MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

and

CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo 915 June, 1986
C.B.I.P. Paper 017

VISUAL ATTENTION IN BRAINS AND COMPUTERS

Anya Hurlbert and Tomaso Poggio

Abstract. Existing computer programs designed to perform visual recognition of objects suffer from a basic weakness: the inability to spotlight regions in the image that potentially correspond to objects of interest. The brain's mechanisms of visual attention, elucidated by psychophysicists and neurophysiologists, may suggest a solution to the computer's problem of object recognition.

This report (which has appeared in *Nature* **321**, pp. 651-652, 12 June 1986) describes research done within the Artificial Intelligence Laboratory and the Center for Biological Information Processing (Whitaker College) at the Massachusetts Institute of Technology. Support for the A. I. Laboratory's research in artificial intelligence is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124. The Center's support is provided in part by the Sloan Foundation and in part by the Whitaker College.

Despite its enormous progress in the last few decades, machine vision is still far from achieving the goal that human vision attains with such speed and reliability -- in David Marr's words, to "know what is where by looking."(Marr, 1976). Recent results in the physiology and psychophysics of visual attention accentuate the gap between machines and humans, and provide a first step to understanding why it is so large and what machines must learn in order to overcome it.

Paradoxically, what appear to be the simplest tasks for humans may be the most difficult for machines. Consider, for example, recognizing your mother in a sketch of her sitting in the kitchen. You could immediately and effortlessly locate her face, match it with your memory, and pronounce it a good or bad likeness. If the sketch were upside-down you could easily right it for a proper view. You would probably expend the most painstaking scrutiny in determining just which feature was slightly off, but even so, your final judgment would be quick. In contrast, a computer, using the most sophisticated face recognition routine, would perform the task slowly and incompletely, because it would not know where to start. Given the location of the two eyes in a sketch cluttered with dark round blobs, the routine could then search for the mouth, nose and chin at the appropriate distances and methodically match each feature to a virtually identical image in its memory. But failing to find the eyes, it could not go on to recognize the face.

The difficulty of the face recognition problem--and, more generally, object recognition -- has called into question one of the main assumptions underlying the construction of a machine that sees as humans do. The assumption holds that the goal of the first stages in vision is solely to determine "where" things are -- that is, to transform the initial image, an array of intensity values, into a map of the scene which records the distance and orientation of each surface point relative to the viewer (the "**2-1/2D** sketch"). In machine vision the **2-1/2D** sketch may serve to guide a mobile robot around an obstacle or to control its manipulations as it picks up a tool. But, like the raw image from which it is computed, the **2-1/2D** sketch is itself simply a large array of numbers. Although it may contain preliminary information for object recognition, by assigning a color or texture to each surface point, it does not tell ``what'' things are. The critical task in object recognition is therefore to *find* the object or its crucial part within an array of intensity values or distances. Until now, many of the attempts to elucidate object recognition (reviewed by Besl and Jain, 1985 and by Harmon et al., 1979, for example) have assumed that the relevant object is already located and isolated in the image.

Unlike machines, humans are adept in spotting the salient features of an object. To understand the mechanisms underlying this ability, psychophysicists have investigated visual attention. Treisman (1983) and Julesz (1984) have demonstrated that humans are extremely efficient in detecting a part of an image that differs in a single aspect from its background. For example, a red dot {\it pops out} in a field of yellow dots, and the same happens for a vertical line in a field of horizontal lines. The time required to detect the unusual item is independent of the number of other items, implying that the search for it occurs in parallel across the entire field. The human visual system obviously possesses a fast, parallel mechanism which can direct attention to salient chunks of the image.[1] Although the possible computational purposes of this mechanism have not been probed by psychophysical experiments, its potential role in object recognition seems critical. For example, in

[1]This mechanism is sometimes called "preattentive." Here, we consider it as part of the entire attention mechanism whose characteristics probably require more complex descriptions than ``"serial" or "parallel."

face recognition the attention mechanism may perform two essential steps: first, to locate "blobs" which could be eyes; and second, to direct processing toward the blobs to verify that they are eyes and thereby to initiate recognition. The role of attention, therefore, may be not only to spotlight distinctive parts of the image but, more importantly, to segment the image into objects or parts of objects, a crucial first step in determining what things are.

