical justification for this choice. For illustration, after selection of the subspace one would obtain the same eigenbasis by solving for the opposite objective function (i.e. minimization of the signal-to-noise ratio or maximization of swiftness) as it leads to the same eigendecomposition.

Generally speaking, dimensionality reduction can only serve to answer question (a), but not question (b): It makes no statement how to represent the remaining relevant information. This can be derived from the fact that any invertible mapping leaves the signal-to-noise ratio unchanged: if $\mathbf{y} = W\mathbf{x}$ maximizes the signal-to-noise ratio, the same is true for any other representation $\mathbf{y} = AW\mathbf{x}$ with $A \in R^{m \times m}$ and $rank(A) = m$. Consequently, the filters or 'receptive fields' of such a representation are not unique but any other basis of the same subspace could be used just as well.

Which reasoning may be used to rank one basis to be preferable above others? This important issue will be addressed in the following. For the sake of clarity, we will from now on ignore the issue of dimensionality reduction and assume that all irrelevant information has been discarded already.

A well known possibility of choosing a unique basis is independent component analysis (ICA)[20] which has been studied extensively in the context of blind source separation. In practice, ICA seeks to minimize some empirical measure of statistical dependency between the filter responses. Thus, the generalized eigenbasis of $C_x(0)$ and $C_x(\tau) + C_x(-\tau)$ is a reasonable choice according to the objective of ICA, because this choice is unique in removing all the considered correlations between different components of $\mathbf{x}$. This particular method is known as **second-order ICA**[4,21] or the 'AMUSE' algorithm as it has been named originally in Ref. 4. In contrast to oriented PCA and slow feature analysis, the motivation underlying ICA is not related to the extremal properties of the generalized eigenvalues. Instead, second-order ICA exploits the uniqueness of simultaneous diagonalization for non-degenerate spectra. The mathematical rule used in SFA to obtain a unique set of basis functions relies on the same objective as AMUSE, namely decorrelation. Without this further assumption, receptive fields with very different shapes would provide a representation that achieves exactly the same "slowness". Furthermore, decorrelation alone is sufficient to determine the receptive fields.

If a basis is chosen only with the goal of maximizing statistical independence, it is not clear whether the filter responses will also have a geometrical meaning in terms of transformations as it would be desirable in the context of invariant image representations. Intuitively, it is clear that the separation of structure from appearance should be useful for object recognition or other visual inference tasks. In other words, we would ideally want a representation $(\vec{s}, \vec{a})$ which represents any image patch $\vec{x}$ in terms of a structure vector $\vec{s}$ and an appearance vector $\vec{a}$. An important aspect of such a new representation $(\vec{s}, \vec{a})$ is that it combines an *invariant* representation of structure given by $\vec{s}$ with an *equivariant* representation of the appearance change specified by $\vec{a}$, which together preserve all the information. Therefore, such representations have been termed a '*split of identity*'[22].

Two different classes of models implementing this idea have been studied: One uses a bilinear generative model, where the two coefficient vectors describing structure and appearance both influence the generation of an image vector in a linear way.[23,24] In the other approach, pairs of basis vectors are used whose coefficients are represented in polar coordinates. A well known example is the energy model of complex cells.[25] As pointed out in Ref. 22, energy and phase of such complex cells implement an *invariant-equivariant* split of identity: The energy is *invariant* against structural changes but varies with contrast in an *equivariant* way (i.e. changes in energy are proportional to changes in contrast). Conversely, the phase is *invariant* against changes in contrast but sensitive to structural changes. Additionally, the complex cell can be designed such that the phase provides an *equivariant* representation of certain geometrical changes. This is the idea behind steerable filter design[3,22] and the approach taken in this paper.

While the analytical design of steerable filters is well understood, we want to do *unsupervised* learning of invariant representations with quadrature pairs. In particular, we will present two algorithms for this task: The first one is an anti-symmetric version of canonical correlation analysis, the other one an extension using a bivector clustering-algorithm.

We now proceed to another interpretation of Eq. 1, which we will use later to study the relationship between representation learning and certain transformations. The following interpretation of Eq. 1 is based on **predictive coding**[26] and **canonical correlation analysis (CCA)**.[12,27]

For a wide-sense stationary process, the two covariance matrices $C_x(0)$ and $C_x(\tau)$ are sufficient to tell how well $\mathbf{x}(t+\tau)$ can be predicted from observing $\mathbf{x}(t)$ on the basis of a linear predictor L_τ in the mean squared error sense $\langle||\mathbf{x}(t+\tau) - L_\tau[\mathbf{x}(t)]||_2^2\rangle$. Assuming without loss of generality that $\langle\mathbf{x}(t)\rangle = 0$, the optimal linear predictor is given by $L_\tau[\mathbf{x}(t)] = C_\mathbf{x}(\tau)C_\mathbf{x}^{-1}(0)$. After a whitening transformation $\mathbf{y}(t) := C_\mathbf{x}^{-1/2}(0)\mathbf{x}(t)$ the form of the linear predictor further simplifies to $L_\tau[\mathbf{y}(t)] = C_\mathbf{y}(\tau)$.

