Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2012-016
CBCL-309

June 21, 2012

Throwing Down the Visual Intelligence Gauntlet

Cheston Tan, Joel Z Leibo, and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Throwing Down the Visual Intelligence Gauntlet

Cheston Tan, Joel Z Leibo, and Tomaso Poggio

McGovern Institute for Brain Research, and the Center for Biological and Computational Learning at MIT

This document is a penultimate draft. The final version was published as a chapter in *Machine Learning for Computer Vision (2012); eds: Cipolla R, Battiato S, Giovanni Maria F. Springer: Studies in Computational Intelligence Vol. 411.*

Abstract

In recent years, scientific and technological advances have produced artificial systems that have matched or surpassed human capabilities in narrow domains such as face detection and optical character recognition. However, the problem of producing truly intelligent machines still remains far from being solved. In this chapter, we first describe some of these recent advances, and then review one approach to moving beyond these limited successes – the *neuromorphic* approach of studying and reverse-engineering the networks of neurons in the human brain (specifically, the visual system). Finally, we discuss several possible future directions in the quest for visual intelligence.

1 Artificial Intelligence: Are We There Yet?

Every few years (and sometimes more often), the world becomes abuzz with excitement over some new technology that, finally, after all these years, promises to fulfill the dream of Artificial Intelligence (AI), first proposed back in 1956 at the dawn of the Age of Computing. The latest of these technologies is *Watson*, a natural language question-answering computer system that, in February 2011, defeated two of the best human contestants ever in the quiz show *Jeopardy!*

In 2007, the buzz-worthy news was that six teams successfully completed the DARPA Grand Challenge [3], a simulated 60-mile urban course designed to test the capabilities of driverless vehicle systems. The overall winner completed the course just over 4 hours, averaging approximately 14 mph. Just two years prior, in 2005, five teams completed the 132-mile off-road desert course, a stunning reversal of the 2004 results, in which the best vehicle only managed to complete an embarrassing 7.36 miles of the 150-mile course [3].

Apart from these and other headline-grabbing milestones like chess-playing computer Deep Blue defeating the reigning world chess champion in 1997, there have also been a host of less-publicized technological advances that have matched (and sometimes even surpassed) human abilities. These include face-detection technology in digital cameras [4], a pedestrian-detection feature in the latest luxury car models [1, 6], and optical character recognition (OCR) technology used by postal services [9] around the world, just to name a few.

Meanwhile, in the realm of computer vision, performance of algorithms on common datasets such as Caltech-101 [2] and PASCAL [8] have improved steadily over the years. While typical human performance on these datasets has not been quantified, it is not unimaginable that computers may reach this performance benchmark in a decade or less.

Given this plethora of advances and achievements that would be utterly jaw-dropping to the early pioneers of computing and AI, should we thus conclude that victory in achieving AI is close at hand? Our answer is a clear ***no***. None of these systems or computers can really be described as intelligent in the way that one would describe a person. Each of these systems performs well in a narrow domain such as categorizing objects or playing chess, but two key characteristics of human intelligence are broadness and flexibility. A typical person is capable of learning not only chess, but hundreds of other games. With sufficient training, a typical person would also be able to become good at any of these.

One might argue that if various artificial systems achieve human-level performance in a sufficient number of these narrow domains, then putting these components together in a single system would result in some semblance of human intelligence. We think otherwise – such a system may fulfill the criterion of broadness, but it would lack flexibility.

We do not mean to downplay the amazing advances in computing that have occurred and are occurring. Computers have changed our lives for the better to an extent that almost no other technology has, and the advances mentioned above will no doubt further these changes in exciting and dramatic ways. The ongoing pursuit of tangible engineering solutions to pressing challenges is an important and worthwhile research agenda. However, the goal of replicating broad and flexible human-like intelligence will not be achieved just by stringing together the solutions to specialized problems.

1.1 A Compass for the Uncharted Journey Towards Intelligence

One approach (arguably the default one) to move beyond the limitations of tackling narrow domains in isolation, is to study the brain – a computational system which we already know exhibits intelligence. However, much of the computation in the brain is still poorly understood. Understanding how the individual functioning of billions of brain cells leads collectively to human intelligence and cognition remains a major challenge.

Nonetheless, a good bet – and the one we make – is that in order to rise to the challenge, we need to understand how the *visual cortex* works, and then reproduce it in computers. Vision is one of the most studied aspects of human cognition, and over one-third of the cerebral cortex (the seat of cognition) is dedicated to visual processing. Replicating intelligence will ultimately require more than just understanding visual cognition, but it is likely to be the best place to start.

This *neuromorphic* approach does not imply the slavish, neuron-by-neuron reproduction of the human visual system. Just like how the marvels of modern flight have been enabled by the scientific understanding of the principles of aerodynamics, we believe that the key to unlocking intelligence is to investigate its principles by studying systems that truly exhibit intelligence (and to validate our understanding of these principles by building testable computational models).

In the rest of this chapter, we first briefly review some of the computational principles underlying visual intelligence in the cortex, and then proceed to sketch the first steps towards replicating these principles

in computational models. Finally, we discuss several suggestions for moving research forward in the direction of true intelligence.

2 The Neuromorphic Approach to Visual Intelligence

By now, there are probably several hundred models of the visual cortex. Most deal with specific visual phenomena (such as visual illusions) or specific parts of the visual cortex. Many of these have yielded useful contributions to neuroscience. However, if the goal is to use these models to guide the engineering of a new generation of systems that approach human visual intelligence, then more comprehensive models addressing a wide range of phenomena and visual areas are needed. Thus, in this section, we review a computational model of visual cortex that fulfills precisely these criteria. This model (or more precisely, this class of models) provides a preliminary but illustrative sketch of how computations performed by the visual cortex are closely linked to the principles underlying visual intelligence (at least for two principles, among the few that are currently known). First, however, we briefly review relevant background knowledge regarding the visual cortex.

