

---

# A Class of Stochastic Models for Invariant Recognition, Motion, and Stereo\*

---

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

## Technical Report 96.1

National Resource Laboratory for the Study of Brain and Behavior  
June 1996

### Abstract

We describe a general framework for modeling transformations in the image plane using a stochastic generative model. Algorithms that resemble the well-known Kalman filter are derived from the MDL principle for estimating both the generative weights and the current transformation state. The generative model is assumed to be implemented in cortical feedback pathways while the feedforward pathways implement an approximate inverse model to facilitate the estimation of current state. Using the above framework, we derive models for invariant recognition, motion estimation, and stereopsis, and present preliminary simulation results demonstrating recognition of objects in the presence of translations, rotations and scale changes.

## 1 INTRODUCTION

A central problem for 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 nature of neural mechanisms underlying this invariance to transformations. Hubel and Wiesel [6] 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 or size over receptive fields of several degrees of visual angle have also been reported in higher visual areas such as IT in the occipitotemporal (ventral) pathway [4]. On the other hand, neurons in the dorsal stream appear to be coding for various types of transformations. For example, cells in the area MSTd have been shown to respond to transformations such as translations, rotations, and expansions/contractions [2]. Thus, the neurobiological data strongly suggest that the visual system factors retinal stimuli into object-centered features and their relative transformations.

We have previously introduced a dynamic Kalman filter based model of visual recognition [10]. The central idea behind this model is that the visual cortex can be regarded as a network that gains enormous efficacies by hierarchically encoding image features and dynamically predicting input stimuli. This model was however susceptible to image plane transformations of previously encoded features. In this paper, we show that the model can be extended by including a first-order component that represents transformations of input features in addition to the zeroth order component that represents object-centered features. This parallels the functional dichotomy exhibited by the dorsal (occipitoparietal) and the ventral (occipitotemporal) pathways of the primate visual cortex [3]. The extended model, described in Section 3, employs two separate but cooperating (hierarchical) networks for representing the zeroth and first order terms of an object model. Together, these networks achieve recognition in the presence of object transformations. One network estimates object identity while the other estimates

---

\*This research was supported by NIH/PHS research grants 1-P41-RR09283 and 1-R24-RR06853-02, and by NSF research grants IRI-9406481 and IRI-8903582.

the relative transformation induced by a new image. The model exploits the existence of reciprocal connections between adjoining cortical areas [3] by instantiating approximate inverse and generative models in the feedforward and feedback connections respectively for dynamic estimation of current state.

The model also suggests relatively straightforward neural mechanisms for motion detection and stereopsis, which we briefly explore in Section 4. We describe preliminary experimental results demonstrating the model’s recognition performance in the presence of translations, rotations and scale changes. In addition, the receptive fields developed by the model when exposed to translating natural image patches qualitatively resemble the receptive fields of simple cells in V1. This suggests that some of these cells, especially those providing input to areas specialized for motion such as MT, may be coding for image transformations such as translations rather than exclusively coding for raw image features such as edges or bars. Thus, the observed invariant response of a complex cell may be a straightforward consequence of the ability of these transformation encoding cells to account for the induced image variations.

## 2 THE APPROACH

The key idea behind the various models discussed in this paper is that for most small transformations, one can approximate a transformed image patch  $\mathbf{I}(\mathbf{x})$  (viewed as an  $n \times 1$  vector) by applying a Taylor series approximation around a reference image patch  $\mathbf{I}(\mathbf{x}_0)$ :

$$\mathbf{I}(\mathbf{x}) \cong \mathbf{I}(\mathbf{x}_0) + \frac{\partial \mathbf{I}(\mathbf{x}_0)}{\partial \mathbf{x}} (\mathbf{x} - \mathbf{x}_0) \quad (1)$$