An important and still open question is: what are the features or primitives that drive attention? Likely candidates are *separable features* which, by definition, can be attended to selectively and are processed independently and in parallel. Pop-out and texture discrimination experiments provide a test for separable features and so far have diagnosed color, line orientation, line ends (terminators) and possibly crossings as candidates.

Conjunction experiments test whether two or more separable features may combine to produce a higher-order primitive. For example, when a green **T** in a field of randomly mixed green **X**s and brown **T**s is the target, it does not pop out, and the time required to detect it increases linearly with the number of background items. Thus the detection of a particular conjunction of color and shape appears to require a search over each item in turn, across the entire field. Conjunction experiments thus reveal another aspect of the attention mechanism, a serial *searchlight* which appears to operate independently of eye movements and does for feature conjunctions what the parallel mechanism does for features.

Until recently, all conjunctions between known separable features had been shown to require the serial searchlight. The recent results of Nakayama and Silverman (1986) reveal a surprising exception to this pattern. In pop-out experiments using fields of small rectangular patterns displayed on a color television monitor, the authors demonstrate that binocular disparity and motion individually behave as separable features. The conjunction of motion and color does not; the search for a pattern of blue upward-moving dots is slow and serial across a field of blue-downward patterns and red-upward patterns. Contrary to this trend, conjunctions of binocular disparity and either color or motion behave as separable features: they are searched for in parallel. (The authors report that when the field splits into two planes, one in front of the other, the search for a conjunction amounts to a pop-out of the unusual item in one plane. Thus we suggest that it may be possible, using a different kind of motion stimulus, to create separate planes of coherent motion and thereby induce a parallel search for motion-color conjunctions.)

The psychophysical studies on separable features coincide with the recent emphasis on functional localisation in visual neurophysiology and neuroanatomy. It is tempting to draw an explicit connection between biology and psychophysics by equating different visual cortical areas with different feature maps; for example, calling V4 the color map, MT the motion map, and V1 the orientation map. The psychophysics would suggest that in a given feature map, at each spatial location there exists a collection of neurons each tuned to a different value of the feature (e.g. red, green or blue for color). Although such an organization has not been demonstrated, the evidence for segregation of functionally similar neurons in distinct cortical areas is steadily accumulating. From this point of view, the results of Nakayama and Silverman have interesting implications for neurons and feature maps: they preclude the existence of neurons tuned for both motion and color, and predict the existence of neurons tuned to a particular combination of binocular disparity and motion and of neurons tuned to disparity and color. The results also suggest that feature maps may be replicated at each of several disparity planes. The prediction of disparity-motion tuned neurons is supported by Maunsell and van Essen's (1983) report of similar neurons in cortical area MT.

Other recent work in visual neurophysiology puts the emphasis on a different aspect of attention. Rather than address computational questions such as, "what are the salient features?" and "how does the attention mechanism work?" or the psychophysical question "is feature processing parallel or serial?", the new class of physiology experiments on alert animals seek to demonstrate the ways in which attention can modulate neuronal responses. In the course of such experiements, insights into the neural circuitry and anatomical location of the attention mechanism have emerged. For example, based on studies of attention-mediated modulation in the inferior parietal lobe (area 7), Lynch et al. (1977) have proposed that neurons there are responsible for directing attention to visual targets.

More recent research has demonstrated the effects of attention at other levels in the visual pathway. Moran and Desimone (1985) have recently shown that, in the monkey, the response of a neuron in V4 or IT to a preferred stimulus (for example, a red horizontal bar) is dramatically reduced when the animal ignores it and instead attends to an ineffective stimulus (such as a green vertical bar) within the same receptive field (which, for IT neurons, may extend at least 12^{o}). The response of the neuron to the preferred stimulus is unaffected when the attended stimulus is outside its receptive field. Thus V4 and IT neurons are able to filter out an irrelevant stimulus when it competes with a relevant stimulus within the same receptive field. V1 neurons do not have this property, and the monkey can not even perform the differential attention task when the two stimuli are close enough to fit within a single receptive field in V1.

