Unsupervised statistical learning in vision: computational principles, biological evidence*

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853, USA
http://kybele.psych.cornell.edu/∼edelman

Nathan Intrator
Institute for Brain and Neural Systems
Box 1843, Brown University
Providence, RI 02912, USA†
http://www.math.tau.ac.il/∼nin/

April 23, 2004

Abstract

Unsupervised statistical learning is the standard setting for the development of the only advanced visual system that is both highly sophisticated and versatile, and extensively studied: that of monkeys and humans. In this extended abstract, we invoke philosophical observations, computational arguments, behavioral data and neurobiological findings to explain why computer vision researchers should care about (1) unsupervised learning, (2) statistical inference, and (3) the visual brain. We then outline a neuromorphic approach to structural primitive learning motivated by these considerations, survey a range of neurobiological findings and behavioral data consistent with it, and conclude by mentioning some of the more challenging directions for future research.

1 Why computer vision should care about unsupervised learning

As the goals of computer vision grow more ambitious, the importance of learning becomes more difficult to deny: nobody wants to have to enter object and scene representations into his or her system by hand. But why should we insist that such learning be, in the first instance, unsupervised?

Because we should not trust our analytical intuitions about the ontology of visual objects. Although the increasing availability of annotated image databases encourages the development of highly sophisticated supervised learning methods that combine linguistic and visual information (Duygulu et al., 2002), the success of such methods is limited by the poverty of the annotations, usually lexical labels. Indeed, a label (such as {cat, forest, grass, tiger}, shown in the work just cited) attached to a picture is both ontologically deficient in that it leaves out a host of possible complementary or alternative labels (Akins, 1996; Smith, 2001), and descriptively deficient in that it falls far short of providing a listener with a clear notion of the scene depicted (Kitcher and Varzi, 2000; Edelman, 2002). Thus, language conceived as a source of supervision for visual learning introduces biases that at best may be irrelevant, and at worst pernicious for any attempt to understand visual perception of objects and scenes. It would be prudent therefore to distrust

*This is an extended abstract of a manuscript in preparation. Address correspondence to the first author at *se37@cornell.edu*.
†On leave from the School of Computer Science, Tel-Aviv University, Tel Aviv 69678, Israel.

it as a source of information for visual learning. This might be just as well: after all, monkeys, whose visual system is so similar to ours, manage to learn to see as well as we do, without resort to anything like language.

Because the situations in which children learn to connect vision to language involve very little, if any, supervision. Despite the dangers inherent in importing linguistic categories into vision, people clearly do so, eventually: an adult's ability to talk about what he or she sees is considerable, at least for artificial scenes and other such structured stimuli.[1] Indeed, learning to talk about the visual world is a key component of cognitive development, which is assessed by tests such as CASL (Carrow-Woolfolk, 1999) that include image-sentence matching tasks. Thus, automatic annotation of images is an appealing goal for computer vision. We should remember, however, that children, at any stage in their development, receive very little explicit supervision. In fact, the so-called whole-object constraint — the two-year old's assumption that a just-heard novel utterance refers to the most salient object in sight (Landau et al., 1988; Markman, 1989) — seems to have been put in place by evolution to make up precisely for this paucity of explicit supervision.[2]

Thus, largely unsupervised learning is the standard setting for human cognitive development. In early visual development, the goals of such learning can be seen as varieties of segmentation:

- *spatial (where to segment):* breaking down scenes into objects, and objects into fragments, that recur in multiple contexts and can be reused; cf. (Edelman et al., 2002b; Edelman and Intrator, 2003);
- *temporal (when to segment):* breaking down sequences of scenes into persistent objects or otherwise capturing their time structure; cf. (Stone, 1996; Cohen and Oates, 1998; Galata et al., 2001).

The remainder of this brief overview focuses on the first issue: learning where to segment.

2 Why computer vision should care about statistics

The philosophical (ontological) and psychological (developmental) considerations just discussed support a radical Empiricist stance, according to which object representations and the features they rely on should be learned, initially in an unsupervised fashion.[3] It is not surprising, therefore, that in search for a conceptual framework for such learning we should turn to the Empiricist philosopher David Hume:

> "All kinds of reasoning consist in nothing but a comparison, and a discovery of those relations, either constant or inconstant, which two or more objects bear to each other." [(Hume, 1740), Part III, Sect. II]
>
> "An experiment loses of its force, when transferr'd to instances, which are not exactly resembling; tho' 'tis evident it may still retain as much as may be the foundation of probability, as long as there is any resemblance remaining." [Part II, Sect. XII]
>
> "…all knowledge resolves itself into probability…" [Part IV, Sect. I]

[1] Natural outdoor scenes are, as poets well know, more likely than not to be ineffable, because the natural languages are so impoverished, ontologically and descriptively, compared to the visual world (Edelman, 2002).

