Development of localized oriented receptive fields by learning a translation-invariant code for natural images*

Rajesh P N Rao†§ and Dana H Ballard‡‖

† The Salk Institute, Sloan Center for Theoretical Neurobiology and Computational Neurobiology Laboratory, 10010 N Torrey Pines Road, La Jolla, CA 92037, USA
‡ Department of Computer Science, University of Rochester, Rochester, NY 14627-0226, USA

Received 4 November 1997

Abstract. Neurons in the mammalian primary visual cortex are known to possess spatially localized, oriented receptive fields. It has previously been suggested that these distinctive properties may reflect an efficient image encoding strategy based on maximizing the sparseness of the distribution of output neuronal activities or alternately, extracting the independent components of natural image ensembles. Here, we show that a strategy for transformation-invariant coding of images based on a first-order Taylor series expansion of an image also causes localized, oriented receptive fields to be learned from natural image inputs. These receptive fields, which approximate localized first-order differential operators at various orientations, allow a pair of cooperating neural networks, one estimating object identity ('what') and the other estimating object transformations ('where'), to simultaneously recognize an object and estimate its pose by jointly maximizing the *a posteriori* probability of generating the observed visual data. We provide experimental results demonstrating the ability of such networks to factor retinal stimuli into object-centred features and object-invariant transformation estimates.

1. Introduction

A central problem faced by the visual system is that of recognizing objects irrespective of transformations such as translations, rotations, and scale changes. Neurophysiological studies in the past few decades have provided some important clues regarding the neural mechanisms underlying this invariance to transformations. Hubel and Wiesel [38] first reported the existence of 'complex' cells in the primary visual cortex whose responses remained invariant to the location of stimuli in their receptive field. Neurons invariant to position and size over receptive fields of several degrees of visual angle have also been reported in higher visual areas such as IT in the ventral occipitotemporal pathway [32]. On the other hand, neurons in the dorsal occipitoparietal stream appear to be coding for various types of transformations, irrespective of stimulus-specific properties. For example, cells in the area MSTd have been shown to respond to transformations such as translations, rotations, and expansions/contractions [22]. Thus, the neurobiological data seem to suggest

[*] This work was performed at the Department of Computer Science, University of Rochester, Rochester, NY, USA.
[§] E-mail: rao@salk.edu
[‖] E-mail: dana@cs.rochester.edu

0954-898X/98/020219+16$19.50

that the visual system factors retinal stimuli into object-centred features and their relative transformations.

It is also known that visual cortical neurons, in particular those in primary visual cortex, possess localized, oriented receptive fields. It was first suggested by Hubel and Wiesel [38] that these neurons could be coding for edges and bars in input images at different orientations. Motivated by the property that natural images possess a $1/f^2$ power spectrum [26], Atick and Redlich [2, 3] provided an information-theoretic explanation of the centre-surround structure of retinal ganglion receptive fields in terms of whitening or decorrelation of outputs at low spatial frequencies and low-pass filtering for noise suppression at high spatial frequencies. Several Hebbian learning algorithms for decorrelation have also been proposed [4, 10, 14, 28, 46, 49, 61, 70], many of which perform principal component analysis (PCA). Although the PCA of natural images produces lower-order components that resemble oriented filters [5, 33], the higher-order components are unlike any known neural receptive field profiles. In addition, the receptive fields obtained are global rather than localized feature detectors.

Recently, it has been shown that a neural network that maximizes the sparseness of the distribution of output activities develops, when trained on natural images, synaptic weights with localized, oriented receptive fields resembling those in primary visual cortex [51]. Similar results have also been obtained using an algorithm based on a linear transform model that attempts to make the outputs of the transform as statistically independent as possible, given the assumption that the cumulative density functions of the outputs can be modelled as sigmoidal functions [13].

These algorithms are all based directly or indirectly on Barlow's principle of *redundancy reduction* [6–9], where the goal is to learn 'feature detectors' whose outputs are as statistically independent as possible, the underlying motivation being that sensory inputs such as images are generally comprised of a set of independent objects or features. An interesting issue not directly addressed by the above learning algorithms is the invariance of these encoded image features to transformations such as translations, rotations or scaling. In particular, one may ask if there exist coding strategies that additionally include the constraint of transformation invariance of object features and provide an alternative explanation of the localized, oriented nature of receptive fields of visual cortical neurons.

In this paper, we answer this question in the affirmative by showing that a set of localized oriented receptive fields arise as a result of a translation-invariant coding strategy based on first-order Taylor series approximations of natural images. The coding strategy gives rise to a pair of cooperating neural networks that jointly maximize the *a posteriori* probability of generating the observed visual data. The first network estimates the identity of an object or feature ('what') and is closely related to the (hierarchical) Kalman filter networks previously studied in [57]. The second network estimates the relative transformations due to object motion ('where'). We show that, when trained on natural images containing small (first-order) translations, neurons in the transformation estimating network develop localized oriented receptive fields approximating first-order differential operators at various orientations, thereby suggesting an alternative functional interpretation of cortical neurons with such receptive fields. Such an interpretation is consistent with the ideas of Koenderink, van Doorn, and others [40, 41, 72]. We verify the efficacy of these learned receptive fields by testing the performance of the network in factoring novel input images containing translated objects into object-centred features and object-invariant translation estimates.

2. Image representation and optimization

Assume that an image, denoted by a vector $\boldsymbol{I}$ of n pixels, can be represented as a linear combination of a set of k basis vectors $U_1, U_2, \ldots, U_k$:

$$\boldsymbol{I} = \sum_{j=1}^{k} U_j r_j. \tag{1}$$

The coefficients r_j denote an internal representation of the image $\boldsymbol{I}$ with respect to the internal model defined by the basis vectors U_j. It is convenient to rewrite the above equation in matrix form as

$$\boldsymbol{I} = U\boldsymbol{r} \tag{2}$$

where U is the $n \times k$ matrix whose columns consist of the basis vectors U_j and $\boldsymbol{r}$ is the $k \times 1$ vector consisting of coefficients r_j. In a neurobiological setting, the values in the ith row of U can be regarded as the strength of the synapses in the ith model neuron while the coefficients r_j denote the pre-synaptic activities received by the neuron.

The key idea behind the model is that one can approximate a new transformed image $\boldsymbol{I}(\boldsymbol{x})$ using a Taylor series expansion around a previously encountered reference image $\boldsymbol{I}$:

$$\boldsymbol{I}(\boldsymbol{x}) = \boldsymbol{I} + \frac{\partial \boldsymbol{I}}{\partial \boldsymbol{x}}\boldsymbol{x} + \text{higher-order terms} \tag{3}$$

where $\boldsymbol{x}$ is an $m \times 1$ vector denoting the relative transformation that the image has undergone. Strictly speaking, if we are concerned only with planar translations, $m = 2$ suffices since one can represent arbitrary translations with a two-component vector. Similarly, rotations within the image plane can be handled using a scalar parameter x ($m = 1$). However, in general, we would like to be able to allow m to be a larger value for several reasons. First, making m larger than, say, 1 or 2 (for representing rotations or translations) causes $\boldsymbol{x}$ to be a distributed representation of the given transformation. Besides being biologically more plausible, distributed representations enjoy several favourable properties [35] such as better generalization, robustness, and resistance to faults in memory and internal noise. Secondly, making m a sufficiently large value endows the network with the ability to handle certain transformations with larger degrees of freedom that may not have been anticipated at design time. Finally, our experimental results suggest that larger values of m may help in the learning process by allowing greater flexibility and stability during the search for a solution to the invariant recognition/pose estimation problem (section 3).

