A Neural Net Investigation of Vertices as Image Primitives in Biederman's RBC Theory

Hany Farid
(farid@gradient.cis.upenn.edu)
Dept. of Computer Science
University of Pennsylvania
Philadelphia, PA 19104-6389
(215) 898-9517

Gregory Provan
(provan@central.cis.upenn.edu)
GRASP Laboratory
University of Pennsylvania
Suite 336C, 3401 Walnut St.
Philadelphia, PA 19104-6228
(215) 898-8549

Thomas Fontaine
(burrow@gradient.cis.upenn.edu)
Dept. of Computer Science
University of Pennsylvania
Philadelphia, PA 19104-6389
(215) 898-4448

Abstract

Neural networks have been used to investigate some of the assumptions made in Biederman's *recognition by components* (RBC) theory of visual perception. Biederman's RBC theory states, in part, that object vertices are critical features for the 2D region segmentation phase of human object recognition. This paper presents computational evidence for Biederman's claim that viewpoint-invariant vertices are critical to object recognition. In particular, we present a neural network model for 2D object recognition using object vertices as image primitives. The neural net is able to recognize objects with as much as 65% mid-segment centered contour deletion, while it is unable to recognize objects with as little as 25% vertex centered deletion. In addition the neural net exhibits shift, scale and partial rotational invariance.

Introduction

Within the computer vision, cognitive science, psychology, and neurophysiology communities there is much debate over what visual primitives, if any, form the basis for visual reasoning. One important theory of visual object recognition, called *recognition by components* (RBC), has been proposed by Biederman [1985]. Some important principles underlying this theory include: (1) a 2D line drawing is sufficient for most unanticipated visual processing independent of depth, color or texture; (2) line drawings can be segmented into distinct regions at points of deep concavity; (3) objects can be represented in 3D as a set of primitive 3D subparts (geons); (4) non-accidental instances of viewpoint-invariant features in the 2D line drawing are sufficient to permit fast access to the qualitative geon-based model of a 3D object.

Biederman's RBC theory is compelling for many reasons, such as the fact that it proposes a small number of 3D primitives, geons, to account for all possible objects[1]. This implies that the required object memory grows sub-linearly with new objects, ensuring a manageable total memory for object recognition, even given the huge variety of objects in the world.

Many researchers are currently investigating the psychological/neurophsiological plausibility of RBC, as well as whether RBC can lead to powerful and efficient computer vision systems. The plausibility of RBC rests on many factors, such as fast qualitative segmentation of 2D images into regions, well-defined mappings of 2D regions to geons, the representational power of the set of 36 geons for 3D object recognition, etc. Biederman has performed extensive psychological studies of the adequacy, robustness to noise and occlusion, etc. of 2D line drawings for visual recognition. For example, he tested the adequacy of 2D vertices to carry the information necessary for region segmentation by contour deletion experiments [Biederman, 1985].

From a computational point of view, although Biederman's RBC theory has been very influential, implementations of this theory within the computer science/computer vision community (e.g. [Biederman, 1992; Bergevin, 1993; Dickinson, 1992]) have not yet tested all important aspects of RBC. In this paper, neural networks are used to replicate some of the experimental results reported by Biederman concerning the adequacy and viewpoint-invariance of vertex information for recognizing objects in 2D line drawings [Biederman, 1985].

Edge vertices play a crucial role in RBC: they contain the information necessary to segment line drawings into distinct regions, to each of which a geon can be matched. There exists substantial evidence that vertices constitute a 2D primitive crucial to identifying the 3D primitives (geons) for a

[1] This geon representation is qualitative only, since it ignores surface texture and other fine metric variations.

particular image. Figure 1 demonstrates the importance of edge vertices for object recognition. This figure shows a line drawing of a flashlight which cannot be recognized given partial evidence from edge mid-segments (top), but can be recognized given partial evidence from edge vertices (bottom).

