

# Localized Receptive Fields May Mediate Transformation-Invariant Recognition in the Visual Cortex\*

Rajesh P. N. Rao and Dana H. Ballard  
Department of Computer Science  
University of Rochester  
Rochester, NY 14627  
{rao,dana}@cs.rochester.edu

## Technical Report 97.2

National Resource Laboratory for the Study of Brain and Behavior  
Department of Computer Science, University of Rochester  
May 1997

### Abstract

Neurons in the visual cortex are known to possess 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 relatively simple neural solution to the problem of transformation-invariant visual recognition also causes localized, oriented receptive fields to be learned from natural images. These receptive fields, which code for various transformations in the image plane, 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 these networks to factor retinal stimuli into object-centered features and object-invariant transformations. The resulting neuronal architecture suggests concrete computational roles for the neuroanatomical connections known to exist between the dorsal and ventral visual pathways.

## 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 during the past several decades have provided some important clues regarding the neural mechanisms underlying this invariance to transformations. Hubel and Wiesel [9] first reported the existence of

---

\*This research was supported by NIH/PHS research grant 1-P41-RR09283.

“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 [6]. 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 [3]. Thus, the neurobiological data strongly suggest that the visual system factors retinal stimuli into object-centered features and their relative transformations.

It is also known that visual cortical neurons, especially those in primary visual cortex, possess localized, oriented receptive fields. It was first suggested by Hubel and Wiesel [9] that these neurons could be coding for edges and bars in input images at different orientations. More 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 [12]. Similar results were also obtained using an algorithm that extracts the independent components of natural images [1]. These algorithms are concerned with the primary task of encoding image features with certain constraints such as sparseness, but do not address the problem of transformation-invariance of these features. Thus, an important question is whether an alternative coding strategy, which achieves object encoding as well as transformation invariance, can also account for the localized, oriented nature of receptive fields of visual cortical neurons.

In this paper, we show that a pair of cooperating neural networks can learn to solve the problem of transformation-invariant recognition by jointly maximizing the *a posteriori* probability of generating the observed visual data. The first network estimates the identity of an object of interest while the second estimates the relative transformations due to object motion. We show that, when trained on natural images, model neurons in the transformation estimating network develop localized oriented receptive fields *tuned towards various transformations*, thus suggesting an alternate functional interpretation of cortical neurons with such receptive fields.

The model described herein extends the previously proposed Kalman filter model of the visual cortex [15] by including a first-order component that represents transformations of input features, in addition to the zeroth order component that represents object-centered features. The functional dichotomy between object recognition and transformation estimation utilized by this extended model parallels the well-known dichotomy between the dorsal and ventral streams in the primate visual cortex [5]. In particular, the architecture of the model suggests concrete computational roles for the neuroanatomical connections known to exist between these two visual pathways [5].

## 2 THE OPTIMIZATION FUNCTION

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

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

The coefficients  $r_j$  denote an internal representation of the image  $\mathbf{I}$  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:

$$\mathbf{I} = U\mathbf{r} \quad (2)$$

where  $U$  is the  $n \times k$  matrix whose columns consist of the basis vectors  $U_j$  and  $\mathbf{r}$  is the  $k \times 1$  vector consisting of coefficients  $r_j$ . In a neurobiological setting, the values in the  $i$ th row of  $U$  can be regarded as the strength of the synapses in the  $i$ th 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  $\mathbf{I}(\mathbf{x})$  using a Taylor series expansion around a previously encountered reference image  $\mathbf{I}$ :<sup>1</sup>

$$\mathbf{I}(\mathbf{x}) = \mathbf{I} + \frac{\partial \mathbf{I}}{\partial \mathbf{x}} \mathbf{x} + \text{higher order terms} \quad (3)$$