The $n \times m$ matrix of partial derivatives $J = \partial \boldsymbol{I}/\partial \boldsymbol{x}$ in equation (3) above is known as the *Jacobian matrix*. One way of approximating the Jacobian J is to simply use a fixed matrix learned from a set of training images [56]. Unfortunately, this does not acknowledge the fact that the Jacobian is a *function of the current reference image*. An alternative method that addresses this concern is to approximate the Jacobian as a linear function of the reference image $\boldsymbol{I}$. Let J_i be the ith column of the Jacobian matrix J. Then, we have the relation:

$$J_i \cong D_i \boldsymbol{I} \tag{4}$$

where D_i is an $n \times n$ matrix whose rows are basis vectors corresponding to the component x_i of the transformation vector $\boldsymbol{x}$. Note that to implement the Jacobian, the jth row of D_i needs to compute an approximation of the differential operator $\partial(\cdot)_j/\partial x_i$, i.e. the output of the operator when applied to image $\boldsymbol{I}$ needs to be $\partial I_j/\partial x_i$ where I_j is the jth pixel of the image. Once again, it is easy to see that the operation in equation (4) can be performed by a set of n linear neurons whose synapses encode the basis vectors forming the n rows of D_i and whose pre-synaptic inputs are given by $\boldsymbol{I}$.

For deriving the estimation and learning rules, it is convenient to use the $n \times nm$ matrix D obtained by concatenating the various transformation basis matrices D_i, i.e. $D = [D_1\ D_2\ \cdots\ D_m]$. Note that in addition to n (the number of pixels in the input image), the size of D depends directly on the dimension m we choose for the transformation vector $\boldsymbol{x}$ and hence on the degree of distributed storage we desire in our representation of transformations. Since D scales as n^2 for each new dimension in $\boldsymbol{x}$, there is clearly a trade-off between how large a matrix D we can afford to learn and how distributed we want our representations to be. For the experiments described in this paper (section 4), we used $m = 12$ when the input images were of size 13×13 ($n = 169$) and $m = 6$ when the inputs were 21×21 ($n = 441$).

Using the definition $D = [D_1\ D_2\ \cdots\ D_m]$, the various equations (4) for $i = 1, \ldots, m$ can be combined into one equation as follows:

$$J = \frac{\partial \boldsymbol{I}}{\partial \boldsymbol{x}} \cong D\mathcal{I} \tag{5}$$

where $\mathcal{I}$ is the $nm \times m$ matrix containing m copies of the reference image $\boldsymbol{I} = U\boldsymbol{r}$:

$$\mathcal{I} = \begin{bmatrix} \boldsymbol{I} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \ldots & \mathbf{0} & \boldsymbol{I} \end{bmatrix}. \tag{6}$$

Thus, in summary, the above arrangement allows one to approximate the Jacobian for an arbitrary image using the basis vectors in D without having to store image-specific Jacobians for each object. As shown in equation (5), the matrix D allows the ith column of the Jacobian matrix to be specified by multiplication of the image with the corresponding submatrix D_i of D (by definition, D_i is simply the columns $n(i-1)+1$ through ni of D).

Our goal is to estimate the coefficients $\boldsymbol{r}$ and the transformation vector $\boldsymbol{x}$ for a given image and, on a longer time scale, learn appropriate basis vectors in U and D directly from the input image stream. For small transformations, one can ignore the higher-order terms in equation (3) and model their effects as stochastic noise:

$$\boldsymbol{I}(\boldsymbol{x}) = \boldsymbol{I} + J\boldsymbol{x} + \boldsymbol{n} \tag{7}$$

$$= U\boldsymbol{r} + D\mathcal{I}\boldsymbol{x} + \boldsymbol{n} \tag{8}$$

where $\boldsymbol{n}$ is assumed to be a Gaussian white noise process with zero mean and unit variance. The above equation can be regarded as implementing a form of *bilinear* generative model for input images (see also [30]). We can now define the following squared-error optimization function:

$$E_1 = (\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x}) \tag{9}$$

$$= (\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - XU\boldsymbol{r})^{\mathrm{T}}(\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - XU\boldsymbol{r}) \tag{10}$$

where the superscript T denotes vector (or matrix) transpose and $X = \sum_{i=1}^{m} x_i D_i$. It can be shown that minimizing E_1 is equivalent to *maximizing the log likelihood* of generating the observed data $\boldsymbol{I}(\boldsymbol{x})$ with respect to the model parameters U, D, $\boldsymbol{r}$, and $\boldsymbol{x}$ (see, for example, [57]).

Without additional constraints, a least-squares optimization function such as E_1 (without the first-order component $D\mathcal{I}\boldsymbol{x}$) generates solutions similar to principal component analysis (PCA), which has been shown to yield poor descriptors of natural image distributions [27, 51]. This motivates us to use additional constraints to limit the range of solutions. For example, one can add to E_1 terms relating to prior distributions for the parameters. Here, we use zero-mean Gaussian distributions for the model priors since these were found to be sufficient for localized receptive field development in the transformation network (see section 4). The resulting optimization function is given by:

$$E = E_1 + \alpha ||\boldsymbol{r}||^2 + \beta ||\boldsymbol{x}||^2 + \gamma ||U||^2 + \lambda ||D||^2 \tag{11}$$

where the operator $|| \cdot ||^2$ denotes the sum of squares of the elements of the vector or matrix argument. Viewed probabilistically, the first term in the above sum corresponds to the negative logarithm of the likelihood of generating the observed visual data while the remaining terms represent the negative logarithms of the prior probabilities of the model parameters [51, 57]. The coefficients α, β, γ, and λ are parameters related to the variances of the prior distributions. Thus, by Bayes theorem, minimizing E is equivalent to maximizing the *a posteriori* probability of generating the observed data $\boldsymbol{I}(\boldsymbol{x})$ (see, for example, [17, 57]).

3. Network dynamics and synaptic learning rules

For the purposes of stability, we minimize E with respect to $\boldsymbol{r}$ and $\boldsymbol{x}$ for fixed values of U and D. The basis vectors U and D are learned on a slower time scale for fixed values of $\boldsymbol{r}$ and $\boldsymbol{x}$. This form of alternating between optimization of parameters can be viewed as implementing an approximate on-line form of the expectation-maximization (EM) algorithm from statistics [12, 20].