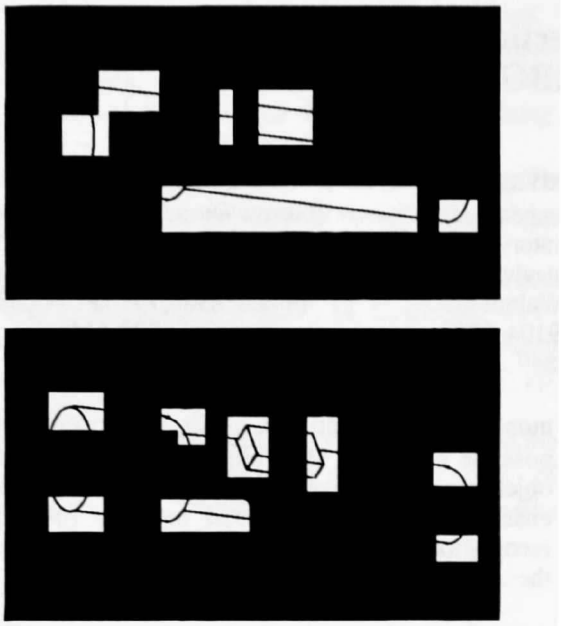

Figure 1. Non recoverable version of an object (top). Recoverable version of an object (bottom). [Biederman, 1985]

From a computational perspective, this paper presents a neural net model which exhibits results similar to Biederman's [1985] experimental demonstration of the importance of object vertices in human 2D object recognition. The neural network architecture presented in this paper is a time-delay, recurrent network, which recognizes simple objects, many of which were originally studied by Biederman [1985; 1987]. The network is able to recognize objects with as much as 65% mid-segment centered contour deletion, while it is unable to recognize objects with more than 25% vertex centered contour deletion. In addition, recognition exhibits invariance to position, scale, and partially to rotation.

Experimental Design

2D object recognition systems utilizing neural networks generally operate on static images (images presented as a *spatial* signal). It has been proposed that object recognition may benefit from considering images as *spatiotemporal* signals [Shastri, 1989]. In this scheme, row i of an NxN image is assimilated by the network at time i ($1 \leq i \leq N$). Recurrent links, as well as multiple links with varying delays between units, are employed to process the temporalized signal. In the networks presented here, three links between units in subsequent layers, with delays of either 1-2-3 or 1-3-5 are utilized. These delays are referred to as the time-delay window. Previous work has shown that the spatiotemporal approach offers advantages such as shift-invariance and inherent retention of local spatial relationships along the temporalized axis [Fontaine, 1992].

Three network models are constructed using spatiotemporal presentation of images in an attempt to develop a network that emphasizes object vertices. The first is a 3-layer fully connected, recurrent net, the second, a 3-layer, partially connected, recurrent net, and the third a 4-layer recurrent net (Figure 2a-c, respectively). Each net contains recurrent links on all hidden and output units, as well as a threshold unit connected to each hidden and output units. The traditional 3-layer nets are trained and tested with both 1-3-5 and 1-2-3 time-delay windows. The 4-layer net uses a 1-2-3 time delay window.

All networks are trained on 100 positive and 100 negative training instances of 30x30 binary images (Figure 3a). For convenience, images are shown here as simple line drawings. The positive training instances consisted of shifted (up to 5 bits or 18% of image size, from the center position) images with either 0, 5, 10, or 15% random contour deletion (Figure 3b-3e). The nets are trained on these perturbed images in order to reduce the chance that the network simply memorize the original image. The negative training instances consists of random bit patterns having the same number of on-bits as the original image (Figure 3f). Training is performed using the Broyden-Fletcher-Goldfrab-Shanno second order learning algorithm [Fletcher, 1980].

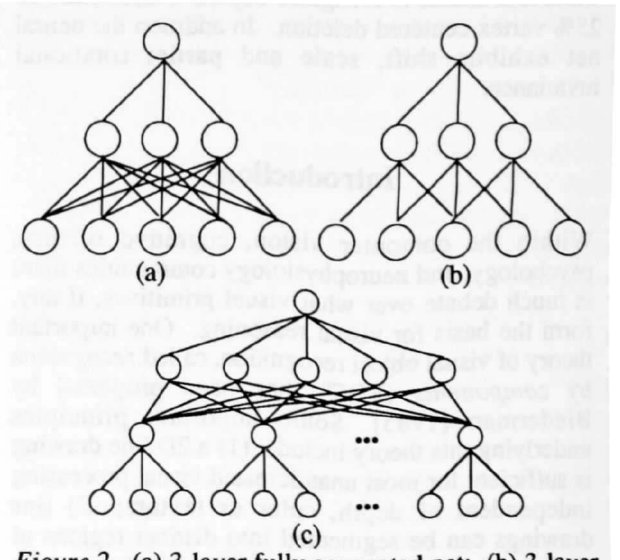

Figure 2. (a) 3-layer fully connected net; (b) 3-layer partially connected net; (c) 4-layer net. All hidden and output units have a recurrent link and a link to a threshold unit (not shown).