A recent psychophysical experiment in humans by Sagi and Julesz (1986) provides an intriguing complement to these physiological results. Sagi and Julesz find that visual attention directed to a random location for an orientation discrimination task enhances the detection of a test flash presented simultaneously within a certain radius of the target. The area of enhancement, which the authors conjecture to be the area covered by the searchlight of attention, varies from 1.5^{o} at 2^{o} eccentricity to about 3^{o} at 4^{o} eccentricity. Interestingly, these areas are likely to be larger than the average receptive field sizes in V1.

The above results imply that attention to one region of an image may involve both suppression of visual processing in irrelevant regions and enhancement of visual processing in relevant regions. Thus attention may indeed be responsible for directing a processing focus to specific locations in the initial steps of recognition. Yet although biological research may have found the key to machine vision, it has yet to describe how it opens the lock.

Computational results suggest that the attention mechanism may be even more complex and powerful than experiments have revealed. Consider again the face recognition problem. Individual features such as eyes or the curved line of nose and mouth can by themselves lead to the hypothesis of a face (see figure 2a). In contrast, as figure 2b shows, features alone cannot be the only cue for recognition. The spatial relationship between the two eye tokens and the closed outer contour can also cue the face recognition process. Ullman (1984) has argued cogently that spatial relations must be computed by a mechanism similar to the serial searchlight of attention.

The unraveling of the full complexity of visual attention will clearly involve computational, psychophysical, and physiological research and in turn will influence not only our understanding of visual perception but also the architecture and the control structure of machine vision systems.

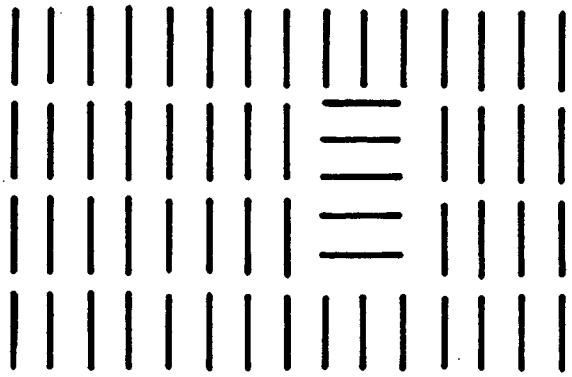

Figure 1. A patch of horizontal line segments "pops out" in a field of vertical line segments.