For a given set of basis vectors D and U, one can minimize E with respect to $\boldsymbol{r}$ and $\boldsymbol{x}$ using gradient descent to obtain the following differential equations for estimating object identity and transformation:

$$\dot{\boldsymbol{r}} = -\frac{k_1}{2}\frac{\partial E}{\partial \boldsymbol{r}} = k_1 (U + XU)^{\mathrm{T}} (\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x}) - k_1 \alpha \boldsymbol{r} \tag{12}$$

$$\dot{\boldsymbol{x}} = -\frac{k_2}{2}\frac{\partial E}{\partial \boldsymbol{x}} = k_2 (D\mathcal{I})^{\mathrm{T}} (\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x}) - k_2 \beta \boldsymbol{x} \tag{13}$$

where $\dot{\boldsymbol{r}}$ and $\dot{\boldsymbol{x}}$ represent the temporal derivatives of $\boldsymbol{r}$ and $\boldsymbol{x}$ respectively, and k_1 and k_2 are positive time constants of the dynamics that determine the rate of descent towards the minima of E. Thus, given a transformed image $\boldsymbol{I}(\boldsymbol{x})$, one needs to compute the *residual error* between the input $\boldsymbol{I}(\boldsymbol{x})$ and its prediction $U\boldsymbol{r} + D\mathcal{I}\boldsymbol{x}$ which was made using the internal model given by U and D. In the case of the object identity estimate $\boldsymbol{r}$, the residual is filtered using the 'feedforward' matrix $(U + XU)^{\mathrm{T}}$ $(= U^{\mathrm{T}}(I + X)^{\mathrm{T}}$, I being the identity matrix) where as in the case of the transformation estimate $\boldsymbol{x}$, the residual is filtered via the matrix $(D\mathcal{I})^{\mathrm{T}}$ $(= J^{\mathrm{T}})$. Note that both the object network and the transformation network use the same residual signal to correct their estimates $\boldsymbol{r}$ and $\boldsymbol{x}$, and both contribute to it. The residual itself can be computed using, for instance, inhibitory feedback of the input. The integration over time required by the differential equations above can be implemented by classical leaky integrate-and-fire neurons (see, for example, [19]).

Figure 1 depicts a neural implementation of the above equations in the form of two parallel but cooperating networks, one estimating object identity ('what') and the other estimating object transformations ('where'). This functional dichotomy between object recognition and transformation estimation is reminiscent of the well known division of labor between the dorsal and ventral streams in the primate visual cortex [24]. An especially favourable property of such an arrangement is that the estimate of object

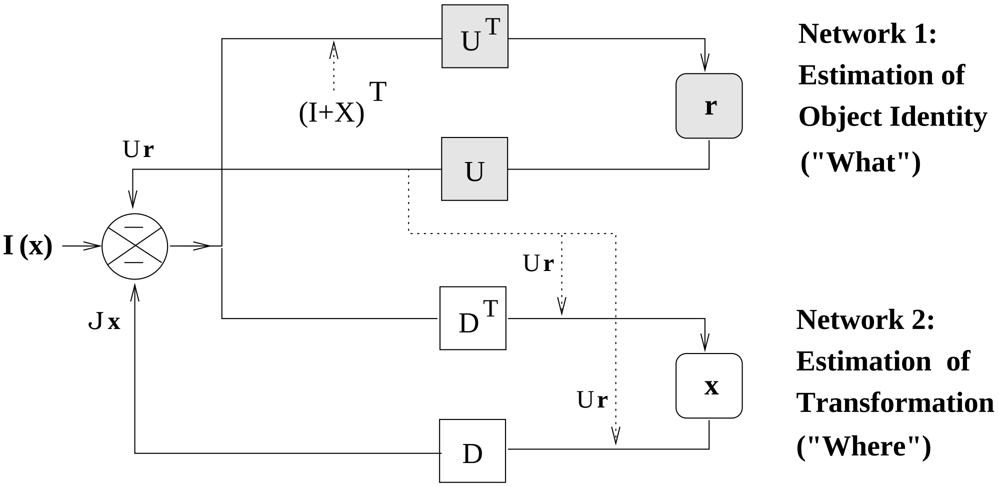
U^T
(I+X)^T
r
Ur
U
I (x)
Network 1: Estimation of Object Identity ("What")
Ur
D^T
Jx
x
Ur
D
Network 2: Estimation of Transformation ("Where")

Figure 1. Network architecture of the model. The shaded portion represents the object identity ('what') network [57] that estimates the zeroth-order component of the input. The bottom unshaded portion of the figure shows the transformation estimating ('where') network that computes the first-order transformations in the input. The '$-$' signs within the circle denote inhibitory feedback for computing of the feedforward residual $(\boldsymbol{I}(\boldsymbol{x})-U\boldsymbol{r}-J\boldsymbol{x})$, where $J = D\mathcal{I}$. The dotted lines indicate connections conveying information between the two otherwise parallel networks.

identity remains stable in the first network as the second network attempts to account for any transformations being induced in the image plane, appropriately conveying the type of transformation being induced in its estimate for $\boldsymbol{x}$. The stability of object identity in the presence of transformations has also been the goal of a number of previously proposed models [31, 34, 43, 44, 50, 53, 69]. Some previous methods for invariance have utilized modifications to the distance metric used for comparing input images to stored templates (such as tangent distance methods [62]) while others have relied on temporal sequences of input patterns undergoing transformations [11, 29, 42, 64, 68, 71]. Unlike the present approach, many of the above methods convey no explicit information regarding the transformation itself or worse, sacrifice information about the transformation in order to achieve invariance.

For specific object and transformation vectors $\boldsymbol{r}$ and $\boldsymbol{x}$, one can minimize E with respect to the object basis matrix U and the transformation basis matrix D, to obtain the following 'learning rules' for adapting the synaptic efficacies represented by these two matrices:

$$\dot{U} = -\frac{c_1}{2}\frac{\partial E}{\partial U} = c_1(I+X)^{\mathrm{T}}(\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x})\boldsymbol{r}^{\mathrm{T}} - c_1\gamma U \tag{14}$$

$$\dot{D} = -\frac{c_2}{2}\frac{\partial E}{\partial D} = c_2(\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x})(\mathcal{I}\boldsymbol{x})^{\mathrm{T}} - c_2\lambda D \tag{15}$$

where $\dot{U}$ and $\dot{D}$ again represent temporal derivatives, c_1 and c_2 are positive time constants that determine the learning rate, and I is the $n \times n$ identity matrix. Note that once again, the residual error $(\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x})$ plays a crucial role in correcting the weights U and D. In addition, both learning rules are Hebbian forms of synaptic adaptation with decay. For instance, in the case of U, the adaptation is proportional to the product of the pre-synaptic activity $\boldsymbol{r}^{\mathrm{T}}$ and the post-synaptic activity $(I + X)^{\mathrm{T}}(\boldsymbol{I}(\boldsymbol{x}) - U\boldsymbol{r} - D\mathcal{I}\boldsymbol{x})$ (see figure 1).

4. Experimental results

The experiments were designed to address two important questions regarding the proposed approach: (i) can the appropriate operators for the differentiation in equation (3) be learned directly from natural images without any additional constraints, and (ii) are these operators learned from natural images sufficient for translation-invariant recognition of novel objects not in the original training set, assuming that the translations induced are comparable to those used during training? A secondary but nevertheless important question was whether the translation estimates themselves remain relatively constant even when different objects are translated. Such object-invariant transformation estimates are extremely desirable since they allow a sensorimotor system to learn sensorimotor mappings that can generalize easily across objects.