where  $\mathbf{x}$  is an  $m \times 1$  vector denoting the relative transformation that the image has undergone and  $J = \frac{\partial \mathbf{I}}{\partial \mathbf{x}}$  is an  $n \times m$  matrix of partial derivatives known as the *Jacobian matrix*. One way of approximating the Jacobian  $J$  is to simply use a fixed matrix  $U'$  learned from a set of training images [14]. Unfortunately, this does not acknowledge the fact that the Jacobian is a *function of the current image*. A better method is to approximate the Jacobian as a linear function of the image  $\mathbf{I}$ . Let  $J_i$  be the  $i$ th column of the Jacobian matrix  $J$ . Then, we have the relation:

$$J_i \cong D_i \mathbf{I} \quad (4)$$

where  $D_i$  is an  $n \times n$  matrix of basis vectors coding for the image transformation corresponding to the component  $x_i$  of the transformation vector  $\mathbf{x}$ . Note that to approximate the Jacobian, the  $j$ th row of  $D_i$  needs to compute an approximation of  $\frac{\partial I_j}{\partial x_i}$ . Once again, it is easy to see that the operation of equation 4 can be performed by a set of  $n$  linear neurons whose synapses encode the basis vectors forming the  $n$  rows of  $D_i$ .

Let  $D$  be the  $n \times nm$  matrix obtained by concatenating the various basis vector matrices  $D_i$  i.e.  $D = [D_1 D_2 \dots D_m]$ . Then, the various equations 4 for  $i = 1 \dots m$  can be re-written as:

$$J = \frac{\partial \mathbf{I}}{\partial \mathbf{x}} \cong D\mathcal{I} \quad (5)$$

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

$$\mathcal{I} = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (6)$$

Note that 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.

<sup>1</sup>Taylor series expansions have previously been used in computer vision for tasks such as motion processing [8] and tracking [2].

Our goal is to estimate the coefficients  $\mathbf{r}$  and the transformation vector  $\mathbf{x}$  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:

$$\mathbf{I}(\mathbf{x}) = \mathbf{I} + J\mathbf{x} + \mathbf{n} \quad (7)$$

$$= U\mathbf{r} + D\mathcal{I}\mathbf{x} + \mathbf{n} \quad (8)$$

where  $\mathbf{n}$  is assumed to be a Gaussian noise process with zero mean and a covariance of one. We can thus define the following squared-error optimization function:

$$E_1 = (\mathbf{I}(\mathbf{x}) - U\mathbf{r} - D\mathcal{I}\mathbf{x})^T (\mathbf{I}(\mathbf{x}) - U\mathbf{r} - D\mathcal{I}\mathbf{x}) \quad (9)$$

$$= (\mathbf{I}(\mathbf{x}) - U\mathbf{r} - XU\mathbf{r})^T (\mathbf{I}(\mathbf{x}) - U\mathbf{r} - XU\mathbf{r}) \quad (10)$$

where  $X = \sum_{i=1}^m x_i D_i$ . It is easy to show that minimizing  $E_1$  is equivalent to *maximizing the log likelihood* of generating the observed data  $\mathbf{I}(\mathbf{x})$  with respect to the model parameters  $U$ ,  $D$ ,  $\mathbf{r}$ , and  $\mathbf{x}$  (see, for example, [15]). We can additionally add to  $E_1$  the terms relating to prior distributions for the parameters. Here, we use zero-mean Gaussian distributions for the model priors (see [12] for other alternatives), yielding the optimization function:

$$E = E_1 + \alpha\|\mathbf{r}\|^2 + \beta\|\mathbf{x}\|^2 + \gamma\|U\|^2 + \lambda\|D\|^2 \quad (11)$$

where the operator  $\|\cdot\|^2$  denotes the sum of squares of the elements of the vector or matrix argument. The coefficients  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\lambda$  are parameters related to the variances of the prior distributions. Minimizing  $E$  is thus equivalent to maximizing the *a posteriori* probability of generating the observed data  $\mathbf{I}(\mathbf{x})$  (see, for example, [15]).