Canonical correlation analysis (CCA)[12] is tightly related to the linear predictor. Precisely, it is equal to the singular value decomposition of the linear predictor $L_\tau[\mathbf{y}(t)] = UDV^T$ if the covariance of the input random variable is 'white' (i.e. $C_{\mathbf{y}(\mathbf{t})} = I_d$). Similar to the maximum variance components found with principal component analysis, CCA can be used to find the components with the maximum correlation coefficient between a pair of vector-valued random variables. Since for Gaussian random variables the mutual information is determined by the correlation coefficient, one can motivate CCA between $\mathbf{y}(t)$ and $\mathbf{y}(t+\tau)$ by an infomax principle across time.

In the general case, the singular value decomposition of the time-lagged covariance matrix returns two *different* orthogonal bases U and V, one for $\mathbf{y}(\mathbf{t})$ and one for $\mathbf{y}(\mathbf{t}-\tau)$ such that the covariance between $U^T\mathbf{y}(\mathbf{t})$ and $V^T\mathbf{y}(\mathbf{t}-\tau)$ becomes diagonal. Since the time-lagged covariance between $\mathbf{y}(\mathbf{t})$ and $\mathbf{y}(\mathbf{t}-\tau)$ equals the linear predictor, one can interpret the two bases as follows: Let $(\mathbf{u}_1, \dots \mathbf{u}_n)$ and $(\mathbf{v}_1, \dots \mathbf{v}_n)$ denote the column vectors of U and V, respectively. Then the linear predictor maps $\mathbf{v}_1 \mapsto \mathbf{u}_1$, $\mathbf{v}_2 \mapsto \mathbf{u}_2$, and so on. In other words, the k-th component $\mathbf{u}_k$ in the input '*predicts*' the k-th component $\mathbf{v}_k$ of the output.

For the problem of invariant representation learning, it is interesting to note that the spectrum of a square matrix can always be decomposed into its symmetric and its anti-symmetric part:

$$C_\mathbf{y}(\tau) = \underbrace{\frac{1}{2}\left(C_\mathbf{y}(\tau) + C_\mathbf{y}^T(\tau)\right)}_{symmetric} + \underbrace{\frac{1}{2}\left(C_\mathbf{y}(\tau) - C_\mathbf{y}^T(\tau)\right)}_{anti-symmetric} \tag{7}$$

These two parts have different geometric meanings. In the next section we show that the *symmetric* part is related to structurally constant features, while the *anti-symmetric* part is related to structural changes such as translations, rotations or other Lie groups. These transforms and their spectral properties have been studied extensively in the context of steerable filter design.[28] As an original contribution of this paper we show how steerable filters are related to the anti-symmetric part of the time-lagged covariance matrix and how this can be used for unsupervised representation learning.

3. PERFECT PREDICTABILITY

In order to assess the differences between the eigenfunctions of the time-lagged covariance matrix in the symmetric and the anti-symmetric case, it is instructive to assume perfect predictability. That is, we assume for now that $\mathbf{y}(t+\tau) = L_\tau\mathbf{y}(t)$ for all t. We will also assume that the instantaneous covariance matrix $C_y(0) = I_d$, as this can always be achieved via the whitening transformation $\mathbf{y}(t) := C_x(0)^{-1/2}\mathbf{x}(t)$.

Perfect predictability imposes strong limitations on the class of possible transformations L_τ: Since $C_y(0) = I_d$ and stationarity imply that $I_d = C_y(0) = L_\tau C_y(0) L_\tau^T = L_\tau L_\tau^T$, we can restrict our discussion to orthogonal transformations. Furthermore, perfect predictability implies that we can write $\mathbf{y}(t) = L_\tau^{t/\tau}\mathbf{y}(0)$ for time-continuous processes.

If we assume, like in 2nd-order ICA, the time-lagged covariance matrix to be symmetric (i.e. $C_y(\tau) = \frac{1}{2}(C_y(\tau) + C_y(\tau)^T)$, then, the only orthogonal transformation L_τ which is continous in τ is the identity $L_\tau \equiv I_d$. This follows from the fact that a symmetric transformation L can always be decomposed into an orthogonal eigenbasis $L = UDU^T$ with $D = diag(D_{11}, \dots, D_{dd})$, and because $LL^T = UD^2U^T = I_d$ implies that $D_{kk}^2 = 1$. Although it is possible to have $\det(L) = 1$ if an even number of the D_{kk} are negative, we can exclude this case by requiring that L_τ can be changed continuously into the identity under the additional constraint of symmetry $L = L^T$ (i.e. we want $\lim_{\tau\to 0}\frac{1}{2}(L_\tau + L_\tau^T) \to I_d$).

In conclusion, a symmetric time-lagged covariance matrix $C_y(\tau)$ and perfect predictability implies that $\mathbf{y}(t) = const$ and hence also $\mathbf{x}(t) = const$. This holds true even in the case of nonlinear feature maps provided that

they are invertible. Conversely, a symmetric $C_y(\tau)$ can differ from the identity matrix only because of noise or imperfect predictability. The anti-symmetric part is not always zero. In contrast to the symmetric part of the covariance matrix, the anti-symmetric part is always sufficient to determine the orthogonal transformation describing the time-evolution under the assumption of perfect predictability.

