IMPLEMENTING STAGES OF MOTION ANALYSIS IN NEURAL NETWORKS

Margaret E. Sereno
Psychology Department
Brown University

Abstract

A neural model is proposed for human motion perception. The goal of the model is to calculate the two-dimensional velocity of elements in an image. Unlike most earlier approaches, the present model is structured in accord with known neurophysiological data. Three distinct stages are proposed. At the first level, units are sensitive to the components of motion that are perpendicular to the orientation of a moving contour. The second level integrates these initial motion measurements to obtain translational motion. The third level uses translational motion measurements to compute general three-dimensional motion such as rotation and expansion. The model shows a high level of performance in solving the measurement of two-dimensional translational motion from local motion information. Most importantly, the present model uses nervous system structure as a natural way to formulate constraints. The psychological implications of staged motion processing are discussed.

Visual motion perception serves many important functions, including the segregation of objects, the estimation of object motion, the control of eye movements, and the estimation of the three-dimensional structure of objects & the environment. The operations responsible for the perception of motion, however, are not well known.

As three-dimensional surfaces move in space, they project light onto the eye, forming a two-dimensional image of the world that changes with time. The visual system must reconstruct a three-dimensional world from this two-dimensional image. This reconstruction can be accomplished by using information about the organization of movement in the changing image. However, the motion of elements in the two-dimensional image (i.e., their speed and direction) is not an inherent property of the image but must be inferred from the varying intensities of the image. Thus, motion analysis is often considered a two-stage process (Hildreth, 1983).

The goal of the first stage is the measurement of two-dimensional motion of elements in an image (i.e., extracting the velocity--speed and direction--of moving elements). To accomplish this goal there must be initial motion detection and measurement by motion sensors, an integration of the initial motion measurements to compute an instantaneous two-dimensional velocity field (the so-called "aperture" problem), and the detection of motion discontinuities. The second stage consists of an interpretation of the three-dimensional structure of surfaces from two-dimensional motion.

I present a neural network model of part of the first stage of motion analysis (i.e., the integration of initial, local measurements to compute a two-dimensional velocity field). The model extracts the true two-dimensional motion of an entire pattern from ambiguous local motion information available at the pattern's component contours. In other words, it solves the "aperture problem" for rigid two-dimensional motion in the plane. Local motion detectors provide ambiguous information because they only measure the component of motion perpendicular to the orientation of a moving contour. A family of possible motions exists that can give rise to the locally detected motion. The aperture problem, then, reduces to the assignment of a unique velocity to an object given only local motion measurements (See Figure 1).

