Grouping Components of Three-Dimensional Moving Objects in Area MST of Visual Cortex

Richard S. Zemel
Carnegie Mellon University
Department of Psychology
Pittsburgh, PA 15213
zemel@cmu.edu

Terrence J. Sejnowski
CNL, The Salk Institute
P.O. Box 85800
San Diego, CA 92186-5800
terry@salk.edu

Abstract

Many cells in the dorsal part of the medial superior temporal (MST) area of visual cortex respond selectively to spiral flow patterns—specific combinations of expansion/contraction and rotation motions. Previous investigators have suggested that these cells may represent self-motion. Spiral patterns can also be generated by the relative motion of the observer and a particular object. An MST cell may then account for some portion of the complex flow field, and the set of active cells could encode the entire flow; in this manner, MST effectively segments moving objects. Such a grouping operation is essential in interpreting scenes containing several independent moving objects and observer motion. We describe a model based on the hypothesis that the selective tuning of MST cells reflects the grouping of object components undergoing coherent motion. Inputs to the model were generated from sequences of ray-traced images that simulated realistic motion situations, combining observer motion, eye movements, and independent object motion. The input representation was modeled after response properties of neurons in area MT, which provides the primary input to area MST. After applying an unsupervised learning algorithm, the units became tuned to patterns signaling coherent motion. The results match many of the known properties of MST cells and are consistent with recent studies indicating that these cells process 3-D object motion information.

1 INTRODUCTION

A number of studies have described neurons in the dorsal part of the medial superior temporal (MSTd) monkey cortex that respond best to large expanding/contracting, rotating, or shifting patterns (Tanaka et al., 1986; Duffy & Wurtz, 1991a). Recently Graziano et al. (1994) found that MSTd cell responses correspond to a point in a multidimensional space of *spiral* motions, where the dimensions are these motion types.

Combinations of these motions are generated as an animal moves through its environment, which suggests that area MSTd could play a role in optical flow analysis. When an observer moves through a static environment, a singularity in the flow field known as the focus of expansion may be used to determine the direction of heading (Gibson, 1950; Warren & Hannon, 1988). Previous computational models of MSTd (Lappe & Rauschecker, 1993; Perrone & Stone, 1994) have shown how navigational information related to heading may be encoded by these cells. These investigators propose that each MSTd cell represents a potential heading direction and responds to the aspects of the flow that are consistent with that direction.

In natural environments, however, MSTd cells are often faced with complex flow patterns produced by the combination of observer motion with other independently-moving objects. These complex flow fields are not a single spiral pattern, but instead are composed of multiple spiral patterns. This observation that spiral flows are local subpatterns in flow fields suggests that an MSTd cell represents a particular regular subpattern which corresponds to the aspects of the flow field arising from a single *cause*—the relative motion of the observer and some object or surface in the scene. Adopting this view implies a new goal for MST: the set of MST responses accounts for the flow field based on the ensemble of motion causes.

An MST cell that responds to a local subpattern accounts for a portion of the flow field, specifically the portion that arises from a single motion cause. In this manner, MST can be considered to be segmenting motion signals. As in earlier models, the MSTd cell responds to the aspects of flow consistent with its motion hypothesis, but here a cell's motion hypothesis is not a heading direction but instead represents the 3-D motion of a scene element relative to the observer. This encoding may be useful not only in robustly estimating heading detection, but may also facilitate several other tasks thought to occur further down the motion processing stream, such as localizing objects and parsing scenes containing multiple moving objects.

In this paper we describe a computational model based on the hypothesis that an MST cell signals those aspects of the flow that arise from a common underlying cause. We demonstrate how such a model can develop response properties from the statistics of natural flow images, such as could be extracted from MT signals, and show that this model is able to capture several known properties of information processing in MST.

2 THE MODEL

The input to the system is a movie containing some combination of observer motion, eye movements, and a few objects undergoing independent 3-D motion. An optical

flow algorithm is then applied to yield local motion estimates; this flow field is the input to the network, which consists of three layers. The first layer is designed after monkey area MT. The connectivity between this layer and the second layer is based on MST receptive field properties, and the second layer has the same connectivity pattern to the output layer. The weights on all connections are determined by a training algorithm which attempts to force the network to recreate the input pattern on the output units. We discuss the inputs, the network, and the training algorithm in more detail below.

2.1 STIMULI

The flow field input to the network is produced from a movie. The various movies are intended to simulate natural motion situations. Sample situations include one where all motion is due to the observer's movement, and the gaze is in the motion direction. Another situation that produces a qualitatively different flow field is when the gaze is not in the motion direction. A third situation includes independent motion of some of the objects in the environment. Each movie is a sequence of images that simulates one of these situations.

The images are created using a flexible ray-tracing program, which allows the simulation of many different objects, backgrounds, observer/camera motions, and lighting effects. We currently employ a database of 6 objects (a block of swiss cheese, a snail, a panther, a fish, a ball, and a teapot) and three different backgrounds. A movie is generated by randomly selecting one to four of the objects, and a background. To simulate one of the motion situations, a random selection of motion parameters follows: a). The observer's motion along (x, z) describes her walking; b). The eyes can rotate in (x, y), simulating the tracking of an object during motion; c). Each object can undergo independent 3-D motion. A sequence of 15 images is produced by randomly selecting 3-D initial positions and then updating the pose of the camera and each object in the image based on these motion parameters. Figure 1 shows 3 images selected from a movie generated in this manner.

We apply a standard optical flow technique to extract a single flow field from each synthetic image sequence. Nagel's (1987) flow algorithm is a gradient-based scheme which performs spatiotemporal smoothing on the set of images and then uses a multi-resolution second-order derivative technique, in combination with an oriented smoothness relaxation scheme, to produce the flow field.

2.2 MODEL INPUT AND ARCHITECTURE

The network input is a population encoding of these optical flow vectors at each location in a 21x31 array by small sets of neurons that share the same receptive field position but are tuned to different directions of motion. The values for each input unit is computed by projecting the local flow vector onto the cell's preferred direction. We are currently using 4 inputs per location, with evenly spaced preferred directions and a tuning half-width of 45°.

This population encoding in the input layer is intended to model the response of cells in area MT to a motion sequence. The receptive field (RF) of each model MT unit is determined by the degree of spatial smoothing and subsampling in the flow