Now You See It, Now You Don't: Towards a Computational Model of Object Detection in Rapidly Presented Images

Maximilian Riesenhuber

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

Objective: To develop a computational model of object detection in rapidly presented visual images, using a three-pronged approach consisting of computational modeling in interaction with human psychophysics and physiological studies in the macaque.

Motivation: During natural viewing, humans constantly make eye movements, on average three per second [18]. Consequently, the visual system must be able to quickly extract relevant information from a visual scene, for instance during object detection tasks such as predator/prey detection. Previous work [7] has shown that humans are able to detect objects (e.g., "a car") even when they appear in a series of images presented at rates of over 8/s. So far, it is unclear how the visual system can perform this rapid high-level cognitive task.

Previous Work: The impressive ability of the human visual system to rapidly perform object detection tasks in briefly presented visual images was first quantitatively investigated in experiments by Potter [7] that showed that the visual system is able to perform object detection in novel images even when they are presented at a rate of over 8/s. While *reaction times*, i.e., including the time required to perform the detection as well as the (button-press) motor response, in that study were on the order of 450ms, a recent EEG study [16] has provided evidence that the visual system can determine within 150ms whether a novel picture contains a member of a certain object class (e.g., an animal). This is on the order of the latency of cells tuned to (views of) complex objects in inferotemporal cortex [6], a brain area thought to be crucial for object recognition. Taken together, these studies indicate that *at least some object detection tasks can be performed in a time interval roughly equal to a single feed-forward pass through the ventral stream.*

Surprisingly, no computational model of object detection has been proposed in the cognitive literature, where it is commonly assumed that "rapidly presented pictures of objects and scenes can be immediately understood" ([1], p. 239), without closer specification of the process that enables this "immediate" scene understanding.

Approach: Given the tight time constraints and the computational complexity of scene analysis, a computationally sensible approach to perform a detection task is to not analyze the whole scene completely but to focus on detecting the object in question in the image, e.g., by comparing candidate image regions to templates of the target object. This is the strategy used in state-of-the-art computer vision algorithms for object detection in cluttered scenes [5, 12, 15, 13]. These computer vision systems generally rely on a global, serial scanning procedure under central control to examine the image. An alternative approach that uses a local, purely bottom-up process in which the image is analyzed in parallel by a hierarchy of increasingly complex feature detectors has very recently been presented with the "HMAX" model of object recognition in cortex by Riesenhuber and Poggio [10]. Preliminary results have been obtained for recognition in scenes with more than one object in the visual field [9]. Building on HMAX, our approach will be to develop a feedforward computational model in which cues act to "tune" the visual system for the task at hand. The model's predictions will be tested in psychophysical (in collaboration with Pawan Sinha and Molly Potter, MIT) and physiological studies (in collaboration with Earl Miller, MIT) — the results of which in turn will serve to refine the model. Special importance will be placed on using the same conditions in model and experiment to allow a quantitative comparison, as we have done successfully in the case of view-tuned cells in inferotemporal cortex [10]. In particular, we will vary cue specificity (identical picture cue, basic level cue, superordinate level cue), distractor/target similarity (i.e., distractors of the same basic level class for identity cues etc.), and scene complexity (single, isolated objects; combinations of isolated objects; cluttered scenes). In the psychophysical part of the project, we plan to run subjects on an RSVP (rapid serial visual presentation) paradigm, while in the physiological studies, monkeys will perform a detection task similar to the one used in [16] on an identity level and a basic level, leveraging the results of an ongoing study where two monkeys have been trained on a cat/dog basic level categorization task [3]. Here, recordings in inferotemporal cortex will allow us to directly look at the time course of the neural response and the detection process, and also to investigate the influence of top-down cue signals on neuronal firing.

Difficulty: While the images used in previous RSVP studies [7, 8, 4] were only loosely controlled for target/distractor similarity and scene complexity, our proposed studies require a finer level of control of these parameters. For identity level trials, we can use a 3D morphing system developed in Poggio lab [14] to generate distractor objects that can be made arbitrarily similar to the target objects. We have already used this technique successfully in modeling studies [11] and physiology experiments [3] on object categorization. In trials involving cues beyond the identity level (e.g., basic or superordinate level), special care must be taken to assure that subjects' categorization schemes agree with the ones assumed in the experiment. In the physiology experiment, it will be challenging to train the monkeys on an additional recognition task using the same stimuli as in the categorization task.

Impact: The question we are studying, how cues on a higher level can influence processing on lower levels, is of fundamental interest not only in vision but in many other areas of cognitive neuroscience where the brain has to filter out the task-relevant component from a stream of incoming signals, such as in speech perception when a voice has to be filtered out among other signals, even other speakers, such as in the "cocktail party" setting. Hence, we believe that the mechanisms we are investigating in the visual system are likely of general relevance also to cognitive processing in other areas of cortex. On a more practical note, insight into how the visual system performs recognition in clutter might also be relevant to developing more powerful computer vision systems.

Research Support: Research at CBCL is supported by ONR, Darpa, NSF, Kodak, Siemens, Daimler, ATR, ATT, Compaq, Honda. M.R. is supported by a McDonnell-Pew Award in Cognitive Neuroscience.

References:

[1] Coltheart, V. In *Fleeting Memories: Cognition of Brief Visual Stimuli* [2].

[2] Coltheart, V., editor (1999). *Fleeting Memories: Cognition of Brief Visual Stimuli*. MIT Press, Cambridge, MA.

[3] Freedman, D., Riesenhuber, M., Shelton, C., Poggio, T., and Miller, E. (1999). In *Soc. Neurosci. Abs.*, **29**, 884.

[4] Intraub, H. (1981). *J. Exp. Psych.: Hum. Percept. Perf.* **7**, 604–610.

[5] Papageorgiou, C., Oren, M., and Poggio, T. (1998). In *Proceedings of the International Conference on Computer Vision, Bombay, India*, 555–562. IEEE, Los Alamitos, CA.

[6] Perrett, D.I., Hietanen, J.K., Oram, M.W., and Benson, P.J. (1992). *Philos. Trans. Roy. Soc. B.* **335**, 23–30.

[7] Potter, M. (1975). *Science* **187**, 565–566.

[8] Potter, M. (1976). *J. Exp. Psych.: Hum. Learn. Mem.* **2**, 509–522.

[9] Riesenhuber, M. and Poggio, T. (1999). *Neuron* **24**, 87–93.

[10] Riesenhuber, M. and Poggio, T. (1999). *Nat. Neurosci.* **2**, 1019–1025.

[11] Riesenhuber, M. and Poggio, T. (1999). AI Memo 1679, CBCL Paper 183, MIT AI Lab and CBCL, Cambridge, MA.

[12] Rowley, H., Baluja, S., and Kanade, T. (1998). *IEEE PAMI* **20**, 23–38.

[13] Schiele, B. and Crowley, J. (1998). Technical Report 453, MIT Media Lab.

[14] Shelton, C. Master's thesis, MIT, (1996).

[15] Sung, K. and Poggio, T. (1998). *IEEE PAMI* **20**, 39–51.

[16] Thorpe, S., Fize, D., and Marlot, C. (1996). *Nature* **381**, 520–522.

[17] Ungerleider, L. and Haxby, J. (1994). *Curr. Op. Neurobiol.* **4**, 157–165.

[18] Yarbus, A. (1967). *Eye Movements and Vision*. Plenum Press, New York.