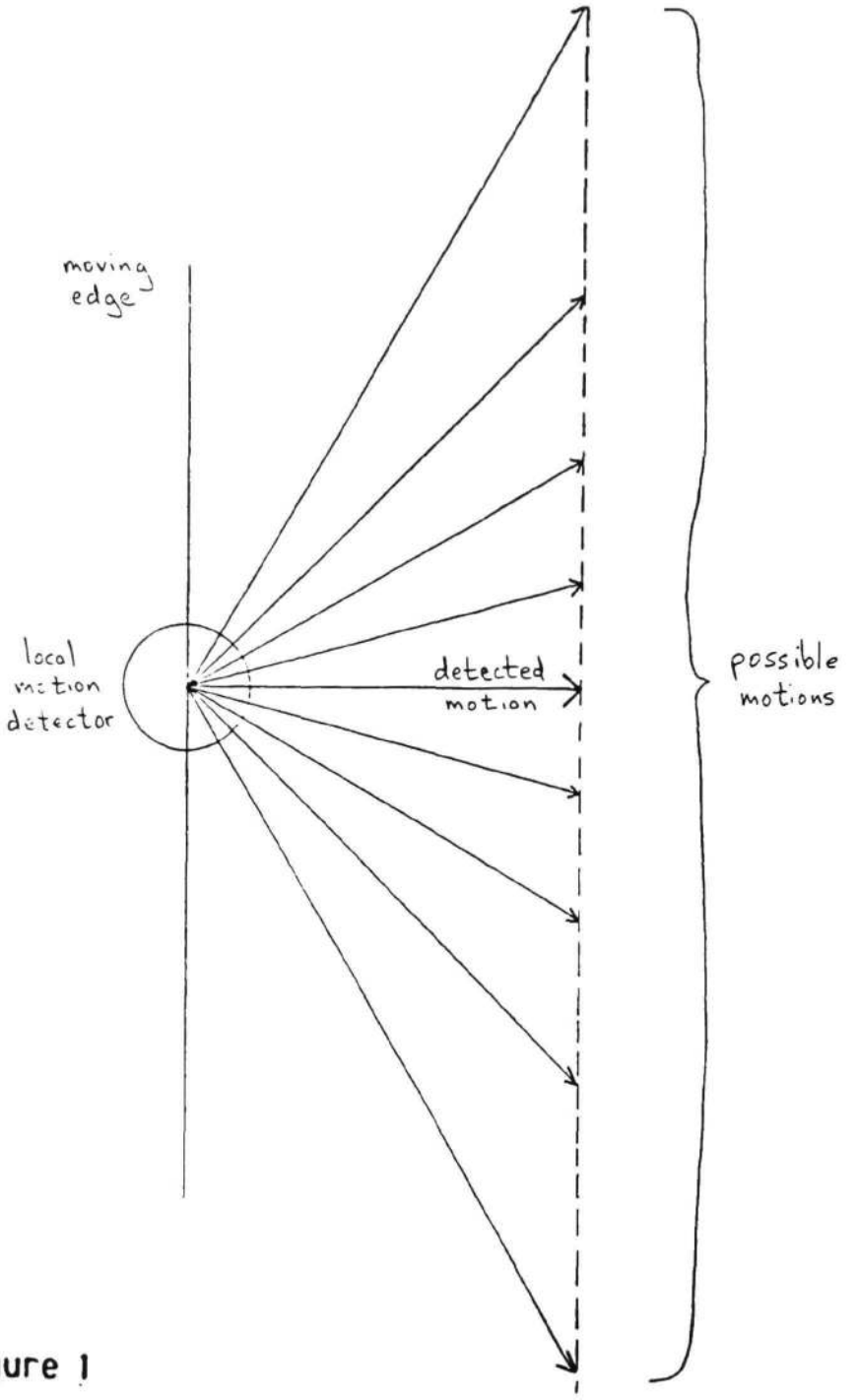


Figure 1

Hildreth (1983) has proposed a computational model for the measurement of two-dimensional motion. In her model, local measurements are obtained from the image and are then combined to compute a unique two-dimensional velocity field by applying constraints to limit the solution. For example, the "smoothness constraint" is based on the observation that objects usually have smooth surfaces. This constraint is implemented by finding the velocity field of least variation. The model works well on simple figures for planar and general three-dimensional motion (e.g., rotation and expansion).

The basic motivation for formulating the present model is to build more structure into the model to enable it to perform transformations on the input data leading to a *unique* solution. This is done by closely adhering to both neurophysiological and psychological data on motion analysis. The model is structured in accord with neurophysiological data because I assume that the nature of the hardware profoundly affects how the problem is solved. The ultimate goal is to integrate the neurophysiological and psychological information to form a more coherent theory of motion perception.

Two ideas about the basic operations involved in motion analysis emerge from the psychological, psychophysical, neurophysiological, and mathematical work on motion. One is that there are primitives of optic flow that are analyzed by specialized neural mechanisms. Work on the mathematics of optic flows demonstrates that any flow field can be decomposed into a linear vector combination of several basic types: translation, rotation, shear, and dilation (Koenderink & Van Doorn, 1976; Longuet-Higgins & Prazdny, 1980). Psychophysical data from adaptation studies have provided evidence for translation, rotation, and expansion sensitive mechanisms (Regan & Beverly, 1978; Regan, 1986). Also, neurophysiological studies in macaque visual cortex (area MST) demonstrate that neurons are sensitive to linear, rotational, and dilational motion (Saito et al., 1986).

The second idea is that the integration of local one-dimensional motion measurements into a full two-dimensional velocity field occurs in several stages. Psychophysical studies demonstrate that one-dimensional motion measurements are combined to compute two-dimensional translational motion (Adelson & Movshon, 1982; Nakayama & Silverman, 1983). Neurophysiological data suggests that the computation of all types of motion in the nervous system does not occur in a single step. A pervasive aspect of the cortical architecture of sensory systems is the presence of multiple topographic representations or maps of sensory surfaces projecting to each other. Several areas involved in motion analysis in the macaque visual cortex include Areas V1, MT, and MST. Area V1 neurons are involved in the analysis of component motion while some MT neuons respond to linear pattern motion (Movshon, Adelson, Gizzi, & Newsome, 1985). As previously noted, a recent study of cells in a visual area (MST) upstream to area MT has discovered neurons that respond selectively to translating, expanding, contracting, and rotating patterns (Saito et al., 1986).

