EMPIRICAL ANALYSES AND CONNECTIONIST MODELING OF REAL-TIME HUMAN IMAGE UNDERSTANDING

IRVING BIEDERMAN, THOMAS W. BLICKLE, GINNY JU, H. JOHN HILTON, AND JOHN E. HUMMEL

UNIVERSITY OF MINNESOTA

In less than 100 msec, humans can accurately interpret images of objects and scenes that have never been experienced previously, or that are extensively degraded, or are viewed from a novel orientation. Recent research and theory (Biederman, 1987a, b, c) suggest that this achievement may be based on a process that decomposes complex visual entities into simple components, typically at regions of matched concavities. Such concavities are almost always produced when shapes are arbitrarily joined (Hoffman & Richards, 1985). The resultant components activate the closest fitting member of a particular set of convex or singly-concave edge-based volumetric primitives, called geons, that are invariant under changes in viewpoint and visual noise and allow objects so represented to possess the same invariance. The geons require only categorical classification of edge characteristics (e.g., straight vs curved; parallel vs nonparallel; vertex type) rather than precise metric specification (e.g., degree of curvature or length of an edge). The latter type of judgments cannot be made with sufficient speed or accuracy by humans to be the controlling processes for real time human object recognition.

The capacity to represent the 10^6 objects that people can rapidly classify derives from an allowance of several viewpoint invariant relations (e.g, TOP-OF, CENTER-CONNECTED) defined for joined pairs of geons such that the same subset of geons represent different objects if they are in different relations to each other. A description of the input consisting of geons + relations is matched against a similar type of description in memory. For example, one kind of lamp can be described as a cylinder CENTERED UNDER THE LARGER END of a cone. Matching is graded in that the activation of a representation will be slower (and of lower maximum value) when image descriptions differ in geons or relations. Geons thus play a role highly analogous to the role played by phonemes in speech perception.

A *Principle of Geon Recovery*, derived from the theory, can account for the major phenomena of object recognition: If an arrangement of two or three geons can be recovered from the image, objects can be quickly recognized even when they are occluded, rotated in depth, novel, extensively degraded, or lacking customary detail, color, and texture.

Empirical Studies of Human Image Understanding

An extensive series of experiments on the perception of briefly presented pictures by human observers has provided empirical support for the theory. In these experiments the subject names or verifies briefly presented (100 msec.) object pictures. Reaction times and errors are the primary dependent variables. Some key results:

1. Simple line drawings showing only the edges of the major geons are identified as rapidly as full color, textured images (Biederman & Ju, 1988). This documents the sufficiency of edge-based descriptions in accounting for the initial activation of a representation of an object.

2. When only two or three geons of a a complex object (such as an airplane or elephant) are visible, recognition can be fast and accurate (though, predictably, not as fast as with the complete image). This supports the derivation of the sufficiency of three geons.

3. Complex objects requiring six or more geons to appear complete are not recognized any more slowly than simple objects (such as a flashlight or cup). This is consistent with a model positing parallel activation of the geons in favor of a serial contour tracing process, such as eyemovements or the kinds of serial routines posited by Ullman (1984).

4. If contour is deleted so that an object's geons cannot be recovered from the image (by deleting cusps for parsing and altering vertices) the object is rendered unrecognizable. If the same or greater amount of contour is deleted but in such a manner that the geons can be recovered through smooth continuation, objects remain identifiable. This result establishes the necessity of the contours posited by RBC.

5. A surprising finding in the previous experiment was the large disruptive effect on error rates and reaction times of interrupting (deleting) contour, such as would be produced when viewing an object behind light foliage, even when the contour could be restored by routines for smooth continuation. This suggests that the routines for contour restoration are not particularly rapid.

6. In the studies described in the previous paragraph, the contour that was removed was removed from every geon in the object. Identification performance is also slowed when objects are missing geons (parts) with the rest of the object intact, such as would occur if the object was partially occluded by a solid surface. According to the theory, the effect of missing or occluded geons is on the matching stage, rather than on the initial determination of the geons.

7. Rotation of the object in the plane slows recognition to a much greater extent than rotation in depth (in contrast to most robot vision models). According to the theory, rotation in the plane affects the TOP-OF relation but the geon descriptions themselves are largely unaffected by rotation in depth.

8. *Complementary* images of objects, in which alternative vertices and edges have been deleted, so that the composite will reveal the original intact image, as illustrated in the upper portion of figure 1, are treated equivalently. This suggests that the memorial representation can be described in terms of geons rather than the precise image features that elicited the geons.

A Connectionist Model of RBC

Hummel, Biederman, Gerhardstein and Hilton (1988) are implementing a connectionist model of RBC as shown in Figure 2. The model is a three layered network which takes as its input an activation vector representing the vertices and cusps in the image of an object as shown in Figure 3. The model gives as output an activation vector representing a geon-based description of the object from which the image was derived. Retinotopic mapping is preserved in all three of the model's layers. Given the spatially specified vertex and edge descriptions in the input vector, a major goal of this effort is to determine: a) if parsing of an image of an object into its constituent geons can be achieved, and b) if the spatial relations among the geons and the global properties of the geons themselves (i.e., parallelism, symmetry), can be derived.

