# A hierarchical machine vision system based on a model of the primate visual system.

Paul M. Gochin (1) and Joseph M. Lubin (2)
(1) Dept. of Psychology, Princeton University, Princeton NJ
(2) Human Information Processing Group and the Dept. of Civil Engineering, Princeton University, Princeton NJ

## Abstract

In this paper, the problem of advancing machine visual pattern recognition capabilities is approached by examining the visual system of the primate. First a model of biological vision is suggested, then an analogous machine vision simulation is developed. The modeling is limited to luminance information (i.e. color, motion and depth are not considered), and biological systems are considered at the network level (biochemical and biophysical details are not simulated). The system architecture consists of a set of invariance transforms (i.e. luminance, spatial and scale) followed by storage using an Adaptive Resonance Theory network.

## Introduction

The undertaking herein described is an effort to construct a biologically constrained framework for visual pattern recognition, at the network level. Thus, functional components of biological systems, not details of biochemical and biophysical construction are considered. The rationale for biological constraint is simply the hypothesis that reverse engineering will provide for more rapid development of adaptive systems than unconstrained experimentation. In addition, the "Neural Network"(NN) modeling may generate hypothesis which can be used as a guide for biological investigations.

The biological model consists of ascribed functions and simplified mechanisms for the subsystems of the hierarchically organized primate visual system (see fig. 1). Current biological theory [Mishkin 1983] suggests that spatial location (dorsal system) and pattern content (ventral system) are processed by somewhat divergent hierarchically organized pathways. The principle pathway for pattern processing is believed to include the retina, lateral geniculate nucleus (LGN), V1 (striate cortex), V2, V4, TEO and Inferior Temporal Cortex (IT) [Desimone & Ungerleider 1989] (note, this is a simplification of known anatomy). IT cortex is believed to be the uppermost level in the hierarchy which is restricted to processing visual form information.

In the context of our model, the function of the retina and LGN is to provide local contrast information at a variety of spatial scales [Kuffler et al. 1984]. Between the LGN and V1 there is a transformation from a "pixel" array to a topographically corresponding array of edge orientation filters [Hubel & Weisel 1962], which we presume serves to reduce the influence of (relatively) high spatial frequency noise and to provide a degree of tolerance for variation in spatial location, viewpoint and orientation. Interactions between V1 and V2 may serve to compensate for a certain amount of lower spatial frequency noise (i.e. partial occlusion of objects) [Grossberg & Mingolla 1985; Heydt & Peterhans 1989]. Our principle interest is in the processing which occurs between V4 and IT, which we hypothesize subserves two functions; 1) scale invariance transformation and 2) storage of visual information.

Our machine vision system is constructed by modifying and combining selected functional components derived from existing theoretical and neurobiological models of the retina [Kuffler et al. 1984; Grossberg 1973], primary visual cortex (V1) [Hubel & Wiesel 1962], scale invariance transformation [Anderson & Van Essen 1987] and categorical storage [Carpenter & Grossberg 1988]. By analogy to the biological findings, our machine retina model makes use of a "center-surround" mechanism to *locally* enhance contrast. For the purposes of the present system, V1 is represented only by "complex cell" type filters. Spatial invariance properties of this architecture are investigated. At this time, theorized functions of V2 (i.e. "pattern completion") have not been incorporated in this system. The final stages, modeling V4 and IT, incorporate mechanisms for dealing with variation in pattern size and stable storage of the resultant representation. For these problems neuroscience has little to suggest, but

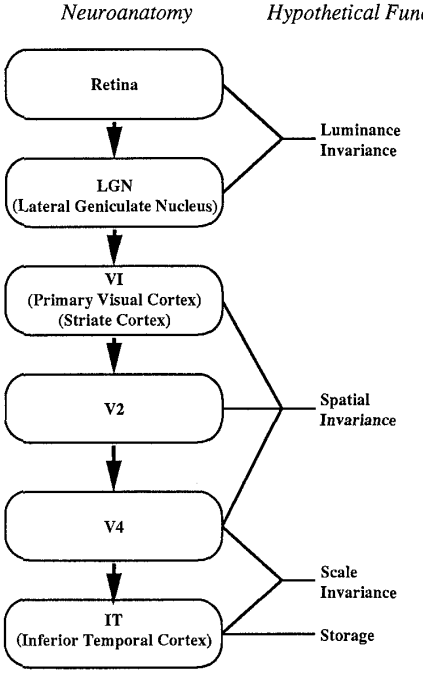


*FIGURE 1. Schematic diagram of primate visual system anatomy and corresponding hypothetical functions from our model.*