As a first step, a model is constructed to solve the aperture problem for rigid motion in the plane (i.e., translation). This is accomplished, first, by using some formal observations on how to uniquely limit the solution and, second, by structuring the model in accord with neurophysiological organization. It is then proposed that this two-dimensional translation information is combined to compute other general motions.

Adelson and Movshon (1982) discuss a solution to unambiguously determine the two-dimensional motion of a pattern given the motion of its local components (See Figure 2). The dashed lines indicate the family of global pattern velocities which are consistent with the locally measured component velocity vector. They note that when at least two nonparallel moving contours belonging to the same pattern are compared, only one vector is common to both one-dimensional families, and it describes the motion of the entire pattern. This vector is the point in velocity space at which the two dashed lines intersect.

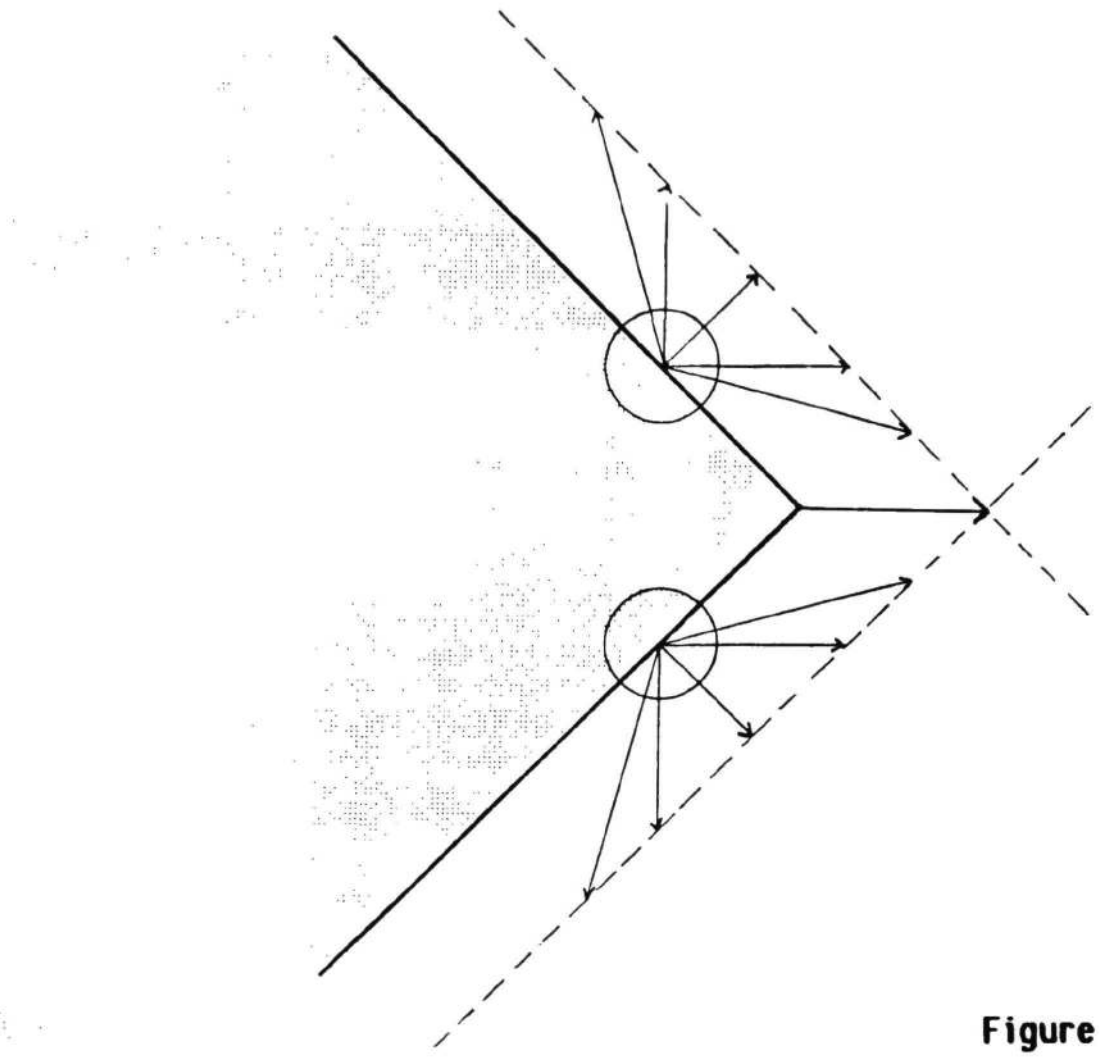


Figure 2

This constraint was implemented in a model (Sereno, 1986) that was structured in accord with the following neurophysiological facts. Some neurons in striate cortex (Area V1) are selective for orientation, speed and direction of edges. However, they only respond to the perpendicular component of motion. Area MT, an area involved in motion analysis, receives a direct topographic projection from V1, is selective for the direction and speed of motion of a stimulus while having little selectivity for spatial structure, and possesses larger receptive fields, indicating spatial summation of its inputs. Moreover, 25% of MT neurons exhibit "pattern" direction selectivity, that is, they are selective for the motion of the pattern as a whole (Movshon, Adelson, Gizzi, & Newsome, 1985).

A "Boltzmann Machine" (Ackley, Hinton, & Sejnowski, 1985) was constructed with an input layer of units representing V1 and output layer of units representing area MT. Each unit is selective for a specific speed and direction of motion (See Figure 3). Specifically, layer V1 contains 32 units (8 directions, 2 speeds and 2 locations) while layer MT contains 24 units (8 directions, 3 speeds and 1 location). V1 units respond only to the component of motion perpendicular to the orientation they are sensitive to; MT units respond to two-dimensional motion.