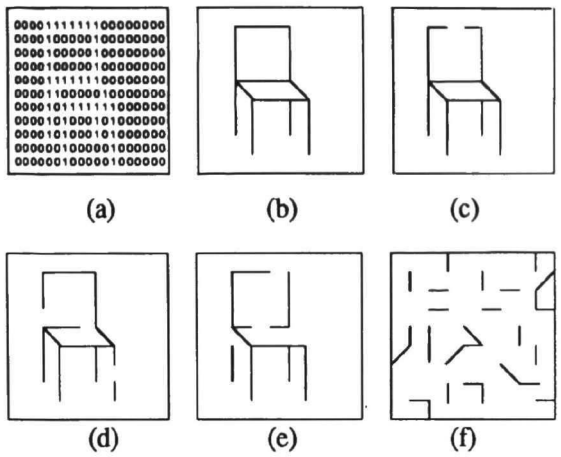

Figure 3. (a) Binary representation of a chair; (b) 0% contour deletion; (c) 5% contour deletion; (d) 10% contour deletion; (e) 15% contour deletion; (f) random bit pattern. Nets are trained on 100 instances of (b-e) and 100 instances (f).

Results

Preliminary experiments indicate that vertical lines are the primary discriminating object feature in the training sets for the traditional 3-layer partially and fully connected nets (Figure 2a and 2b) with either a 1-3-5 or 1-2-3 time-delay window. That is, the network is able to correctly identify objects with complete non-vertical line deletion, while the network is unable to correctly identify objects having high degrees of vertical line deletion (Figure 4).

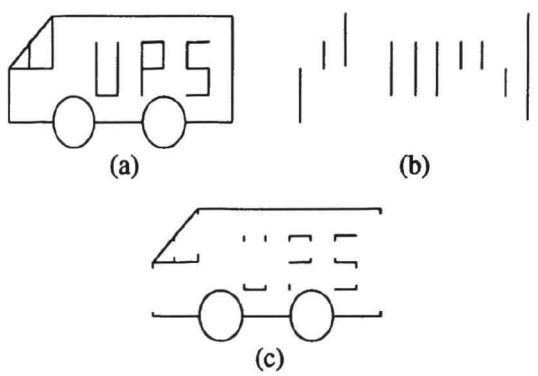

Figure 4. (a) Original image: truck; (b) truck with complete non-vertical line deletion; (c) truck with high degree of vertical line deletion. The fully and partially connected networks are able to correctly identify (a) and (b) but not (c).

Further experiments reveal that vertices are the primary discriminating object feature in the training sets for the 4-layer net (Figure 2c). The experiments consist of training five networks to recognize 1 of 5 objects: chair, cup, lamp, house, and teddy bear (Figure 5). The first three objects (chair, cup, and lamp) are selected for their similarity to images used in Biederman's human experiments. The house is selected because of its increased detail and the teddy bear is selected for its lack of straight edges.

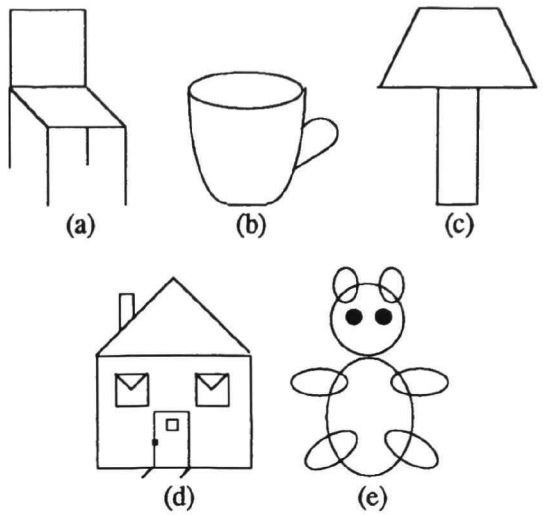

Figure 5. Five networks were trained to recognize one of five objects: (a) chair; (b) cup; (c) lamp; (d) house; (e) teddy bear.