rather may gain from theoretical insight. The problem of scale invariance is approached by using a controlled multi-scale projection onto a fixed size storage space, constrained by biological considerations. Of the existing NN models of pattern storage, Adaptive Resonance Theory [Carpenter & Grossberg 1988] seems the most biologically plausible, at least for IT cortex [Gochin 1990].

### The biological model

The central hypothesis is that at the highest level of the ventral pathway of the visual system, believed to be IT cortex [Mishkin 1982], information is stored as a statistically formed prototype, but one which is constructed from information passed through a set of invariance filters. This approach provides an alternative, at least up to the level of IT cortex, to Marr's theory [Marr & Nishihara 1978] that visual information is represented by spatially invariant primitives which are in some fashion bound together. With the approach herein proposed there is no binding problem.

**Luminance invariance:** At the first stages, the retina and LGN, from the standpoint of this model, luminance information is passed through a transform which provides information invariant to absolute magnitude, while accentuating spatial discontinuities. This is achieved by a center-surround architecture where the presence of light has opposite effects on the center and surrounding inputs to retinal ganglion cells (which provide the output from the retina to the brain).

**Local spatial invariance:** In the next two stages (V1, V2) the invariance transformations are of a spatial nature. It has been shown that between the retina and V1 a transform occurs where individual neurons become sensitive to the orientation of edges. For the class of neurons know as "complex cells" there additionally is a tolerance for the location of the edge, so long as it falls within a range referred to as the receptive field (RF). From the standpoint of our model, both of these factors provide a significant degree of invariance with respect to location on the retina (translational invariance) and the point in space from which the pattern is viewed (viewpoint invariance). Since the neurons are relatively broadly tuned for orientation, a limited degree of rotational invariance is also provided.

We hypothesize that more extensive tolerance for rotation is achieved beyond IT cortex and hence this topic will not be dealt with here. A biologically plausible mechanism has been suggested which could provide more extensive invariance to translation [Anderson & Van Essen 1987], however, the eccentricity related restriction of spatial sampling in the retina (discussed further below) suggests that such a mechanism would be quite limited. Another form of spatial tolerance falls in the category of "pattern completion", hypothesized to be computed between V1 and V2. This topic has been considered in detail by others [Grossberg & Mingola 1985; Heydt & Peterhans 1989] and will not be treated here.

**Scale invariance:** The final invariance transform to be considered here is spatial scale, which we hypothesize occurs between V4 and IT. If an object fails to fall within, or is so large it cannot be contained within the visual field, the problem is resolved by moving the head or eyes. However, movement towards or away from an object to change the size of the image falling on the retina is generally impractical, and primate eyes are not equipped with "zoom" lenses. Thus we propose that moderate variations in spatial scale are resolved by computing a transformation. Recognition of objects too large for the visual field probably requires processing beyond IT. Furthermore, we hypothesize that recognition which requires high spatial frequency analysis over a large area of the visual field may also require additional processing.

The organization of the visual system suggests some constraints on the transform, in particular, the spatial sampling at the center of vision (the fovea) is much denser than at the periphery. While the graded *average* increase in receptive field size with eccentricity from the fovea is frequently discussed, the equally important fact that there is considerable range of RF sizes at each eccentricity [Wilson & Sherman 1976] is overlooked. We propose that at least one function of this range is to facilitate sampling of the visual field at a number of spatial scales without aliasing. In essence, a number of separate maps are hypothesized to exist at different spatial scales, each of which is comprised of the same number of sampling points, has uniform spatial sampling within each map, and is centered around the fovea. High spatial resolution maps would be restricted to limited eccentricity from the fovea while lower spatial frequency maps would include the foveal and in addition more eccentric regions, the extent of eccentricity being greatest for the lowest frequency maps. The required distribution of RF sizes to support this hypothesis have been observed in visual cortex from V1 to V4 [Desimone & Schein 1987]. We propose, that multiple spatial scale maps from V4 converge at some point between V4 and IT onto a single map in IT cortex. Uncontrolled, this convergence of input would produce mayhem, however we further hypothesize that only one scale map is activated at a time, due to the influence of a heuristic, attentionally controlled selection system.