An orthogonal matrix is in general neither symmetric nor anti-symmetric and its singular value decomposition cannot reveal any pattern because the singular values are all equal to one. However, if the singular value decomposition is applied only to the anti-symmetric part of an orthogonal matrix, then all non-zero singular values show up in pairs: The corresponding left eigenvectors $\mathbf{u}_j, \mathbf{u}_{j+1}$ being mapped onto the same but interchanged right eigenvectors $\mathbf{v}_j = \mathbf{u}_{j+1}$ and $\mathbf{v}_{j+1} = \mathbf{u}_j$. Moreover, if one represents the orthogonal transform L_τ using either the left or the right eigenbasis of its anti-symmetric part $L_\tau - L_\tau^T$, then the corresponding matrix $U^T L_\tau U$ becomes block-diagonal with two-by-two blocks or ± 1 on the main diagonal (see Eq. 10 below).

For any transformation L which is similar to an orthogonal transform, having a representation W such that WLW^{-1} is block diagonal implies that the energy for each block is invariant. For orthogonal transforms it is always possible to find a block diagonal representation such that each block is at most two-dimensional. The corresponding pairs of basis functions can be seen as generalized *quadrature pairs* or *steerable filters*[3,28] of the considered transform L.

Two orthogonal filters are called a 'quadrature pair' if they are related to each other via the Hilbert transform which induces a 90 degree phase shift within each frequency band. Simple examples are $\sin(x)$ and $\cos(x)$ or slightly more generally $\sin(k(x + \Delta x))$ and $\cos(k(x + \Delta x))$. The important role of quadrature pairs in the context of invariant representations shall be illustrated briefly for one-dimensional translations. Let $f(x)$ describe an arbitrary one-dimensional pattern defined over the real axis. This can be shifted via the transform $\mathcal{L}_{\Delta x}[f(x)] := f(x - \Delta x)$. Using the definition of the standard scalar product $\langle g(x), h(x) \rangle = \int g(x)\, h(x) dx$, it is well known that the *power spectrum*

$$P(k) = \langle \sin(kx), \mathcal{L}_{\Delta x}[f(x)] \rangle^2 + \langle \cos(kx), \mathcal{L}_{\Delta x}[f(x)] \rangle^2 \tag{8}$$

is invariant under the (Lie group) transformation L for all k. This finding is a direct consequence of the fact that $\{\exp(ikx) : k \in \mathbb{R}\}$ are the eigenfunctions of L and the modulus of all eigenvalues is equal to one. The real and imaginary part of these eigenfunctions are exactly the sine and cosine functions that show up as 'quadrature pairs' in Eq. 8. $P(k)$ reflects the energy within each subspace in the span of the quadrature pair at frequency k. The fact that $P(k)$ does not depend on Δx means that one can think of L as being a direct sum of individual operators $L = \bigoplus_k L_k$, each of which acts independently on mutually orthogonal subspaces. Conversely, whenever the transformation of a pattern can be decomposed into such elementary transformations, the energy conservation can be made explicit by using quadratic feature maps.

The distinct properties of sines and cosines are easily assessed by noting that any translated sinusoidal function of a given frequency can be perfectly represented by a linear superposition of a pair of sine and cosine functions of same frequency:

$$\sin(kx + \Delta x) = a(\Delta x) \sin(kx) + b(\Delta x) \cos(kx) \tag{9}$$

This property is called *steerability*[3,28–30] and the coefficients as a function of the transformation parameter Δx are called the *steering functions*. The steering functions themselves are again sine and cosine functions for which the sum of squares $[a(\Delta x)]^2 + [b(\Delta x)]^2 = k$ is constant.

4. UNSUPERVISED LEARNING OF GENERALIZED QUADRATURE PAIRS

Now we focus on the issue whether and how we can find two-dimensional subspaces that stay invariant under the action of interesting image transformations. In particular, we explain how the action of certain Lie-groups can be described as a set of rotations within mutually orthogonal, two-dimensional, and real-valued subspaces. To this end, we start with considering the special orthogonal group $SO(n)$. Its elements are the orthogonal

matrices with determinant one, and it forms a real compact Lie group of dimension n(n-1)/2. Moreover, for any orthogonal matrix R it is possible to find a factorization $R = OGO^T$, where O is orthogonal and G has the form

$$G = \begin{pmatrix} G_1 & & & & & \\ & \cdots & & & & \\ & & G_m & & & \\ & & & \pm 1 & & \\ & & & & \cdots & \\ & & & & & \pm 1 \end{pmatrix}, \tag{10}$$

where the G_k are 2×2 rotation matrices as defined below in Eq. 12. The subspace spanned by those column vectors in O which belong to the eigenvalues ± 1 correspond to dimensions that do not change under the transformation R. For the construction of invariant representations we want to exclude these dimensions and hence we assume the following generative model:

$$\mathbf{x}(t) = AGA^{-1}\mathbf{x}(t-\tau) \tag{11}$$

where $G = \bigoplus_{k=1}^{n} G_k$ is a block-diagonal matrix with 2×2 dimensional rotation matrices

$$G_k = \begin{pmatrix} \cos(\Delta\phi_k) & -\sin(\Delta\phi_k) \\ \sin(\Delta\phi_k) & \cos(\Delta\phi_k) \end{pmatrix} \tag{12}$$

on the main diagonal. A is invertible but otherwise arbitrary and plays the role of the mixing matrix. The set of all matrices $\mathcal{M} = \{M : M = AGA^{-1}\}$ defines the class of all possible transforms that can be addressed with this generative model. As discussed above, this set contains all matrices which are similar to an (interesting) orthogonal matrix.