[2] Conceptual background supplied by a story can help older pre-schoolers (Booth and Waxman, 2002), but this kind of interaction with an adult is an exception, not the rule, for most children in most cultures.

[3] The dispreferred alternative is the Rationalist stance, according to which the features are preordained in the form of some alphabet and the object representations are composed from them. This conceptual seepage from linguistics has led to many dead ends in computer vision, such as the syntactic pattern recognition (Fu, 1976) and "geon"-based representations (Dickinson et al., 1997). Rationalism has not fared any better in linguistics either; see, e.g., (Culicover, 1999; Edelman and Christiansen, 2003; Tomasello, 2003; Postal, 2004).

Hume's realization of the central and crucial role of *statistical inference* in knowledge generation (that is, learning) has been developed by many others, including his contemporary Thomas Bayes, the pioneering statisticians Karl Pearson and Ronald A. Fisher, and the neurobiologist Horace B. Barlow. Their combined insights led to the modern applications of inference to vision and other senses (Barlow, 1990; Knill and Richards, 1996).

The conception of visual learning as inference is naturally complemented by the emerging view of perception as statistical *decision making*, stated cogently in the following passage by the originator of the ecological theory of perception, the psychologist J. J. Gibson:

> "...the percept is always a wager. Thus uncertainty enters at *two* levels, not merely one: the configuration may or may not indicate an object, and the cue may or may not be utilized at its true indicative value." [from (Gibson, 1957)]

To put this insight to work, one needs to decide how to represent uncertainty. The arguments for unsupervised learning stated above suggest that that imposing a low-dimensional (i.e., tractable) parametric model by fiat is not prudent, as it is difficult to "guess" what the best model is. Instead, we should seek fundamental rules for dealing with the data at hand, based on some initial information extracted from data in an unsupervised manner. This achieves the necessary dimensionality reduction through the extraction of features, which are essential for later use in detection or recognition tasks.

A simple and intuitive exposition of the kind of statistical reasoning that is directly relevant to feature extraction in vision is found in Barlow's notion of *suspicious coincidence*: two events (e.g., visual features), A and B, should be lumped together and treated as a unit if their probability of their joint occurrence is much higher than the product of their individual probabilities, $P(A, B) \gg P(A)P(B)$ (Barlow, 1959; Barlow, 1989; Barlow, 1990). A modified version of this approach, which is related to Bayesian inference and to Minimum Description Length learning (Rissanen, 1987; Geman, 1996), is outlined later, in section 4.

3 Why computer vision should care about the brain

Paraphrasing the opening line of *Anna Karenina*,[4] successful perceptual systems — of which those embedded in living organisms are invariably good examples — are all alike in their adoption of the same general principles of biological information processing. Some of these principles are outlined next.

3.1 What you will find in the brain

Anatomically, in all the visual areas, as in the entire neocortex in general, information is processed by the same few kinds of cells, arranged in the same laminar/columnar structure (Braitenberg, 1977; Gilbert, 1988). The uniformity of the cortex is not limited to its anatomy: functional studies reveal a limited repertoire of computational mechanisms, of which tuned **receptive fields** (RFs) are probably the most ubiquitous one. In neurophysiology, the RF of a cell is defined as the part of the visual field in which a stimulus must appear to elicit a response from the cell (Kuffler and Nicholls, 1976). Together with the specification of the preferred stimulus of the cell, this constitutes a useful first approximation of its input-related function. For a complete characterization of the cell's function, its context sensitivity (induced by lateral and descending links) and its projective field should also be specified. The characteristics of the receptive fields of cortical cells and their interconnection patterns (such as the **map-like projections** between cortical areas) constrain

[4] "Happy families are all alike; every unhappy family is unhappy in its own way."

the kind of information processing that can be supported by the cortex (Edelman, 1995b; Edelman, 1995a; Phillips and Singer, 1997; Edelman, 1999). A tutorial describing possible uses of the distributed RF-based representations found in biological information processing systems appears in (Pouget et al., 2000).

