A Developmental Model of a Gating Mechanism which Leads to Formation of Lateral Cortical Connectivity Patterns

L. Shams  J. Fiser  C. von der Malsburg

University of Southern California
Department of Computer Science
Center for Neural Engineering
HEDCO Neurosciences Bldg.
Los Angeles, CA 90089-2520

Abstract
We present a model inspired by classical conditioning for learning to predict of the visual stimuli in the next time frame. This process of prediction optimization leads to the development of an iso-orientation connectivity pattern as well as a gating mechanism based on motion information. We use gray-level sequences of real visual scenes as retinal input and V1-type orientation tuned filters as features upon which the predictory scheme operates. The sequences contain objects moving in plane, and the network learns to predict the response of the orientation selective cells for the entire visual field namely the moving portion, the background, as well as the points of discontinuity. We argue that translation which was learned in our model is only one example of a general framework where processing a visual transformation does not require object specific feature detection.

Introduction
The essential problem of object recognition in higher level vision is how to discard certain aspects of a visual input while detecting other features in order to identify that the input belongs to a given category among many possible ones. One traditional computational approach to this problem is to represent the object under a given condition (such as pose, size, illumination), and then expand this representation by attaching additional information about the object under different conditions. This approach leads to the problem of how identification of an object is possible from a condition/view which has never been experienced before (Logothetis & Sheinberg, 1996). In this paper we suggest an alternative approach to object recognition based on data on visual development. The basic idea is that the developing visual system first detects typical transformations of the visual inputs (such as translation, looming, occlusion), and their effects on the visual input. Since these transformations are objet independent they can
apply to any new input, moreover, they can be hardwired in the nervous system prior to
learning specific objects. Applying a transformation might involve developing a
representation for new "latent" dimensions, such as speed in 3 dimensions for
transforming a looming image. Nevertheless, the result is that an object representation
has to be developed in this transformed input space facilitating invariant object
perception. We present a scheme which develops circuitry for such a transformation
based on real image sequences, and sets the stage for learning complex features of objects
from lower dimensional input.

**Translation as a sample transformation**

To demonstrate the basic concept we selected the transformation of translation. When an
image of an object is shifting on the retina the visual input is changing in every moment.
Extracting features from this input for object representation is difficult, unless the effects
of changes in the input due to shifting can be discarded. This can be done if a predictory
scheme can tell where a given feature will be and what changes might be anticipated
based on its present condition. Such a predictory scheme for shift can be developed
based on motion detection. If an area of the visual field is coherently moving in one
direction with a given speed, the position of feature A in the next time frame can be
anticipated with high certainty. Thus information unrelated to shift can be extracted from
the input.

Two points about this scheme are worth noting. First, learning a transformation does
not require identification of the object in advance. In our example the only requirement is
existence of a circuitry that can detect common motion. There is evidence in the
physiological, psychophysical and child development literature that the circuitry and the
functionality of motion detection emerges very early preceding complex object
recognition abilities (Burkhalter, Bernardo and Charles, 1993; Katz & Callaway, 1992;
Spelke, 1994).

The second point is that our learning strategy is different from the traditional scheme
of learning shift invariant object recognition (Földiák, 1991; Fukushima, 1980; Hinton,
1987) In the traditional scheme every object has to appear in each position during
training where it to be identified later. This scheme still requires for every new type of
transformation to show each object in all possible condition (subsampled). In our case
only the transformation has to be learned with a sufficient number of examples. Thus we
break down the learning process into learning first some "meta-knowledge" which can be
applied across the set of inputs, and only then we attempt to encode individual inputs.
This paper presents an example for the first step.

**The learning model**

We simulate a recurrent network of V1 type cells with lateral connections via a one-
layer feedforward network where the number of input and output units are equal, and
each unit in the output layer is an abstraction of the same unit in the input layer. The
forward connections from an input unit to its corresponding output unit represents a
feedback loop, and connections from an input unit to other output units represent lateral connections (Fig. 1).

Figure 1. Deriving the model architecture. A) Schematic representation of a V1 patch - with full connectivity. Cylinders symbolize hypercolumns, where each hypercolumn contains a set of local Gabor filters at each possible orientation and spatial frequency range. Links represent recurrent and horizontal interactions (shown only for the central hypercolumn). B) Abstraction of the recurrent network of Panel A by a feedforward network (only for the central hypercolumn). The top cylinder represents the central hypercolumn of the lower layer in time t+1. Recurrent link is represented by a link from the central column in the lower layer to the top column. Other feedforward links represent the horizontal connections of Panel A. A separate layer of motion detectors (shown in the middle) interacts with the feedforward connections. Elements of the motion layer extract local motion information and interact in a gating manner with the feedforward links. C) Multiplicative modulation of each feedforward connection by local motion inputs. There are multiple links between each input and output element (only one link is shown). The local motion units neighboring the feedforward link create a unique initial connection pattern on each feedforward link. Learning acts on the feedforward links based on the result of the multiplicative interaction between the result of the Gabor filter, and the set of motion units. During training only links with ecologically viable combinations of motion pattern will survive.