4.1. Development of localized receptive fields

In order to answer question (i) above, we tested the learning algorithm for D ('where' network) on a set of natural image patches. One of the natural images used for extracting the training patches is shown in figure 2(A). The natural images were low-pass filtered with a 5×5 Gaussian kernel to attenuate the effects of image noise. For the first experiment, the receptive field size (equal to the image patch size) was set to 13×13 pixels. The box labelled RF1 in the figure depicts the relative size of the receptive field with respect to the natural image.

Training inputs $\boldsymbol{I}(\boldsymbol{x})$ were obtained by translating randomly-selected reference image patches ($= \boldsymbol{I}$ in equation (3)) horizontally or vertically in one of four random directions by 2 pixels. The object estimate $\boldsymbol{r}$ for each reference patch was clamped during the translation and thus the object network output remained fixed at $\boldsymbol{I}$. An estimate $\boldsymbol{x}$ for the relative translation between $\boldsymbol{I}$ and $\boldsymbol{I}(\boldsymbol{x})$ was obtained by allowing equation (13) to converge to a stable value. This translation estimate was then fixed for the current and next 10 reference image patches translated in the same direction, and for each of these 11 translations, equation (15) was used to modify the basis matrix D. The network parameters were set as follows: $m = 12$, $k_2 = 0.2$, $\beta = 0.008$, $\lambda = 0.0005$. The learning rate parameter c_2 was initialized to 0.4 and decreased after every 400 input presentations by dividing with 1.008.

Figure 2(B) shows intensity-coded images of 16 of a set of 169 learned basis vectors (rows of the basis matrix D_i, $i = 1$) after convergence. Bright regions are positive values ('excitatory synapses'), dark regions denote negative values ('inhibitory synapses'). As exemplified by these 16 basis vectors, each of the 169 vectors in the matrix was found to be oriented in the same direction (diagonal in this case) and each was localized to the region centred on its respective pixel location in the image. Selected vectors in the other matrices D_i for $i = 2, \ldots, 12$ are shown in figure 2(C). Each row of three images represents three of the 169 rows of each D_i, $i = 2, \ldots, 12$. Once again note the iso-orientation of the filters for any particular i, localized to their respective positions. By learning copies of such iso-orientation 'derivative' filters within the rows of D_i, the network is able to convolve an image with such filters, thereby satisfying equation (5).

The results of learning were remarkably robust to image patch size and natural image samples. Figure 3 shows the results for a receptive field size of 21×21. The network parameters in this case were identical to the 13×13 case with the exception of $m = 6$. Once again, the basis vectors are localized and oriented in space. For each i, the basis vector forming the jth row of D_i was found to have converged to an approximation of the differential operator $\partial(\cdot)_j/\partial x_i$ as required by the model (equation (5)).

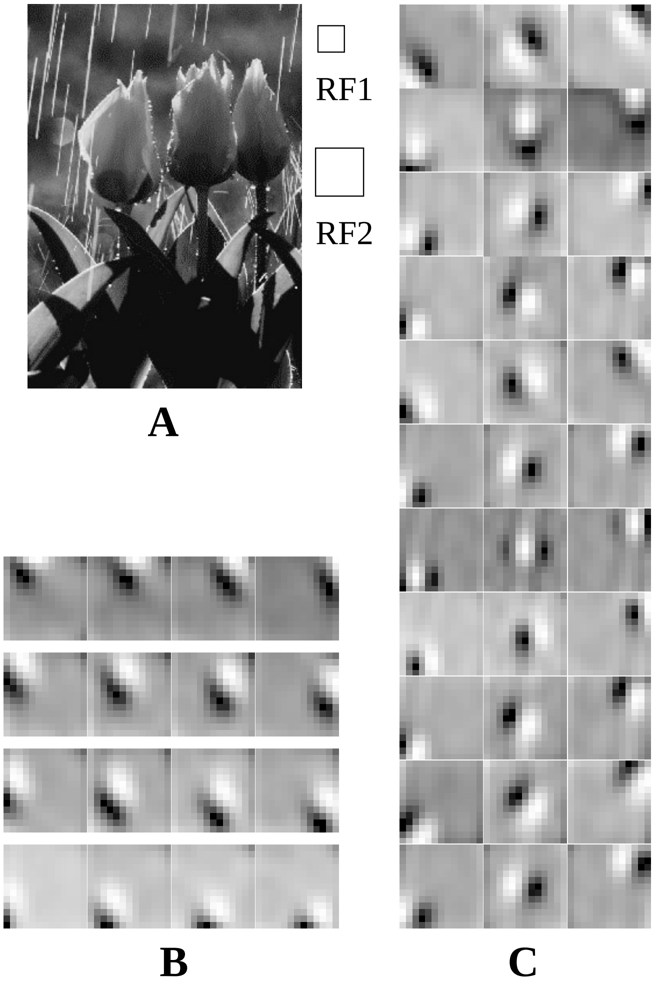

Figure 2. Localized receptive fields for translation estimation learned from natural images (RF size 13 × 13). (A) A natural image used for training. A randomly selected image patch was translated in one of four directions and the translation estimate $\boldsymbol{x}$ thus obtained was then fixed for the current and next 10 image patches, each of which was translated in the same direction. For each of these 11 translations, equation (15) was used to train D. The boxes labelled RF1 and RF2 depict the relative size of the receptive fields of size 13×13 and 21×21 as compared to the natural image. (B) Intensity-coded images of 16 of the 169 learned basis vectors (rows of D_i, $i = 1$). Bright regions are positive values (excitatory synapses), dark regions are negative values (inhibitory synapses). These vectors appear to be tuned towards diagonal translations. Note that these receptive fields are all at the same orientation but localized to their respective positions in the image array. (C) A selected set of learned basis vectors for $i = 2, \ldots, 12$: each row of three images represents three of the 169 rows of each D_i. Once again note the iso-orientation of the filters for any particular i, localized at their respective positions.