Neurophysiologial data is available which supports this theory. First, receptive field sizes in V4, at any given retinal eccentricity, vary over a wide range [Desimone & Schein 1987]. Second, neurons from cortical areas V1 to V4 are arranged in maps which topographically correspond with the layout of the retina, but no such organization appears to occur for neurons in IT cortex [Gross et al. 1972]. Rather, in anesthetized animals IT receptive fields are reported to be large compared to other visual areas and to always include the fovea. In awake animals there is evidence to suggest that the size of IT receptive fields can change in correspondence with changes in the animal's attention [Richmond et al. 1983, Moran & Desimone 1985]. Third, IT neuron responses show invariance to pattern size [Schwartz et al. 1983].

**Storage of visual information:** Lesions of temporal cortex in monkeys and humans cause deficits in visual pattern recognition and memory [Gross 1973]. Neurophysiological evidence suggests that plasticity in V1 is generally restricted to a critical developmental period [Weisel 1982]. There is biochemical evidence suggestive of a graded increase in neural plasticity from V1 to IT [Mishkin & Appenzeller 1987]. There is evidence from single neuron recordings for long term changes in activity of IT neurons [Myashita 1988]. All of this evidence suggests that visual information may be stored in IT. We suggest that Adaptive Resonance Theory (ART) is a reasonable mechanism for the storage process. The rationale is based on a combination of the fact that ART is an unsupervised, stable learning mechanism and a correspondence of ART functional components with neurobiological observations [Gochin 1990].

### The machine vision implementation

**Luminance contrast enhancement module:** In order to fulfill the objective of local contrast enhancement, a center-surround architecture with shunting inhibition was implemented using a form of Grossberg's "short term memory" equation as the kernel (see Equation 1) [Grossberg 1973]. The term $x_i$ represents the excitatory activity of the central neuron, while $x_k$ are the neurons in the surround which provide inhibitory input. The term $I_i$ is the bottom-up input from photoreceptors. The constant $B$ is the upper limit on

the activity level. As suggested by Grossberg, sigmoidal functions of activation, $f(w)$, are used because of their combined properties of noise suppression (Quenching) at very low activity levels, enhanced representation of moderately low activity, quasi-linear behavior at intermediate and compression of high activity.

$$\dot{x}_i(t) = (B - x_i(t))\, f(I_i(t)) - x_i(t) \sum_{k \neq i} f(x_k(t)) \qquad (1)$$

The dynamical behavior of this class of equation is shown in figure 2. An example of the effects of this network on an image are shown in figure 3. Note the elimination of low level noise and the accentuation of areas of highest contrast (i.e. most rapid spatial change in luminance). For purposes of scale invariance transformation, maps of this type are computed over a range of spatial scales with retina-like restrictions on the extent of high spatial frequency maps. The larger receptive fields are achieved by summing over larger center and surround areas.

**Spatial invariance module:** While it is assumed that large variations in the location of a pattern must be resolved by a heuristic search mechanism, small variations, and various forms of spatial noise may be resolved by local spatial processing. For the present, we have been studying a simplified model of V1 "complex cell" behavior. The salient features are orientation selectivity and limited translational invariance. The kernel for each node of the simplest model was:

$$x_{ij} = \sum_{k=-W}^{W} a_k \left[ \left[ \sum_{m=-H}^{H} I_{i+k,j+m} - \tfrac{1}{2} \left[ \sum_{m=-H}^{H} I_{i+1+k,j+m} + \sum_{m=-H}^{H} I_{i-1+k,j+m} \right] \right] \right]^{+} \qquad (2)$$

In this model, input to a node is comprised of an excitatory center strip 1 pixel wide and 2H+1 pixels high, and two equal sized inhibitory lateral strips (scaled to have a net maximal influence equal to the single central strip). The com-