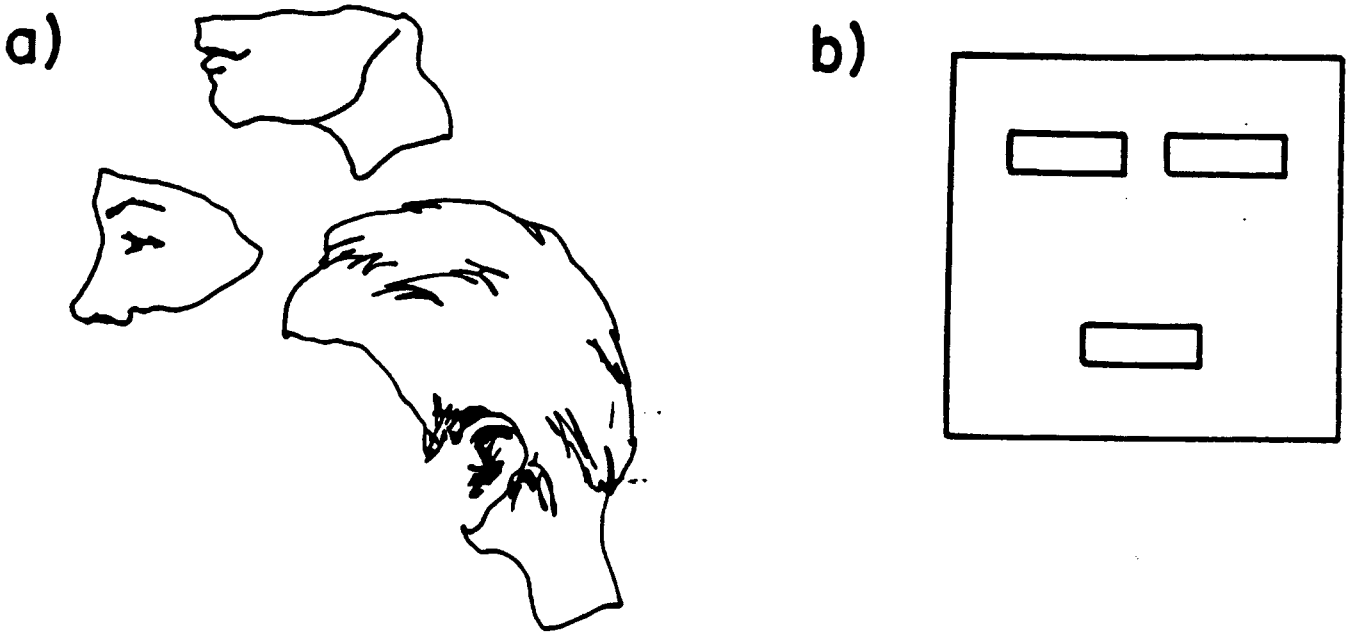

Figure 2. In (a) each separate set of "face" features is sufficient to suggest the hypothesis of a face. In (b) it is the spatial relation between features and not the features themselves that cue recognition to the face hypothesis.

Reading list

Barlow, H. B. *Proc. Roy. Soc. Lond. B*, **212**, 1--35 (1981).

Barrow, H.G. and Tenenbaum, J. M., in *Computer Vision Systems*, Eds. Hanson, A. and Riseman, E., Academic Press, New York (1978).

Bergen, J. R., and Julesz, B., *Nature*, **303**, 696--698 (1983).

Besl, P.J. and Jain, R., *Computing Surveys*, **17**, 74-145 (1985).

Bledsoe, W.W., "Man-machine facial recognition," Report PRI-22, Panoramic Res., Palo Alto (1966).

Braitman, D.J., *Brain Res.*, **307**, 17-2-28 (1984).

Bushnell,C., M. E. Goldberg and D. L. Robinson , *J. Neurophysiol.*, **46**, 755--772 (1981).

Crick, F., *Proc. Natl. Acad. Sci. USA*, **81**, 4586--4590 (1984).

Drumheller, M. and Poggio, T., *Proceedings of IEEE Conference in Robotics*, San Francisco (1986).

Garner, W.R., *The Processing of Information and Structure*, Lawrence Erlbaum, Potomac, MD (1974).

Haenny, P.E., Maunsell, J. H. R., and Schiller, P.H., *Perception*, **13**, A12 (1984).

Harmon, L.D., Kahn, M.K., Lasch, R. and Ramig, P.F., *Pattern Recognition*, **13**, 97-110 (1979).

Hillis, D., *Artificial Intelligence Lab. Memo*, No. **646**, MIT, Cambridge, MA (1981).

Hurlbert, A. and Poggio, T., *Trends Neurosci.*, **8**, 309-311, (1985)

Kanade, T., "Picture processing and recognition of human faces," PhD thesis, Dept. of Information Science, Kyoto University, (1973).

Kelly, M.D. "Visual Identification of people by computer," Stanford, Computer Science Dept. Report No. CS168 (1970).

Koch C. and Ullman, S., *Human Neurobiology*, **4**, 219-227, (1985).

Julesz, B., *Trends Neurosci.*, **7**, 41--48 (1984).

Lynch, J.C., Mountcastle, V.B., Talbot, W.H. and Yin T.C.T., *J. Neurophysiol.*, **40**, 362 (1977).

Marr, D., *Phil. Trans. R. Soc. Lond. B*, **275**, 483--524 (1976).

Minsky, M., *Proc. IRE*, **49**, 8-30 (1961).