### 3 NETWORK DYNAMICS AND SYNAPTIC LEARNING RULES

For the purposes of stability, we minimize  $E$  with respect to  $\mathbf{r}$  and  $\mathbf{x}$  for fixed values of  $U$  and  $D$ . The basis vectors  $U$  and  $D$  are learned on a slower time scale for fixed values of  $\mathbf{r}$  and  $\mathbf{x}$ . This form of alternating between minimization of parameters can be viewed as implementing a variant of the Expectation-Maximization (EM) algorithm from statistics.

For a given set of basis vectors  $D$  and  $U$ , we can minimize  $E$  with respect to  $\mathbf{r}$  and  $\mathbf{x}$  using gradient descent:

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

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

Thus, given a transformed image  $\mathbf{I}(\mathbf{x})$ , one needs to compute the *residual error* between the input  $\mathbf{I}(\mathbf{x})$  and the prediction  $U\mathbf{r} + D\mathcal{I}\mathbf{x}$ . In the case of the object identity estimate  $\mathbf{r}$ , the residual is filtered using the “feedforward” matrix  $(U + XU)^T$  where as in the case of the transformation

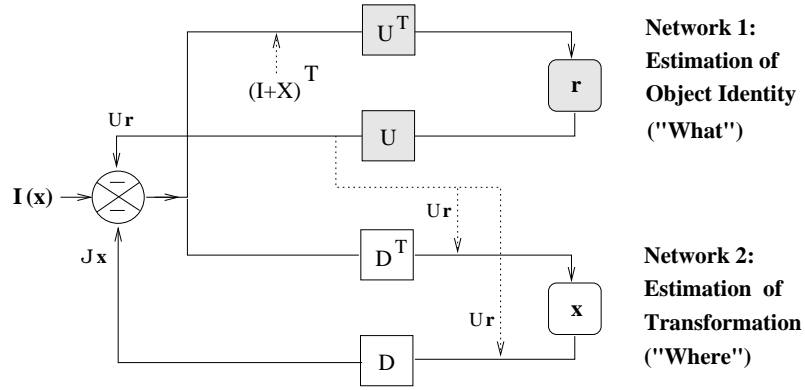


Figure 1: **Network Architecture of the Model.** The shaded portion represents the object identity (“What”) network [15] that estimates the zeroth order component of the input. The bottom unshaded portion of the figure shows the transformation estimating (“Where”) network that computes the first order transformations in the input. The “-” signs within the circle denote inhibitory feedback for computing of the feedforward residual  $(\mathbf{I}(\mathbf{x}) - U\mathbf{r} - J\mathbf{x})$ , where  $J = D\mathcal{I}$ . The dotted lines indicate connections conveying information between the two otherwise parallel networks. These connections suggest a similar computational role for the neuroanatomical connections known to exist between the dorsal and ventral visual pathways [5].

estimate  $\mathbf{x}$ , the residual is filtered via the matrix  $(D\mathcal{I})^T$ . Note that both the object network and the transformation network use the same residual signal to correct their estimates  $\mathbf{r}$  and  $\mathbf{x}$ , and both contribute to it. The residual itself can be readily computed using, for instance, inhibitory feedback of the input. 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”). An especially favorable property of such an arrangement is that the estimate of object 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  $\mathbf{x}$ . Such a property has also been the goal of some previously proposed models such as [7, 11, 13].

For specific object and transformation vectors  $\mathbf{r}$  and  $\mathbf{x}$ , 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 these two synaptic weight matrices:

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

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

where  $I$  is the  $n \times n$  identity matrix. Note that once again, the residual error  $(\mathbf{I}(\mathbf{x}) - U\mathbf{r} - D\mathcal{I}\mathbf{x})$  plays a crucial role in correcting the weights  $U$  and  $D$ .