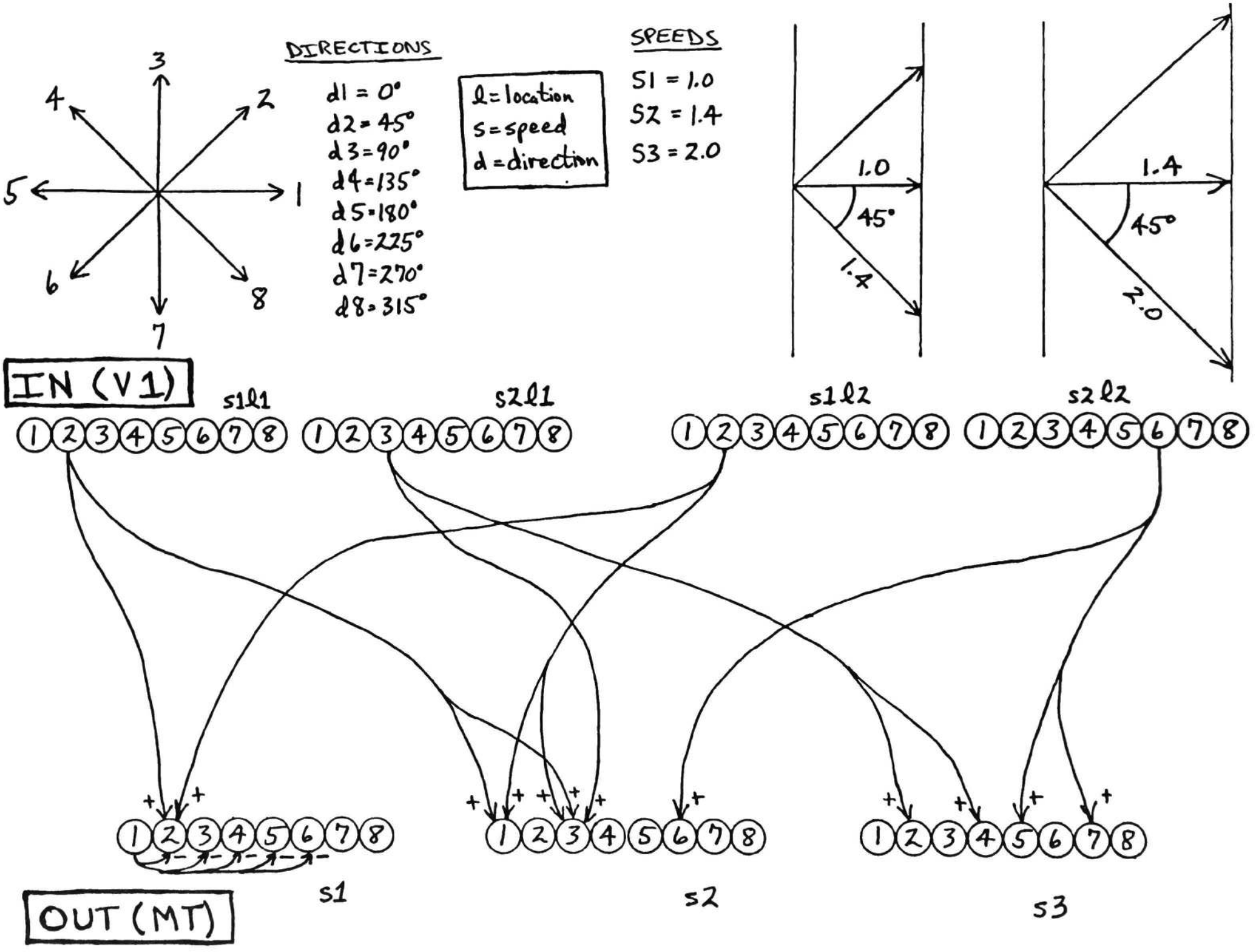

DIRECTIONS
d1 = 0°
d2 = 45°
d3 = 90°
d4 = 135°
d5 = 180°
d6 = 225°
d7 = 270°
d8 = 315°
l = location
s = speed
d = direction
SPEEDS
S1 = 1.0
S2 = 1.4
S3 = 2.0
1.0
45°
1.4
1.4
45°
2.0
IN (V1)
s1l1
s2l1
s1l2
s2l2
OUT (MT)
s1
s2
s3

Figure 3

The formal solution described above was hardwired into the system by having each V1 unit project to the family of pattern velocities in the output layer that could describe the true motion underlying its response. With this predefined connectivity, when a number of differently oriented line segments belonging to the same moving pattern are input to the system, a gradient descent algorithm results in the system changing to a configuration in which the activity of the output unit describing the pattern motion is selectively enhanced. Figure 4 presents an example of a pattern of line segments moving across the two sets of input units (See Figure 4). After 20,000 iterations, the output unit describing the pattern velocity is driven to an "on" state 100% of the time. In addition, a motion illusion (the Split Herringbone Illusion) is presented to the model. The alternating columns of lines actually move in opposite directions while the perceived motion is perpendicular to these directions, consistent with an "intersection of constraints" solution. After 20,000 iterations, the perceived direction is selectively enhanced.

These results demonstrate that the intersection of constraints described above can be realized in a two-layered neural network. The specific implementation makes a testable neural prediction about how the first layer of neurons (area V1) projects to the second layer of neurons (MT) to transform the neural response from selectivity for one-dimensional