2.1 Anatomy and Physiology of the Primary Visual Cortex (V1)

Neural recordings from the *primary visual cortex* (also known as *V1*) of cats in the early 1960s by Nobel laureates[1] David Hubel and Torsten Wiesel yielded the observation of so-called *simple cells* responding to edges of a particular orientation. Hubel and Wiesel also described another class of cells with more complicated responses which came to be called *complex cells*. In the same publication [15], they hypothesized that (at least some of) the complex cells may be receiving their inputs from the simple cells.

The simple cells' receptive fields[2] contain oriented "on" regions in which presenting an appropriately-oriented edge stimulus excited the cells, and "off" regions for which presenting a stimulus suppresses neural activity. These classical Gabor-like receptive fields can be understood by noting that they are easily built from a convergence of inputs from cells in the *lateral geniculate nucleus* (LGN), a brain structure from which V1 receives strong inputs. The simple cells respond only when receiving simultaneous inputs from several LGN cells with receptive fields arranged along a line of the appropriate orientation. Fig. 1a is a reproduction of Hubel and Wiesel's original drawing from their 1962 publication illustrating the appropriate convergence of LGN inputs.

In contrast to simple cells, Hubel and Wiesel's complex cells respond to edges with particular orientations, but notably have no "off" regions where stimulus presentation reduces responses. Most complex cells also have larger receptive fields than simple cells, i.e. an edge of the appropriate orientation will stimulate the cell when presented anywhere over a larger region of space. Hubel and Wiesel noted that the complex cell fields could be explained by a convergence of inputs from simple cells. Fig. 1b reproduces their scheme.

[1] One half of the 1981 Nobel Prize in Physiology or Medicine was jointly awarded to Hubel and Wiesel. The other half was awarded to Roger Sperry.

[2] A cell's *receptive field* is a region of visual space in which the presence of a stimulus will affect the activity of that cell.

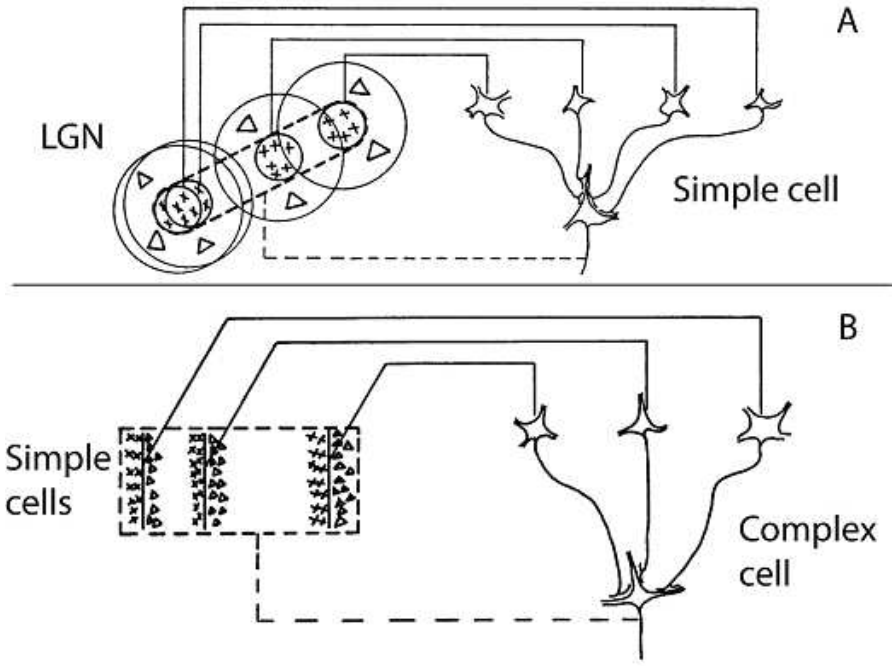


Figure 1: Construction of V1 simple and complex cell receptive fields via convergent inputs. (Adapted from Hubel and Wiesel 1962)

Following Hubel and Wiesel, we say that the simple cells are tuned to a particular preferred feature. This tuning is accomplished by weighting the LGN inputs in such a way that a simple cell fires when the inputs arranged to build the preferred feature are co-activated. In contrast, the complex cells' inputs are weighted such that the activation of any of their inputs can drive the cell by itself. So the complex cells are said to pool the responses of several simple cells. As information about the stimulus passes from LGN to V1, its representation increases in selectivity; patterns without edges (such as sufficiently small circular dots of light) are no longer represented. Then, as information passes from simple cells to complex cells, the representation gains tolerance to the spatial position of the stimulus. Complex cells downstream from simple cells that respond only when their preferred feature appears in a small window of space, now represent stimuli presented over a larger region.

2.2 Hubel-Wiesel Models: Successive Tuning and Pooling

Beyond V1, visual cortex is broadly organized into two parallel processing streams, a *ventral* stream mostly involved in analysis of shape information, and a *dorsal* stream mostly involved in analysis of motion and location [23]. Both streams are organized hierarchically, with receptive field sizes and preferred feature complexity increasing along the way from their starting point in V1 to subsequent areas (see Fig. 2). The present chapter focuses on the ventral stream, but see [18] for related models of the dorsal stream.