3.2 What you won't find in the brain

Most of the favorite representational primitives and computational tools employed in computer vision are not to be found in any brain. Some examples are:

- for the input representation: no distinct, individually addressable pixels;
- for the output representation: no inherently sharp boundaries between regions (although an *ad hoc* boundary may emerge if a decision based on a hard criterion is forced upon the system);
- for the computation in general: no random-access memory and no dynamic binding of values to variables (except in situations where humans must resort to scrutiny, such as explicit reasoning of the kind found in IQ tests);[5]
- for the statistical computation in particular: no distinct bins and no precise integer operations; in particular, if a neurally implemented counter is incremented, its neighbors would receive a boost too, all by variable quantities that include a certain amount of noise.

Contrary to a possible hasty conclusion, all this is actually good for you: the limitations imposed by biology should prevent the designer of a visual system from violating the Principle of Least Committment (Marr, 1976). In the present context, such a violation would result from resorting too early to discrete, categorically labeled representations, in which important distributed, graded information about the stimulus is irretrievably lost (Edelman, 1999).

4 Unsupervised statistical learning in vision: putting the theory to work

The purpose of this extended abstract is to pull together many different strands of thinking about perceptual learning (and to gather in one place many of the relevant literature sources). Accordingly, we devote the remainder of the available space to outlining the principles on which we believe work on visual learning should be based, rather than describing a specific project that implements such principles.[6]

4.1 The principles

The world as its own representation. It took the computer vision community several decades to realize that attempting to reconstruct the visual world in the form of a detailed general-purpose 3D geometrical representation is a futile undertaking (Barrow and Tenenbaum, 1993; Edelman, 1999). The efforts expended along the way could have been saved if the following early observation were heeded:

> "The primary function of perception is to keep our internal framework in good registration with that vast external memory, the external environment itself." [*Machine perception: what makes it so hard for computers to see?,* (Reitman et al., 1978), p.72]

[5] Note that this restriction on the use of dynamic λ-binding has serious implications for any model of cognitive function that has the power of the Turing Machine; see (Edelman and Intrator, 2003) for a discussion of this issue.

[6] See (Edelman, 1999; Edelman et al., 2002b; Edelman and Intrator, 2003) for such details.

Using the world as its own representation (O'Regan, 1992) and extracting information from it, as necessary, through the action of feature detectors (Barlow, 1979) based on the notion of receptive fields (Edelman, 1995a; Edelman, 1999), is a biologically inspired approach to vision in general that is also computationally tractable.

Active vision. A system that relies on the external world to supply information on demand must be capable of task-dependent gaze/attention control, as suggested by the proponents of active vision (Aloimonos et al., 1988; Edelman, 1995b). Crucially, this design choice obviates the need for an absolute translation invariance of the representations: if memory is not an issue, it is cheap enough (and much more feasible) to recruit many location-specific representations than to construct a single invariant one.

A generative learning model. The considerations expressed in the preceding sections suggest that the generative goal, aiming at capturing the probability distribution relevant to the task at hand — specifically, the joint probability $P(x_1, x_2, \ldots, x_n)$ of all the relevant measurement variables $\{x_i\}$ — is more appropriate for the initial stages of unsupervised perceptual learning than classification. This is because the labels needed for the latter may not be relied upon or not available at all, and because the very distinction between different classes may not be valid.[7] Note that the ultimate goal of any learning system — capturing the joint probability of distal events in the world — can only be approximated by working with what any such system must work, namely the proximal measurements, or features.[8] A realistic formulation of the learning problem must therefore include the search for the best features (Intrator, 1992; Intrator, 1993; Intrator and Gold, 1993) alongside the search for the best approximation of their joint probability.

Sparse coding. Unsupervised statistical learning is beset by a paradox: statistics can only be computed over a set of candidate primitive descriptors if these are identified in advance, yet the identification of the candidates requires prior statistical data (Barlow, 2001; Edelman et al., 2002b). To circumvent this issue, one may choose to learn direct (that is, localized rather than distributed) representations. This can be done by enforcing sparsity, a condition in which only a few detectors (out of the very many existing ones) fire for any given stimulus (Gardner-Medwin and Barlow, 2001). In a way, this amounts to making redundancy of the stimulus set explicit, rather than trying to reduce it, as would be the case for a highly distributed representation in which each unit is utilized to the maximum possible degree and made to participate in the coding of as many stimuli as possible (Barlow, 2001). The redundancy that goes along with sparsity facilitates the creation of new coincidence detectors on the fly for specific tasks.