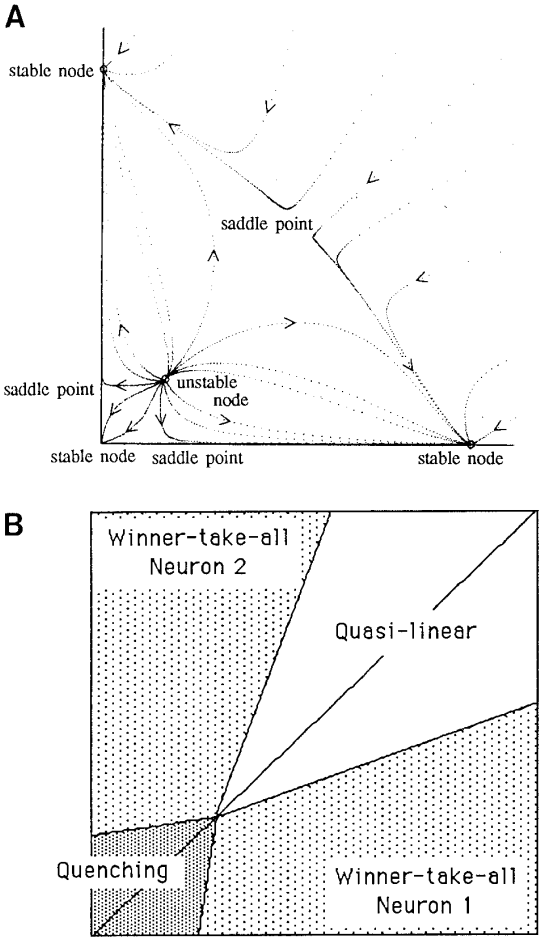


*FIGURE 2. Figure 2a depicts a phase portrait for a numerically simulated two-neuron shunting on-center off-surround short term memory network [Grossberg 1973, Lubin & Gochin 1990]. The phase plot describes the temporal evolution of the neurons' activities with each axis representing the activation of a single neuron. The geometry of the portrait is presented in terms of three common dynamical systems landmarks: unstable or source nodes, saddle points, and stable or sink nodes. The functional implications for information processing are shown in figure 2b. An early system bifurcation created a quenching region in the lower left corner. This region quenches all activity that originates within it, that is, the trajectory falls to the origin signifying that the activities of both neurons decay to zero. A winner-take-all (WTA) region is associated with each neuron wherein all trajectories evolve to the state where one neuron is strongly active while the second is fully suppressed. The more linear the intermediate region of the sigmoidal signal function, the better these networks are able to encode partial contrast. This characteristic represents a slow winner-take-all process and is depicted in the figure as a quasi-linear subdivision of the WTA regions.*

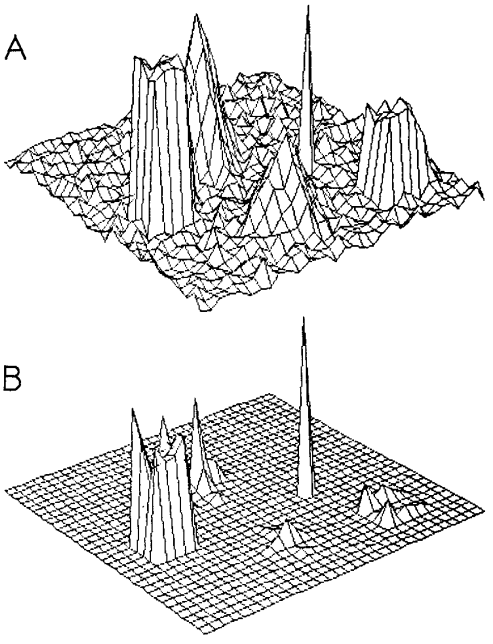


*FIGURE 3. Surface plots of a 2-D image, where surface height represents the intensity level. (A) Original 32 by 32 pixel image (B) Activation levels of a 32 x 32 array of center-surround nodes, where each node receives bottom up input from one pixel and is inhibited by the 8 nearest neighbor nodes.*