The second term in the above sum is simply the product of the  $n \times k$  image Jacobian with the  $k \times 1$  vector  $(\mathbf{x} - \mathbf{x}_0)$ , which describes the relative transformation that the image has undergone. We assume, for simplicity, that  $\mathbf{x}_0 = \mathbf{0}$  and use  $\Delta \mathbf{I}$  to denote the difference  $\mathbf{I}(\mathbf{x}) - \mathbf{I}(\mathbf{0})$ . We can then model the above equation using the following stochastic model (cf. [10]):

$$\Delta \mathbf{I}(t) = U(t)\mathbf{x}(t) + \mathbf{n}_d(t) \quad (2)$$

where  $U$  is a set of generative weights approximating the Jacobian and  $\mathbf{n}_d$  is a zero-mean Gaussian noise process with covariance  $\Sigma_d(t)$ . For deriving the learning rules, it is convenient to view the  $n \times k$  matrix  $U$  as an  $nk \times 1$  vector  $\mathbf{u} = [U_1 U_2 \dots U_n]^T$  where  $U_i$  denotes the  $i$ th row of  $U$ . The weights  $\mathbf{u}$  are modeled by the dynamic system  $\mathbf{u}(t+1) = \mathbf{u}(t) + \mathbf{n}_u(t)$  where  $\mathbf{n}_u$  has mean  $\bar{\mathbf{u}}_u(t)$  and covariance  $\Sigma_u(t)$ . The behavior of the transformation state  $\mathbf{x}$  is modeled as the dynamic system  $\mathbf{x}(t+1) = f(V(t)\mathbf{x}(t)) + \mathbf{n}(t)$  where  $f$  is a (possibly nonlinear) vector-valued activation function,  $V$  is a set of “prediction” weights representing the *state transition matrix*, and  $\mathbf{n}(t)$  is a Gaussian noise process with mean  $\bar{\mathbf{n}}(t)$  and covariance  $\Sigma(t)$ . Given this stochastic model, the goal of a cortical module processing a local image patch is to estimate, as best as possible, the generative weights  $\mathbf{u}$  and the current transformation state  $\mathbf{x}$ .<sup>1</sup>

We use an optimization criterion based on the Minimum Description Length (MDL) principle [13] for optimally estimating the model parameters  $\mathbf{u}$  and  $\mathbf{x}$ . Assume that we have already computed estimates  $\bar{\mathbf{x}}$  and  $\bar{\mathbf{u}}$  based on prior data with covariances  $E[(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T] = M$  and  $E[(\mathbf{u} - \bar{\mathbf{u}})(\mathbf{u} - \bar{\mathbf{u}})^T] = S$ . Given data  $\mathcal{D}$  and model parameters  $\mathcal{M}$ , the MDL principle advocates minimizing the cost function:

$$J(\mathcal{M}, \mathcal{D}) = (\Delta \mathbf{I} - U\mathbf{x})^T \Sigma_d^{-1} (\Delta \mathbf{I} - U\mathbf{x}) + (\mathbf{x} - \bar{\mathbf{x}})^T M^{-1} (\mathbf{x} - \bar{\mathbf{x}}) + (\mathbf{u} - \bar{\mathbf{u}})^T S^{-1} (\mathbf{u} - \bar{\mathbf{u}}) + \alpha \mathbf{u}^T \mathbf{u} + \beta \mathbf{x}^T \mathbf{x} \quad (3)$$

The above equation, which can be regarded as a form of weighted least squares, results from applying Shannon’s *optimal coding theorem* and using multivariate Gaussians for coding the various error terms (cf. [13]). In particular, the first term represents the data modeling error, the second and third represent errors in the estimation of the state and the weights, and the last two terms are MDL model terms that act as regularizers and penalize data overfitting.

<sup>1</sup>For the sake of brevity, we omit the estimation of the prediction weights  $V$  in this exposition but interested readers are referred to [10].