motion to selectivity for two-dimensional motion. The projection, consequently, produces MT units with a wider range and higher cut-off of preferred speeds than V1 units, a finding consistent with existing neurophysiological data (Van Essen, 1985). Another important aspect of the model is that it predicts that two-dimensional motion measurements result from the integration of one-dimensional motion measurements from nearby spatial locations.

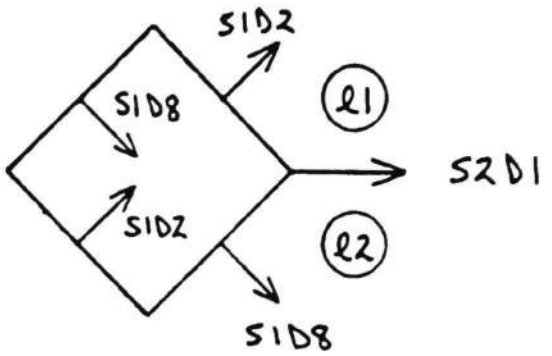
To summarize, a positive aspect of the model is that it is neurally-based with the result that it produces one solution to a given input. No post hoc assumptions or constraints are needed to limit the solution. However, a major limitation of the model is limited to the discrete values of speed and direction of movement to which the input units are sensitive. A neurally plausible solution to this problem of representing intermediate values of speed and direction is to let the information be carried by an ensemble code. This requires that individual units have continuous valued activities. For example, a speed or direction that lies exactly in between the values of 2 units can be represented by activity in each unit that is 1/2 the maximum activity. Such a representation, however, cannot be implemented on a Boltzmann Machine because the units cannot have continuous valued activity. However, it is not difficult to show that the intersection of constraints illustrated in Figure 2 amounts to a solution of a set of linear equations and hence, it can be solved with linear methods that permit continuous-valued output. Therefore, a second model was constructed using a simple linear associator with error correction, such as that used by Anderson (1983) (See Figure 5).