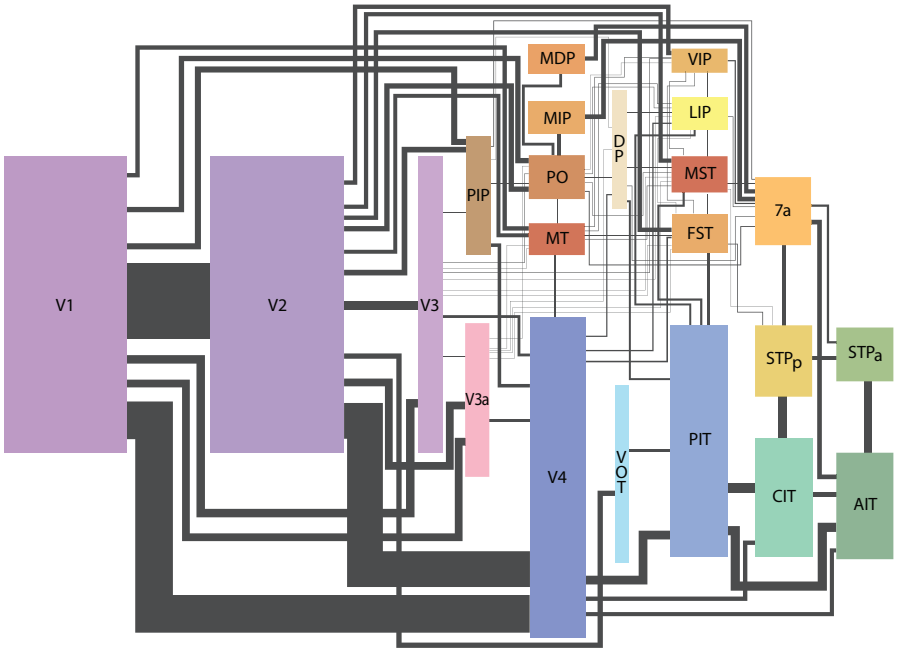


Figure 2: Areas of the visual cortex. Each rectangle represents a visual area; the size is proportional to cortical surface area. The lines connecting the areas have a thickness proportional to the estimated number of fibers in the connection. Ventral stream areas are in the bottom half; dorsal stream areas are in the upper half. (Reprinted from [38])

At the top of the ventral hierarchy, the cells in *inferotemporal* (IT) cortex respond selectively to highly complex stimuli, and invariantly over several degrees of visual angle. Hierarchical models inspired by the work of Hubel and Wiesel [10, 13, 22, 25, 28, 30, 32, 37, 39], henceforth termed ***H-W models*** ([28, 30, 32] were previously known as ***HMAX***), seek to achieve similar selectivity and invariance properties by subjecting visual inputs to successive – and often alternating – tuning and pooling operations, just like how V1 simple cells are tuned to specific patterns of LGN inputs and complex cells pool the responses of simple cells with similar orientation selectivity. A major algorithmic claim made by these H-W models is that the repeated application of AND-like tuning is the source of the selectivity in IT cortex. Likewise, repeated application of OR-like pooling produces invariant responses.

H-W models nicely illustrate the close correspondence between visual intelligence and the visual cortex. Through careful scientific study, the idealized tuning and pooling operations in H-W models have been distilled from the jumble of neural activity in the visual cortex. These operations are, not surprisingly, associated with the principles of selectivity and invariance that underlie object recognition – an important component of visual intelligence.

2.3 Consistency with Experimental Results at Multiple Levels

H-W models were constructed based on the dual principles of selectivity and invariance, derived initially from the study of V1 but also subsequently found in other visual areas. The key test of any model, however, is the consistency of its predictions with phenomena beyond those used in its construction.

In that regard, recent H-W models are consistent with experimental findings at multiple levels of analysis, from the computational to the biophysical level. Being consistent across all these levels is a high bar and an important one. It is relatively easy to develop models that just explain one phenomenon or one illusion or one visual area, but they remain just that: a model of that one thing. Such narrow models are not useful for guiding future research on the general problems of vision and intelligence.