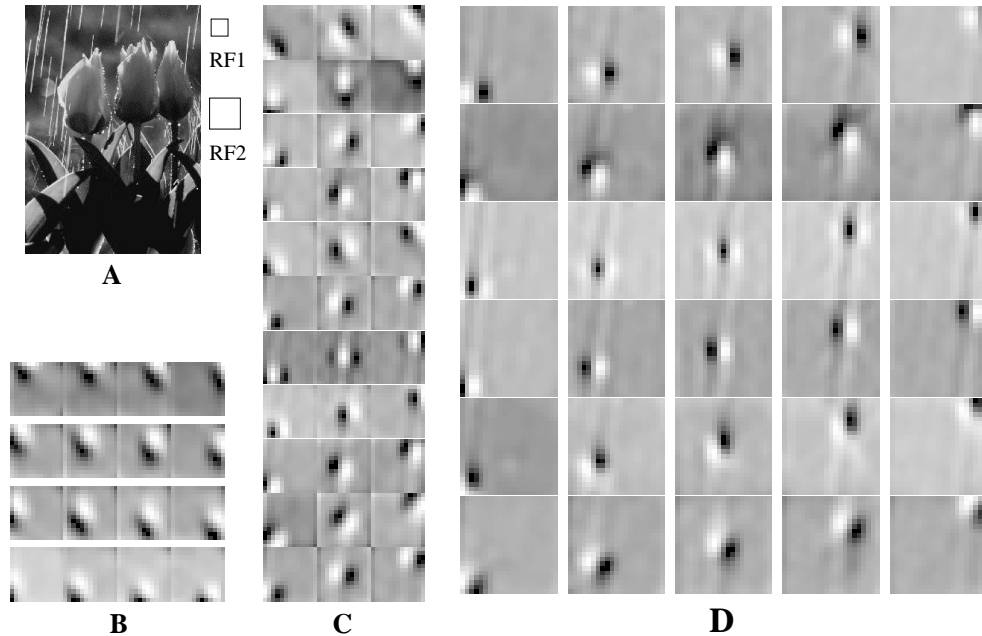


Figure 2: **Localized Receptive Fields for Translation Learned from Natural Images.** (A) shows a natural image used for training. A randomly selected image patch was translated in one of 4 directions by 2 pixels with respect to an original reference patch (= **D**). The estimate  $\mathbf{x}$  thus obtained was then fixed for the next 10 image patches translated in the same direction, and Equation 15 was used to train  $D$  ( $k_2 = 0.2$ ,  $\beta = 0.008$ ,  $\lambda = 0.0005$ .  $c_2$  was initialized to 0.4 and decreased at each iteration). (B) shows 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  $13 \times 13$  receptive fields (strictly speaking, projective fields) are all at the same orientation but at different image locations. 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. (C) shows 3 of the learned basis vectors for  $i = 2, \dots, 12$ : 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 different positions. (D) The results of learning were remarkably robust to image patch size and natural image samples, as shown here for a receptive field size of  $21 \times 21$ ,  $i = 1, \dots, 6$ . RF1 and RF2 depict the relative size of the receptive fields in (B & C) and D as compared to the natural image.

## 4 EXPERIMENTAL RESULTS AND CONCLUSIONS

Figure 2 shows the localized, oriented basis vectors for translation learned from natural image patches, at two different scales. In both cases, for each  $i$ , the basis vector forming the  $j$ th row of  $D_i$  converged to an approximation of  $\frac{\partial I_i}{\partial x_i}$  as required by the model. After convergence, these “natural” basis vectors were tested in a transformation-invariant recognition task (Figure 3). This figure illustrates what is perhaps the most important property of the model: the transformation estimates  $\mathbf{x}$  remain approximately the same for different objects transformed in an identical manner. This independence and decoupling of the transformation estimates  $\mathbf{x}$  from object estimates  $\mathbf{r}$  is crucial for learning general sensory-motor routines (for example, grasping a cup) that can be uniformly applied across objects without regard to object specific features (for example, texture or color of the cup) that are irrelevant to motor programming. Furthermore, when a transformation estimate  $\mathbf{x}$  is