In the following we show how A and G can be inferred from a time series $(\mathbf{x}(0), \mathbf{x}(\tau), \mathbf{x}(2\tau), \ldots \mathbf{x}(n\tau))$ generated according to Eq. 11. We first define $Q := (AA^T)^{-1/2}A$ in order to obtain the factorization $A = (AA^T)^{1/2}Q$ where $(AA^T)^{1/2}$ is symmetric positive definite and Q orthogonal. Inserting this factorization into Eq. 11 we get

$$\mathbf{x}(t) = (AA^T)^{1/2}QGQ^T(AA^T)^{-1/2}\mathbf{x}(t-\tau) \tag{13}$$

Note that AA^T can be estimated from the zero-time lag covariance matrix as it can be shown that $E[\mathbf{x}(t)\mathbf{x}^T(t)] = \frac{1}{2}AA^T$. Next we define two whitened data matrices $X := (AA^T)^{-1/2}(\mathbf{x}(0), \mathbf{x}(\tau), \mathbf{x}(2\tau), \ldots \mathbf{x}((n-1)\tau))$ and $Y := (AA^T)^{-1/2}(\mathbf{x}(1), \mathbf{x}(\tau), \mathbf{x}(2\tau), \ldots \mathbf{x}(n\tau))$ of dimension $d \times n$ which only differ by the presence or absence of $\mathbf{x}(0)$ and $\mathbf{x}(n\tau)$, respectively. Thus we can write compactly

$$Y = QGQ^T X \tag{14}$$

Given that XX^T is invertible we can multiply from the right with the pseudo-inverse $X^- = X^T(XX^T)^{-1}$ yielding

$$YX^- = QGQ^T\,. \tag{15}$$

This equation is different from an eigenvalue decomposition of YX^- only by the fact that G is block diagonal rather than diagonal. For each block it holds that

$$(G_k - G_k^T)^2 = \begin{pmatrix} 0 & -\sin(\Delta\phi_k) \\ \sin(\Delta\phi_k) & 0 \end{pmatrix}^2 = \begin{pmatrix} \sin^2(\Delta\phi_k) & 0 \\ 0 & \sin^2(\Delta\phi_k) \end{pmatrix} \tag{16}$$

so that $(G - G^T)^2$ is diagonal with $\sin^2(\Delta\phi_k)$ on the main diagonal for block 'k'. Hence, whenever the $\Delta\phi_k$ are mutually different, we can identify Q up to a rotation within each block by computing the eigenvalue decomposition of

$$(YX^- - (YX^-)^T)^2 = Q(G - G^T)^2Q^T\,. \tag{17}$$

In other words, we are able to find an orthogonal basis U for which each pair of column vectors $(\mathbf{u}_{2k-1}, \mathbf{u}_{2k})$ spans the same invariant subspace as the corresponding pair of column vectors $(\mathbf{q}_{2k-1}, \mathbf{q}_{2k})$ of the original matrix Q such that $U^TQ = K_\psi$. The block-diagonal rotation matrix K_ψ accounts for the ambiguity of a global phase