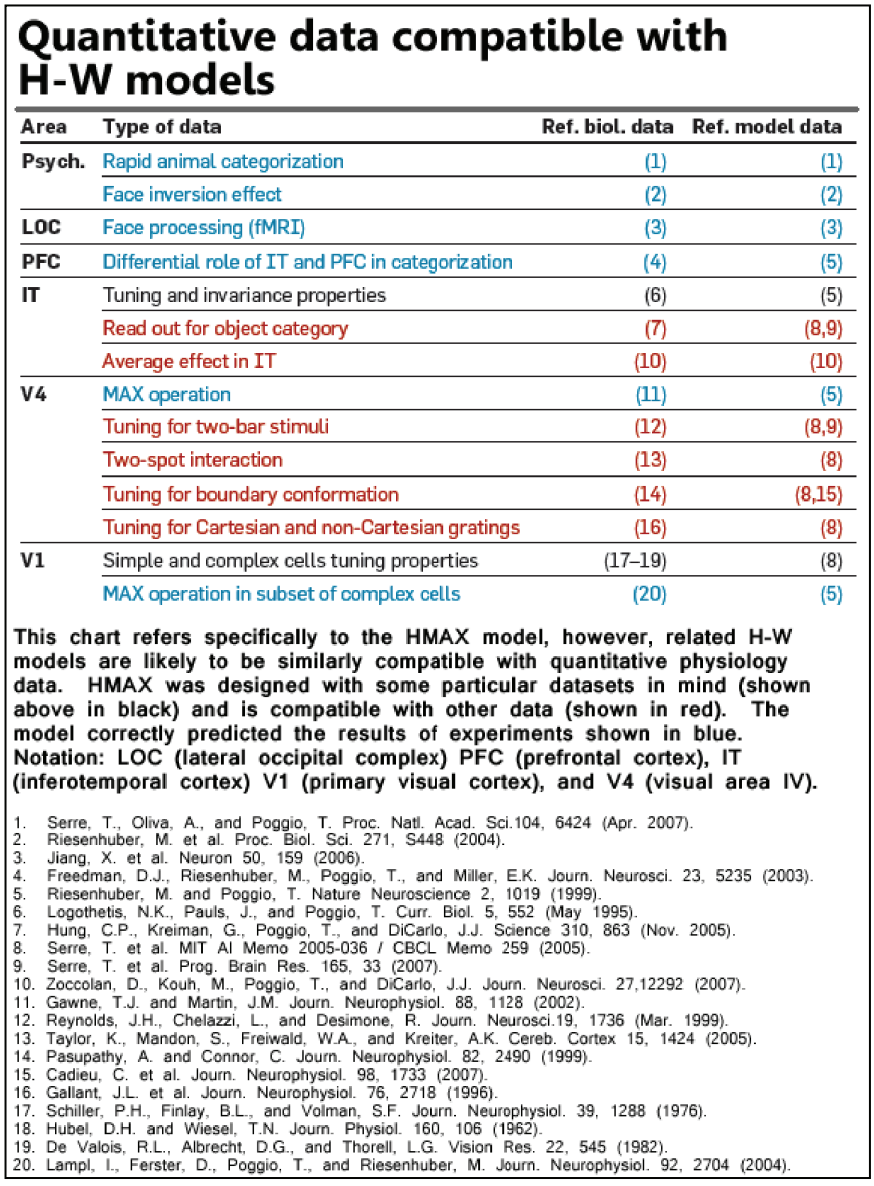
Predictions of H-W models (in particular, the HMAX model [28, 30, 32]), have generally fared well when checked against experimental results. In this section, we briefly review evidence from electrophysiology and psychophysics in relation to this class of models. A more comprehensive list of comparisons between experimental and model data can be found in Fig. 3 (also see [29, 30] for more details).

Beginning in the primary visual cortex, an electrophysiology study in cats uncovered evidence that the brain employs an OR-like pooling operation, predicted by Riesenhuber and Poggio [28] as the mechanism by which complex cell receptive fields are built from simple cells [20]. Further evidence for this operation was independently found in area V4 (a visual area that receives inputs from V1) in rhesus monkeys [14], confirming that this key operation exists in multiple visual areas.

In addition, single-unit electrophysiology experiments in area V2 revealed that neurons in this area are sensitive to combinations of orientations [11], consistent with the AND-like tuning operation in the model [28]. Furthermore, a quantitative fit was established between H-W model units and the firing rates of V4 neurons evoked by a library of stimuli. This model was fit with data from a subset of cells, and could generalize to predict the responses of other cells to novel stimuli [12].

H-W models have been particularly influential in the study of IT cortex, the visual area situated at the top of the hierarchy of shape-processing areas. It is possible to decode object category information invariantly to translation and scaling from the responses of a population of IT cells [16]. The top-most layer of an H-W model also supports similar decoding (see Fig. 4). It is this observation that populations of IT cells and H-W model units both provide useful representations for decoding stimulus information which motivates much of the interest in IT cortex from the computer vision and neuroscience communities.

Importantly, H-W models demonstrate human performance levels on certain psychophysical tasks. Human observers can categorize scenes containing a particular prominent object, such as an animal or a vehicle, after only 20 ms of exposure. Old EEG experiments in humans, but also especially new data obtained with a “read-out” information decoding technique, show category information in the neural population of IT of the macaque at 80ms after a stimulus is seen. These experimental results establish a lower bound on the latency of visual categorization decisions made by the human visual system, and suggest that categorical decision-making can be implemented by a feedforward information processing mechanism like an H-W model [19, 21, 33, 34, 36]. Serre et al. showed that a specifc H-W model does indeed reach human-level performance on this task [30]. Many of the individual images on which the



Quantitative data compatible with H-W models

Area	Type of data	Ref. biol. data	Ref. model data
Psych.	Rapid animal categorization	(1)	(1)
	Face inversion effect	(2)	(2)
LOC	Face processing (fMRI)	(3)	(3)
PFC	Differential role of IT and PFC in categorization	(4)	(5)
IT	Tuning and invariance properties	(6)	(5)
	Read out for object category	(7)	(8,9)
	Average effect in IT	(10)	(10)
V4	MAX operation	(11)	(5)
	Tuning for two-bar stimuli	(12)	(8,9)
	Two-spot interaction	(13)	(8)
	Tuning for boundary conformation	(14)	(8,15)
	Tuning for Cartesian and non-Cartesian gratings	(16)	(8)
V1	Simple and complex cells tuning properties	(17–19)	(8)
	MAX operation in subset of complex cells	(20)	(5)

This chart refers specifically to the HMAX model, however, related H-W models are likely to be similarly compatible with quantitative physiology data. HMAX was designed with some particular datasets in mind (shown above in black) and is compatible with other data (shown in red). The model correctly predicted the results of experiments shown in blue. Notation: LOC (lateral occipital complex) PFC (prefrontal cortex), IT (inferotemporal cortex) V1 (primary visual cortex), and V4 (visual area IV).

1. Serre, T., Oliva, A., and Poggio, T. Proc. Natl. Acad. Sci.104, 6424 (Apr. 2007).
2. Riesenhuber, M. et al. Proc. Biol. Sci. 271, S448 (2004).
3. Jiang, X. et al. Neuron 50, 159 (2006).
4. Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. Journ. Neurosci. 23, 5235 (2003).
5. Riesenhuber, M. and Poggio, T. Nature Neuroscience 2, 1019 (1999).
6. Logothetis, N.K., Pauls, J., and Poggio, T. Curr. Biol. 5, 552 (May 1995).
7. Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. Science 310, 863 (Nov. 2005).
8. Serre, T. et al. MIT AI Memo 2005-036 / CBCL Memo 259 (2005).
9. Serre, T. et al. Prog. Brain Res. 165, 33 (2007).
10. Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J.J. Journ. Neurosci. 27,12292 (2007).
11. Gawne, T.J. and Martin, J.M. Journ. Neurophysiol. 88, 1128 (2002).
12. Reynolds, J.H., Chelazzi, L., and Desimone, R. Journ. Neurosci.19, 1736 (Mar. 1999).
13. Taylor, K., Mandon, S., Freiwald, W.A., and Kreiter, A.K. Cereb. Cortex 15, 1424 (2005).
14. Pasupathy, A. and Connor, C. Journ. Neurophysiol. 82, 2490 (1999).
15. Cadieu, C. et al. Journ. Neurophysiol. 98, 1733 (2007).
16. Gallant, J.L. et al. Journ. Neurophysiol. 76, 2718 (1996).
17. Schiller, P.H., Finlay, B.L., and Volman, S.F. Journ. Neurophysiol. 39, 1288 (1976).
18. Hubel, D.H. and Wiesel, T.N. Journ. Physiol. 160, 106 (1962).
19. De Valois, R.L., Albrecht, D.G., and Thorell, L.G. Vision Res. 22, 545 (1982).
20. Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. Journ. Neurophysiol. 92, 2704 (2004).

Figure 3: Summary of comparisons between data from biological experiments and from the HMAX model. (Adapted from [31])

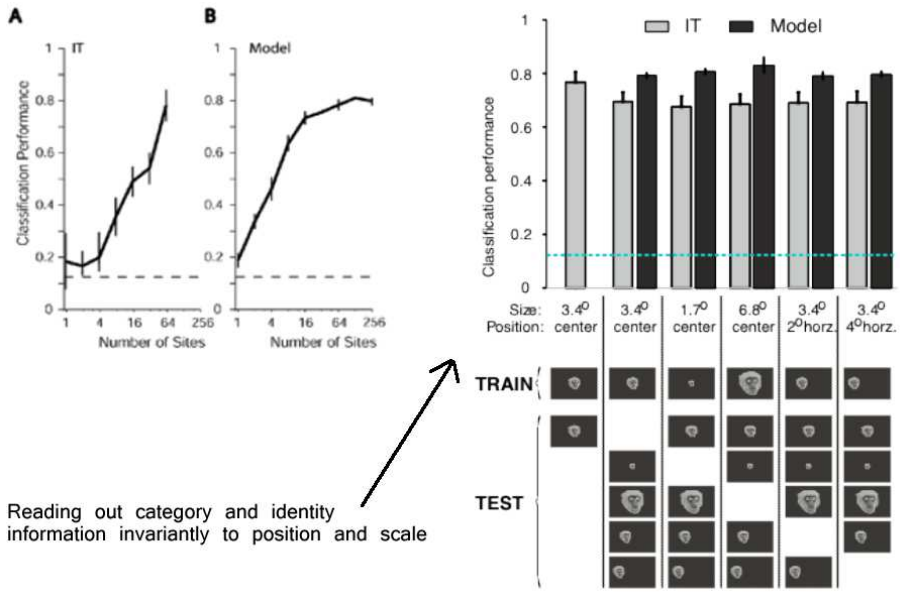


Figure 4: Comparison of position- and scale-invariant decoding of stimulus category from populations of IT and model units. Both generalize fairly well when tested on novel positions and scales (TEST), after training on objects of a specific position and scale (TRAIN). See [16, 29] for more detaisl. (Adapted from [16, 29])

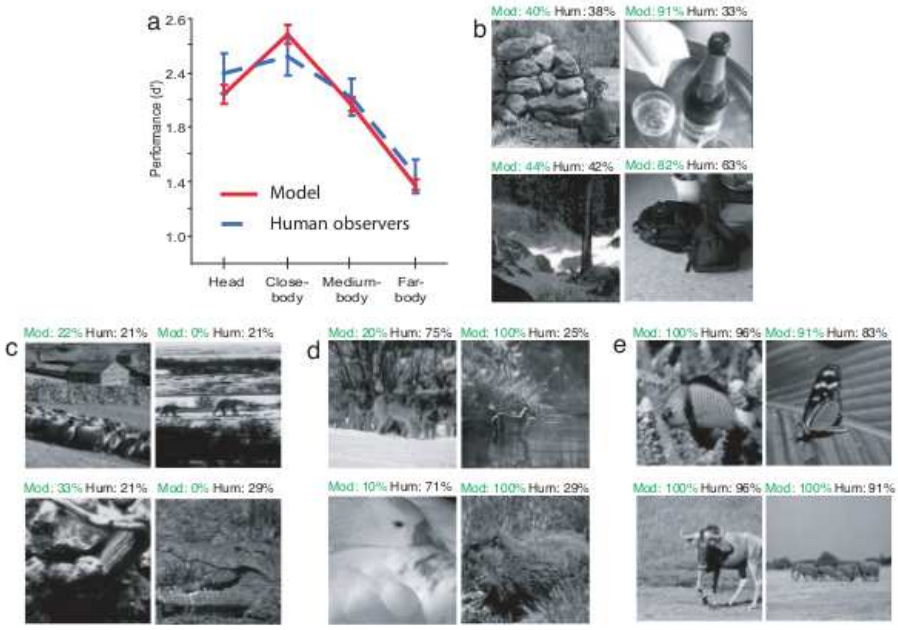


Figure 5: Comparison between the model and humans on the task of categorizing rapidly-presented scenes as either containing an animal or not. (a) The model and humans exhibit a similar pattern of performance. (b–e) Examples of classifications by the model and human observers. The percentages above each thumbnail correspond to the number of times the image was classified as containing an animal by the model or by humans. Common false alarms (b) and misses (c) for the model and human observers. (d, e) Examples of animal images for which the agreement between the model and human observers is poor (d) and good (e). Overall correlations between human and model classification are 0.71, 0.84, 0.71 and 0.60 for the “Head”, “Close-body”, “Medium-body” and “Far-body” images respectively. (Reprinted from [30]. Copyright 2007, National Academy of Sciences, USA)

model failed the task were also the most difficult for humans to discriminate – suggesting a deep correspondence between the model’s mechanisms and those implemented by human visual cortex (Fig. 5).

2.4 From Neuroscience Models to Engineering Applications

While H-W models clearly have a long way to go before they begin to approach human levels of broad visual intelligence, they have nonetheless shown success beyond the realm of neuroscience by proving useful for computer vision.

Recently, one H-W model based on the motion-processing dorsal stream of visual cortex has been found to match human performance on the task of recognizing and annotating mouse behaviors in video clips [17]. Many biological experiments dealing with genetically-modified strains of mice require

laborious human annotation of many hours of video in order to quantitatively analyze effects of the various genetic modifications performed in the quest for understanding diseases such as autism and Parkinson's disease. The neuromorphic model was developed into a trainable, general-purpose system capable of automatically analyzing complex mouse behaviors, performing on par with humans and outperforming a current commercial, non-neuromorphic system [5].

Furthermore, the adherence to neuromorphism has not significantly disadvantaged H-W models in terms of performance on standard computer vision datasets. At various points in time, these models have matched or surpassed state-of-the-art systems on datasets such as CalTech101 [24, 32] and Labeled-Faces-in-the-Wild (LFW) [26, 27]. These results reinforce the belief that ultimately, the quest to understand the key properties of biological intelligence will be essential in producing truly intelligent artificial systems.

3 What's Next in the Quest for Visual Intelligence?

Thus far, we have argued for the neuromorphic approach to tackling the challenge of achieving human-level visual intelligence, and then reviewed the successes of this approach up to this point. In this final section, we briefly describe some suggestions as to how computational neuroscience and computer vision should proceed from this point onwards.

3.1 Going Beyond "What is Where"

Much of computer vision research today is geared towards the "what" and "where" problems. Specifically, "what" problems include the detection and categorization of objects and actions, while "where" problems include localization, segmentation and tracking. However, as alluded to earlier, although determining "what is where" in an image is an important part of vision, visual intelligence is much more than that.

Take, for instance, the task of connecting an Ethernet cable to a laptop computer. Visually, this simple task consists of several sub-tasks, including finding potential locations on the surface of the laptop where the cable's connector might fit, as well as determining which orientation the connector should be held at. Certainly, these sub-tasks include traditional "what" and "where" problems (e.g. segmentation of the laptop; detecting and localizing a region that matches the connector shape), but the greater challenge is determining precisely what these sub-tasks should be.

3.2 From Perception to Abstraction

Another example of how visual intelligence goes beyond "what is where" involves abstract visual concepts. Take, for example, the concepts of "peaceful" or "crowded"; images can be classified (albeit somewhat subjectively) as being "peaceful" or not, and being "crowded" or not[3]. Yet, these abstract

[3]Determining the images for which these classification tasks do not make sense in the first place is also part of visual intelligence

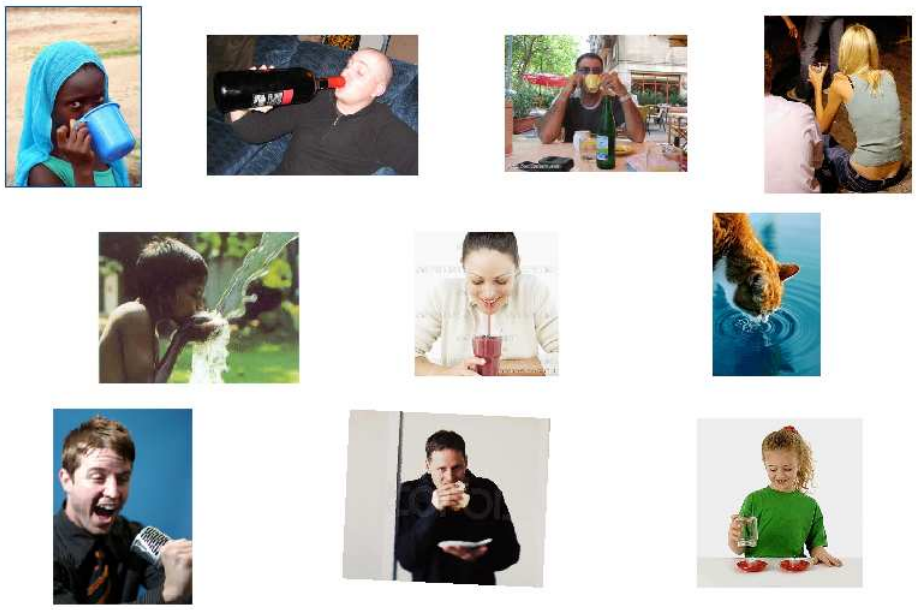


Figure 6: All of these images contain *drinking*. Credit: Shimon Ullman

notions are not so much about specific objects and their locations, but rather about properties of the image as a whole.

Even specific objects themselves have abstract properties. Every object has a physical size in terms of pixels, but the property of *conceptual size* (i.e. "largeness") depends on several factors, including *inferred size* (by determining the corresponding real-world scale of the depicted scene, if applicable), as well as comparisons to typical sizes of other instances from the same object category.

People are also good at identifying somewhat abstract actions from single static images. Shimon Ullman gives the example of identifying *drinking* (see Fig. 6). Current systems can probably recognize that these images contain liquids, cups, glasses, bottles, hands, mouths, etc. However, such systems do not encode the higher level knowledge that one form of drinking consists of using a cup to bring liquid towards a mouth, or that alternative forms of drinking are also valid.

3.3 A Loose Hierarchy of Visual Tasks

The "what" and "where" problems can be considered to be predominantly perceptual. It is clear that visual intelligence spans a variety of problems, ranging from these perceptual ones, to the more abstract or cognitive ones described earlier. It may be useful to conceptualize or organize the challenge of visual intelligence into a loose hierarchy of visual tasks, with the "easy" perceptual tasks at the bottom, and

the most abstract tasks at the top[4]. If the Turing Test [35] is a yardstick for general human intelligence, then this collection of visual tasks can be considered to be a measure of visual intelligence – a ***Visual Turing Test***, so to speak.

Such a big-picture view of things may be important for guiding research, particularly if the goal is to create a single system that can accomplish all the tasks. For instance, focusing on a single task such as object categorization may lead to certain algorithms or approaches being considered state-of-the-art. However, such algorithms, by virtue of being extremely good at one thing, could inadvertently turn out to be poorly-suited for other tasks – falling into a local minimum, in some sense[5]. By laying out the bigger picture of tasks to be performed, research efforts are more likely to be directed towards approaches that are more general.

3.4 It's Time to Try Again – the MIT Intelligence Initiative

Finally, we conclude this chapter by describing efforts in our lab and others at MIT to once again attempt to tackle the problem of intelligence, visual and otherwise – the MIT Intelligence Initiative [7].

The problem of intelligence – the nature of it, how the brain generates it and how it could be replicated in machines – is arguably one of the deepest and most important problems in science today. Philosophers have studied intelligence for centuries, but it is only in the last several decades that key developments in a broad range of science and engineering fields have opened up a thriving "intelligence research" enterprise, making questions such as these approachable: How does the mind process sensory information to produce intelligent behavior – and how can we design intelligent computer algorithms that behave similarly? What is the structure and form of human knowledge – how is it stored, represented and organized?

Many of us at MIT believe that the time has come for a new, fresh attack on these problems. The launching off point will be a new integration of the fields of cognitive science, which studies the mind, neuroscience, which studies the brain, and computer science and artificial intelligence, which develop intelligent hardware and software. These fields grew up together in the 1950s, but drifted apart as each became more specialized. In the 21st century, they are re-converging as a result of new advances that allow studies of the brain and mind to inform the design of intelligent artifacts and vice versa. Hence, for the original, daring goals of understanding and replicating intelligence, perhaps it is time to try again.

4 Acknowledgments

The authors wish to thank the members of the Center for Biological and Computational Learning (CBCL), including former members whose work and ideas have contributed immensely to this chapter. This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL).

[4] The idea being that, roughly speaking, only the more "intelligent" systems should be able to perform tasks near the top of the hierarchy

[5] These algorithms are of course still very valuable for solving specific problems.

This research was sponsored by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05-C-7262). Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation.

References

[1] A pedestrian detection system that stops a car automatically. URL http://articles.economictimes.indiatimes.com/2011-02-27/news/28638493_1_detection-system-volvo-collision-warning-system

[2] Caltech 101. URL http://www.vision.caltech.edu/Image_Datasets/Caltech101/

[3] DARPA Grand Challenge. URL http://en.wikipedia.org/wiki/DARPA_Grand_Challenge

[4] Digital Camera Face Recognition: How It Works. URL http://www.popularmechanics.com/technology/how-to/4218937

[5] HomeCageScan 2.0. URL http://www.cleversysinc.com/products/software/homecagescan

[6] Night View Assist: How night becomes day. URL http://www.daimler.com/dccom/0-5-1210218-1-1210320-1-0-0-1210228-0-0-135-7165-0-0-0-0-0-0-0.html

[7] The MIT Intelligence Initiative. URL http://isquared.mit.edu/

[8] The PASCAL Visual Object Classes Homepage. URL http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[9] USPS Awards Parascript Contract for OCR to Support Automated Parcel Bundle Sorting at USPS Facilities throughout the United States. URL http://money.msn.com/business-news/article.aspx?feed=PR&Date=20110601&ID=13713512/

[10] Amit, Y., Mascaro, M.: An integrated network for invariant visual detection and recognition. Vision Research **43**(19), 2073–2088 (2003). DOI 10.1016/S0042-6989(03)00306-7. URL http://dx.doi.org/10.1016/S0042-6989(03)00306-7

[11] Anzai, A., Peng, X., Essen, D.V.: Neurons in monkey visual area V2 encode combinations of orientations. Nature Neuroscience **10**(10), 1313–1321 (2007). URL http://www.nature.com/neuro/journal/vaop/ncurrent/full/nn1975.html

[12] Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., Poggio, T.: A model of V4 shape selectivity and invariance. Journal of Neurophysiology **98**(3), 1733 (2007). URL http://jn.physiology.org/content/98/3/1733.short

[13] Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36**(4), 193–202 (1980). DOI 10.1007/BF00344251. URL http://www.springerlink.com/content/r6g5w3tt54528137

[14] Gawne, T.J., Martin, J.M.: Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. Journal of Neurophysiology **88**(3), 1128 (2002). URL http://jn.physiology.org/content/88/3/1128.short

[15] Hubel, D., Wiesel, T.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology **160**(1), 106 (1962). URL http://jp.physoc.org/content/160/1/106.full.pdf

[16] Hung, C.P., Kreiman, G., Poggio, T., DiCarlo, J.J.: Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. Science **310**(5749), 863–866 (2005). DOI 10.1126/science.1117593. URL http://www.sciencemag.org/cgi/content/abstract/310/5749/863

[17] Jhuang, H., Garrote, E., Yu, X., Khilnani, V., Poggio, T., Steele, A., Serre, T.: Automated home-cage behavioural phenotyping of mice. Nature Communications **1**(6), 1–9 (2010). URL http://www.nature.com/ncomms/journal/v1/n6/abs/ncomms1064.html

[18] Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. International Conference on Computer Vision (ICCV) **11**, 1–8 (2007). URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4408988

[19] Keysers, C., Xiao, D., Földiák, P., Perrett, D.: The speed of sight. Journal of Cognitive Neuroscience **13**(1), 90–101 (2001). URL http://www.mitpressjournals.org/doi/abs/10.1162/089892901564199

[20] Lampl, I., Ferster, D.: Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. Journal of Neurophysiology **92**(5), 2704 (2004). URL http://jn.physiology.org/content/92/5/2704.short

[21] Li, F., VanRullen, R., Koch, C., Perona, P.: Rapid natural scene categorization in the near absence of attention. Proceedings of the National Academy of Sciences of the United States of America **99**(14), 9596 (2002). URL http://www.pnas.org/content/99/14/9596.short

[22] Mel, B.W.: SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally Inspired Approach to Visual Object Recognition. Neural Computation **9**(4), 777–804 (1997). DOI 10.1162/neco.1997.9.4.777. URL http://dx.doi.org/10.1162/neco.1997.9.4.777 http://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.4.777

[23] Mishkin, M., Ungerleider, L.G., Macko, K.A.: Object vision and spatial vision: Two cortical pathways. Trends in Neurosciences **6**, 414–417 (1983)

[24] Mutch, J., Lowe, D.: Multiclass Object Recognition with Sparse, Localized Features. In: 2006 IEEE Conference on Computer Vision and Pattern Recognition, pp. 11–18. IEEE (2006). DOI 10.1109/CVPR.2006.200. URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1640736

[25] Perrett, D., Oram, M.: Neurophysiology of shape processing. Image and Vision Computing **11**(6), 317–333 (1993). URL http://linkinghub.elsevier.com/retrieve/pii/0262885693900115

[26] Pinto, N., DiCarlo, J.J., Cox, D.D.: Establishing Good Benchmarks and Baselines for Face Recognition. In: IEEE European Conference on Computer Vision, Faces in 'Real-Life' Images Workshop (2008). URL http://hal.archives-ouvertes.fr/inria-00326732/

[27] Pinto, N., DiCarlo, J.J., Cox, D.D.: How far can you get with a modern face recognition test set using only simple features? In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2591–2598. IEEE (2009). DOI 10.1109/CVPR.2009.5206605. URL http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5206605

[28] Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience **2**(11), 1019–1025 (1999). DOI 10.1038/14819

[29] Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., Poggio, T.: A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex. CBCL Paper #259/AI Memo #2005-036 (2005). URL http://en.scientificcommons.org/21119952

[30] Serre, T., Oliva, A., Poggio, T.: A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America **104**(15), 6424–6429 (2007). URL http://cat.inist.fr/?aModele=afficheN&cpsidt=18713198

[31] Serre, T., Poggio, T.: A neuromorphic approach to computer vision. Communications of the ACM **53**(10), 54–61 (2010). URL http://portal.acm.org/citation.cfm?id=1831425

[32] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust Object Recognition with Cortex-Like Mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 411–426 (2007). URL http://portal.acm.org/citation.cfm?id=1263421&dl=

[33] Thorpe, S., Fabre-Thorpe, M.: Seeking categories in the brain. Science **291**(5502), 260 (2001). URL http://www.sciencemag.org/content/291/5502/260.short

[34] Thorpe, S., Fize, D., Marlot, C.: Speed of processing in the human visual system. Nature **381**(6582), 520–2 (1996). DOI 10.1038/381520a0. URL http://www.ncbi.nlm.nih.gov/pubmed/8632824

[35] Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950)

[36] VanRullen, R., Koch, C.: Visual selective behavior can be triggered by a feed-forward process. Journal of Cognitive Neuroscience **15**(2), 209–217 (2003). URL http://www.mitpressjournals.org/doi/abs/10.1162/089892903321208141

[37] Wallis, G., Rolls, E.T.: A model of invariant object recognition in the visual system. Progress in Neurobiology **51**, 167–194 (1997). URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.880&rep=rep1&type=pdf

[38] Wallisch, P., Movshon, J.: Structure and Function Come Unglued in the Visual Cortex. Neuron **60**(2), 195–197 (2008). URL http://linkinghub.elsevier.com/retrieve/pii/s0896-6273%2808%2900851-9

[39] Wersing, H., Körner, E.: Learning optimized features for hierarchical models of invariant object recognition. Neural Computation **15**(7), 1559–88 (2003). DOI 10.1162/089976603321891800. URL http://www.mitpressjournals.org/doi/abs/10.1162/089976603321891800