NIPS*91 Post-Conference Workshop on

Self-Organization and Unsupervised Learning in Vision

December 6–7, 1991 in Vail, Colorado

Workshop Chair: **Jonathan A. Marshall**, Ph.D.

Department of Computer Science, CB 3175, Sitterson Hall
University of North Carolina, Chapel Hill, NC 27599-3175, U.S.A.
919-962-1887, marshall@cs.unc.edu

Substantial neurophysiological and psychophysical evidence suggests that visual experience guides or directs the formation of much of the fine structure of animal visual systems. Simple unsupervised learning procedures (e.g., Hebbian rules) using winner-take-all or local k-winner networks have been applied with moderate success to show how visual experience can guide the self-organization of visual mechanisms sensitive to low-level attributes like orientation, contrast, color, stereo disparity, and motion. However, such simple networks lack the more sophisticated capabilities needed to demonstrate self-organized development of higher-level visual mechanisms for segmentation, grouping/binding, selective attention, representation of occluded or amodal visual features, resolution of uncertainty, generalization, context-sensitivity, and invariant object recognition.

A variety of enhancements to the simple Hebbian model have been proposed. These include anti-Hebbian rules, maximization of mutual information, oscillatory interactions, intraneuronal interactions, steerable receptive fields, pre- vs. post-synaptic learning rules, covariance rules, addition of behavioral (motor) information, and attentional gating. Are these extensions to unsupervised learning sufficiently powerful to model the important aspects of neurophysiological development of higher-level visual functions?

Some of the specific questions that the workshop will address are:

- Does our visual environment provide enough information to direct the formation of higher-level visual processing mechanisms?
- What kinds of information (e.g., correlations, constraints, coherence, and affordances) can be discovered in our visual world, using unsupervised learning?
- Can such higher-level visual processing mechanisms be formed by unsupervised learning? Or is it necessary to appeal to external mechanisms such as evolution (genetic algorithms)?
- Are there further enhancements that can be made to improve the performance and capabilities of unsupervised learning rules for vision?
- What neurophysiological evidence is available regarding these possible enhancements to models of unsupervised learning?
- What aspects of the development of visual systems must be genetically pre-wired, and what aspects can be guided or directed by visual experience?
- How is the output of an unsupervised network stage used in subsequent stages of processing?
- How can behaviorally relevant (sensorimotor) criteria become incorporated into visual processing mechanisms, using unsupervised learning?

This 2-day informal workshop brings together researchers in visual neuroscience, visual psychophysics, and neural network modeling. Invited speakers from these communities will briefly discuss their views and results on relevant topics. In discussion periods, we will examine and compare these results in detail.

The workshop topic is crucial to our understanding of *how animal visual systems got the way they are*. By addressing this issue head-on, we may come to understand better the factors that shape the structure of animal visual systems, and we may become able to build better computational models of the neurophysiological processes underlying vision.

Program: Workshop on Self-Organization and Unsupervised Learning in Vision

Friday morning, December 6, 7:30–9:30 a.m.

Daniel Kersten, Department of Psychology, University of Minnesota.
"Environmental structure and scene perception: Perceptual representation of material, shape, and lighting"

David C. Knill, Center for Research in Learning, Perception, and Cognition, University of Minnesota.
"Environmental structure and scene perception: The nature of visual cues for 3-D scene structure"

DISCUSSION

Edward M. Callaway, Department of Neurobiology, Duke University.
"Development of clustered intrinsic connections in cat striate cortex"

Michael P. Stryker, Department of Physiology, University of California at San Francisco.
"Problems and promise of relating theory to experiment in models for the development of visual cortex"

DISCUSSION

Friday afternoon, December 6, 4:30–6:30 p.m.

Joachim M. Buhmann, Lawrence Livermore National Laboratory.
"Complexity optimized data clustering by competitive neural networks"

Nicol G. Schraudolph, Department of Computer Science, University of California at San Diego.
"The information transparency of sigmoidal nodes"

DISCUSSION

Heinrich H. Bülthoff, Department of Cognitive and Linguistic Sciences, Brown University.
"Psychophysical support for a 2D view interpolation theory of object recognition"

John E. Hummel, Department of Psychology, University of California at Los Angeles.
"Structural description and self organizing object classification"

DISCUSSION

Saturday morning, December 7, 7:30–9:30 a.m.