The nets are then presented with corresponding images having either 25, 45, or 65% contour deletion centered either at the mid-segment or vertices (Figure 6). Similar to those results obtained by Biederman with human subjects, the networks are better able to recognize objects having mid-segment as opposed to vertex centered contour deletions (Table I). It is clear from these results that object vertices are critical to object recognition in the 4-layer recurrent network.

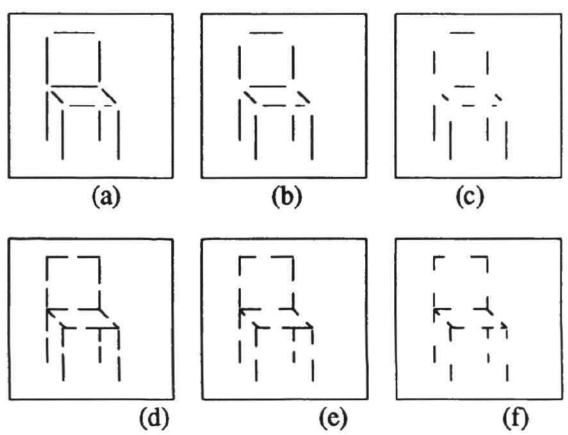

Figure 6. Chair with vertex and mid-segment contour deletion. (a-c) 25%, 45%, 65% vertex; (d-f) 25%, 45%, 65% mid-segment.

Table I. Recognition patterns of objects having mid-segment and vertex centered contour deletions.

Object[1]	Contour Deletion	Locus of Contour Deletion	
		Mid-segment[2]	Vertex[2]
Chair	25%	2.645	2.670
(3.006)	45%	2.585	25.112
	65%	2.679	64.363
Cup	25%	7.818	93.224
(3.348)	45%	11.347	92.864
	65%	10.476	94.560
Lamp	25%	1.725	1.889
(1.992)	45%	2.484	98.090
	65%	5.878	75.675
House	25%	4.003	4.979
(2.501)	45%	5.854	10.253
	65%	5.636	32.325
Teddy Bear[3]	25%	3.164	3.508
(2.909)	45%	3.937	10.531
	65%	7.316	11.866

(1) Object name and score (see footnote 2) of prototypical image.
(2) The values given are the score = $[0.5\sum_i (O_i - T_i)^2 * 10^3]$, where O_i is the output value, and T_i is the target value. The score is representative of the deviation of the test image from the prototypical image. Therefore a low value represents a high degree of match between the prototypical and test image.
(3) Vertices are considered to be where curved lines intersect one another.

Table II. Results of shifting, scaling, and rotation on recognition of a chair.

Object	Score[1]
chair	3.006
shift left	3.132
shift up	1.764
shift down/right	6.869
shift up/left	3.233
scale 50%[2]	6.313
scale 80%	6.383
scale 120%	2.073
rotate 15°[3]	2.813
cup	96.818
lamp	87.590
truck	102.660
house	102.374
random lines[4]	78.573
random vertices[5]	89.003

(1) Score as computed in Table I. Low value represents high degree of match between prototypical and test image.
(2) The net is not able to correctly classify images scaled less than 20%.
(3) The net is not able to correctly classify images with more than 45° rotation.
(4) Randomly drawn horizontal and vertical lines with no vertices.
(5) Randomly drawn vertices. Each vertex was a 3 bit vertical segment connected to a 3 bit horizontal segment.

In order to determine if the trained networks exhibit position, scale and rotationally invariant characteristics, the network trained to recognize a chair is further tested. Scaled, shifted, and rotated chairs are presented to the net. The net is able to successfully identify all such objects as well as distinguish the chair from other objects (e.g. cup, lamp, truck, house, and random patterns) (Table II).

Related Work

This work complements three other systems designed to evaluate the computational feasibility of RBC [Biederman, 1992; Bergevin, 1993; Dickinson, 1992]. The PARVO system [Bergevin, 1993] takes noise-free line drawings and extracts geons from segmented regions to do object recognition. OPTICA [Dickinson, 1992] focuses on the segmentation of regions from noise-free line drawings, but does not perform full recognition. The system by Biederman, et. al. [1992] extracts geons from line drawings, but, similar to [Dickinson, 1992] does not recognize objects. Biederman's system is most similar to ours, in that it uses a neural network architecture; it differs by not explicitly identifying viewpoint-invariant features: analysis of the neural network's hidden layers is required to make such features explicit. In contrast to these systems, our system is able to explicitly identify a viewpoint-invariant image feature (vertices), and correctly identify objects in noisy (contour deleted) images.