In the simple linear associative model, the same neurophysiological assumptions hold, except that learning can occur. This means that the connection weights are modifiable. The matrix, A, of modifiable synaptic weights describes the projection of the input layer of neurons to the output layer. The vectors f and g represent the activities across the input and output layers, respectively. Learning occurs when pairs of these vectors, one pair per pattern, are associated to form the connectivity matrix. To do this, two assumptions are made: The first assumption is that a neuron's activity results from the linear summation of its input. That is, the activity of each neuron, in the second layer, is determined by the activity of its inputs weighted by their connection strengths. Second, the matrix of connection strengths is constructed according to the generalized Hebbian rule for connectivity modification which asserts that synaptic strength is porportional to pre- and postsynaptic cell activity. This learning rule is used with error correction in which the difference between the true association,

Boltzmann Machine Model

Diamond

% "on" after 20,000 iterations	Speed and Direction
35%	S1 D2
43%	S1 D8
100%	S2 D1
38%	S2 D3
42%	S2 D7
0%	other units



Split Herringbone

% "on" after 20,000 iterations	Speed and Direction
20%	S1 D2
20%	S1 D4
20%	S2 D1
100%	S2 D3
21%	S2 D5
0%	other units

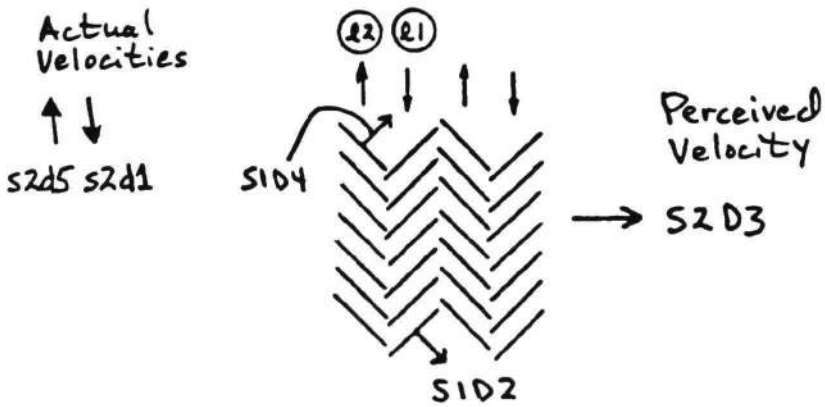


Figure 4

Two assumptions of the linear associative model:

f_i = vector of input layer neuron activities representing component velocities for the ith pattern

g_i = vector of output layer neuron activities representing pattern velocities for the i^{th} pattern

g_i' = vector of output layer neuron activities that results when a pattern, f_i, is input to the system

1) Neurons take a linear summation of their input:

$$g_i' = A \quad f_i$$

2) Learning Rule: Synaptic strength is proportional to the product of pre-synaptic and post-synaptic activities:

$$\Delta A = \sum_{i=1}^{n} g_i \quad f_i^T$$

Error Correction Procedure:

$$\Delta A = k \quad (g_i - g_i') \quad f_i^T$$

ΔA is learned and added to the developing A connectivity matrix:

$$A_{t+1} = A_t + \Delta A$$

Figure 5

g, and the actual association, g', is learned and added to the developing A connectivity matrix.

To teach the model, different patterns moving at different velocities are input to the system. For each pattern, a vector, f, describing component velocities and a vector, g, describing pattern velocities are associated using error correction.

After learning is completed, the matrix is tested. The output of each stored input is computed. That is, each f is input to the system to get an output g'. The output g' is then compared to the true association g by taking the cosine between them. If the vectors are the same, the cosine will equal 1. The system is then tested with nonassociated vector pairs to see how well the system generalizes to new stimuli.