The model's first layer is organized into 182 hexagonally arranged edge and vertex detectors at three spatial scales (20' [N=138], 40' [N=37], and 80' [N=7]). The receptive fields of adjacent detectors overlap and, together, span the central 4° of the visual field. Each of the detectors has 26 nodes for expressing the various vertices (Y, Arrow, Tangent Y, L, T), the edge types comprising them (straight, curved, or cusp) at one of eight orientations.

The next (or hidden) layer is organized into 119 100-node clusters, each of which is termed a geon field. Each geon field receives input from seven contiguous feature detectors at a given scale and passes input, 1:1, to a corresponding geon field in the upper layer. Each detector is mapped to its seven contiguous geon fields at the appropriate scale. The set of seven detectors

mapped to their middle and upper layer geon fields is termed a *column*. The representation in the lower and upper layers have been designed a priori to express general assumptions about edge coding and RBC's object representations, respectively. As the model is being trained to recognize objects through back propagation, the representation in the middle layer will emerge as a function of the constraints of the mapping between the lower and upper layers.

The columns all have identical connection matrices and no connections exist between columns. In this manner, the response of the system will be identical independent of where an object happens to fall in the visual field. A significant economy in the number of connections results from this columnar organization. In total the model contains 17,932 nodes (4,732 in the first layer [26 nodes per feature detector X 182 detectors], 12,000 in the middle layer [100 per geon field for 120 geon fields], and 1,200 in the top layer [10 per geon field for 120 geon fields]). With the system fully interconnected this would result in 71,184,000 connections. But the columnar restriction results in only 2,304,000 connections, a savings of 96.76% in the number of connections.

The upper layer codes distributed representations of geons by locally representing geon attributes, such as whether the cross section is straight or curved. In this layer relations among the geons are represented implicitly in terms of the spatial relations among the patterns of activation representing those geons. The distributed coding of geons at this layer produces an object-centered representation.

REFERENCES

Biederman, I. (1987a). Recognition-by-Components: A Theory of Human Image Understanding. Psychological Review, 94, 115-147.

Biederman, I. (1987b). Matching Image Edges to Object Memory. In Proceedings of the First International Conference on Computer Vision, IEEE Computer Society, 384-392. London, England, June, 1987.

Hoffman, D. D., & Richards, W. (1985) Parts of recognition. Cognition, 18, 65-96.

Ullman, S. (1984). Visual routines. Cognition, 18, 97-159.

Figure 1. Complementary images of a single object. When viewed separately, these images are treated equivalently.

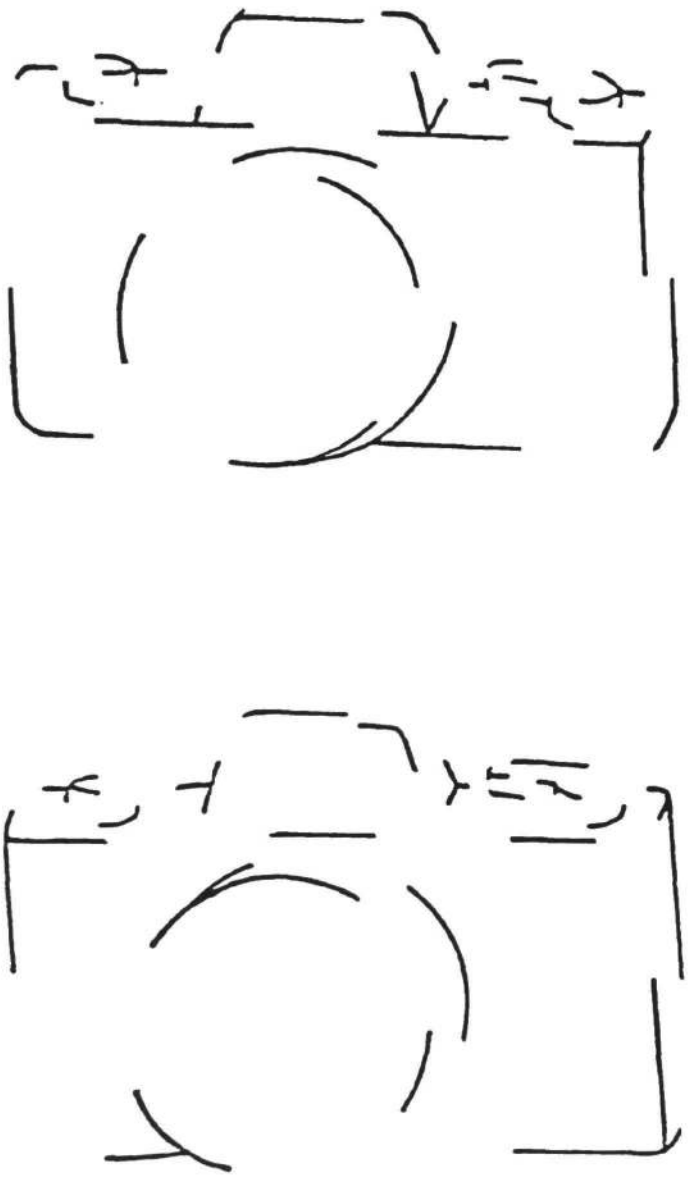

Figure 2. The connectionist model of object recognition.