$J$  can be minimized in two ways: (a) by adapting the estimate for weights  $\mathbf{u}$ , and (b) by adapting the estimate for state  $\mathbf{x}$ . In order to ensure stability, we adapt the weights  $\mathbf{u}$  at a much slower rate than the process estimating current state  $\mathbf{x}$ . We first derive the learning rule for modifying  $\mathbf{u}$ . Notice that  $(\Delta\mathbf{I} - U\mathbf{x}) = (\Delta\mathbf{I} - X\bar{\mathbf{u}})$  where  $X$  is the  $n \times nk$  matrix given by:

$$X = \begin{bmatrix} \mathbf{x}^T & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}^T & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{x}^T \end{bmatrix} \quad (4)$$

Differentiating  $J$  with respect to  $\mathbf{u}$  and setting the result to 0, we obtain, after some algebraic manipulation, the following update rule for the mean of the optimal stochastic weight vector at time  $t$  (using  $\mathbf{x} = \bar{\mathbf{x}}$ ):

$$\hat{\mathbf{u}} = \bar{\mathbf{u}} + P_u \bar{X}^T \Sigma_d^{-1} (\Delta\mathbf{I} - \bar{X}\bar{\mathbf{u}}) - \alpha P_u \bar{\mathbf{u}} \quad (5)$$

where  $\bar{\mathbf{u}}(t) = \hat{\mathbf{u}}(t-1) + \bar{\mathbf{u}}_u(t-1)$  and  $\bar{X}$  is the matrix obtained by replacing  $\mathbf{x}$  with  $\bar{\mathbf{x}}$  in the definition of  $X$  above. The covariance matrix  $P_u$  is updated according to:

$$P_u(t) = (S^{-1}(t) + \bar{X}(t)^T \Sigma_d^{-1}(t) \bar{X}(t) + \alpha I)^{-1} \quad (6)$$

where  $S(t) = P_u(t-1) + \Sigma_u(t-1)$ .

In a similar manner, we can derive the update rules for the mean  $\hat{\mathbf{x}}$  and covariance  $P$  of the optimal transformation state  $\mathbf{x}$  for a given set of synaptic weights  $U$  at time  $t$ :

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + P U^T \Sigma_d^{-1} (\Delta\mathbf{I} - U\bar{\mathbf{x}}) - \beta P \bar{\mathbf{x}} \quad (7)$$

$$P = (M^{-1} + U^T \Sigma_d^{-1} U + \beta I)^{-1} \quad (8)$$

where  $\bar{\mathbf{x}}(t) = f(V(t-1)\hat{\mathbf{x}}(t-1)) + \bar{\mathbf{u}}(t-1)$  and  $M(t) = \frac{\partial f}{\partial \mathbf{x}} P(t-1) \left( \frac{\partial f}{\partial \mathbf{x}} \right)^T + \Sigma(t-1)$  with the partial derivative being evaluated at  $\mathbf{x} = \hat{\mathbf{x}}(t-1)$ . The partial derivatives arise from a first-order Taylor series approximation to the activation function  $f$ . The above set of equations can be regarded as implementing a form of the *extended Kalman filter* [7].

One last issue that needs to be addressed is how the matrix  $U^T$  required by Equation 7 is implemented in the model. If we assume that the feedback weights implement the generative model as given by  $U$ , a relatively easy solution is to use a feedforward matrix  $W$  to approximate  $U^T$ . We show in [10] that a learning rule for  $W$  similar to Equation 5 leads to  $\bar{W} = \bar{U}^T$  asymptotically, even when initialized to arbitrary random values. The bottom unshaded portion of Figure 1 summarizes the basic architecture of a transformation estimating module.

### 3 INVARIANT RECOGNITION

The previous section sketched a method for estimating the transformation state  $\mathbf{x}$  with reference to a prior image  $\mathbf{I}(0)$ . We have previously shown [10] that the prior image  $\mathbf{I}_0 = \mathbf{I}(0)$  can itself be encoded in a separate network with its own set of feedforward and feedback weights.<sup>2</sup> This involves a stochastic imaging model of the form:

$$\mathbf{I}_0(t) = U_0(t)\mathbf{r}(t) + \mathbf{n}_{bu}(t) \quad (9)$$

where  $U_0$  is a set of generative weights and  $\mathbf{r}$  is an internal state vector coding object identity, and  $\mathbf{n}_{bu}$  is a zero-mean Gaussian noise process for the ‘‘bottom-up’’ input with covariance  $\Sigma_{bu}$ . This generative model leads to estimation algorithms for  $U_0$  and  $\mathbf{r}$  similar to those derived for  $U$  and  $\mathbf{x}$  in the previous section:

$$\hat{\mathbf{u}}_0 = \bar{\mathbf{u}}_0 + P_{u_0} \bar{R}^T \Sigma_{bu}^{-1} (\mathbf{I}_0 - \bar{R}\bar{\mathbf{u}}_0) - \alpha P_{u_0} \bar{\mathbf{u}}_0 \quad (10)$$

$$\hat{\mathbf{r}} = \bar{\mathbf{r}} + P_r \bar{U}_0^T \Sigma_{bu}^{-1} (\mathbf{I}_0 - \bar{U}_0 \bar{\mathbf{r}}) - \beta P_r \bar{\mathbf{r}} \quad (11)$$

<sup>2</sup>A related encoding technique for this purpose is the Helmholtz machine [1].

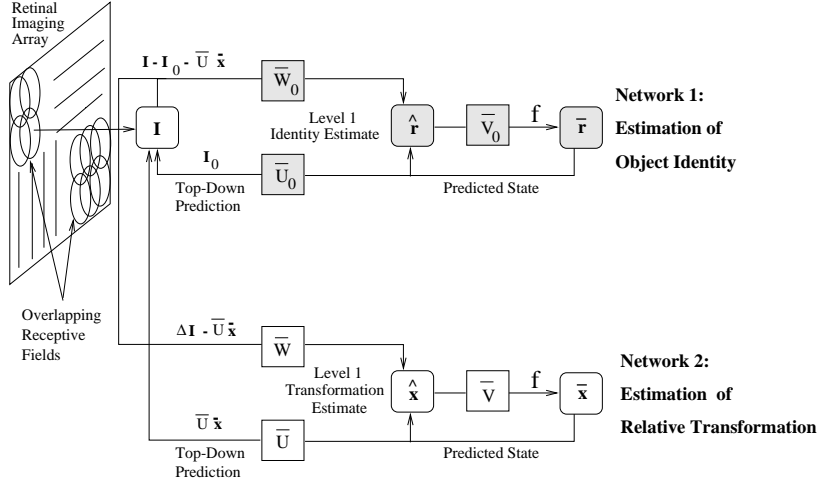


Figure 1: **Architecture of the model.** The shaded portion is part of the hierarchical architecture introduced in [10] for estimating the zeroth order component of the input. The bottom unshaded portion of the figure shows the extension to the architecture that captures the first order transformations in the input.

In the above,  $\bar{\mathbf{u}}_0(t) = \hat{\mathbf{u}}_0(t-1) + \bar{\mathbf{u}}_{u0}(t-1)$  and  $\bar{\mathbf{r}}(t) = f(V_0(t-1)\hat{\mathbf{r}}(t-1)) + \bar{\mathbf{r}}(t-1)$ . The state covariance matrices  $P_{u0}$  and  $P_r$  are updated in a manner similar to that for  $P_u$  and  $P$  respectively. Once again, the feedforward pathway is assumed to implement the matrix  $U_0^T$  necessary for estimating  $\mathbf{r}$  i.e.  $\bar{W}_0$  is approximately  $\bar{U}_0^T$ . The matrix  $\bar{R}$  is defined in a manner similar to  $\bar{X}$  in the previous section.

In summary, the model for invariant recognition consists of two cooperating networks (Figure 1), one that estimates object identity  $\mathbf{r}$  as given by the reference image  $\mathbf{I}_0$  and another that estimates the relative transformation  $\mathbf{x}$  between the current image  $\mathbf{I}$  and the reference image  $\mathbf{I}_0$ . An especially attractive 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 [5, 8, 9].