One simulation will be described to illustrate the performance of the system. For this simulation, direction sensitive units are placed every 15 degrees and have bandwidths of 90 degrees (peak response tapers off to 0, 45 degrees on either side of the peak direction). There are 17 peak directions (spanning 180 degrees) and 8 peak speeds (spanning 30 degrees/sec). Since each unit is sensitive to both a speed and a direction, a total of 136 units (136 speed/direction combinations) are available at each location. In this simulation, the system learns on 50 patterns and is then tested on these 50 patterns and on 50 new patterns. The patterns are composed of 1 to 3 line segments positioned at different angles relative to each other. Some example patterns are shown in Figure 6 (See Figure 6). Each pattern is moved at a different velocity.

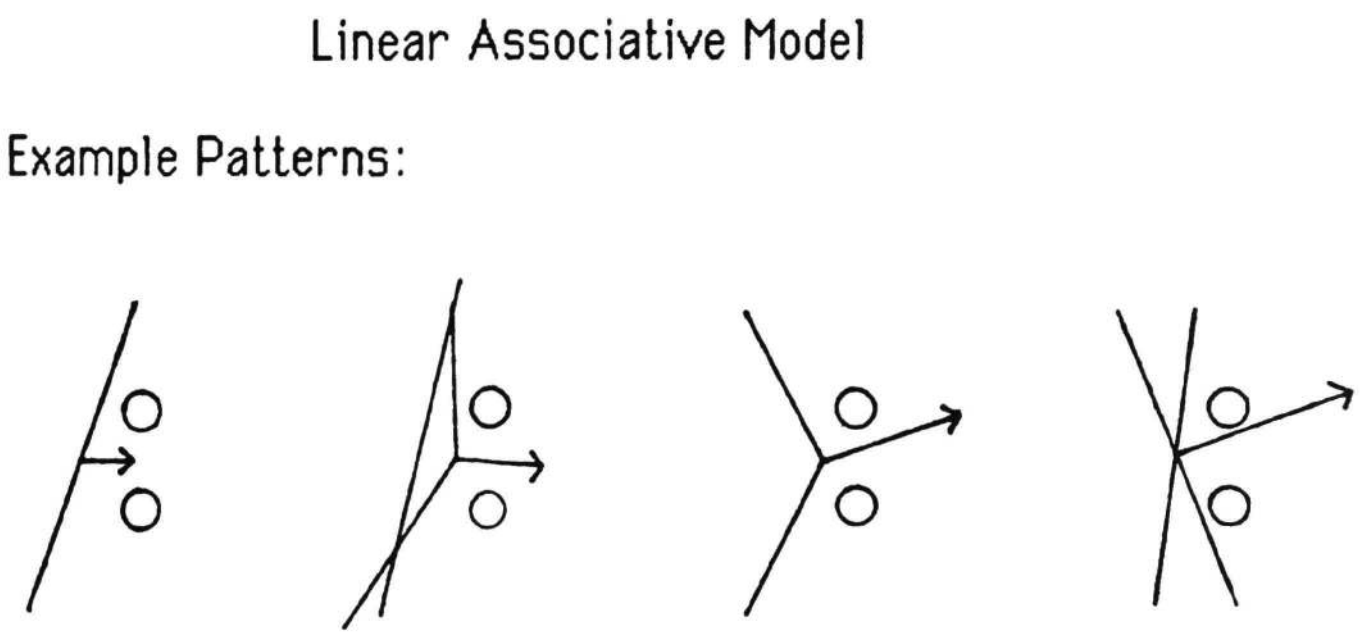


Figure 6

After 15 associations per vector pair, the system reaches stable performance and is tested. The mean cosine between the true association that the system learns, g, and the actual association that the system produces, g', is equal to .98. This represents very good performance. Moreover, the mean cosine for the new, nonassociated vectors is equal to .97. This also represents very good performance and demonstrates that the system is able to generalize quite well to stimuli it has never seen before.