Allan Dobbins, Computer Vision and Robotics Laboratory, McGill University.
"Local estimation of binocular optic flow"

Alice O'Toole, School of Human Development, The University of Texas at Dallas.
"Recent psychophysics suggesting a reformulation of the computational problem of structure-from-stereopsis"

DISCUSSION

Jonathan A. Marshall, Department of Computer Science, University of North Carolina at Chapel Hill.
"Development of perceptual context-sensitivity in unsupervised neural networks: Parsing, grouping, and segmentation"

Suzanna Becker, Department of Computer Science, University of Toronto.
"Learning perceptual invariants in unsupervised connectionist networks"

Albert L. Nigrin, Department of Computer Science and Information Systems, American University.
"Using presynaptic inhibition to allow neural networks to perform translational invariant recognition"

DISCUSSION

Saturday afternoon, December 7, 4:30–7:00 p.m.

Jürgen Schmidhuber, Department of Computer Science, University of Colorado.
"Learning non-redundant codes by predictability minimization"

Laurence T. Maloney, Center for Neural Science, New York University.
"Geometric calibration of a simple visual system"

DISCUSSION

Paul Munro, Department of Information Science, University of Pittsburgh.
"Self-supervised learning of concepts"

Richard Zemel, Department of Computer Science, University of Toronto.
"Learning to encode parts of objects"

DISCUSSION

WRAP-UP, 6:30–7:00

Invited Speakers

Suzanna Becker, Department of Computer Science, University of Toronto

"Learning perceptual invariants in unsupervised connectionist networks"

A major goal in developing unsupervised learning algorithms is to find an objective for the learning which can provide a single unifying principle that explains many progressively more complex stages of perceptual processing. We have proposed an objective for perceptual learning which is based on this goal, and causes units to learn simple invariant properties of the input distribution. The learning procedure is derived from a continuous version of the Gmax algorithm proposed by Pearlmutter & Hinton (1986). We have generalized this algorithm to apply to a set of competing units, which learn to detect different invariant properties of the input patterns by forming a mixture model of the input distribution. This method has been demonstrated on the problem of learning to detect disparity in random dot stereograms; the competing units each learn to become tuned to a different disparity. In addition to the many spatially invariant features in visual images, there are many temporally invariant features of the sensory input that the model should be able to capture. The acoustic signal generated by a particular speaker is relatively constant in overall pitch, timbre etc. As one moves about in the world, the visual field flows by in characteristic patterns of expansion, dilation and translation. And independently moving rigid objects are invariant with respect to shape, texture and many other features, up to very high level properties such as the object's identity.

This work was done jointly with Geoffrey Hinton.

Joachim M. Buhmann, Lawrence Livermore National Laboratory

"Complexity optimized data clustering by competitive neural networks" (Joachim Buhmann and Hans Kühnel)

Data clustering is a complex optimization problem with applications ranging from vision and speech processing to data transmission and data storage in technical as well as in biological systems. We propose a clustering strategy which explicitly reflects the tradeoff between simplicity and precision of a data representation. The resulting clustering algorithm *jointly optimizes* distortion errors and complexity costs. A maximum entropy estimation of the clustering cost function yields an optimal number of clusters, their positions and their occupation probabilities. This algorithm for complexity optimized clustering is implemented by an artificial neural network with *winner-take-all* connectivity. Our approach establishes a unifying framework for different clustering methods like K-means clustering, fuzzy clustering, entropy-constrained vector quantization or topological feature maps, and competitive neural networks.

Heinrich Bülthoff, Department of Cognitive and Linguistic Sciences, Brown University

"Psychophysical support for a 2D view interpolation theory of object recognition"

Edward M. Callaway, Department of Neurobiology, Duke University

"Development of clustered intrinsic connections in cat striate cortex"

Allan Dobbins, Computer Vision and Robotics Laboratory, McGill University

"Local estimation of binocular optic flow"