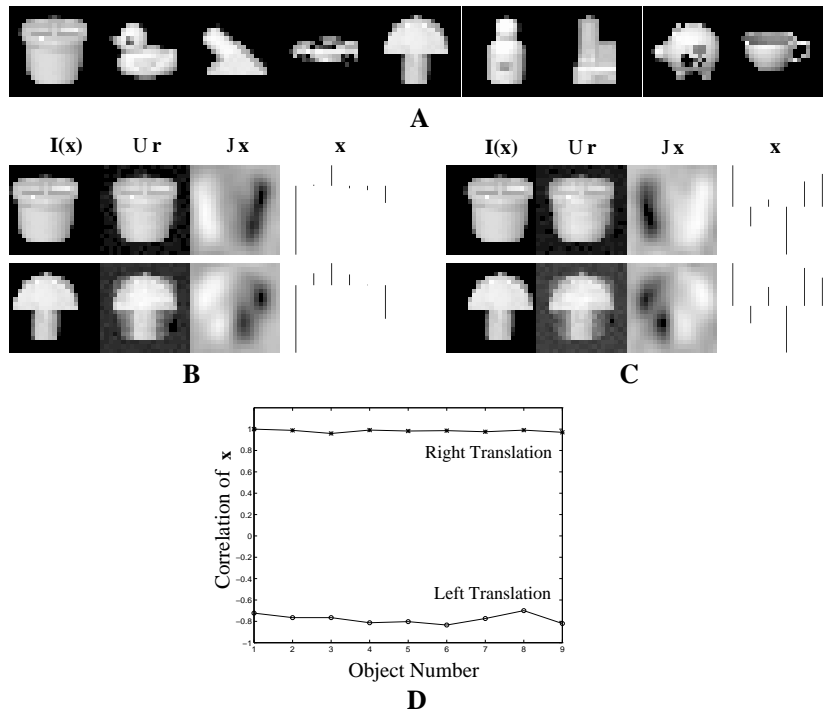
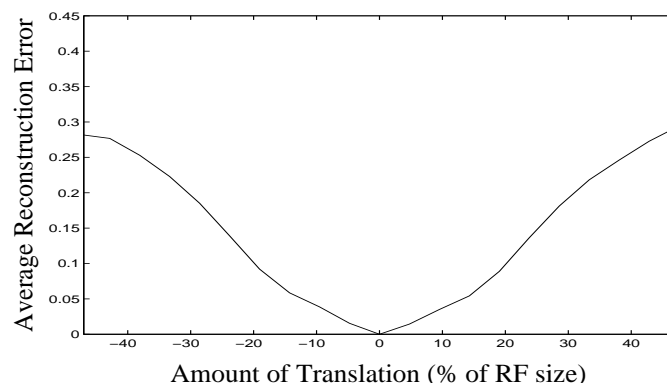


Figure 3: **Translation-Invariant Recognition.** (A) shows the images (of size  $21 \times 21$ ) used for training the object identity network, with  $\mathbf{x} = \mathbf{0}$  ( $k_1 = 0.2$ ,  $\alpha = 0.008$ ,  $\gamma = 0.0005$ .  $c_1$  was initialized to 0.4 and decreased at each iteration). (B) & (C) show the response of the networks (Figure 1) to two different objects translated leftward and rightward respectively. The natural basis vector matrix  $D$  from Figure 2 (D) was used. Note that in each case, the translated image is factored into the original image ( $U\mathbf{r}$ ) and a shift ( $J\mathbf{x}$ ). The transformation vector  $\mathbf{x}$  (upward bar = positive value, downward bar = negative value) is approximately the same for the given translation even though different objects were translated. (D) The plot at the top is the correlation between the right translation vector  $\mathbf{x}_1$  for object 1 and the right translation vectors for all other objects in the training set. The plot at the bottom is the correlation between  $\mathbf{x}_1$  and *left* translation vectors for all training objects. The high (and relatively constant) positive correlations among right translation vectors and the high (and relatively constant) negative correlations with the left translation vectors supports the hypothesis that transformation estimates in the model are object-invariant.