For learning the basis vectors U in the ‘what’ network, Olshausen and Field [51] have previously demonstrated that an algorithm similar to equation (14) for learning U but with a non-Gaussian prior distribution on $\boldsymbol{r}$ (equation (12)) produces localized oriented basis vectors within the columns of U. We have recently shown [58] that a rectified Gaussian prior as in equation (11) (but with a full covariance matrix instead of α) leads to a lateral

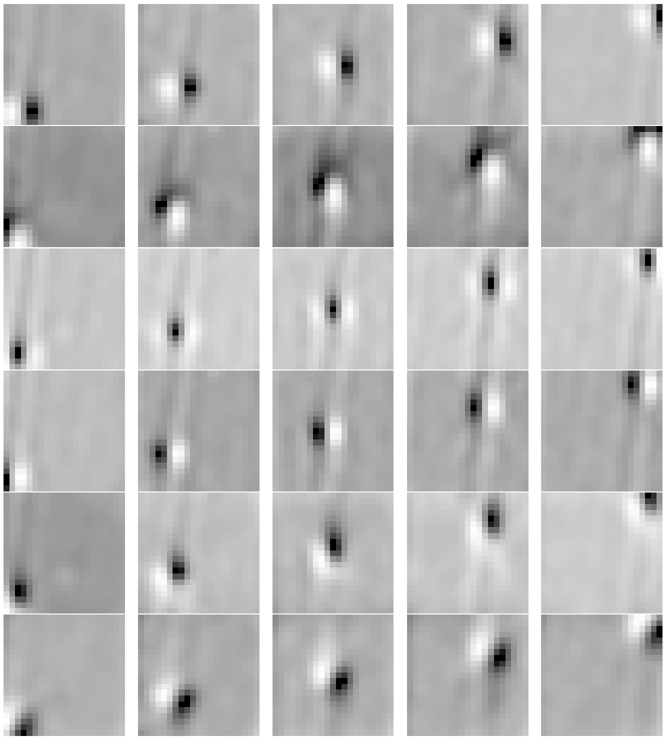

Figure 3. Receptive fields for translation estimation (RF size 21 × 21). The results of learning were found to be remarkably robust to variations in image patch size and natural image samples. Shown here are the results obtained for a receptive field size of 21 × 21. Each of the six rows shows five of the 441 basis vectors forming the rows of a given D_i, for $i = 1, \ldots, 6$. All basis vectors for a given D_i are oriented in the same direction, each being localized to its respective position within the image array. Learning copies of such iso-orientation differential operators within the rows of D_i allows the network to perform the differentiation in the Taylor series expansion of equation (5).

inhibition term in the dynamics of $\boldsymbol{r}$ [28] which in turn allows localized and oriented receptive fields to be developed (rectification is implemented by imposing the constraint of non-negativity on the components of $\boldsymbol{r}$) . In the next section, we utilize the learning rule for U as given in equation (14) to test the efficacy of the learned translation basis vectors D_i in mediating translation-invariant recognition of novel objects.

4.2. Translation-invariant pattern recognition

The basis vectors described in the previous section were tested in a simple translation-invariant pattern recognition task involving a small set of novel man-made objects. Figure 4(A) shows images (of size 21 × 21) that were used to train the object identity ('what') network. For each image, the object estimate $\boldsymbol{r}$ was allowed to converge according to equation (12) and the weights U were subsequently modified according to equation (14). The transformation estimate during the object learning process was set to $\boldsymbol{x} = \mathbf{0}$. The network parameters were set as follows: $k = 15$, $k_1 = 0.3$, $\alpha = 0.008$, $\gamma = 0.0005$. The learning rate parameter c_1 was initialized to 0.4 and decreased gradually by dividing with 1.0008 at each iteration (one sweep through the training images) for 5000 iterations.

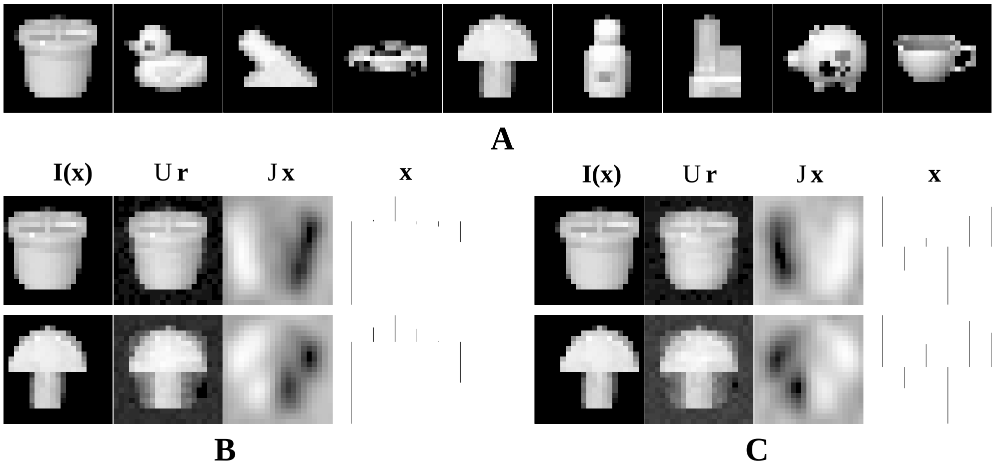

Figure 4. Example of translation-invariant pattern recognition. (A) Images (of size 21 × 21) used for training the object identity ('what') network. (B) and (C) Response of the trained networks (figure 1) to two different objects translated leftward (B) and rightward (C) by two pixels. Note that in each case, the translated image is factored into the original image ($U\boldsymbol{r}$) and a shift ($J\boldsymbol{x}$). This allows the 'what' network to recognize the object while the 'where' network simultaneously computes the pose or relative transformation (in this case, a translation) that the original object has undergone. An especially attractive property of such an arrangement is that the transformation vector $\boldsymbol{x}$, shown here as a histogram (upward bar = positive value, downward bar = negative value), is approximately the same for a given translation even though different objects were translated.

Figures 4(B) and 4(C) show the response of the two networks (figure 1) after training, to two different objects translated leftward and rightward respectively by two pixels. The learned basis vector matrix D from figure 3 was used in the 'where' network. Note that in each case, the network was able to factor the translated image into the original image ($U\boldsymbol{r}$) and a shift ($J\boldsymbol{x}$). This allows the 'what' network to recognize the object in spite of the translation. Although the translation involved in this particular example is admittedly quite small (section 5 discusses the issue of larger transformations), it serves to illustrate an important difference between the proposed approach and some previous approaches to invariant recognition (for example, [31]): the invariance to transformations is not achieved at the cost of sacrificing important information regarding the transformations themselves. Rather, an estimate of the current transformation is obtained simultaneous with the invariant estimation of object identity. In addition, as seen in the figure, the transformation vector $\boldsymbol{x}$ (upward bar = positive value, downward bar = negative value) is approximately the same for the given translation even though different objects were translated.