Neurons in the middle temporal (MT) visual cortical area of primates are direction-selective and many exhibit strong direction- and speed-dependent interactions between receptive field (RF) center and surround (Allman et al., 1985). In addition MT neurons are binocular-disparity selective. Since any smooth vector field is locally diffeomorphic to a constant vector field, over a sufficiently small neighbourhood the variation of velocity is small. Obversely, if the spatial rate of change of the velocity field is large it is desirable to allow estimates of the flow field be from neighborhoods as large as possible. Two reasons are offered: pooling estimates increases signal to noise performance, and redundancy is reduced by representing the flow field by a smaller number of vectors. I will argue that ensembles of MT neurons with different RF size and in which center and surround have similar velocity and disparity tuning are estimating the spatial neighbourhood over which the flow field is well-approximated as constant. For example, suppression of center by surround implies that the field is constant over a larger area than the RF center whereas lack of suppression implies either that the field is spatially limited or rapidly space-varying. Using as large a neighbourhood as the evidence warrants allows adaptive spatial pooling of afferents to increase signal to noise performance and optimize descriptive length of the flow field.

John E. Hummel, Department of Psychology, University of California at Los Angeles

"Structural description and self organizing object classification"

How can a self organizing neural network learn to classify two or more images as the same object? Generalization across examples (in this case, images) – a traditional strength of neural networks (NNs) – cannot provide a satisfactory answer to this question, particularly if the goal is to understand human recognition competence. An alternative approach is to use an object's image to activate a viewpoint invariant structural description of the object (e.g., specifying its parts and their interrelations).

Once the structural description is active, it is relatively straight forward to use it as a basis for self organizing object classification.[1] Two problems confronting structural description as a basis for self organizing object classification are discussed. First, representing viewpoint invariant structural descriptions in NNs requires a way to represent attribute bindings (or conjunctions) dynamically, i.e., temporarily, so that the same units can be used in multiple conjunctions. For example, unless the representation of an object can be bound dynamically to a representation specifying its location, it will be necessary to posit a separate representation of the object for each location in which it can reside. Such an approach would fail to classify different views of the object as equivalent. The second problem concerns whether and how neural structures capable of activating viewpoint invariant structural descriptions might self organize. A potential solution is offered for the first problem (dynamic binding). In regard to the second, it is speculated that traditional self organization based solely on the properties of the input domain (in this case, 2-D images of 3-D objects) will prove insufficient to derive the structures necessary to achieve viewpoint invariant structural descriptions.

[1] Hummel, J. & Biederman, I. (1990). "Dynamic binding in a neural network for shape recognition." Technical report 90-5, Image Understanding Laboratory, University of Minnesota.

Daniel Kersten, Center for Research in Learning, Perception, & Cognition, University of Minnesota

"Environmental structure and scene perception: Perceptual representation of material, shape, and lighting" (Daniel Kersten and David C. Knill)

A major problem faced by any visual system is the limited capacity which it has available for the transmission and representation of visual information. This concern has led some to conjecture that functional characteristics of neurons in the early visual system develop or self-organize to provide statistically optimal coding of image information. We argue that high level perceptual representations may also reflect statistical regularities in 3-D scenes. In particular, it may be advantageous for visual systems to construct distinct, but interacting representations of surface shape, material and lighting precisely because these scene properties are statistically independent in the world. We will demonstrate a number of recently discovered visual illusions that illustrate interactions between depth, shape, material and lighting in the human visual system.

David C. Knill, Center for Research in Learning, Perception, & Cognition, University of Minnesota

"Environmental structure and scene perception: The nature of visual cues for 3-D scene structure" (David C. Knill and Daniel Kersten)

Regularities, both lawful and statistical, in scene attributes are at the heart of a number of so-called image cues for 3-D scene structure. The visual system must adapt (ontogenetically or phylogenetically) to this structure in order to use these regularities in extracting information about scene structure from images. While many self-organizing systems have been modelled which adapt to regularities in image structure, none have been developed which adapt to regularities in scene structure. We discuss the problem of adapting to scene structure in learning to use image cues, presenting results of our work with two types of contour, those projected from surface markings and those projected from shadow boundaries, as examples.

Laurence T. Maloney, Center for Neural Science, New York University

"Geometric calibration of a simple visual system"

A visual system is geometrically calibrated if its estimates of the spatial properties of a scene are accurate: straight lines are judged straight, angles are correctly estimated, collinear line segments are perceived to fall on a common line. This paper describes calibration methods for a model visual system whose photoreceptors are initially at unknown locations. The methods can also compensate for optical distortions that are equivalent to remapping of receptor locations (e.g. spherical aberration). The methods work by comparing visual input across eye/head movements; they require no explicit feedback, and no knowledge about the particular contents of a scene. This work has implications for development and calibration in biological visual systems.