## 4 MOTION AND STEREO

### 4.1 Motion Estimation

There are two possible ways of obtaining  $\Delta\mathbf{I}$ . First, as shown in the previous section, we may use the reference image  $\mathbf{I}_0$  predicted by the object identity network and compute the difference  $\Delta\mathbf{I}$  with respect to this predicted reference image. This allows *motion estimation* relative to current top-down prediction.<sup>3</sup> An alternate strategy for bottom-up *motion detection* is suggested by recent neurobiological studies [11] which indicate that parvo- and magnocellular inputs from the LGN may converge indirectly (via layer  $4C\alpha$  and  $4C\beta$  cells) onto individual pyramidal cells in layer 4B of area V1 of the primate visual cortex. Layer 4B cells form the major source of input from V1 to the area MT which has been implicated in motion analysis. Since the magnocellular pathway is known to conduct signals at a much faster rate than the parvocellular pathway [3], parvocellular signals  $\mathbf{I}_P$  impinging on 4B pyramidal cells will always lag behind their magnocellular counterparts  $\mathbf{I}_M$  by some constant offset. Thus, we can

<sup>3</sup>The weights  $\hat{W}$  and  $\hat{V}$  together with  $f$  can be regarded as implementing a bank of separable *spatiotemporal filters*:  $\hat{W} = \hat{U}^T$  realizes the spatial weighting functions for the inputs while the prediction weight matrix  $\hat{V}$  together with  $f$  determines the temporal weighting function based on the past history of inputs. The estimate for the state  $\mathbf{x}$  thus conveys the responses of this bank of spatiotemporal filters at a given instant in time.

once again approximate the changes in the two corresponding signals via:

$$\mathbf{I}_M \cong \mathbf{I}_P + \frac{\partial \mathbf{I}_P}{\partial \mathbf{x}} \mathbf{x} \quad (12)$$

Defining  $\Delta \mathbf{I} = \mathbf{I}_M - \mathbf{I}_P$ , we can apply the algorithms from Section 2 for estimating the Jacobian (via a set of weights  $\hat{U}$ ) and the relative motion vector  $\mathbf{x}$ .

## 4.2 Stereopsis

Rather than modeling two temporally differing channels  $\mathbf{I}_M$  and  $\mathbf{I}_P$  from the same eye as above, we can model two spatially differing binocular channels  $\mathbf{I}_L$  and  $\mathbf{I}_R$  from the left and the right eye respectively. As before, we can then approximate one of these channels (say  $\mathbf{I}_L$ ) using a Taylor series expansion around the other:

$$\mathbf{I}_L \cong \mathbf{I}_R + \frac{\partial \mathbf{I}_R}{\partial \mathbf{x}} \mathbf{x} \quad (13)$$

Defining  $\Delta \mathbf{I} = \mathbf{I}_L - \mathbf{I}_R$ , we can apply the algorithms from Section 2 for estimating the Jacobian via a set of weights  $\hat{U}$ . In this case, the state estimate for  $\mathbf{x}$  represents a distributed encoding of *stereo disparity* (see, for example, Figure 2F).

## 5 EXPERIMENTAL RESULTS AND CONCLUSIONS

Figure 2 shows some preliminary experimental results for invariant recognition using a small set of man-made objects.<sup>4</sup> The results suggest that many apparently non-linear responses of cells may in fact be explained by feedback and cooperation within the larger context of networks that the cell is directly or indirectly connected to. For example, the invariant response of a “complex” cell appears highly non-linear when viewed in isolation, but this invariance can also be explained by considering the responses of cells in a companion network that accounts for the first order terms. As the input shifts, the zeroth order response of the complex cell remains unchanged because the deviation can be modeled by a first order variation as depicted in the example in Figure 2F. This explanation is further supported by the emergence of simple cell-like receptive fields in the network when exposed to translating natural image patches (Figure 3).

The experiments described in this paper used single-level networks. However, the model lends itself naturally to a hierarchical estimation scheme (see [10] for an example). In this more general scheme, expectation based top-down signals are fed back from a higher level module operating at a larger spatial scale. These are dynamically combined with bottom-up signals to produce reliable estimates of recognition and transformation state at each hierarchical level. Top-down feedback becomes especially desirable in the presence of occlusions and noisy channels since it allows the model to rely on higher level estimates when the bottom-up input becomes noisy, and vice versa. The hierarchical structure also counters the well-known *aperture* problem in motion estimation by integrating information from larger spatial extents at higher levels and feeding back the estimates to lower levels. A natural consequence of such a scheme is a gradual increase in receptive field sizes as one ascends the hierarchical object identity/transformation networks, in many ways similar to the increase in receptive field sizes found in successively higher areas in the ventral/dorsal visual pathways [12].