plete kernel for each node is the sum of a set of such inputs spanning i+/-W, restricted such that only activity greater than 0 for each location is summed. In the current model the ranges of i and j were 1 to 32, H = 1 and W = 1. A kernel of this type is computed for each of 4 orientations (0, 45, 90 & 135 degrees), and for each required spatial scale. The term $a_k$ allows for a falloff with eccentricity from the receptive field center. In simulations, when all $a_k = 1$ there was an unacceptable loss of spatial information, with a gaussian falloff in $a_k$, a degree of invariance was provided without as drastic a loss of spatial information [see Daugman 1985]. The importance of spatial invariance, within the context of our system is described in the section "Categorical storage" below. In future simulations this simple model will be replaced with a version utilizing shunting inhibition for competition between a number of factors (i.e. lateral inhibitory strips, sub-receptive field loci, orientation). Furthermore, rather than use input pixel values for the $I$ term they would be replaced by inputs from units of Grossberg's Boundary Contour System [Grossberg & Mingolla 1985].

**Scale invariance:** The scale invariance transformation is essentially a combination of two components, an appropriate set of samples of the visual field and a mechanism to control selection from that set. The computation of the required set, is described in the two preceding sections. We have not, as yet, completed a heuristic network implementation of the proposed attention-based control system. To date we have only explored the interaction between the simplest ART storage module and an algorithmic scale control mechanism, which will be described in the following section. A schematic diagram of the scale invariance transform is shown in figure 4.

**Categorical storage:** For reasons outlined earlier, Adaptive Resonance Theory (ART) has been selected as the mechanism for storage and retrieval of information. Briefly, ART is a modular, expandable, two-layer competitive learning network which makes use of top-down influences and some global interactions to stabilize and control the learning. In principle, each layer of an ART system is comprised of a set of nodes which have inhibitory interconnections and which can be described by an equation similar to (1) above. An important result of this competition is a normalization such that the

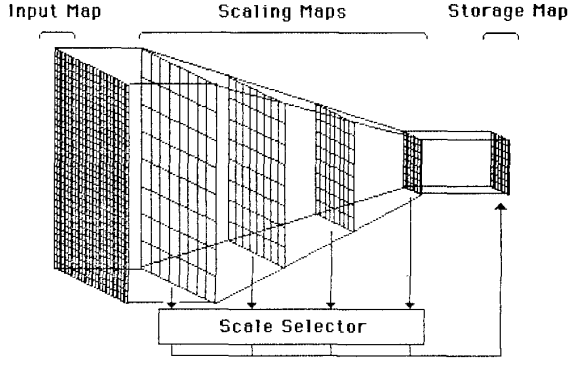


*FIGURE 4. Schematic diagram of the image pyramid-style scale invariance transform. The input consists of a 32x32 pixel pattern, which is transformed into an 8x8 pixel pattern at all scales. In the current simulation, pixels in the input pattern which map onto the same output pixel are ORed together. The 8x8 pattern is treated as a 64 bit binary input to the ART network.*

total activity of the network is constant, independent of the number of active nodes in the input pattern. The bottom layer, referred to as F1, projects to the second layer, F2, through a set of trainable weights. These weights change in accord with a system of differential equations which incorporate a Hebb law in conjunction with a decay term [Carpenter & Grossberg 1988]. As a result of the F1 input conditioned by a set of weights, and competition in F2, a set of F2 nodes are activated which best represent the bottom-up F1 input. The F2 nodes then feed back to F1 through another set of weights. When this feedback occurs, simultaneously there is a generalized change in the gain of the entire F1 layer, such that only F1 neurons which are activated by bottom-up input and by the F2 feedback are able to maintain their activity. In addition, a global comparison is made between the current F1 activity level and the original level produced by the bottom-up input. The ratio of these factors is referred to as the vigilance level (the value selected for this parameter determines how precise categorization will be). If the current ratio is below the selected vigilance level a reset occurs. This results in temporary inactivation of the currently active F2 nodes, and the global gain in F1 is reset so that the bottom-up input is once again sufficient to activate F1 nodes. The process repeats until a set of F2 nodes are accessed which are appropriate to meet the bottom-up and top-down requirements. At this time the network "resonates" in its current state, which permits time for the relatively slow, long term memory changes to occur.