This work bears some similarities to recent research in object recognition using neural networks. Spirkovska and Reid [1992] use networks for object recognition which display position, scale and rotation invariance. The major difference is that they hand-code the invariants into the network, whereas the work presented here attempts to "learn" the invariants necessary for recognition, and does not assume them a priori. Soucek [1992] presents some other examples of neural-net-based object recognition systems which are also scale and translation invariant. In addition, the use of neural networks for scale and translation invariant pattern recognition,

such as handwriting recognition, is widespread (e.g. [Fontaine, 1992; Fukushima, 1983]).

Discussion

We have described a neural net analysis of one aspect of Biederman's RBC theory, the strong dependence of the object recognition process on the use of vertices and weaker dependence on edge mid-sections as region segmentation cues. These experiments provide experimental confirmation of Biederman's claims of the viewpoint-invariance of vertices in 2D line drawings being critical to object recognition. Further work is needed to prove that RBC is a computationally feasible explanation for the human visual system. Here, it is simply shown that line junctions are an important invariant property of 2D line drawings, since recognition fails for more than 25% vertex deletion.

From the perspective of 2D object recognition, vertices act as an image primitive in the network presented here. Thus, object recognition exhibits scale, shift, and partial rotational invariance. Achieving rotational invariance has proven to be a hard problem in the area of 2D and 3D object recognition, so it is not surprising that the neural net presented here does not exhibit complete invariance to rotation.

Future Work

We intend to extend this work in order to achieve higher degrees of rotational invariance for 2D and 3D object recognition. One approach currently being considered is the determination to what extent vertices may yield invariant cues for achieving invariance under perspective transformation. Our future work includes studying how neural networks can be used to "learn" invariant properties other than object vertices and geometric properties known analytically (e.g. conics, sets of line and points, etc.) [Forsyth, 1991]. This extension is concerned with the fact that few examples of geometric invariant are currently known [Forsyth, 1991]. Since such image properties are clearly crucial to any efficient wide-ranging recognition system, the library of invariant features needs to be extended.

References

Bergevin, R., and Levine, M. 1993. Generic Object Recognition: Building and Matching Coarse Descriptions from Line Drawings. *IEEE Trans. Patt. Anal. Machine Intell*, 15(1):19-36.

Biederman, I. 1985. Human Image Understanding: Recent Research and a Theory. *Computer Vision, Graphics and Image Processing*, 32:29-73.

Biederman, I., and Blickle, T. 1987. The Perception of Objects with Deleted Contours. State University of Buffalo. Unpublished.

Biederman, I., Hummel, J., Gerhardstein, P., and Cooper, E. 1992. From Image Edges to Geons to Viewpoint invariant Object. *Proc. SPIE Applications of AI X: Machine Vision and Robotics*, 570-578.

Dickinson, S., Pentland, A., and Rosenfeld, A. 1992. 3-D Shape Recovery Using Distributed Aspect Matching. *IEEE Trans. Patt. Anal. Machine Intell*, 14(2):174-198.

Fletcher, R. 1980. *Practical Methods of Optimization.*: J. Wiley.

Fontaine, T., and Shastri, L. 1992. Character Recognition Using a Modular Spatiotemporal Connectionist Model, Technical Report MS-CIS-92-94, Dept. of Computer Science, University of Pennsylvania.

Forsyth, D., Mundy, J., Zisserman, A., Coelho, C., Heller, A., and Rothwell, C. 1991. Invariant Descriptors for 3D Object Recognition and Pose. *IEEE Trans. Patt. Anal. Machine Intell*, 13:971-991.

Fukushima, K., Miyake, S., and Ito, T. 1983. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:826-834.

Shastri, L. 1989. Personal communication.

Soucek, B. 1992. *Fast Learning and Invariant Object Recognition.* J. Wiley.

Spirkovska, L., and Reid, R. 1992. Higher Order Neural Networks in Position, Scale and Rotation Invariant Object Recognition. *Fast Learning and Invariant Object Recognition*, ed. by B. Soucek. J. Wiley.