4.2 An outline of the resulting approach

In the light of the preceding discussion, the emerging solution to the problem of unsupervised learning in vision is to bootstrap the system by allowing very fast learning of direct, sparse representations that can serve as the basis for computing statistics (Edelman et al., 2002b). Distributed patterns of activations over such a basis can then support the processing of many more stimuli than those represented directly (Edelman and Intrator, 2003). To make a better use of memory, the system may engage in continued statistically driven recycling of less useful representational units and allocation of new ones.

[7] If at least some classification data are available, versatile representations can be learned in a semi-supervised fashion, as described in (Intrator and Edelman, 1996; Intrator and Edelman, 1997).

[8] An analysis of the conditions under which proximal measurements convey veridical information about distal objects appears in (Edelman, 1999).

One prominent theory of visual cortical plasticity compatible with this approach is BCM, according to which vectors of synaptic weights seek to become orthogonal (in the input space) to frequently occurring events, and non-orthogonal to events occurring with low probability (Bienenstock et al., 1982). A generalization of the BCM theory that addresses the problem of extracting statistically significant features from multidimensional data has been formulated by (Intrator and Cooper, 1992). This approach defines an event as a peak in the input probability distribution; a suspicious event is then signalled by the occurrence of a peak away from the origin, in a low-dimensional projection of the input space.[9] The BCM rule for synaptic weight modification effectively seeks projections along which the probability density deviates maximally from a Gaussian distribution.[10] The RFs of units trained with the BCM rule are thus tuned to the detection of low-dimensional, statistically prominent structure in the high-dimensional measurement space.

The probabilistic line of reasoning suggests that sensory coding is "... the process of preparing a representation of the current sensory scene in a form that enables subsequent learning mechanisms to be versatile and reliable" (Barlow, 1990). Specifically, a representation is useful for learning if it includes records of recurring and co-occurring events. As noted by Barlow, a convenient substrate for such a representation is provided by Selfridge's Pandemonium (Selfridge, 1959). In Barlow's Probabilistic Pandemonium, the response strength of a feature-detector demon would be proportional to $-\log P$, where P is the probability of occurrence of the feature the demon detects. Similarly, in the BCM theory, the response of a feature detector becomes proportional to the inverse of the posterior probability of occurrence of the event to which it is tuned, up to some saturation limit (Intrator and Cooper, 1992; Intrator, 1996). Although the difficulty of coming up with independent features and with monitoring the statistics of occurrence of each of them should not be underestimated (Barlow, 1994), this is certainly a worthy goal for any perceptual system, because of the ability it would confer to learn and reason in an informed and principled manner.[11]

4.3 Supporting behavioral and neurobiological evidence

A full survey of the relevant data would run to a book length; here, we highlight a few chosen behavioral and neurobiological findings.[12]

Human subjects use statistical cues for unsupervised learning of object "parts." Behavioral studies that exerted control over temporal (Fiser and Aslin, 2002) and spatial (Fiser and Aslin, 2001; Edelman et al., 2002a) conditional probabilities of components of structured stimuli show that subjects are attuned to such cues and can use them to learn stimulus parts in an unsupervised fashion.

The ability of human subjects to tolerate stimulus translation is limited in a manner that suggests reliance on active gaze control for learning structural relations. (Dill and Edelman, 2001) found that performance in a same/different discrimination task using articulated animal-like 3D shapes was fully transferred across retinal location if local cues were diagnostic, but not if the decision had to be based on relative

[9] The peak around 0 is considered noise; a suspicious event is simply a sharp peak (ideally, a δ-function) in the joint probability that rises above the noise given by the marginal probabilities.

[10] Due to the Central Limit Theorem, most projections are Gaussian, and thus can be described completely by their covariance matrix (second-order statistics).

[11] While coding the inverse of posterior probability of events facilitates rapid creation of "suspicious coincidence" detectors, it also represents an optimal coding scheme in the sense that events that occur with low probability evoke a stronger neural response, thus reducing neural energy dissipation (Intrator, 1996; Cooper et al., 2004).

[12] A review of the computational models of the neuronal mechanisms behind these data is beyond the scope of this paper.

location of various fragments. This suggests that the human visual system treats local cues and structural information differently, relying for the former on multiple feature detectors replicated over the visual field, and for the latter – on fixation control coupled with the use of visual space as its own representation (Edelman and Intrator, 2003).