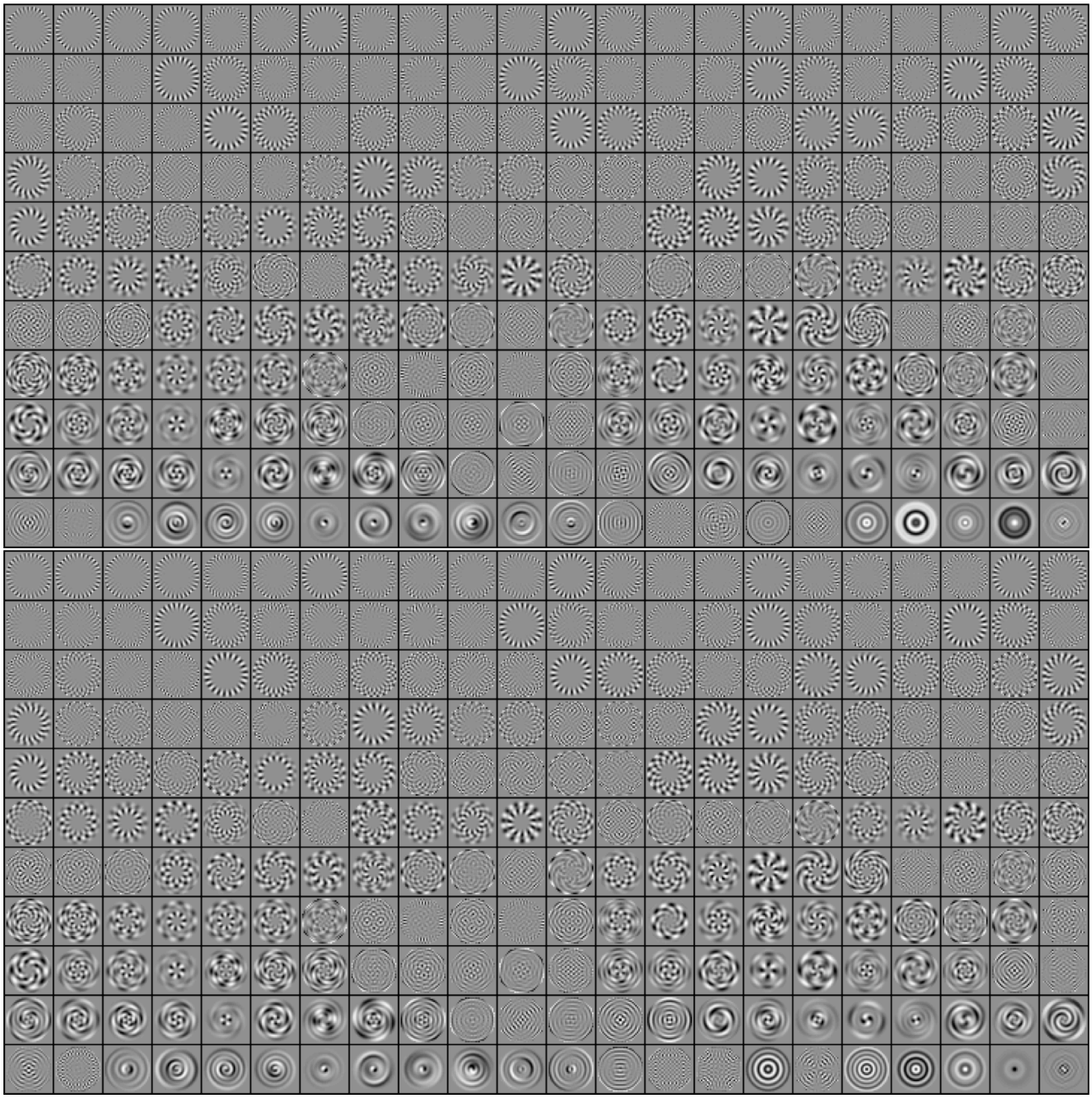

Figure 1. Steerable basis learned with ACA. For each generalized quadrature pair, the two basis functions are at the same location within the upper and the lower box, respectively. The quadrature pairs are sorted with respect to the size of the corresponding eigenvalues (which corresponds to angular frequency).

shift. Note that G can be identified as well because (via Eq. 16) the $\Delta\phi_k$ are determined by the arc sine of the square root of the eigenvalues.

To sum up, let us recall that the time-lagged covariance matrix after whitening equals the optimal linear predictor. In classical CCA, a different left and right basis is used to diagonalize this matrix. Our algorithm, which we refer to as *Anti-Symmetric Correlation Analysis (ACA)*, proceeds by diagonalizing the squared anti-symmetric part $L_A^2 := \frac{1}{2}(L - L^T)^2$ of a linear transformation L.

In contrast to 2-nd order ICA, ACA assumes that the independent sources are not merely vectors, but rather two-dimensional subspaces spanned by generalized quadrature pairs. Therefore, it can be seen as a *2-nd order subspace ICA*.[31]

The *anti-symmetric* part of the time-lagged covariance matrix appears to be more appropriate to model common structural changes in the data: Orthogonal transforms in pixel space do not change the contrast but the amplitude ratios between different basis functions which directly corresponds to structural changes in an image. In the next section, we present an example which shows that ACA can be used to learn a complete basis of rotation invariant, steerable filters.

5. LEARNING A ROTATION INVARIANT, STEERABLE BASIS

For illustration, we use the ACA algorithm for learning a steerable basis consisting of quadrature pairs that can be steered w.r.t rotations in the image plane. To this end we construct a rotation operator for 32×32 patches based on the discrete two-dimensional Fourier basis where each basis vector is mapped onto its rotated version. We use the first 484 basis vectors out of a total of 1024. In order to avoid distortions due to the square shape of the patches all pixels are set to zero outside the circle of 16 pixel radius. Using the eigenvalue decomposition of the anti-symmetric part of the rotation operator, as explained in the previous section, we obtain a complete steerable basis. The result is shown in Fig. 1.

Steerability with respect to rotations is in principle very similar to the case of translation discussed above: if the image is represented in polar coordinates (r, ϕ) with the origin at the center of rotation, then the eigenfunctions of the rotation operator $\mathcal{R}_{\Delta\phi} f(r, \phi) = f(r, \phi + \Delta\phi)$ can be written as $g(r)\exp(i\phi)$. Analogous to the power spectrum in Eq. 8, the squared absolute values $P_\circ(k)$ of these eigenfunctions are invariant under rotations and can similarly be expressed in terms of their real and imaginary part

$$P_\circ(k) = \langle g(r)\sin(k\phi), \mathcal{R}_{\Delta\phi}[f(r,\phi)]\rangle^2 + \langle g(r)\cos(k\phi), \mathcal{R}_{\Delta\phi}[f(r,\phi)]\rangle^2 . \tag{18}$$

6. ROBUST LEARNING OF QUADRATURE PAIRS FROM SEQUENCES OF NATURAL IMAGES VIA BIVECTOR CLUSTERING

There are multiple different sorts of changes in sequences of natural images which cannot all be described by the same transformation. For instance, if the image is translated to the left for some period and then translated to the right for another period, the anti-symmetric part of the corresponding covariance matrices add to zero $\frac{1}{2}(C_x(\tau) - C_x(\tau)^T) + \frac{1}{2}(C_x(-\tau) - C_x(-\tau)^T) = \frac{1}{2}(C_x(\tau) - C_x(\tau)^T) + \frac{1}{2}(C_x(\tau)^T - C_x(\tau)) = 0$. Therefore, we will now modify the ACA algorithm in order to cope with temporal sequences which contain a large variety of different transforms.