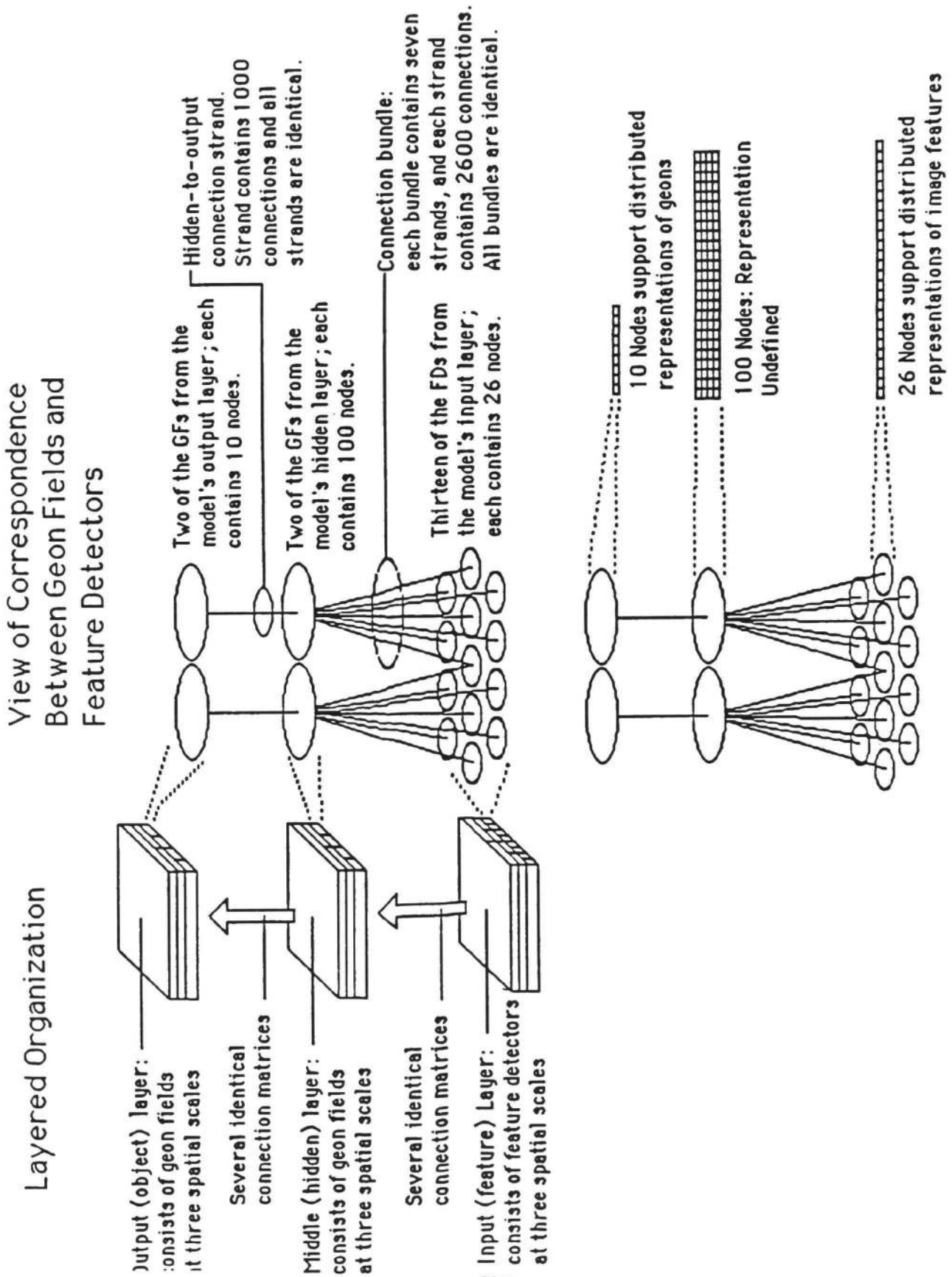

Figure 3. An example of the coding of a single object, a flashlight. The first two values by each vertex give the X and Y coordinates. The next value is the scale at which that vertex would be detected, the first letter provides the vertex type (A = Arrrow, F = Fork, L = L, T = T. The following letters and numbers specify the edge type (S = straight, C = Curved) and orientation of that edge (eight values). Cusps are designated with a K.