Response properties of neurons in the monkey visual cortex point to connections between attention/gaze control and receptive field structure. In primates, attention can be steered overtly (through gaze control) and covertly, with the eyes maintaining steady fixation. The earliest evidence for the role of attention in shaping the receptive fields of visual neurons came from the cortical area V4 (Moran and Desimone, 1985). Since then, a wealth of evidence for the cortical mechanisms of active vision became available; the relevant findings include gaze-dependent gain fields in V4 (Pouget and Sejnowski, 1997; Connor et al., 1997) and sensitivity to translation of neuronal responses in areas ranging from the inferotemporal (IT) cortex (Op de Beeck and Vogels, 2000; DiCarlo and Maunsell, 2003) down to the primary visual area V1, where neurons have been found that are gated by top-down signals (Vidyasagar, 1998), or that are tuned for gaze direction and that track the Bayesian probability of stimulus appearance (Sharma et al., 2003).

The detectors for various visual qualities in the brain form sparsely active, map-like ensembles. Since the first reports of a map-like arrangement of low-level (orientation, etc.) feature detectors in V1 (Hubel and Wiesel, 1962), much evidence has been accumulated that attests to the presence of ensemble representations in areas V4 and IT (Pasupathy and Connor, 2002; Fujita et al., 1992; Kobatake and Tanaka, 1994; Wang et al., 1998; Op de Beeck et al., 2001; Tsunoda et al., 2001). The topographic characteristics of these representations are exemplified by the tendency of nearby neurons to prefer similar stimuli (Wang et al., 2000), and more so following repeated exposure (Erickson et al., 2000). Importantly, the distributed code is sparse, both in V1 (Field, 1994; Vinje and Gallant, 2000) and in IT (Young and Yamane, 1992; Rolls and Tovee, 1995). Moreover, the code is sparser for more familiar objects, which activate, in the prefrontal cortex, fewer neurons than novel ones; these neurons are also more narrowly tuned (Rainer and Miller, 2000).

Rapid, experience-driven plasticity is ubiquitous in the brain. Experience-dependent plasticity is found throughout the visual cortex, from V1 (Gilbert, 1994) to IT (Sakai and Miyashita, 1991; Kobatake et al., 1992). Having the monkey engage in categorization is very effective in shaping neuronal selectivity (Kobatake et al., 1998; Sigala and Logothetis, 2002), but learning can also be driven by mere exposure. Indeed, the response properties of IT neurons can be modified by a single block of 10 trials, with a 5-second total exposure (Tovee et al., 1996). Some neurons become more responsive to the stimulus shown, others less (Rolls et al., 1989), as expected if deallocation (and not only allocation) is taking place. In general, the time course of neuronal selectivity parallels that of the behavioral manifestations of learning (Messinger et al., 2001). The neuronal characteristics of learning also conform to quantitative models that combine activity-dependent plasticity with synaptic normalization, such as the Hebb/Oja and the BCM rules (Fregnac et al., 1988; Clothiaux et al., 1991; Bear and Malenka, 1994; Abbott and Nelson, 2000).

5 Concluding remarks

In lieu of conclusions, we discuss what we perceive as the more promising approaches and interesting challenges on the road toward better representation methods and unsupervised statistical learning in vision.

5.1 Representation: using distributed patterns of activations of "what+where" units

A representational scheme that is cognizant of the philosophical, computational, psychological and neurobiological considerations mentioned earlier distinguishes between the "what" and the "where" aspects of the sensory input, and lets the latter serve as the scaffolding holding the would-be objects in place (Edelman, 1999; Edelman, 2002). According to this scheme, the "what" entities (the would-be objects) are coded by their similarities to an ensemble of familiar reference shapes (Duvdevani-Bar and Edelman, 1999). At the same time, the "where" aspects of the object/scene structure are represented by the spatial distribution of the receptive fields of the detectors tuned to the ensemble members (Edelman and Intrator, 2003). Functionally, this amounts to the use of visual space as its own representation (O'Regan, 1992). In computer vision, a fruitful approach to representation that relies jointly on local photometry ("what") and global configuration or geometry ("where") of objects has been developed by Perona and his colleagues (Burl et al., 1998; Fei-Fei et al., 2003). The statistical theory of shapes thus represented is being vigorously developed; some of the entry points into that literature can be found in (Kendall, 1984; Carne, 1990; Le and Kendall, 1993).

5.2 Learning: using local statistical mechanisms

How should such representations be acquired? Our approach to the unsupervised statistical learning of ensemble representations has been outlined in section 4.2. To conform to the constraints imposed by the neurobiological data, the mechanisms that implement such learning must be local, in two senses. First, the accounting (e.g., the computation of a figure of merit) that drives the learning must operate incrementally rather than in a batch mode (Fei-Fei et al., 2003; Duygulu et al., 2002), because brains are not equipped with data registers where many numbers can be accumulated before they are acted upon.[13] Second, the action (namely, the synaptic modification) that is to be carried out on the basis of the statistical rule must be local to the neuron at which its criteria are evaluated (Edelman et al., 2002b), because biology does not provide for the "transportation" of the required numbers across the cortex.