In order to see how to modify ACA, it is instructive to explain its geometric meaning first: The anti-symmetric covariance matrix can be interpreted as a representation of the average bivector with respect to a particular basis constructed from pairs of consecutive data points in the temporal sequence. Bivectors are a special case of multi-vectors which are used in geometric algebra[32,33] or Grassmann algebra which allow one to deal conveniently with multi-dimensional subspaces rather than uni-dimensional vectors only. Geometrically, bivectors are a basis-free representation of two-dimensional subspaces or, more specifically, a rotation within that subspace, which can be defined via the totally anti-symmetric wedge product between two plain vectors $S(\mathbf{a}, \mathbf{b}) = \mathbf{a} \wedge \mathbf{b} = -\mathbf{b} \wedge \mathbf{a}$. Then, the 'orientation' of the resulting bivector S corresponds to the subspace spanned by the two plain vectors $\mathbf{a}, \mathbf{b}$, and its length equals $|\mathbf{a}||\mathbf{b}|\sin(\phi)$ where ϕ denotes the angle between $\mathbf{a}$ and $\mathbf{b}$. In particular, the bivector is zero if there is no structural change at all (i.e. if $\mathbf{a} \propto \mathbf{b}$).

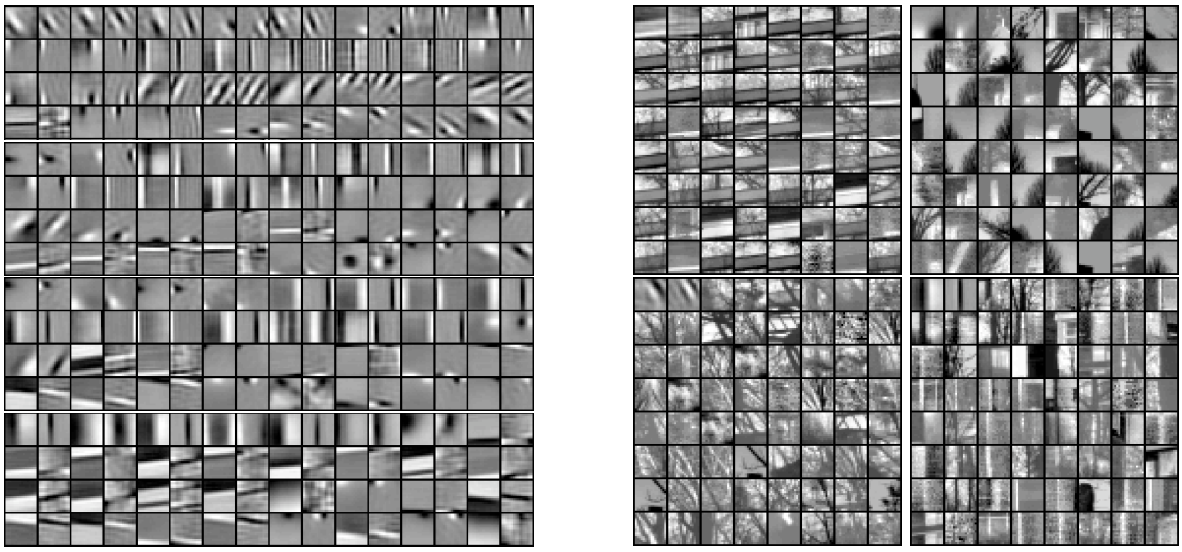

Figure 2. Left: Image basis obtained with the bivector clustering algorithm. The basis functions are ordered such that the quadrature pairs are grouped next to each other within each row. Furthermore, the quadrature pairs are grouped into four clusters of different spatial frequencies (rows 1-4, 5-8, 9-12, 13-16) and ordered with respect to orientation within each cluster. Right: Random samples taken from four different clusters. The quadrature pair representing the cluster centroid is plotted in the upper left corner for each cluster. There is a clear perceptual difference between the different clusters as can be assessed by eye.

In order to avoid a lengthy introduction to the concepts of geometric algebra, we will here only use the fact that for a fixed basis, the bivector coefficients can be defined as anti-symmetrized dyads $\mathbf{a} \wedge \mathbf{b} = \mathbf{a}\mathbf{b}^T - \mathbf{b}\mathbf{a}^T$ using common matrix algebra. Also, every linear combination of anti-symmetrized dyads again defines a bivector with respect to the same basis. Therefore, the anti-symmetric part of the time-lagged covariance matrix can be interpreted as the average bivector constructed from pairs of consecutive data points:

$$C_y(\tau) - C_y(\tau)^T = \frac{1}{N-1} \sum_{k=1}^{N-1} \mathbf{x}(k\tau) \wedge \mathbf{x}((k+1)\tau) = \frac{1}{N-1} \sum_{t=1}^{N-1} \mathbf{x}((k+1)\tau)\mathbf{x}(k\tau)^T - \mathbf{x}(k\tau)\mathbf{x}((k+1)\tau)^T \,. \quad (19)$$

The crucial point here is that the time evolution between two consecutive time steps is modeled by a rotation in the two-dimensional subspace rather than by a difference vector. Since changes in the length of an image patch vector do only change the contrast but not the structural content, it is plausible to model structural changes with bivectors which are insensitive to changes in the length but increase with the size of the rotation angle as $\sin(\phi)$.