An interesting possibility is to allow interactions between corresponding levels of the two otherwise parallel hierarchical networks. This is especially desirable in light of the fact that the Taylor series approximation is generally not valid for large transformations at the lowest level. By modeling the differences not just at the lowest level (as in this paper) but also at the higher levels, the Taylor series approximations can once again be used to account for transformations at successively larger scales. Indeed, there exists considerable cross-talk and anatomical connections between the dorsal and ventral streams [3]. Investigation of such mutually-interconnected hierarchical networks constitutes an active direction of future research.

---

<sup>4</sup>Since these preliminary simulations used static images, we modeled the prediction step simply as  $\bar{\mathbf{x}}(t) = \hat{\mathbf{x}}(t - 1)$  and likewise for  $\mathbf{r}$ .

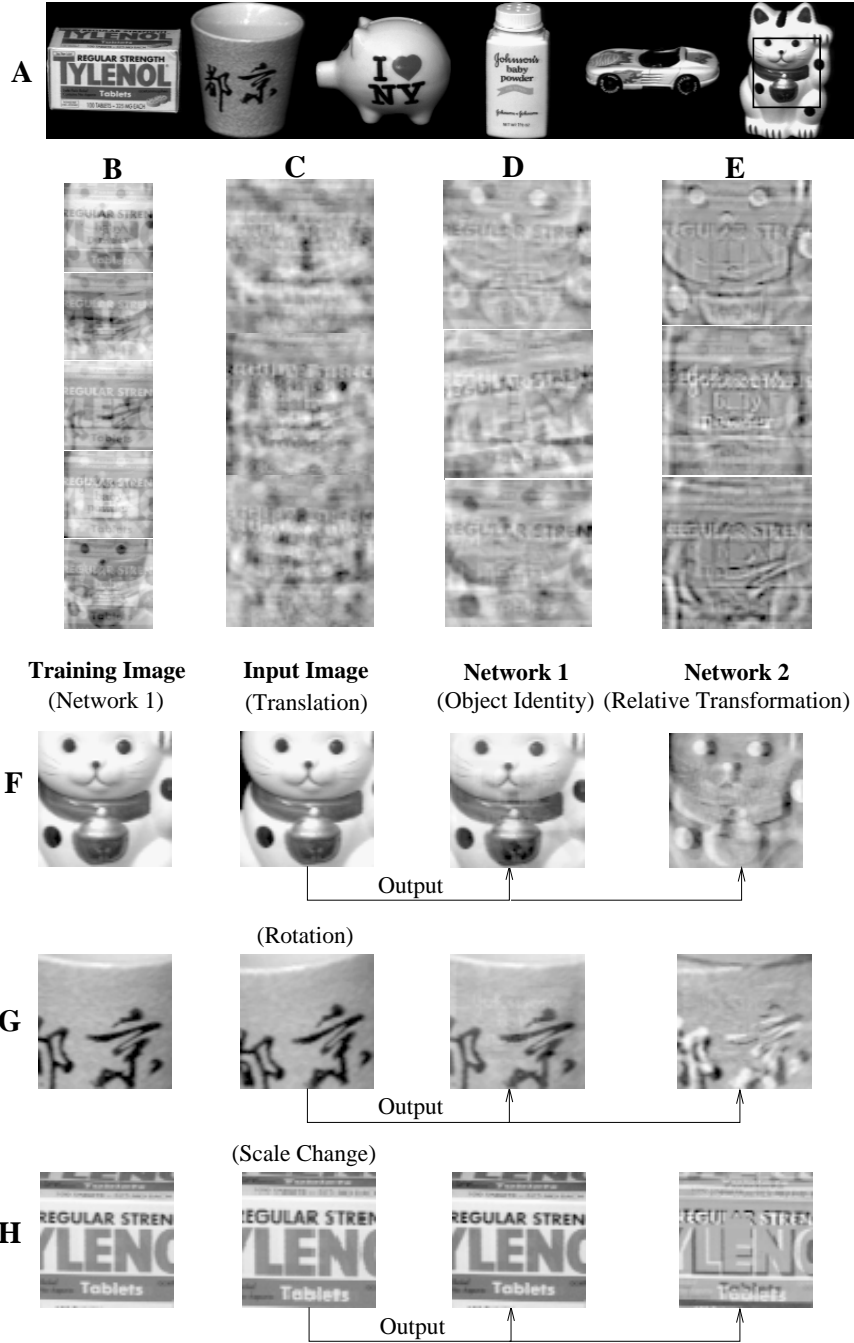


Figure 2: **Experimental Results.** (A) shows the objects used for training a pair of single-level cooperating networks. The box on the extreme right image shows the size of the receptive field. (B) shows the five feedforward synaptic weight vectors (as given by the rows of  $\widehat{W}$ ) for the zeroth order (object identity) network. (C) through (E) show three of the learned weight vectors (enlarged here for clarity) developed by the network for translation (5 pixels in 8 directions), rotation ( $10^\circ$  clockwise/counterclockwise), and expansion/contraction (10% scale change) respectively. (F) through (H) show some examples of typical recognition behavior, where an input image is factored into an object identity component in network 1 and the relative transformation in network 2. The two images at the extreme right are the reconstructed top-down predictions  $\widehat{U}_0 \bar{\mathbf{r}}$  and  $\widehat{U} \bar{\mathbf{x}}$  respectively.

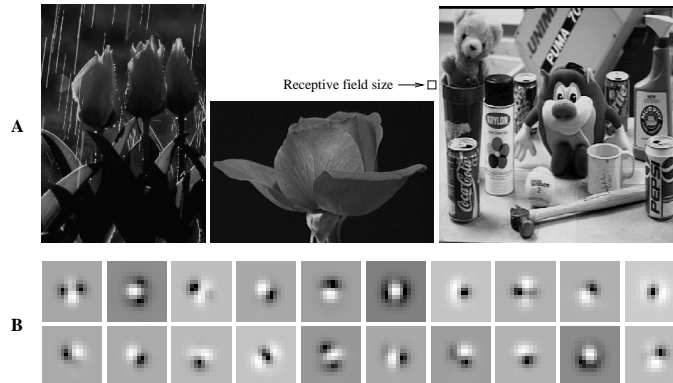