5.3 Challenges ahead

In summary, we propose that statistical learning in high-level computer vision can and should benefit from progress in the understanding of biological vision on all levels: computational theory, algorithms and implementation. A neuromorphic approach to the twin problems of object structure representation (Edelman and Intrator, 2003) and learning (Edelman et al., 2002b) that we outlined here is only a small step in that direction. Some of the more challenging issues that still need to be overcome are dealing with cluttered scenes, supporting dynamic reallocation of units, and integrating the treatment of bottom-up, lateral and top-down interactions within the same computational framework.

References

Abbott, L. F. and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature Neuroscience*, 3 Supp:1178 – 1183.

Akins, K. (1996). Of sensory systems and the 'aboutness' of mental states. *Journal of Philosophy*, XCIII:337–372.

[13] Incremental, trial by trial learning may also alleviate the credit assignment problem that arises when batch learning and a global figure of merit are used.

Aloimonos, J. Y., Weiss, I., and Bandopadhay, A. (1988). Active vision. *Intl. J. Computer Vision*, 2:333–356.

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *The mechanisation of thought processes*, pages 535–539. H.M.S.O., London.

Barlow, H. B. (1979). The past, present and future of feature detectors. In Albrecht, D., editor, *Recognition of Pattern and Form*, volume 44 of *Lecture Notes in Biomathematics*, pages 4–32. Springer, Berlin.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30:1561–1571.

Barlow, H. B. (1994). What is the computational goal of the neocortex? In Koch, C. and Davis, J. L., editors, *Large-scale neuronal theories of the brain*, chapter 1, pages 1–22. MIT Press, Cambridge, MA.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12:241–253.

Barrow, H. G. and Tenenbaum, J. M. (1993). Retrospective on "Interpreting line drawings as three-dimensional surfaces". *Artificial Intelligence*, 59:71–80.

Bear, M. F. and Malenka, R. C. (1994). Synaptic plasticity: LTP and LTD. *Curr. Opin. Neurobiol.*, 4:389–399.

Bienenstock, E., Cooper, L., and Munro, P. W. (1982). Theory for the development of neural selectivity: orientation specificity and binocular interaction in visual cortex. *J. of Neuroscience*, 2:32–48.

Booth, A. E. and Waxman, S. R. (2002). Word learning is 'smart': evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, 84:B11–B22.

Braitenberg, V. (1977). *On the texture of brains*. Springer-Verlag, New York.

Burl, M. C., Weber, M., and Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry. In *Proc. 4^{th} Europ. Conf. Comput. Vision, H. Burkhardt and B. Neumann (Eds.), LNCS-Series Vol. 1406–1407, Springer-Verlag*, pages 628–641.

Carne, T. K. (1990). The geometry of shape spaces. *Proc. Lond. Math. Soc.*, 61:407–432.

Carrow-Woolfolk, E. (1999). *Comprehensive Assessment of Spoken Language (CASL)*. AGS Publishing, Circle Pines, MN.

Clothiaux, E. E., Cooper, L. N., and Bear, M. F. (1991). Synaptic plasticity in visual cortex: Comparison of theory with experiment. *Journal of Physiology*, 66:1785–1804.

Cohen, P. R. and Oates, T. (1998). A dynamical basis for the semantic content of verbs. In *The Grounding of Word Meaning: Data and Models Workshop, AAAI-98*, pages 5–8.

Connor, C. E., Preddie, D. C., Gallant, J. L., and Van Essen, D. C. (1997). Spatial attention effects in macaque area V4. *J. of Neuroscience*, 17:3201–3214.

Cooper, L. N., Intrator, N., Blais, B. S., and Shouval, H. Z. (2004). *Theory of Cortical Plasticity*. World Scientific, Singapore.

Culicover, P. W. (1999). *Syntactic nuts: hard cases, syntactic theory, and language acquisition*. Oxford University Press, Oxford.

DiCarlo, J. J. and Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology*, 89:3264–3278.

Dickinson, S., Bergevin, R., Biederman, I., Eklundh, J., Munck-Fairwood, R., Jain, A., and Pentland, A. (1997). Panel report: The potential of geons for generic 3-d object recognition. *Image and Vision Computing*, 15:277–292.

Dill, M. and Edelman, S. (2001). Imperfect invariance to object translation in the discrimination of complex shapes. *Perception*, 30:707–724.

