COMMITTED COMPLEX CELLS

Elie Bienenstock, Anastasia Anishchenko, Stuart Geman
Departments of Applied Mathematics, Physics and Neuroscience
Brain Science Program
Brown University, Providence RI

The notion of a Complex Cell, henceforth CC, has been central to our understanding of the physiology of sensory systems for almost half a century. In essence, a CC implements a *disjunction of conjunctions* of elementary stimuli. Consider a classical CC in V1. Such a CC is active whenever any of a collection of stimuli is present in its receptive field: a disjunction. Each one of these disjuncts is an elongated bar, *i.e.*, the conjunction of several elementary stimuli—small dark or light spots in precise (collinear) retinal positions.

The notion of a CC, in addition to being a natural interpretation for a wealth of experimental single-cell data, plays a central role in many models of invariant perception. The original proposal of Hubel and Wiesel has been elaborated into sophisticated hierarchical models of sensory perception. In these models, the complexity, i.e., roughly, the amount of invariance in the response of a cell, increases gradually from low levels to higher levels.

In the HMAX model (Riesenhuber and Poggio,1999), which is a recent, physiologically realistic, model of sensory perception built along these principles, the disjunction operation at any layer of the hierarchy is implemented in the form of a MAX calculation, whereby a cell's response is equal to the activity of its most strongly active input. The conjunction at any level is implemented as a filtering operation using activation variables of the level below. Although the issue to be addressed in the present paper is a general one, we shall for clarity focus the discussion on the HMAX model.

Our main point is that the HMAX model, like other CC-based feedforward models, fails to account for the ability of natural vision systems to deal with images that are highly cluttered and/or contain *several* complex objects. Scene interpretation under conditions such as prevail in natural environments is a challenging computer-vision problem, due to the fact that objects present themselves in a variety of poses, illuminations, states of contiguity and occlusion. The task is to segment *and* recognize, and the challenge is in achieving an adequate articulation of these two operations, the former being classically a low-level problem, while the second belongs to high-level vision. Much of current computer-vision literature is concerned with the chicken-and-egg problem of which, from segmentation and recognition, should come first: for many natural images, it is impossible to perform segmentation in the absence of high-level models yet it is also impossible to recognize objects without at the same time segmenting them from their background and from each other.

We study the performance of the HMAX model (as specified in Riesenhuber and Poggio,1999) on an object-recognition task where the stimuli are realistic grey-level images of chairs in various poses and renditions (synthetic images from 3-D models). We show that the model performs well on images containing a single object, but that performance degrades rapidly as the number of objects in the image increases. The model generally ceases to function when the number of objects in the image reaches 3 or 4.

While it has been argued that natural vision systems might deal with such situations by resorting to selective-attention mechanisms, we believe that such strategies are ineffective for natural images where pervasive ambiguity—resulting from contiguity and mutual occlusion of objects—creates, as mentioned above, a chicken-and-egg problem.

We suggest an alternative solution, in the form of a mechanism for expressing which parts of an image belong with each other—or are likely to belong with each other.

Consider, in the HMAX model, two CC's, *a* and *b*, coding, respectively, for a horizontal bar and a vertical bar, and feeding into a higher-level cell *c*, which codes for an L-junction. The invariance of *c*'s response is inherited from *a* and *b*, but this comes at a price: there is no way for *c* to know whether the stimuli that activate *a* and *b* overlap and thus actually form an L-junction. We would like to have it both ways: *a* and *b* should be invariant and should at the same time tell *c* something specific about their activating stimuli. This is not contradictory: what needs to be transmitted by *a* is merely its (tentative) *commitment* to one of the disjuncts that activate it, say S_a. Based on the commitment of *a* to a disjunct S_a and of *b* to a disjunct S_b, *c* will assess whether or not S_a and S_b overlap. If they do, the scene indeed contains an L-junction.

The above is an outline of a proposed *Committed Complex Cell*, henceforth CCC. Commitment can be implemented by various biological mechanisms, a likely candidate being fast synaptic plasticity. As mentioned, commitment at a given level allows the next higher level in the hierarchy to decide whether or not parts should be grouped together, based on the amount of overlap between them. Biologically, overlap of stimuli can be assessed by firing synchrony, and more generally a departure from conditionally independent firing given the rates of firing.

A CCC is also a CC, and a CCC-based model of invariant perception looks somewhat like a regular CC-based hierarchical model—e.g. of the HMAX type—where activity propagates from lower to higher levels. Computing with CCC's however differs from computing with CC's in two important ways. First, the commitment variable is used in grouping computations. This is implemented in a simple graph-theoretic framework, an alternative to explicitly modeling firing synchrony. Second, the existence of a commitment variable associated with each CCC is used in a mixed bottom-up/top-down computation, in order to resolve the ambiguity inherent in complex images.

Reference:

Riesenhuber, M. and Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.