In order to learn image bases which are adapted to the structural changes in sequences of natural images, we would like to search for 'typical' subspaces optimally aligned with the clusters of the most common rotation planes in the bivector space. To this end, we use a simple Grassmann clustering algorithm[34] which is based on the k-means algorithm and uses the projection F-norm to define a distance between any pair of subspaces:

$$d([U],[V]) := 2^{-1/2}||UU^T - VV^T||_{Fro} \equiv \frac{1}{\sqrt{2}}\text{trace}\left\{(UU^T - VV^T)^T(UU^T - VV^T)\right\} \,. \quad (20)$$

The subspaces $[U]$ and $[V]$ are represented by matrices $U \in R^{d\times m_1}$ and $V \in R^{d\times m_2}$ with $U^TU = I_{m_1}$ and $V^TV = I_{m_2}$. That is the columns of U and V define the bases of the two subspaces, and UU^T and VV^T represent projections onto these subspaces.

For the two-dimensional subspace spanned by two consecutive image patch vectors, the representing matrix U can be obtained from the sum $\mathbf{s}(k\tau) := \mathbf{x}((k+1)\tau) + \mathbf{x}(k\tau)$ and the difference vector $\mathbf{t}(k\tau) := \mathbf{x}((k+1)\tau) - \mathbf{x}(k\tau)$.

If the image patch vectors are normalized, the sum and difference vectors $\mathbf{s}(k\tau), \mathbf{t}(k\tau)$ are orthogonal, and hence $U = \left[\frac{\mathbf{s}}{||\mathbf{s}||}, \frac{\mathbf{t}}{||\mathbf{t}||}\right]$ can be used to compute the distance of the rotation from the cluster centroids. Note, that the covariance between $\mathbf{s}(k\tau)$ and $\mathbf{t}(k\tau)$ equals the anti-symmetric part of the time-lagged covariance matrix $C_{\mathbf{x}}(\tau)$:

$$\langle \mathbf{st} \rangle = \langle (\mathbf{x}((k+1)\tau) + \mathbf{x}(k\tau))(\mathbf{x}((k+1)\tau) - \mathbf{x}(k\tau))^T \rangle = C_{\mathbf{x}}(\tau) - C_{\mathbf{x}}(\tau)^T \tag{21}$$

This equality elucidates the deeper relationship between ACA, which searches for the eigensubspaces of the *average* bivector, and the bivector clustering algorithm used in this section.

Figure 2 (left) shows the image basis obtained with the bivector clustering algorithm applied to a sequence of 48000 image patches (16×16 pixels) of a movie sequence taken from the van Hateren movie data base as described in.[36] The basis functions are ordered such that the quadrature pairs are grouped next to each other within each row. Furthermore, the quadrature pairs are grouped into four clusters of different spatial frequencies (rows 1-4, 5-8, 9-12, 13-16) and ordered with respect to orientation within each cluster.

Each quadrature pair constitutes the centroid of a cluster of image patches. The r.h. in Fig. 2 shows random samples taken from four different clusters. For comparison, the quadrature pair is plotted in the upper left corner for each cluster. There is a clear perceptual difference between the different clusters as can be assessed by eye.

7. CONCLUSION

We disentangled the task of unsupervised invariant representation learning into two distinct aspects: *dimensionality reduction* and finding an *invariant-equivariant* split of identity. Also, we showed that the influential idea of using 'slowness' or 'temporal stability' can be used only to address the problem of dimensionality reduction but does not serve as a criterion to decide how to represent the relevant subspace.

While 2nd-order ICA allows one to choose a unique basis, it has been shown in Ref. 37 that the linear generative model provides only a poor match to the true statistics of the data. Therefore, we studied the ideal case of perfectly predictable transformations in order to get insights about what kind of statistics we can expect to find in natural images. Following the idea of steerable filters, we argued for decorrelating the data into two-dimensional subspaces rather than searching for one-dimensional eigenspaces. We presented a generative model and an anti-symmetric version of Canonical Correlation Analyis, (ACA), which allows one to determine these subspaces uniquely for non-degenerate spectra.

Each two-dimensional subspace found with this method can be interpreted as a generalized quadrature pair, which instantiates the idea of an invariant-equivariant split of identity. The amplitude encodes only for contrast and the phase only for structural changes. In particular, we demonstrated that ACA can be used to learn a basis of steerable filters from a sequence of a rotating white noise pattern.

Finally, we extended the algorithm in order to cope with more complex data statistics, by clustering bivectors corresponding to transformations in the data, and using the cluster-means as basis functions. We applied this algorithm to movie sequences of natural images, and obtained a basis of generalized quadrature pairs. We demonstrated that this basis separates the data into perceptually distinct clusters.

8. ACKNOWLEDGMENTS

This work is supported by the German Ministry of Education, Science, Research and Technology through the Bernstein award (BMBF; FKZ: 01GQ0601).

REFERENCES

1. D. Kersten, "Predictability and redundancy of natural images," J Opt Soc Am A **4**, pp. 2395–2400, Dec 1987.
2. A. Hyvärinen, Hurri, and Väyrynen, "Bubbles: a unifying framework for low-level statistical properties of natural image sequences," J. Opt. Soc. Am. A **20**(7), pp. 1237–1252, 2003.
3. W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," IEEE Trans. Pattern Anal. Mach. Intell. **13**(9), pp. 891–906, 1991.
4. L. Tong, R.-W. Liu, V. Soon, and Y.-F. Huang, "Indeterminacy and identifiability of blind identification.," IEEE Trans. on Circuits and Systems **38**, pp. 499–509, 1991.