Figure 3: **Receptive fields for translation from natural images.** (A) shows the natural images used for training a pair of single-level cooperating networks. A given image patch was translated in 8 different directions by 1 pixel with respect to an original reference patch. (B) shows the receptive fields (synaptic weights as given by  $\widehat{W}$ ) developed by the first-order transformation estimating network. These resemble the oriented wavelet-like receptive field profiles of simple cells in V1 that have been previously modeled as Gabor functions or difference of Gaussians.

## References

- [1] P. Dayan, G.E. Hinton, R.M. Neal, and R.S. Zemel. The Helmholtz machine. *Neural Computation*, 7:889–904, 1995.
- [2] 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.
- [3] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.
- [4] 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.
- [5] 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.
- [6] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, 195:215–243, 1968.
- [7] P.S. Maybeck. *Stochastic Models, Estimation, and Control (Vols. I and II)*. New York: Academic Press, 1979.
- [8] 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.
- [9] 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.
- [10] R.P.N. Rao and D.H. Ballard. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation* (in press). Copy available at <ftp://ftp.cs.rochester.edu/pub/u/rao/papers/dynrec.ps.Z>, 1996.
- [11] A. Sawatari and E.M. Callaway. Convergence of magno- and parvocellular pathways in layer 4B of macaque primary visual cortex. *Nature*, 380:442–446, 1996.
- [12] D.C. Van Essen. Functional organization of primate visual cortex. In A. Peters and E.G. Jones, editors, *Cerebral Cortex*, volume 3, pages 259–329. Plenum, 1985.
- [13] R.S. Zemel. *A Minimum Description Length Framework for Unsupervised Learning*. PhD thesis, Department of Computer Science, University of Toronto, 1994.