The constancy of translation estimates across objects was verified for the objects in the test set as shown in figure 5. The upper plot is the correlation (normalized vector dot product) between the translation estimate vector $\boldsymbol{x}_1$ for object 1 and the translation estimate vectors for all other objects in the training set for a rightward translation. The lower plot shows the correlation between $\boldsymbol{x}_1$ and translation estimates for all training objects for a leftward translation. The high and relatively constant positive correlations among right-translation vectors for different objects (the upper plot in figure 5) support the hypothesis that transformation estimates in the model are relatively object-invariant. This is further supported by the relatively constant and high negative correlations between

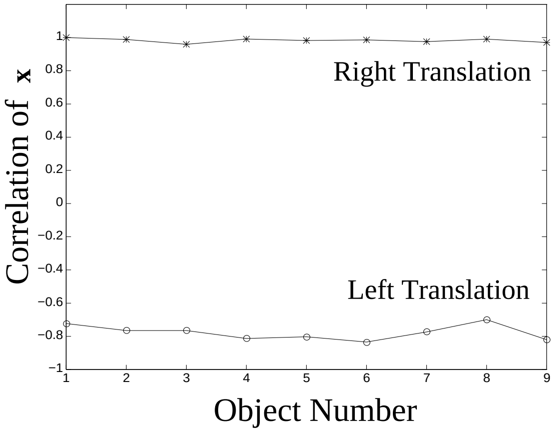

Figure 5. Constancy of estimated translation across objects. Upper plot: the correlation between the estimated translation vector $\boldsymbol{x}_1$ for object 1 and the estimated translation vectors for all other objects in the training set for a *rightward* translation. Lower plot: the correlation between $\boldsymbol{x}_1$ and *leftward* translation vectors for all training objects. Note the high (and relatively constant) positive correlations among rightward translation vectors and the high (and relatively constant) negative correlations with the leftward translation vectors, supporting the hypothesis that the transformation estimates are relatively object-invariant.

rightward translation vector $\boldsymbol{x}_1$ and the left translation vectors for each object (the lower plot in figure 5). The high negative values for these correlations demonstrate that the estimates for rightward translations are sufficiently different from the estimates for leftward translation, thereby allowing reliable discrimination between rightward and leftward translations.

The independence and decoupling of the transformation estimates $\boldsymbol{x}$ from object estimates $\boldsymbol{r}$ is especially important for learning general sensory-motor routines (for example, grasping a cup) that can be uniformly applied across objects without regard to object specific visual features (for example, patterns/writing on the cup or the colour of the cup) that are usually irrelevant to motor programming. Furthermore, when a transformation estimate $\boldsymbol{x}$ is used to drive a motor routine such as a saccadic eye movement, the resulting 'efference copy' [16] of the motor signal can be used to update the transformation estimate $\boldsymbol{x}$ [39]. This updating of internal spatial representations by intended movements has been observed in the parietal cortex [23] and has inspired numerous models based on the notion of 'gain fields' [55, 60, 73]. The work presented here suggests a possible neural mechanism for converting the raw retinal information to spatial location estimates, which can subsequently be utilized to program motor actions and which can in turn be modulated by eye movements and other intended motor activities.

5. Discussion

The experimental results indicate that a set of localized and oriented basis vectors for estimating first-order image translations can be learned directly from natural images using a Taylor series based approach to encoding images. Such an approach allows a pair of cooperating neural networks, one estimating object identity ('what') and the other estimating object transformations ('where'), to simultaneously recognize an object and estimate its pose by jointly maximizing the posterior probability of generating the input data. The property of relative invariance of the object estimate in the presence of translations suggests a possible explanation for the invariant response of a complex cell to stimuli placed in different locations within its receptive field [38]. Although the invariant response of the cell appears highly nonlinear when viewed in isolation, this invariance can also be explained by considering the responses of cells in a companion network that account for the first-order terms. Thus, as the input is displaced to different locations in the receptive field, the zeroth-order response of the complex cell remains unchanged whenever the displacement can be modelled by a first-order variation using the transformation estimating cells in the companion network.

An obvious drawback of using a first-order Taylor series for approximating image transformations is the limitation to relatively small transformations of image features. Although larger transformations can be estimated by a first-order model up to a certain degree of accuracy (figure 6), the error in image reconstruction gradually increases due to the insufficiencies of a first-order approximation. A possible solution is to use a hierarchical multiscale estimation scheme. Black and Jepson [15] demonstrate the viability of such an approach in the context of an image pyramid scheme. A neurally plausible alternative is to use a hierarchical network, such as the hierarchical Kalman filter model proposed in [57], wherein higher levels operate over larger spatial scales than lower ones and maintain more global, more abstract, and coarser estimates. In such an approach, a given transformation is represented in a hierarchical and distributed fashion within the various levels. A hierarchical structure is especially desirable since it counters the well known *aperture problem* [1] in motion estimation by allowing information from larger

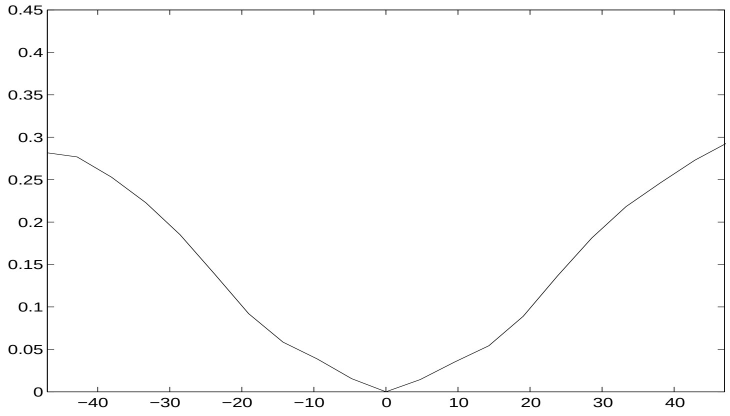

Figure 6. The effect of large translations on image reconstruction error. The basis vectors from figure 3, which were learned from two pixel translations of natural image patches, were tested for larger transformations. A set of 100 randomly selected natural image patches (vectors $\boldsymbol{I}$ normalized to length 1) were translated in two different directions (leftwards and rightwards), and the values of the optimization function E, representing the image reconstruction error, were used to plot the average reconstruction error over the 100 translated patches as a function of the amount of image translation (in percentages of the receptive field (RF) size). Negative percentages denote leftward translations while positive percentages denote rightward translations. The graph indicates that despite being trained on only two pixel translations, the basis vectors can nevertheless represent larger translations reaching up to ±30–35% of RF size, after which the image reconstruction error may reach too high a value for the purpose of transformation-invariant recognition. This motivates the need for hierarchical, multiscale approaches and Lie group-based methods for transformation estimation (see section 5).

spatial extents at higher levels to disambiguate the local lower-level estimates. A natural consequence of such a scheme is a gradual increase in receptive field size as one ascends the hierarchical object identity and transformation networks, similar to the increase in receptive field size found in successively higher areas of the ventral/dorsal visual pathways [24]. Assigning such a correspondence between the model and neuroanatomical structure leads to the possibility of testable predictions. For example, the dotted lines in figure 1 indicate connections conveying information between the object identity and transformation networks. These connections suggest a similar computational role for the neuroanatomical connections known to exist between the dorsal and ventral visual pathways [24].

Another attractive approach to extending the present method to larger transformations of image features is to view the learned differential operators D_i as generators of *Lie transformation groups*. In such an approach, the transformed images $\boldsymbol{I}(\boldsymbol{x})$ are assumed to be generated from a reference image $\boldsymbol{I}$ using a matrix exponential:

$$\boldsymbol{I}(\boldsymbol{x}) = \exp\left(\sum_{i=1}^{m} x_i D_i\right)\boldsymbol{I}. \tag{16}$$

Note that equation (7) is just a first-order approximation to the above equation (see also equation (10)). Thus, in principle, the operators D_i learned in this paper using a first-order approximation could be directly used in equation (16) above to handle larger transformations. We are currently evaluating this promising strategy for transformation invariance [59]. A number of other authors have previously explored the general application of Lie group theory to visual perception [21, 25, 36], computer vision [66, 67] and image processing [48].