used to drive a motor routine such as a saccadic eye movement, the resulting “efference copy” of the motor signal can be used to update the transformation estimate  $\mathbf{x}$  [10]. This updating of internal spatial representations by intended movements has been observed in the parietal cortex [4] and has inspired numerous models based on the notion of “gain fields” [16]. The work presented here suggests a possible neural mechanism for converting the raw retinal information to spatial location estimates, which can be modulated by eye movements and other motor activities.

The results presented here involved small translations of retinal stimuli. Other transformations such as scaling, rotation, swing, and tilt can also be handled if these are included in the training data [14]. Larger transformations can be handled by the model up to a certain degree of accuracy (Figure 4) but the error  $E$  in image reconstruction gradually increases due to the insufficiencies of a first-order Taylor series approximation. Fortunately, this problem can be addressed using a *hierarchical estimation* scheme (such as in [15]), wherein higher levels operate over larger spatiotemporal



**Figure 4: The Effect of Large Transformations on Image Reconstruction Error.** The basis vectors from Figure 2 (D), 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  $\mathbf{I}$  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  $\pm 30$ - $35\%$  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 methods for transformation estimation (see text).

scales than lower ones (see also [2]). In such a scheme, top-down signals from a higher level module are fed back and integrated with bottom-up signals to produce reliable estimates of object identity and object transformation at each hierarchical level. A given transformation is represented in a hierarchical and distributed fashion within the various levels. The higher levels maintain more global, more abstract, and coarser estimates than the lower ones. The hierarchical structure is especially desirable since it counters the well-known *aperture* problem in motion estimation by allowing information from larger spatial extents at higher levels to disambiguate 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/transformation networks, in many ways similar to the increase in receptive field size found in successively higher areas in the ventral/dorsal visual pathways. Our current efforts include investigating such networks and elucidating their strengths and weaknesses.

## References

- [1] A.J. Bell and T.J. Sejnowski. The ‘independent components’ of natural scenes are edge filters. Submitted to *Vision Research*, 1996.
- [2] M.J. Black and A.D. Jepson. Eigenttracking: Robust matching and tracking of articulated objects using a view-based representation. In *Proc. of ECCV*, pages 329–342, 1996.
- [3] C.J. Duffy and R.H. Wurtz. Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large field stimuli. *Journal of Neurophysiology*, 65:1329–1345, 1991.

- [4] J. Duhamel, C.L. Colby, and M.E. Goldberg. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, 255:90–92, 1992.
- [5] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.
- [6] C.G. Gross, C.E. Rocha-Miranda, and D.B. Bender. Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Neurophysiology*, 35:96–111, 1972.
- [7] G.E. Hinton. A parallel computation that assigns canonical object-based frames of reference. In *7th International Joint Conference on Artificial Intelligence*, pages 683–685, 1981.
- [8] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
- [9] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, 195:215–243, 1968.
- [10] M.I. Jordan and D.E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.
- [11] B.A. Olshausen, C.H. Anderson, and D.C. Van Essen. A multiscale dynamic routing circuit for forming size- and position-invariant object representations. *Journal of Computational Neuroscience*, 2:45–62, 1995.
- [12] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.
- [13] W. Pitts and W.S. McCulloch. How we know universals: the perception of auditory and visual forms. *Bulletin of Mathematical Biophysics*, 9:127–147, 1947.
- [14] R.P.N. Rao and D.H. Ballard. 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, June 1996.
- [15] R.P.N. Rao and D.H. Ballard. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4):721–763, 1997.
- [16] D. Zipser and R.A. Andersen. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, 331:679–684, 1988.