This work was supported in part by a grant from the National Eye Institute EY08266.

Jonathan A. Marshall, Department of Computer Science, University of North Carolina at Chapel Hill

"Development of perceptual context-sensitivity in unsupervised neural networks: Parsing, grouping, and segmentation"

Context is an important determinant of all perceptual experience. I will describe a combination of neurally plausible excitatory and inhibitory learning rules that contributes to the development of context-sensitive grouping and segmentation mechanisms. In the network, some lateral inhibitory connection weights weaken, allowing some neurons to become coactivated. Other lateral inhibitory weights remain strong and enforce contextual constraints on allowable combinations of activations. The network chooses a near-optimal representation of multiple superimposed patterns, by simultaneous distributed activation of multiple neurons. The network implements a form of credit assignment; it even suppresses the activation of some of the most-excited neurons when necessary to achieve maximally context-sensitive, global input representations.

Paul Munro, Department of Information Science, University of Pittsburgh

"Self-supervised learning of concepts"

A mathematical system, self-supervised learning (SSL) is presented that describes a form of learning by high-order concept units (C-units) that learn to become sensitive to categories of stimuli associated by finding some feature that they share. Implicit in the SSL model is the assumption that each C-unit receives input from at least two information channels, or "banks." Under SSL, each C-unit becomes very selective with respect to one of the channels. The preferred patterns in the training bank serve as a "seed" for concept formation, and an associative process causes patterns from the other banks to drive the C-unit to the extent that they are correlated with the seed in the environment.

Albert L. Nigrin, Department of Computer Science and Information Systems, American University

"Using presynaptic inhibition to allow neural networks to perform translational invariant recognition"

Previous neural networks have not been able to achieve translational invariant recognition (TIR) of objects using only neural elements. To achieve TIR, networks such as the neocognitron need far too many connections to be practical while networks such as ART need to preprocess their inputs using standard algorithms. We believe that current networks are incapable of TIR because of the manner by which competition is implemented. In current networks *nodes* to compete for activation. We believe it is more advantageous for the *output* of nodes to compete (for the right to activate other nodes). When this is done, TIR is accomplished with ease.

We will present a multistage network which will reduce the translational uncertainty (where the object is in a field) of a one dimensional object. The network presented can then easily be generalized to more dimensions. The input field will be labeled $F^{(1)}$ and higher level fields will be labeled $F^{(2)}$, $F^{(3)}$ etc. Assume that the size of $F^{(1)}$ is n. Also, assume that the maximum size object presented to $F^{(1)}$ is of size less or equal to m, where $m < n$. Since the object can be centered at any of $(n - m)$ locations, the translational uncertainty is $(n - m)$.

Signals from $F^{(1)}$ to $F^{(2)}$ will reproduce the object at $F^{(2)}$; however since $F^{(2)}$ will contain only $n/2$ cells, the translational uncertainty of the object at $F^{(2)}$ will be only $n/2 - m$. Similarly, signals from $F^{(2)}$ to $F^{(3)}$ will reproduce the object at $F^{(3)}$. $F^{(3)}$ will contain only half as many cells as $F^{(2)}$ so again the translational uncertainty will be reduced – this time to $n/4 - m$. By using a $\log_2(n)$ number levels, the translational uncertainty can be reduced to zero.

Small numbers of cells and connections are needed to implement this architecture. The number of cells needed is only O(n). The total number of connections needed are O($n\ log_2\ n$). This follows since at most O(n) connections are needed between each level in the hierarchy and there are $\log_2(n)$ fields in the architecture,

After the object has been centered, it can be recognized using a number of different neural network architectures such as SONNET, ART, or the neocognitron. We will present some of the advantages and disadvantages of each of those classifying networks. Finally, we will show how size invariant recognition can also be achieved if the input to the architecture is provided by a properly designed boundary contour system.