The experiments presented herein involved two-dimensional translations of image stimuli, but other transformations such as scaling and rotation within the image plane can be accommodated if these are included in the training data [56]. Three-dimensional transformations such as rotations in depth can be handled by training the ‘what’ network on a sufficient number of views of an object, as suggested by Bülthoff, Edelman, Poggio and co-workers [18, 47, 54], and allowing the ‘where’ network to account for the intermediate poses. This is consistent with Tarr and Pinker’s multiple-views-plus-transformation (MVPT) theory of recognition [65].

Taylor series expansions have previously been used in computer vision for tasks such as optic flow computation [37] where it is assumed that image brightness varies smoothly in space and time. Motivated by the Taylor series approach to optic flow computation, Black and Jepson have independently arrived at an approach similar to that being proposed herein but within the context of principal component analysis (PCA), using a hard-wired rather than a learned set of differential operators [15]. The limitations of PCA in modelling natural image distributions are well known [27, 51]. PCA is suitable only when the data are well described by Gaussian clouds. It additionally constrains its basis vectors to be mutually orthogonal. Recent work by Field [27] and others strongly suggests that natural images form a highly non-Gaussian distribution that cannot be described satisfactorily by orthogonal basis vectors. Also, the number of basis vectors in PCA has to be less than the dimensionality of the input space which means that overcomplete representations cannot be learned (see [45, 52, 63] for arguments regarding the need for overcomplete representations). Perhaps more importantly, PCA can only capture linear pairwise statistical dependences. However, natural scenes are rife with higher-order statistical structure that cannot be accounted for by linear pairwise statistics [51]. The approach presented herein allows the flexibility of tailoring a possibly overcomplete set of non-orthogonal basis vectors to match input distributions by allowing one to choose appropriate prior distributions for the parameters.

An interesting extension of the idea of expanding a transformed image into a spatial Taylor series is to use a *spatiotemporal* Taylor series to capture object motion attributes in both space and time. In such an approach, the transformation estimates $\boldsymbol{x}$ indicate both the spatial transformations induced as well as the time duration of the transformation. Thus, motion attributes such as velocity and direction selectivity, and more general transformations involving looming, receding or rotating stimuli would be conjointly coded by the various components of the $\boldsymbol{x}$ vector. Preliminary attempts at modelling the spatiotemporal properties of cortical neurons using a zeroth-order spatiotemporal 'what' network have been promising [58]. Extending such an approach to spatiotemporal 'what' and 'where' networks remains a subject of ongoing investigations.

Acknowledgments

We would like to thank Dan Ruderman and an anonymous reviewer for helpful comments and suggestions. We are especially grateful to Dan Ruderman for pointing out the close connection between our approach and Lie group theory. This research was supported by NIH/PHS research grant 1-P41-RR09283.

References

[1] Adelson E H and Movshon J A 1982 Phenomenal coherence of moving visual patterns *Nature* **300** 523–25
[2] Atick J J 1992 Could information theory provide an ecological theory of sensory processing *Network: Comput. Neural Syst.* **3** 213–51
[3] Atick J J and Redlich A N 1992 What does the retina know about natural scenes? *Neural Comput.* **4** 196–210
[4] Atick J J and Redlich A N 1993 Convergent algorithm for sensory receptive field development *Neural Comput.* **5** 45–60
[5] Baddeley R J and Hancock P J B 1991 A statistical analysis of natural images matches psychophysically derived orientation tuning curves *Proc. R. Soc.* B **246** 219–223
[6] Barlow H B 1961 Possible principles underlying the transformation of sensory messages *Sensory Communication* ed W A Rosenblith (Cambridge, MA: MIT Press) pp 217–34
[7] Barlow H B 1972 Single units and cognition: A neurone doctrine for perceptual psychology *Perception* **1** 371–94
[8] Barlow H B 1989 Unsupervised learning *Neural Comput.* **1** 295–311
[9] Barlow H B 1994 What is the computational goal of the neocortex? *Large-Scale Neuronal Theories of the Brain* ed C Koch and J L Davis (Cambridge, MA: MIT Press) pp 1–22
[10] Barrow H G 1987 Learning receptive fields *Proc. IEEE 1st Int. Conf. on Neural Networks (San Diego, CA, June 1987)* ed M Caudill and C Butler (Piscataway, NJ: IEEE) vol 4, pp 115–21
[11] Barrow H G and Bray A J 1992 A model of adaptive development of complex cortical cells *Artificial Neural Networks 2* ed I Aleksander and J Taylor (Amsterdam: Elsevier Science) pp 881–4
[12] Baum L E, Petrie T, Soules G and Weiss N 1970 A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains *Ann. Math. Stat.* **41** 164–71
[13] Bell A J and Sejnowski T J 1997 The 'independent components' of natural scenes are edge filters *Vision Res.* **37** 3327–38
[14] Bienenstock E L, Cooper L N and Munro P W 1982 Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex *J. Neurosci.* **2** 32–48
[15] Black M J and Jepson A D 1996 Eigentracking: Robust matching and tracking of articulated objects using a view-based representation *Proc. 4th Eur. Conf. on Computer Vision (ECCV '96, Cambridge, UK, April 1996)* ed B Buxton and R Cipolla (Berlin: Springer) vol 1, pp 329–42
[16] Brooks V B 1986 *The Neural Basis of Motor Control* (Oxford: Oxford University Press)
[17] Bryson A E and Ho Y C 1975 *Applied Optimal Control* (New York: Wiley)
[18] Bülthoff H H, Edelman S Y and Tarr M J 1995 How are three-dimensional objects represented in the brain? *Cereb. Cortex* **5** 247–60
[19] Cowan J D 1995 Neural networks: the early days *Advances in Neural Information Processing Systems 2* ed D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 893–900

[20] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *J. R. Stat. Soc.* B **39** 1–38

[21] Dodwell P C 1983 The Lie transformation group model of visual perception *Percept. Psychophys.* **34** 1–16

[22] Duffy C J and Wurtz R H 1991 Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large field stimuli *J. Neurophysiol.* **65** 1329–45

[23] Duhamel J, Colby C L and Goldberg M E 1992 The updating of the representation of visual space in parietal cortex by intended eye movements *Science* **255** 90–2

[24] Felleman D J and Van Essen D C 1991 Distributed hierarchical processing in the primate cerebral cortex *Cereb. Cortex* **1** 1–47

[25] Ferraro M and Caelli T M 1994 Lie transformation groups, integral transforms, and invariant pattern recognition *Spatial Vis.* **8** 33–44

[26] Field D J 1987 Relations between the statistics of natural images and the response properties of cortical cells *J. Opt. Soc. Am.* A **4** 2379–94

[27] Field D J 1994 What is the goal of sensory coding? *Neural Comput.* **6** 559–601

[28] Földiák P 1990 Forming sparse representations by local anti-Hebbian learning *Biol. Cybern.* **64** 165–70

[29] Földiák P 1991 Learning invariance from transformation sequences *Neural Comput.* **3** 194–200