Minsky, M., and S. Papert, *Perceptrons*, MIT Press ,Cambridge, Massachusetts, (1969).

Moran, J. and Desimone, R., *Science*, **229**, (1985).

Mountcastle, V. B., Andersen, R. A. and Motter, B.C. , *J. Neurosc.*, **1**, 11, 1218-1235 (1981).

Nakayama, K. and Silverman, G.H., *Nature*, **320**, 264 (1986).

Newsome, W. T. and Wurtz, R.H., *Soc. Neur. Abstr.*, **7**, 732 (1981).

Poggio, T., *Physical and Biological processing of Images*, Eds. Braddick, O.J. and Sleigh, A.C., Springer Verlag, 128-153 (1983).

Poggio,T., Torre, V. and Koch, C., *Nature*, **317**, 314-319, (1985).

Posner, M. I., *Quart. J. Exp. Psychol .*, **32**, 3--25 (1980).

Posner, M.I., Y. Cohen, and R. D. Rafal, *Phil. Trans. R. Soc. Lond. B*, **298**, 187--198 (1982) .

Posner, M. I., C. R. R Snyder, and B. J. Davidson, *J. Exp. Psychol.: General*, **109**, 160--174 (1980).

Sagi, D., and B. Julesz, *Perception*, **A23** , (1984).

Sagi, D., and B. Julesz, *Nature*, in press.

Treisman, A., *J. Exp. Psychol.: H.P. and P.*, **8**, 194-214 (1982).

Treisman, A., *Physical and Biological Processing of Images*, O.J.Braddick and A.C. Sleigh, 316-325, Editors. Springer Verlag , Berlin (1983).

Treisman, A., and G. Gelade, *Cog. Psychol.*, **12**, 97--136 (1980).

Treisman, A., and Schmidt, H., *Cog. Psychol.*, **14**, 107--141 (1982).

Tsal, Y. , *J. Exp. Psychol.: H.P. P.*, **9**, 523--530 (1983).

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

Van Essen D. C., and J. Maunsell, *Trends Neurosci.*, **6**, 370--375 (1983).

Wurtz, R. H., Goldberg, M. E., Robinson, D. L., *Progress in Psychobiology and Physiological Psychology*, **9**, 43--83 (1980).

Zeki, S.M., *Nature*, **274**, 423--428 (1978).

This blank page was inserted to preserve pagination.

CS-TR Scanning Project
Document Control Form

Date : 10/26/95

Report # AIM-915

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 7 (18 IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☒ DOD Form ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages(by page number):_____

Photographs/Tonal Material (by page number):_____

Other (note description/page number):

Description :	Page Number:
IMAGE MAP:	(1-7) UN# TITLE PAGE, 1-7
	(8-18) SCAN CONTROL, DOD, TRGTS (3)

Scanning Agent Signoff:

Date Received: 10/26/95 Date Scanned: 11/6/95 Date Returned: 11/9/95

Scanning Agent Signature: Michael W. Cobb

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER MIT AIM 915	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER AD-A183849
4. TITLE (and Subtitle) "Visual attention in brains and computers"		5. TYPE OF REPORT & PERIOD COVERED A. I. Memo
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Anya Hurlbert and Tomaso Poggio		8. CONTRACT OR GRANT NUMBER(s) DARPA DACA76-85-C-0010 DARPA/ONR N00014-85-K-0124
9. PERFORMING ORGANIZATION NAME AND ADDRESS Artificial Inteligence Laboratory 545 Technology Square Cambridge, MA 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, VA 22209		12. REPORT DATE September 1986 13. NUMBER OF PAGES 7
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Information Systems Arlington, VA 22217		15. SECURITY CLASS. (of this report) unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

visual recognition
face recognition
parallel-serial routines
attention

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Existing computer programs designed to perform visual recognition of objects suffer from a basic weakness: the inability to spotlight regions in the image that potentially correspond to objects of interest. The brain's mechanisms of visual attention, elucidated by psychophysicists and neurophysiologists, may suggest a solution to the computer's problem of object recognition.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**