5. K. Diamantaras and S.-Y. Kung, "Cross-correlation neural network models," IEEE TRANSACTIONS ON SIGNAL PROCESSING **42**, pp. 3218–3223, Nov 1994.
6. K. Nordberg, G. Granlund, and H. Knutsson, "Representation and learning of invariance," in Image Processing, 1994. Proceedings. ICIP-94, **2**, pp. 585–589, IEEE International Conference, Nov 2006.
7. P. Perona, "Deformable kernels for early vision," IEEE Trans. Pattern Anal. Mach. Intell. **17**(5), pp. 488–499, 1995.
8. S. Becker, "Mutual information maximization: models of cortical self-organization," Network: Comput. Neural Syst. **7**, pp. 7–31, Feb 1996.
9. H. Knutsson and M. Borga, "Learning visual operators from examples: A new paradigm in image processing," iciap **00**, p. 58, 1999.
10. C. Burges, J. Platt, and S. Jana, "Extracting noise robust features from audio data," in Proc. Int. Conf. Acoustics, Speech, Signal Processing, pp. 1021–1024, IEEE, (Cambridge, MA), 2002.
11. L. Wiskott and T. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances.," Neural Computation **14**(4), pp. 715–770, 2002.
12. H. Hotelling, "Relations between two sets of variates," Biometrika **28**, pp. 321–377, 1936.
13. M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," Automation and Remote Control **25**, pp. 821–837, 1964.
14. B. Schölkopf, A. Smola, and M. Klaus-Robert, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Comput. **10**(5), pp. 1299–1319, 1998.
15. S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K. Müller, "Invariant feature extraction and classification in kernel spaces," 2000.
16. A. Bray and D. Martinez, "Kernel-based extraction of slow features: Complex cells learn disparity and translation invariance from natural images," in Advances in Neural Information Processing Systems, S. Becker, S. Thrun, and K. Obermayer, eds., **15**, MIT Press, (Cambridge, MA), 2003.
17. P. Földiak, "Learning invariance from transformation sequences," Neural Comput. **3**(2), pp. 194–200, 1991.
18. J. Stone, "Learning perceptually salient visual parameters using spatiotemporal smoothness constraints," Neural Comput. **8**(7), pp. 1463–1492, 1996.
19. C. Kayser, K. Kording, and P. Konig, "Learning the nonlinearity of neurons from natural visual stimuli," Neural Comput **15**(8), pp. 1751–9, 2003.
20. A. Hyvärinen, J. Karhunen, and E. Oja, Independent Component Analysis, John Wiley & Sons, 2001.
21. L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," Physical Review Letters **72**(23), pp. 3634–3637, 1994.
22. G. H. Granlund and H. Knutsson, Signal Processing for Computer Vision, Kluwer Academic, Dordrecht, 1995.
23. J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," Neural Comput **12**, pp. 1247–1283, Jun 2000.
24. R. P. N. Rao and D. H. Ballard, "Development of localized oriented receptive fields by learning a translation-invariant code for natural images.," Network: Computation in Neural Systems **9**(2), pp. 219–234, 1998.
25. E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion.," J. of the Optical Society of America A **2**(2), pp. 284–299, 1985.
26. P. Elias, "Predictive coding (part 1+2)," IEEE Transactions on Information Theory **1**(1), pp. 16–33, 1955.
27. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," Neural Comput **16**, pp. 2639–2664, Dec 2004.
28. Y. Hel-Or and P. C. Teo, "Canonical decomposition of steerable functions," J. Math. Imaging Vis. **9**(1), pp. 83–95, 1998.
29. E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," IEEE transactions on informations theory **38**(2), 1992.
30. H. Knutsson, R. Wilson, and G. Granlund, "Anisotropic nonstationary image estimation and its applications: Part i–restoration of noisy images," IEEE Trans. Communications **31**(3), pp. 388–397, 1983.
31. G. Nolte, A. Ziehe, F. Meinecke, and K.-R. Müller, "Analyzing coupled brain sources: distinguishing true from spurious interaction," in Advances in Neural Information Processing Systems, J. P. Yair Weiss, Bernhard Schlkopf, ed., **18**, MIT Press, (Cambridge, MA), 2006.
32. D. Hestenes, Space-Time Algebra, Gordon and Breach, New York, 1966.
33. C. Doran and A. Lasenby, Geometric Algebra for Physicists, Cambridge, Cambridge, 2003.
34. P. Gruber and F. Theis, "Grassmann clustering," in Proceedings of the European Signal Processing Conference, (EUSIPCO 2006), (Florence, Italy), September 2006.
35. T. Kohonen, S. Kaski, and H. Lappalainen, "Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM," Neural Computation **9**(6), pp. 1321–1344, 1997.
36. J. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," Proc R Soc Lond B Biol Sci. **265**(1394), pp. 1724–1726, 1998.
37. M. Bethge, "Factorial coding of natural images: How effective are linear model in removing higher-order dependencies?," J. Opt. Soc. Am. A **23**, pp. 1253–1268, June 2006.