To obtain a finer performance measure, a calculation was made to determine one value of speed and one value of direction for each pattern. A weighted average was taken in which each unit's preferred speed or direction was weighted by its activation level (See Figure 7). The mean difference between the weighted average for the real direction (g) and the reconstructed direction (g') for old patterns was 3.0 degrees while the mean difference for new patterns was 4.2 degrees. The mean difference between weighted averages for real and reconstructed speeds for old patterns was 1.1 degrees per second compared to 1.6 degrees per second for new patterns.

Weighted Average Calculation

$$\text{pattern speed} = \Sigma_i (r_i * s_i) / \Sigma_i r_i$$

$$\text{pattern direction} = \Sigma_i (r_i * d_i) / \Sigma_i r_i$$

where i = unit number
r = activation level of unit
s = speed to which unit is most sensitive
d = direction to which unit is most sensitive

Figure 7

In sum, the model shows excellent performance for extracting two-dimensional translational motion from one-dimensional motion information.

The present model is then extended to handle the two-dimensional projected velocity of objects moving in depth (e.g., in rotating and expanding objects). Again, the model is constructed taking into account the relevant neurophysiological data. Saito et al. (1986), for example, describe three classes of directionally selective cells with large receptive fields (about 35 degrees compared to a mean of about 6 degrees for MT cells) in area MST, an area which receives a direct projection from MT. One class of cells is sensitive to translation in the plane, a second class (size-change cells) is selective for expanding or contracting patterns, and a final class (rotation cells) is selective for rotating patterns (clockwise or counterclockwise) in the frontoparallel plane, or rotating patterns in depth. A common feature of these neurons is that they respond to appropriate patterns anywhere in their large receptive fields at the expense of

being able to precisely signal information about location. Saito et al. (1986) argue that these cells are sensitive to "whole events" of visual motion because they integrate elemental motion signals from MT cells.

These data suggest that the visual system utilizes several distinct stages for motion analysis. In an analogous fashion, the present model takes the output of a second layer that responds to two-dimentional linear motion and feeds it into a third layer that responds to motion of rotation, dilation, or contraction.

The proposed model will be tested using complex motion (the combination of simpler motions). Moreover, the model will be introduced to moving patterns which give rise to illusory perception such as the rotating spiral illusion. In this illusion, a rotating spiral appears to expand or contract. The three layers of the present model result in the extraction of elemental motion which can then be combined in an ensemble code to compute the perceived two-dimensional motion.

The obvious advantage of such a model is that it makes use of the structure of the nervous system as a natural way to constrain the model. Consequently, it can provide insight into the sequential processes involved in motion analysis.

References

Adelson, E.H. & Movshon, J.A. (1982). Phenomenal coherence of moving visual patterns. *Nature*, 300, 523-525.

Ackley, D.H., Hinton, G.E., & Sejnowski, T.J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9, 147-169.

Anderson, J.A. (1983). Cognitive and psychological computation with neural models. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13, 799-815.

Hildreth, E.C. (1983). *The measurement of visual motion.* Cambridge: MIT Press.

Koenderink, J.J. & Van Doorn, A.J. (1976). Local strructure of movement parallax of the plane. *Journal of the Optical Society of America*, 66, 717-723.

Longuet-Higgens, H.C. & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London B*, 208, 385-397.

Movshon, J.A., Adelson, E. H., Gizzi, M.S., & Newsome, W.T. (1985). The analysis of moving visual patterns. In C. Chages, R. Gattas, and C.G. Gross (Eds.), *Pattern Recognition Mechaisms.* Rome: Vatican Press.

Nakayama, K. & Silverman, G.H. (1983). Perception of moving sinusoidal lines. *Journal of the Optical Society of America*, 72.

Regan, D. (1986). Visual processing of four kinds of relative motion. *Vision Research*, 26, 127-145.

Regan, D. & Beverly, K.I. (1978). Looming detectors in the human visual pathway. *Vision Research*, 18, 415-421.

Saito, H., Yukie, M., Tanaka, K., Hikosaka, K. Fukada, Y., & Iwai, E. (1986). Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey, *Journal of Neuroscience*, 6, 145-157.

Sereno, M.E. (1986). A neural model for the measurement of visual motion. *Journal of the Optical Scociety of America, 3,* p. 72.

Van Essen, D.C. (1985). Functional organization of primate visual cortex. In A. Peters & E.G. Jones (Eds.), *Cerebral Cortex: Vol. 3: Visual cortex.* New York: Plenum Publishing Corporation.