[30] Freeman W T and Tenenbaum J B 1997 Learning bilinear models for two-factor problems in vision *Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR, San Juan, Puerto Rico, June 1997)* (Los Alamitos, CA: IEEE Computer Society) pp 554–60

[31] Fukushima K 1980 Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position *Biol. Cybern.* **36** 193–202

[32] Gross C G, Rocha-Miranda C E and Bender D B 1972 Visual properties of neurons in inferotemporal cortex of the macaque *J. Neurophysiol.* **35** 96–111

[33] Hancock P J B, Baddeley R J and Smith L S 1992 The principal components of natural images *Network: Comput. Neural Syst.* **3** 61–70

[34] Hinton G E 1981 A parallel computation that assigns canonical object-based frames of reference *Proc. 7th Int. Joint Conf. on Artificial Intelligence (Vancouver, August 1981)* vol 2, pp 683–5

[35] Hinton G E, McClelland J L and Rumelhart D E 1986 Distributed representations *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* vol 1 (Cambridge, MA: MIT Press)

[36] Hoffman W C 1977 An informal historical description of the “LTG/NP” *Cahiers de Psychologie* **20** 139–50

[37] Horn B K P and Schunck B G 1981 Determining optical flow *Artificial Intell.* **17** 185–203

[38] Hubel D H and Wiesel T N 1968 Receptive fields and functional architecture of monkey striate cortex *J. Physiol.* **195** 215–43

[39] Jordan M I and Rumelhart D E 1992 Forward models: Supervised learning with a distal teacher *Cogn. Sci.* **16** 307–54

[40] Koenderink J J 1988 Operational significance of receptive field assemblies *Biol. Cybern.* **58** 163–71

[41] Koenderink J J and van Doorn A J 1987 Representation of local geometry in the visual system *Biol. Cybern.* **55** 367–75

[42] Kohonen T, Kaski S and Lappalainen H 1997 Self-organizing formation of various invariant-feature filters in the adaptive-subspace SOM *Neural Comput.* **9** 1321–44

[43] Konen W, Maurer T and Von der Malsberg C 1994 A fast dynamic link matching algorithm for invariant pattern recognition *Neural Networks* **7** 1019–30

[44] LeCun Y, Boser B, Denker J S, Henderson B, Howard R E, Hubbard W and Jackel L D 1989 Backpropagation applied to handwritten zip code recognition *Neural Comput.* **1** 541–51

[45] Lewicki M S and Sejnowski T J 1998 Learning nonlinear overcomplete representations for efficient coding ed M I Jordan, M J Kearns and S A Solla *Advances in Neural Information Processing Systems 10* (Cambridge, MA: MIT Press) to appear

[46] Linsker R 1988 Self-organization in a perceptual network *Computer* **21** 105–17

[47] Logothetis N K, Pauls J, Bülthoff H H and Poggio T 1994 View-dependent object recognition by monkeys *Curr. Biol.* **4** 401–14

[48] Nordberg K 1994 Signal representation and processing using operator groups *Linköping Studies in Science and Technology, Dissertations* No 366, Department of Electrical Engineering, Linköping University, Sweden

[49] Oja E 1989 Neural networks, principal components, and subspaces *Int. J. Neural Syst.* **1** 61–8

[50] Olshausen B A, Anderson C H and Van Essen D C 1995 A multiscale dynamic routing circuit for forming size- and position-invariant object representations *J. Comput. Neurosci.* **2** 45–62

[51] Olshausen B A and Field D J 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* **381** 607–9

[52] Olshausen B A and Field D J 1997 Sparse coding with an overcomplete basis set: A strategy employed

by V1? *Vision Res.* **37** 3311–25

[53] Pitts W and McCulloch W S 1947 How we know universals: the perception of auditory and visual forms *Bull. Math. Biophys.* **9** 127–47

[54] Poggio T and Edelman S 1990 A network that learns to recognize 3D objects *Nature* **343** 263–6

[55] Pouget A and Sejnowski T J 1995 Spatial representations in the parietal cortex may use basis functions *Advances in Neural Information Processing Systems 7* ed D S Touretzky and T K Leen (Cambridge, MA: MIT Press) pp 157–164

[56] Rao R P N and Ballard D H 1996 A class of stochastic models for invariant recognition, motion, and stereo *Technical Report* 96.1, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester, NY

[57] Rao R P N and Ballard D H 1997 Dynamic model of visual recognition predicts neural response properties in the visual cortex *Neural Comput.* **9** 721–63

[58] Rao R P N and Ballard D H 1997 Efficient encoding of natural time varying images produces oriented space-time receptive fields *Technical Report* 97.4, National Resource Laboratory for the study of Brain and Behavior, Department of Computer Science, University of Rochester, NY

[59] Rao R P N and Ruderman 1998 Learning Lie transformation groups for invariant visual perception, in preparation

[60] Salinas E and Abbott L F 1996 Transfer of coded information from sensory to motor networks *J. Neurosci.* **15** 6461–74

[61] Sanger T D 1989 Optimal unsupervised learning in a single-layer linear feedforward neural network *Neural Networks* **2** 459–73

[62] Simard P, LeCun Y and Denker J 1993 Efficient pattern recognition using a new transformation distance *Advances in Neural Information Processing Systems V* (San Mateo, CA: Morgan Kaufmann) pp 50–8

[63] Simoncelli E P, Freeman W T, Adelson E H and Heeger D J 1992 Shiftable multiscale transforms *IEEE Trans. Information Theory* **38** 587–607

[64] Stone J V 1996 Learning perceptually salient visual parameters using spatiotemporal smoothness constraints *Neural Comput.* **8** 1463–92

[65] Tarr M J and Pinker S 1989 Mental rotation and orientation-dependence in shape recognition *Cogn. Psychol.* **21** 233–82

[66] Tsao T R, Shyu H J, Libert J M and Chen V C 1991 A Lie group approach to a neural system for three-dimensional interpretation of visual motion *IEEE Trans. Neural Networks* **2** 149–55

[67] Van Gool L, Moons T, Pauwels E and Oosterlinck A 1995 Vision and Lie's approach to invariance *Image Vision Comput.* **13** 259–77

[68] Wallis G, Rolls E and Földiák P 1993 Learning invariant responses to the natural transformations of objects *Proc. 1993 Int. Joint Conf. on Neural Networks (Nagoya, Japan, October 1993)* (Piscataway, NJ: IEEE) vol 2, pp 1087–90

[69] Webber C J 1991 Self-organization of position- and deformation-tolerant neural representations *Network: Comput. Neural Syst.* **2** 43–61

[70] Williams R J 1985 Feature discovery through error-correction learning *Technical Report* 8501, Institute for Cognitive Science, University of California at San Diego

[71] Wiskott L 1998 Learning invariance manifolds, submitted

[72] Young R A 1985 The Gaussian derivative theory of spatial vision: Analysis of cortical cell receptive field line-weighting profiles *General Motors Research Publication* GMR-4920

[73] Zipser D and Andersen R A 1988 A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons *Nature* **331** 679–84