The input to the network is a small patch in the visual field represented by a grid of 3x3 "hypercolumns" (set of detectors selective for a given scale and orientation). The model illustrates how learning is performed for each hypercolumn in the output layer. We use a set of simple cell-like filters selective for three orientations and one spatial frequency at every sampling locus of the visual field (a Gabor-jet) to represent the hypercolumns in V1. We assume that these cells are fully connected to all of their retinotopic neighbors initially. The input to the network consists of nine jets along with motion information for each of the nine coordinates. As retinal input we used sequence of frames taken by a camera, containing various objects moving horizontally to the right at a fixed rate. Each gray scale frame was complemented by a binary motion map. For every jet at (x,y) in frame i of a sequence the goal of learning is to predict the values of the same jet in the next time unit (captured in frame i+1). Thus, we used grids extracted from frame i as input and the jets corresponding to their center extracted from frame i+1 as target for all frames of each sequence. The network was trained using the Widrow-Hoff gradient method.
We did not build in any a priori knowledge enforcing any binding among the units of the same jet or between jets. The only a priori constraint was the multiplicative interaction among the Gabor units and the motion units transforming the network into a sigma-pi network (Mel & Koch, 1990; Rumelhart, Hinton and McClelland, 1986) (Fig 1C). Each feedforward connection is gated by 5 motion inputs (4 neighbors' and its own) leading to a total of 25 possible motion configurations. The weights of the motion synapses are all fixed to one, and only the weights of connections going to the output units are adaptive. The output of unit j is computed as follows:

$$o_j = \sum_{w} w_k * x_k = \sum_{w} w_k * x_k \prod_{i} \mu_{ik}$$

where k is one of the 32 possible combinations of 5 binary motion variables and

$$\mu_{ik} = \begin{cases} m, & \text{if } k = 1 \\ n, & \text{if } k = 0 \end{cases}$$

m and n are motion and no-motion detectors, respectively, and wijk is the weight of the connection going from ith Gabor input unit which is multiplied by kth motion input aggregate.

Results

As mentioned in the previous section, each feedforward connection from an orientation selective unit in the input layer to an output unit (representing another orientation selective unit) is multiplied by a unique set of motion inputs. Thus, each connection can be interpreted as the Gabor filter at a certain coordinate and a unique motion configuration. Out of the 1100 connection weights corresponding to the 3 orientation filters at 9 coordinates and several motion configurations for each, only a small subset of them remain after learning. The only connections which survived were those of the left neighbor jet and the center jet (or the feedback loops), and among those only those connecting the same orientation in the input and output, and within those only the subset of motion configurations which make the correct prediction of the output filter (or the filter in the next frame) due to the motion status of the visual coordinates in the scene. In other words, the only connections coming from neighboring filters were those of the left neighbor because the motion is to the right, and the surviving connections in this set were only those whose motion configuration indicated that the left neighbor is moving (therefore, replacing the center jet in the next time unit). The rest of the surviving connections belonged to the center coordinate, and in here again only connections linking the same orientation in input and output survived, but this was constrained by the motion configuration again where only those connections with non-moving center (self) or left neighbor coordinates sustained.

Discussion

As we alluded in the previous section, the task of learning to predict the next cell response, involved two distinct but related dynamics. One is that the best predictors of the output of a cell with a given orientation tuning are cells with the same orientation tuning, leading to iso-orientation lateral connectivity pattern. The second dynamics is the gating of such iso-orientation connections based on motion information. If the motion is to the right the only relevant Gabor filter information in predicting the output in the next time unit is that of itself and that of the left neighbor, and not those of the up,
right and down neighbors. In our simulation the network learned that among these relevant iso-orientation connections when to use the information extracted from the left neighbor and when to use that of the same coordinate; a choice controlled by motion information gating. In short, we showed how the motion in one direction and one speed can lead to development of lateral connections, and a proper gating based on spatiotemporal information. This readily extends to all four motion directions in plain, and more velocities but simply introducing a more elaborate layer of motion detectors. Therefore, this scheme can clearly be used in predicting translatory stimuli. In order to extend the paradigm to a 3D motion, depth information --which is again a quite early-developed capacity--should be incorporated into the modulatory part of the network. Processing sequences containing looming or fading stimuli can potentially lead to development of connections among cells with same orientation tuning but of varying spatial frequency, potentially useful for achieving size invariant prediction and/or recognition. The predictive paradigm is not limited theoretically to predicting moving stimuli, but in principal any stimuli which is undergoing one of more visual transformation. In theory, features useful for dealing with illumination variance may be learned if the sequences of the scenes contain such transformations. We also postulate that such features learned through the process of prediction optimization of the general scenes can serve as primitive features used in object-oriented processing such as segmentation and recognition. The iso-orientation connectivities learned in our simulation, for example, can contribute to the contour integration and segmentation tasks, and the possible iso-orientation-intra-spatial frequency connections expected to be learned in the looming objects case may be used for size-invariant object recognition. In our model, nonetheless, we assumed the motion information accurately computed and ready to be used. We acknowledge that the computation of the global motion direction involves addressing the local aperture problem, and may not be locally available with the level of the accuracy which we have used.

References:


