On the Prospects for Building a Working Model of the Visual Cortex

Thomas Dean
Brown University, Providence, RI
tld@cs.brown.edu

Glenn Carroll and **Richard Washington**
Google Inc., Mountain View, CA
{gcarroll,rwashington}@google.com

Abstract

Human-level visual performance has remained largely beyond the reach of engineered systems despite decades of research and significant advances in problem formulation, algorithms and computing power. We posit that significant progress can be made by combining existing technologies from machine vision, insights from theoretical neuroscience and large-scale distributed computing. Such claims have been made before and so it is quite reasonable to ask what are the new ideas we bring to the table that might make a difference this time around. From a theoretical standpoint, our primary point of departure from current practice is our reliance on exploiting time in order to turn an otherwise intractable unsupervised problem into a locally semi-supervised, and plausibly tractable, learning problem. From a pragmatic perspective, our system architecture follows what we know of cortical neuroanatomy and provides a solid foundation for scalable hierarchical inference. This combination of features provides the framework for implementing a wide range of robust object-recognition capabilities.

In July of 2005, one of us (Dean) presented a paper at AAAI entitled "A Computational Model of the Cerebral Cortex" (Dean 2005). The paper described a graphical model of the visual cortex inspired by David Mumford's computational architecture (1991; 1992; 2003). At that same meeting, Jeff Hawkins gave an invited talk entitled "From AI Winter to AI Spring: Can a New Theory of Neocortex Lead to Truly Intelligent Machines?" drawing on the content of his popular book *On Intelligence* (2004). A month later at IJCAI, Geoff Hinton gave his Research Excellence Award lecture entitled "What kind of a graphical model is the brain?"

In all three cases, the visual cortex is cast in terms of a generative model in which ensembles of neurons are modeled as latent variables. All three of the speakers were optimistic regarding the prospects for realizing useful, biologically-inspired systems. In the intervening two years, we have learned a great deal about both the provenance and the practical value of those ideas. The history of the most important of these is both interesting and helpful in understanding whether these ideas are likely to yield progress on some of the most significant challenges facing AI.

Are we poised to make a significant leap forward in understanding computer and biological vision? If so, what are the main ideas that will fuel this leap forward and are they new or recycled? How important is the role of Moore's law in our pursuit of human-level perception? The full story delves into the role of time, hierarchy, abstraction, complexity, symbolic reasoning, and unsupervised learning. It owes much to insights that have been discovered and forgotten at least once every other generation for many decades.

Since J.J. Gibson (1950) presented his theory of ecological optics, scientists have followed his lead by trying to explain perception in terms of the invariants that organisms learn. Peter Földiák (1991) and more recently Wiskott and Sejnowski (2002) suggest that we learn invariances from temporal input sequences by exploiting the fact that sensory input tends to vary quickly while the environment we wish to model changes gradually. Gestalt psychologists and psychophysicists have long studied spatial and temporal grouping of visual objects and the way in which the two operations interact (Kubovy & Gepshtein 2003), and there is certainly ample evidence to suggest concrete algorithms.

The idea of hierarchy plays a central role in so many disciplines it is fruitless to trace its origins. Even as Hubel and Weisel unraveled the first layers of the visual cortex, they couldn't help but posit a hierarchy of representations of increasing complexity (1968). However, direct evidence for such hierarchical organization is slim to this day and machine vision has yet to actually learn hierarchies of more than a couple layers despite compelling arguments for their utility (Ullman & Soloviev 1999).

Horace Barlow (1961) pointed out more than forty years ago that strategies for coding visual information should take advantage of the statistical regularities in natural images. This idea is the foundation for work on sparse representations in machine learning and computational neuroscience (Olshausen & Field 1996).

By the time information reaches the primary visual cortex (V1), it has already gone through several stages of processing in the retina and lateral geniculate. Following the lead of Hubel and Wiesel, many scientists believe that the output of V1 can be modeled as a tuned band-pass filter bank. The component features or *basis* for this representation are called *Gabor filters* — mathematically, a Gabor filter is a two-dimensional Gaussian modulated by a complex sinu-

soid — and are tuned to respond to oriented dark bars against a light background (or, alternatively, light bars against a dark background). The story of why scientists came to this conclusion is fascinating, but the conclusion may have been premature; more recent work suggests Gabor filters account for only about 20% of the variance observed in the output of V1 (Olshausen & Field 2005).

We're sure that temporal and spatial invariants, hierarchy, levels of increasingly abstract features, unsupervised learning of image statistics and other core ideas that have been floating around for decades must be part of the answer, but machine vision still falls far short of human capability in most respects. Are we on the threshold of a breakthrough and if so what will push us through the final barriers?

Temporal and Hierarchical Structure

The primate cortex serves many functions and most scientists would agree that we've discovered only a small fraction of its secrets. Let's suppose that our goal is to build a computational model of the ventral visual pathway, the neural circuitry that appears to be largely responsible for recognizing *what* objects are present in our visual field. A successful model would, among other things, allow us to create a video search platform with the same quality and scope that Google and Yahoo! provide for web pages. Do we have the pieces in place to succeed in the next two to five years?

In many areas of science and engineering, time is so integral to the description of the central problem that it can't be ignored. Certainly this is the case in speech understanding, automated control, and most areas of signal processing. By contrast, in most areas of machine vision time has been considered a distraction, a complicating factor that we can safely ignore until we've figured out how to interpret static images. The prevailing wisdom is that time will only make the problem of recognizing objects and understanding scenes more difficult. A similar assumption has influenced much of the work in computational neuroscience, but now that assumption is being challenged. There are a number of proposals that suggest time is an essential ingredient in explaining human perception (Földiák 1991; Dean 2006; Hawkins & George 2006; Wiskott & Sejnowski 2002). The common theme uniting these proposals is that the perceptual sequences we experience provide essential cues that we exploit to make critical discriminations. Consecutive samples in an audio recording or frames in a video sequence are likely to be examples of the same pattern undergoing changes in illumination, position, orientation, etc. The examples provide exactly the variation required to train models able to recognize patterns invariant with respect to the observed transformations.

There is an even more compelling explanation when it comes to learning hierarchies of spatial and temporal features. Everyone agrees, despite a lack of direct evidence, that the power of the primate cortex to learn useful representations owes a great deal to its facility in organizing concepts in hierarchies. Hierarchy is used to explain the richness of language and our extraordinary ability to quickly learn new concepts from only a few training examples.

It seems likely that learning such a hierarchical representation from examples (input and and output pairs) is at least as hard as learning polynomial-size circuits in which the subconcepts are represented as bounded-input boolean functions. Kearns and Valiant showed that the problem of learning polynomial-size circuits (in Valiant's *probably approximately correct* learning model) is infeasible given plausible cryptographic limitations (1989).

However, if we are provided access to the inputs and outputs of the circuit's internal subconcepts, then the problem becomes tractable (Rivest & Sloan 1994). This implies that if we had the "circuit diagram" for the visual cortex and could obtain labeled data, inputs and outputs, for each component feature, robust machine vision might become feasible. Instead of labels, we have knowledge — based on millennia of experience summarized in our genes — enabling us to transform an otherwise intractable unsupervised learning problem (one in which the training data is unlabeled) into a tractable semi-supervised problem (one in which we can assume that consecutive samples in time series are more likely than not to have the same label).

This property called *temporal coherence* is the basis for the optimism of several researchers that they can succeed where others have failed.[1] If we're going to make progress, temporal coherence has to provide some significant leverage. It is important to realize, however, that exploiting temporal coherence does not completely eliminate complexity in learning hierarchical representations. Knowing that consecutive samples are likely to have the same label is helpful, but we are still left with the task of segmenting the time series into subsequences having the same label, a problem related to learning HMMs (Freund & Ron 1995).

Learning Invariant Features

It's hard to come up with a trick that nature hasn't already discovered. And, while nature is reluctant to reveal its tricks, decades of machine vision researchers have come up with their own. Whether or not we have found neural analogs for the most powerful of these, a pragmatic attitude dictates adapting and adopting them, where possible. David Lowe (Lowe 2004) has developed an effective algorithm for extracting image features called SIFT (for scale invariant feature transform). The algorithm involves searching in scale space for features in the form of small image patches that can be reliably recovered in novel images. These features are used like words in a dictionary to categorize images, and each image is summarized as an unordered collection of such *picture* words.

The basic idea of using an unordered collection of features to support invariance has been around for some time. It even has a fairly convincing story in support of its biological

[1] Or, at least, it is one half of the basis for betting on the success of this approach, and the half on which we have concentrated. The other half relies on the fact that machines, like humans, can intervene in the world to resolve ambiguity and distinguish cause from correlation. Intervention can be as simple as exploiting parallax to resolve accidental alignments between distinct but parallel lines.

plausibility (Riesenhuber & Poggio 1999).[2] However, balancing invariance (which encourages false positives) against selectivity (which encourages false negatives) requires considerable care to get right. For instance, one approach argues that overlapping features which correspond to the receptive fields of cortical neurons avoid false positives (Ullman & Soloviev 1999); another approach provides greater selectivity by taking into account geometric relationships among the picture words (Sudderth *et al.* 2005).

Lowe's trick of searching in scale space has an analog in finding spatiotemporal features in video (Laptev & Lindeberg 2003). Finding corresponding points in consecutive frames of a video is relatively easy for points associated with patches that stand out from their surrounding context. We can exploit this fact to track patches across multiple frames. This method identifies features that are persistent in time. In addition, it learns to account for natural variation in otherwise distinctive features. There are cells in the retina, lateral geniculate and primary visual cortex whose receptive fields span space and time and are capable, in theory, of performing this sort of tracking (Dayan & Abbott 2001). Exactly how these spatiotemporal receptors are used to extract shape, distinguish figure from ground and infer movement, is unknown, but clearly here is another place where time plays a key role in human perception.

Why the ventral visual pathway?

Our focus is on V1 and the ventral visual pathway, that part of V2 responsible for identifying *what* is in the visual field. The motivation is that this seems to be the sweet spot for relatively uncontested knowledge concerning brain function and understanding of what is being represented and how it is being computed. By way of contrast, the dorsal visual pathway — responsible for tracking the location and motion of objects in our visual field (Goodale & Westwood 2004) — is less well-studied and presents a more complicated picture. As an example, positional information appears to be affected by motion (Whitney *et al.* 2003) and determined relative to a primary object (Chafee, Averbeck, & Crowe 2007). Saccades and head movements tend to change the point of view and must somehow be integrated with the retinotopically mapped information flowing through the lateral geniculate nuclei. Both of these present us with problems in how to integrate disparate information sources, including mixing retinotopic with non-retinotopic information, a challenging problem for which we presently lack any clear solution (see Rolls and Stringer (2007) for an interesting start).

While we believe V1 and the ventral pathway provide useful clues for developing artificial perceptual systems, we are neither so naïve nor so ignorant as to be unaware that the brain still holds many secrets, and our model does not even account for all the currently extant data. As mentioned above, Olshausen and Field (Olshausen & Field 2005) give the somewhat pessimistic estimate that we presently understand only about 20% of V1's functional behavior. We simply don't know what the other 80% of the computation is, whether it is important, or what it might be useful for. On the other side of the coin, we know that attention plays a significant role in visual perception (Reynolds & Chelazzi 2004), but we are assuming we can make progress without detailed understanding of human attentional mechanisms. A similar situation pertains at the level of neuroanatomy. Our model incorporates a version of feedforward and feedback connections, but does not presently include lateral connections (Lund, Angelucci, & Bressloff 2006). Again, we believe we can achieve some modicum of success without lateral connections, but we await experimental results before venturing a definitive answer. The take away on this is that we expect to adapt our model in response to shortcomings exposed by experimentation, but we are aware both of the gaps in our knowledge of the brain and the discrepancies between our model and what is presently known about the visual cortex.

Will big ideas or big iron win the race?

Compared with previous approaches, there is one other advantage we have allowing us to consider models of realistic scale: increased computing power and the where withal to take advantage of it. The human primary visual cortex (V1) consists of approximately 10^9 neurons and 6×10^9 connections (Colonnier & O'Kusky 1981). Whereas we have relatively poor data for modeling individual neurons, despite the press for the IBM / EPFL Blue Brain Project, we are better positioned with respect to the *aggregate* behavior of thousands of neurons. Most of the serious computational brain models aim at the level of a *cortical hyper-column*, a structural unit consisting of 10^4–10^5 neurons (Mountcastle 1998). If we assign one processor per hyper-column, a computing cluster with 10^3 processor cores and accompanying communications capacity can simulate on the order of 10^8 neurons. This would be about 10% of V1, and somewhat beyond the reach of most academic labs, but Google and several other industrial labs can field resources at this scale and beyond. Working at a smaller scale would risk confounding effects introduced by the scale with effects of the model itself. Working at a realistic scale allows us to focus on the model. Moreover, deploying resources at scale allows us to turn around experiments in minutes or hours, as opposed to days or weeks. This means we can iterate over alternatives, which will surely be necessary, and explore the space of solutions in a way not practical for an under-resourced system.

Simulating the brain to achieve human-level sensory and cognitive abilities is just starting to make the transition from tantalizing possibility to practical, engineered reality. Making that transition will take both good ideas – including venerable old ideas and some new ones – and heavy-duty computing power. While we believe the inclusion of time is an essential element of a solution, and our model offers promise of success, we are aware that more may be necessary. The extant data on brain function is notably sparse, and our model makes no attempt to take all known brain function into account, e.g., attentional mechanisms. What we do have is a plausible model — biologically inspired though certainly not biologically accurate — and the tools

[2] Serre et al (Serre *et al.* 2007) describe a system achieving state-of-the-art performance in object recognition by combining biological clues and techniques from machine vision.

to evaluate and improve it. While many of the ideas have been around for some time, the infrastructure to quickly and convincingly evaluate them has been lacking. Robust, high-performance distributed computing hardware and software doesn't make you smarter, but it does allow you to reach a little further and *that* could make the crucial difference.

References

Barlow, H. B. 1961. Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., ed., *Sensory Communication*. Cambridge, MA: MIT Press. 217–234.

Chafee, M. V.; Averbeck, B. B.; and Crowe, D. A. 2007. Representing spatial relationships in posterior parietal cortex: Single neurons code object-referenced position. *Cerebral Cortex*.

Colonnier, M., and O'Kusky, J. 1981. Number of neurons and synapses in the visual cortex of different species. *Revue Canadienne de Biologie* 40(1):91–99.

Dayan, P., and Abbott, L. F. 2001. *Theoretical Neuroscience*. Cambridge, MA: MIT Press.

Dean, T. 2005. A computational model of the cerebral cortex. In *Proceedings of AAAI-05*, 938–943. Cambridge, Massachusetts: MIT Press.

Dean, T. 2006. Learning invariant features using inertial priors. *Annals of Mathematics and Artificial Intelligence* 47(3-4):223–250.

Földiák, P. 1991. Learning invariance from transformation sequences. *Neural Computation* 3:194–200.

Freund, Y., and Ron, D. 1995. Learning to model sequences generated by switching distributions. In *COLT: Proceedings of the Workshop on Computational Learning Theory*, 41–50. Morgan Kaufmann Publishers.

Gibson, J. J. 1950. *Perception of the Visual World*. Boston: Houghton Mifflin.

Goodale, M. A., and Westwood, D. A. 2004. An evolving view of duplex vision: separate but interacting cortical pathways for perception and action. *Current Opinion in Neurobiology* 14:203–211.

Hawkins, J., and Blakeslee, S. 2004. *On Intelligence*. New York: Henry Holt and Company.

Hawkins, J., and George, D. 2006. Hierarchical temporal memory: Concepts, theory and terminology.

Hubel, D. H., and Wiesel, T. N. 1968. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology* 195:215–243.

Kearns, M., and Valiant, L. G. 1989. Cryptographic limitations on learning boolean functions and finite automata. In *Proceedings of the Twenty First Annual ACM Symposium on Theoretical Computing*, 433–444.

Kubovy, M., and Gepshtein, S. 2003. Perceptual grouping in space and in space-time: An exercise in phenomenological psychophysics. In Behrmann, M.; Kimchi, R.; and Olson, C. R., eds., *Perceptual Organization in Vision: Behavioral and Neural Perspectives*. Mahwah, NJ: Lawrence Erlbaum. 45–85.

Laptev, I., and Lindeberg, T. 2003. Space-time interest points. In *Proceedings of the ninth IEEE International Conference on Computer Vision*, volume 1, 432–439.

Lee, T. S., and Mumford, D. 2003. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America* 2(7):1434–1448.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2):91–110.

Lund, J. S.; Angelucci, A.; and Bressloff, P. C. 2006. The contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Prog. Brain Ressearch* 154:93–121.

Mountcastle, V. B. 1998. *Perceptual Neuroscience: The Cerebral Cortex*. Cambridge, MA: Harvard University Press.

Mumford, D. 1991. On the computational architecture of the neocortex I: The role of the thalamo-cortical loop. *Biological Cybernetics* 65:135–145.

Mumford, D. 1992. On the computational architecture of the neocortex II: The role of cortico-cortical loops. *Biological Cybernetics* 66:241–251.

Olshausen, B. A., and Field, D. J. 1996. Natural image statistics and efficient coding. *Computation in Neural Systems* 7(2):333–339.

Olshausen, B. A., and Field, D. J. 2005. How close are we to understanding V1? *Neural Computation* 17:1665–1699.

Reynolds, J. H., and Chelazzi, L. 2004. Attentional modulation of visual processing. *Annual Review of Neuroscience* 27:611–647.

Riesenhuber, M., and Poggio, T. 1999. Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2(11):1019–1025.

Rivest, R. L., and Sloan, R. 1994. A formal model of hierarchical concept learning. *Information and Computation* 114(1):88–114.

Rolls, E. T., and Stringer, S. M. 2007. Invariant global motion recognition in the dorsal visual system: a unifying theory. *Neural Computation* 19:139–169.

Serre, T.; Wolf, L.; Bileschi, S.; Riesenhuber, M.; and Poggio, T. 2007. Object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(3):411–426.

Sudderth, E. B.; Torralba, A. B.; Freeman, W. T.; and Willsky, A. S. 2005. Describing visual scenes using transformed dirichlet processes. In *Advances in Neural Information Processing Systems 18*.

Ullman, S., and Soloviev, S. 1999. Computation of pattern invariance in brain-like structures. *Neural Networks* 12:1021–1036.

Whitney, D.; Goltz, H. C.; Thomas, C. G.; Gati, J. S.; Menon, R. S.; and Goodale, M. A. 2003. Flexible retinotopy: Motion-dependent position coding in the visual cortex. *Science* 302(5646):878–881.

Wiskott, L., and Sejnowski, T. 2002. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation* 14(4):715–770.