The ART model we currently have in operation is the highly simplified "fast-learning" ART-1 version [Carpenter & Grossberg 1988]. In this version inputs are only binary and the F2 layer competition is modeled as winner-take-all, thus only one node can be active at a time. Making these simplifications it becomes unnecessary to construct a simulator using differential equations. Also, competition need not be implemented in F1 since a binary input is already maximally contrast enhanced. Weight changes in the projection from F1 to F2 are normalized for pattern size by the equation:

$$z_{ij} = \frac{L}{L - 1 + |X|} \tag{3}$$

where $z_{ij}$ is a weight from F1 node i to F2 node j, $L$ is a constant greater than 1 and $|X|$ is the total F1 activity (i.e. the total number of active F1 nodes). All weights from F2 to F1 are initially set to 1, then reset to 0 in those cases where the F2 node is active but the F1 node was not.

In an effort to study the behavior of scale-invariant search, where all scales are always evaluated, the search was directly coupled to the F1-F2 projection. That is, the decision as to which F2 node was the winner was not only dependent on the usual weighted F1 input, but in addition, a determination was made as to which of all available scales produced the most active F2 node. During learning the appropriate scale was fixed. Once learning was complete, with all input patterns tested, the optimal scale was always identified by the network and correct categorization always occurred.

In the absence of invariance transforms, ART networks are extremely poor at generalization, tending to form new categories or recode old ones dramatically. For example, given the letter "T" as an input pattern, a shift of the entire pattern down and to the right by one pixel will result in a spatial overlap of only one pixel. It is not possible to have an ART-1 network, which receives such bit image patterns, encode both patterns as the same item (except in the extreme case where vigilance is very low). Resolution of this particular problem lies with the model of spatial invariance described above. However, experimentation with the lumi-

nance and spatial invariance transforms awaits our completion of an ART simulator which can deal with continuous instead of binary inputs.

### Bibliography


Anderson, CH, Van Essen, DC. 1987. Proc. Natl. Acad. Sci. 84:6297-6301.

Carpenter, GA, Grossberg, S. 1988. IEEE Computer 21:77-88.

Daugman, JG. 1985. JOSA 2:1160-1169.

Desimone, R, Schein, SJ. 1987. J. Neurophysiol. 57:835-868.

Desimone, R, Ungerleider, LG. 1989. Handbook of Neuropsychology, V2 pp. 267-297, Elsevier.

Gochin, PM. 1990. Proc. of IJCNN 1:I77-I80.

Gross, CG, Rocha-Miranda, CE, Bender, DB. 1972. 35:96-111.

Gross, C.G. 1973. In: R. Jung, (Ed.) Handbook of Sensory Physiology, Vol. 7, Part 3B,Berlin: Springer Verlag, pp. 451-482.

Grossberg, S. 1973. Stud. Appl. Math. 52:217-257.

Grossberg S, Mingolla E. 1985. Perception and Psychophysics, 38:141-171.

Heydt R von der, Peterhans E. 1989. J. Neurosci. 9, 1731-1748.

Hubel, DH, Wiesel, TN. 1962. J. Physiol. 160:106-154.

Kuffler, SW, Nicholls, JG, Martin, RA. 1984. From Neuron to Brain. Sinauer.

Lubin JM, Gochin PM, 1990. Bifurcation in Shunting On-Center Off-Surround Neural Networks, Princeton University Robotics and Expert Systems Laboratory Technical Report HIP/RESL-90-02.

Marr, D, Nishihara, HK. 1978. Proc. R. Soc. Lond. 200:269-294.

Mishkin, M. 1982. Phil. Trans. R. Soc. Lond., B298:85-95.

Mishkin, M, Appenzeller, T. 1987. Sci. American.

Mishkin, M, Ungerleider, LG. 1983. Trends Neurosci. 6:414-417.

Miyashita, Y. Nature, 1988, 335:817-820.

Moran, J., and Desimone, D. Science, 1985, 229:782- 784.

Richmond, BJ, Wurtz, RH, Sato, T.1983. J. Neurophysiol. 50:1415-1432.

Schwartz, E.L., Desimone, R., Albright, T.D., and Gross, C.G. Proc. Nat. Acad. Sci., 1983, 80:5776-5778.

Wiesel, TN. 1982. Nature 299:583-591.

Wilson, JR, Murray Sherman, S. 1976. 39:512-533.



**Acknowledgements:** This work was supported by NIH grant EY-06085 (PMG) and through a grant from the James S. McDonnell Foundation to the Human Information Processing Group at Princeton University.