Duvdevani-Bar, S. and Edelman, S. (1999). Visual recognition and categorization on the basis of similarities to multiple class prototypes. *Intl. J. Computer Vision*, 33:201–228.

Duygulu, P., Barnard, K., de Freitas, J., and Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. ECCV-2002*.

Edelman, S. (1995a). Receptive fields for vision: from hyperacuity to object recognition. Unpublished manuscript, available online.

Edelman, S. (1995b). Vision reanimated. Unpublished manuscript, available online.

Edelman, S. (1999). *Representation and recognition in vision*. MIT Press, Cambridge, MA.

Edelman, S. (2002). Constraining the neural representation of the visual world. *Trends in Cognitive Sciences*, 6:125–131.

Edelman, S. and Christiansen, M. H. (2003). How seriously should we take Minimalist syntax? A comment on Lasnik. *Trends in Cognitive Science*, 7:60–61.

Edelman, S., Hiles, B. P., Yang, H., and Intrator, N. (2002a). Probabilistic principles in unsupervised learning of visual structure: human data and a model. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 19–26, Cambridge, MA. MIT Press.

Edelman, S. and Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science*, 27:73–109.

Edelman, S., Intrator, N., and Jacobson, J. S. (2002b). Unsupervised learning of visual structure. In Bülthoff, H. H., Wallraven, C., Lee, S.-W., and Poggio, T., editors, *Proc. 2nd Intl. Workshop on Biologically Motivated Computer Vision*, volume 2525 of *Lecture Notes in Computer Science*, pages 629–643. Springer.

Erickson, C. A., Jagadeesh, B., and Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in monkeys. *Nature Neuroscience*, 3:1143–1148.

Fei-Fei, L., Fergus, R., and Perona, P. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. ICCV-2003*.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, 6:559–601.

Fiser, J. and Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 6:499–504.

Fiser, J. and Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 28:458–467.

Fregnac, Y., Schulz, D., Thorpe, S., and Bienenstock, E. (1988). A cellular analogue of visual cortical plasticity. *Nature*, 333:367–370.

Fu, K. S. (1976). Tree languages and syntactic pattern recognition. In Chen, C. H., editor, *Pattern Recognition and Artificial Intelligence*, pages 257–291. Academic Press, New York.

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346.

Galata, A., Johnson, N., and Hogg, D. C. (2001). Learning Variable Length Markov Models of behaviours. *Computer Vision and Image Understanding*, 81:398–413.

Gardner-Medwin, A. R. and Barlow, H. B. (2001). The limits of counting accuracy in distributed neural representations. *Neural Computation*, 13:477–504.

Geman, S. (1996). Minimum Description Length priors for object recognition. In *Challenging the frontiers of knowledge using statistical science (Proc. JSM'96)*.

Gibson, J. J. (1957). Survival in a world of probable objects. *Contemporary Psychology*, 2:33–35.

Gilbert, C. D. (1988). Neuronal and synaptic organization in the cortex. In Rakic, P. and Singer, W., editors, *Neurobiology of Neocortex*, pages 219–240. Wiley, New York, NY.

Gilbert, C. D. (1994). Neuronal dynamics and perceptual learning. *Current Biology*, 4:627–629.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160:106–154.

Hume, D. (1740). *A Treatise of Human Nature*. Available online.

Intrator, N. (1992). Feature extraction using an unsupervised neural network. *Neural Computation*, 4:98–107.

Intrator, N. (1993). Combining Exploratory Projection Pursuit and Projection Pursuit Regression. *Neural Computation*, 5:443–455.

Intrator, N. (1996). Neuronal goals: Efficient coding and coincidence detection. In Amari, S., Xu, L., Chan, L. W., King, I., and Leung, K. S., editors, *Proceedings of ICONIP: Progress in Neural Information Processing*, volume 1, pages 29–34, Hong-Kong. Springer.

Intrator, N. and Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17.

Intrator, N. and Edelman, S. (1996). How to make a low-dimensional representation suitable for diverse tasks. *Connection Science*, 8:205–224.

Intrator, N. and Edelman, S. (1997). Learning low dimensional representations of visual objects with extensive use of prior knowledge. *Network*, 8:259–281.

Intrator, N. and Gold, J. (1993). Three-dimensional object recognition in gray-level images: the usefulness of distinguishing features. *Neural Computation*, 5:61–74.

Kendall, D. G. (1984). Shape manifolds, Procrustean metrics and complex projective spaces. *Bull. Lond. Math. Soc.*, 16:81–121.