Alice O'Toole, School of Human Development, The University of Texas at Dallas

"Recent psychophysics suggesting a reformulation of the computational problem of structure-from-stereopsis"

I will try to bring together some recent and diverse psychophysical data that suggest that current computational models of structure-from-stereopsis do not provide an adequate model of human stereopsis. Among these phenomena are (1) post-fusional latency (Gillam et al., 1984, 1988), (2) the disparity threshold elevations seen with intervening depth features (Fahle & Westheimer, 1988), (3) Rogers & Caganello's (1989) work on 2nd derivative detection of depth, and (4) related work by Mitchison & Westheimer on depth saliency. Combined, these observations suggest changes in the computational theory of structure-from-stereopsis, and hence, the usefulness of the traditionally used constraints. While some of the principles these data suggest have been used in Lehky & Sejnowski's (1990) model of stereoacuity, their implications for models of structure-from-stereopsis have not been considered.

Jürgen Schmidhuber, Department of Computer Science, University of Colorado

"Learning non-redundant codes by predictability minimization"

I propose a novel, general, and useful principle for unsupervised learning of distributed non-redundant representations of inputs or input sequences. With a given set of representational units, each unit tries to react to the environment such that it minimizes its predictability by an adaptive predictor that sees all the other units. This encourages each unit to filter 'abstract entities' out of the environmental input such that these entities are independent of those upon which the other units focus. I discuss various simple yet potentially powerful implementations of the principle which aim at finding binary factorial codes (Barlow, 1989), i.e., codes where the probability of the occurrence of a particular input is simply the product of the probabilities of the corresponding code symbols. None of these implementations requires the explicit maximization of the determinant of the output covariance matrix. Unlike the latter method, however, and unlike the more conventional linear methods for principal component analysis they can remove not only linear but also non-linear output redundancy. Methods for finding non-redundant codes automatically embed Occam's razor for finding codes using a minimal number of units. The final part of this paper describes an entirely local algorithm based on predictability minimization that can learn unique representations of extended sequences. Experiments show that algorithms based on the principle of predictability minimization are practically feasible.

Nicol G. Schraudolph, Department of Computer Science, University of California at San Diego

"The information transparency of sigmoidal nodes"

We propose an information-theoretic optimality criterion for neural networks which suggests the use of sigmoid activation functions when signalling power is limited. Furthermore, the framework provides for a principled comparison of these activation functions under both near-optimal and marginal (near-saturated) conditions, and allows calculation of an optimal weight vector size for weight normalization schemes.

Michael P. Stryker, Department of Physiology, University of California at San Francisco

"Problems and promise of relating theory to experiment in models for the development of visual cortex"

The promise of helping to understand neurobiology through the use of mathematical or computational models is often unfulfilled because of two difficulties: (1) relating elements of the model to features of the biology, and (2) relating parameters of the model to existing or feasible biological measurements. These difficulties, and strategies for their amelioration, will be illustrated by reference to a model of cortical ocular dominance column development (Miller, Keller, & Stryker, 1989, Science 245:605-615) as well as three other recent models of brain development or function. I will suggest that while we lack methods for testing the biological relevance of models with more than a very few parameters, more general formulations of some models may reduce the number of parameters needed.

Richard Zemel, Department of Computer Science, University of Toronto

"Learning to encode parts of objects"

Object recognition can be greatly facilitated by grouping data features into subsets that are likely to belong to a single object. This segmentation is often performed based on a fixed vocabulary of object parts. Many generic part vocabularies have been developed, but these often fail to capture the salient part structure in a particular domain. The alternative of hand-tailoring a part vocabulary to a domain is labor-intensive. Our goal is to develop an unsupervised procedure that enables a network to learn a good set of parts for a given domain.

We present a method of training a network to develop encodings of the parts of objects. The method is based on the idea that for any image containing a few objects, we would like a small set of active units to capture the underlying part structure. We use an objective function based on minimum description length to enable an autoencoder network to form these sparse, distributed codes for images. Preliminary experiments demonstrate that this learning algorithm successfully discovers the part structure of simple images. In one experiment, each image in a training set contains a square and a rectangle, in various positions. The network learns the natural part encoding of each image – just two hidden units are active, each responding to a particular part-position conjunction. The network thus discovers that the two dimensions of shape and position underlie the input space. We plan to further develop our algorithm to learn the part structure of more complicated scenes.

This is joint work with Geoffrey Hinton.