Kitcher, P. and Varzi, A. (2000). Some pictures are worth $2^{\aleph_0}$ sentences. *Philosophy*, 75:377–381.

Knill, D. and Richards, W., editors (1996). *Perception as Bayesian Inference*. Cambridge University Press, Cambridge.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, 71:856–867.

Kobatake, E., Tanaka, K., and Tamori, Y. (1992). Long-term learning changes the stimulus selectivity of cells in the inferotemporal cortex of adult monkeys. *Neuroscience Research*, S17:237.

Kobatake, E., Wang, G., and Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophysiol.*, 80:324–330.

Kuffler, S. W. and Nicholls, J. G. (1976). *From neuron to brain*. Sinauer, Sunderland, MA.

Landau, B., Smith, L. B., and Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3:299–321.

Le, H. and Kendall, D. G. (1993). The Riemannian structure of Euclidean shape spaces: a novel environment for statistics. *The Annals of Statistics*, 21:1225–1271.

Markman, E. (1989). *Categorization and naming in children*. MIT Press, Cambridge, MA.

Marr, D. (1976). Early processing of visual information. *Phil. Trans. R. Soc. Lond. B*, 275:483–524.

Messinger, A., Squire, L. R., Zola, S. M., and Albright, T. D. (2001). Neuronal representations of stimulus associations develop in the temporal lobe during learning. *Proceedings of the National Academy of Science*, 98:12239–12244.

Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784.

Op de Beeck, H. and Vogels, R. (2000). Spatial sensitivity of Macaque inferior temporal neurons. *J. Comparative Neurology*, 426:505–518.

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, 4:1244–1252.

O'Regan, J. K. (1992). Solving the real mysteries of visual perception: The world as an outside memory. *Canadian J. of Psychology*, 46:461–488.

Pasupathy, A. and Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, 5:1332–1338.

Phillips, W. A. and Singer, W. (1997). In search of common foundations for cortical computation. *Behavioral and Brain Sciences*, 20:657–722.

Postal, P. M. (2004). *Skeptical linguistic essays*. Oxford University Press, New York.

Pouget, A. and Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, 9:222–237.

Pouget, A., Zemel, R. S., and Dayan, P. (2000). Information processing with population codes. *Nature Review Neuroscience*, 1:125–132.

Rainer, G. and Miller, E. K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. *Neuron*, 27:179–189.

Reitman, W., Nado, R., and Wilcox, B. (1978). Machine perception: what makes it so hard for computers to see? In Savage, C. W., editor, *Perception and cognition: issues in the foundations of psychology*, volume IX of *Minnesota studies in the philosophy of science*, pages 65–87. University of Minnesota Press, Minneapolis, MN.

Rissanen, J. (1987). Minimum description length principle. In Kotz, S. and Johnson, N. L., editors, *Encyclopedia of Statistic Sciences*, volume 5, pages 523–527. J. Wiley and Sons.

Rolls, E. T., Baylis, G. C., Hasselmo, M. E., and Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. *Exp. Brain Res.*, 76:153–164.

Rolls, E. T. and Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. of Neurophysiology*, 73:713–726.

Sakai, K. and Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, 354:152–155.

Selfridge, O. G. (1959). Pandemonium: a paradigm for learning. In *The mechanisation of thought processes*. H.M.S.O., London.

Sharma, J., Dragoi, V., Tenenbaum, J. B., Miller, E. K., and Sur, M. (2003). V1 neurons signal acquisition of an internal representation of stimulus location. *Science*, 300:1758–1763.

Sigala, N. and Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415:318–320.

Smith, B. (2001). Fiat objects. *Topoi*, 20:131–148.

Stone, J. V. (1996). Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8:1463–1492.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Harvard University Press, Cambridge, MA.

Tovee, M. J., Rolls, E. T., and Ramachandran, V. S. (1996). Rapid visual learning in neurones of the primate temporal visual cortex. *NeuroReport*, 7:2757–2760. Mooney faces.

Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4:832–838.

Vidyasagar, T. R. (1998). Gating of neuronal responses in macaque primary visual cortex by an attentional spotlight. *Neuroreport*, 9:1947–1952.

Vinje, W. E. and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287:1273–1276.

Wang, G., Tanifuji, M., and Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neurosci. Res.*, 32:33–46.

Wang, Y., Fujita, I., and Murayama, Y. (2000). Neuronal mechanisms of selectivity for object features revealed by blocking inhibition in inferotemporal cortex. *Nature Neuroscience*, 3:807–813.

Young, M. P. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1331.