Prior Knowledge in Support Vector Kernels
Bernhard Schölkopf, Patrice Simard, Alex Smola, Vladimir Vapnik

We explore methods for incorporating prior knowledge about a problem at hand in Support Vector learning machines. We show that both invari(cid:173) ances under group transfonnations and prior knowledge about locality in images can be incorporated by constructing appropriate kernel functions.

************************************

A Revolution: Belief Propagation in Graphs with Cycles
Brendan J. Frey, David MacKay

Until recently, artificial intelligence researchers have frowned upon the application of probability propagation in Bayesian belief net(cid:173) works that have cycles. The probability propagation algorithm is only exact in networks that are cycle-free. However, it has recently been discovered that the two best error-correcting decoding algo(cid:173) rithms are actually performing probability propagation in belief networks with cycles.

************************************

Dynamic Stochastic Synapses as Computational Units
Wolfgang Maass, Anthony Zador

In most neural network models, synapses are treated as static weights that change only on the slow time scales of learning. In fact, however, synapses are highly dynamic, and show use-dependent plasticity over a wide range of time scales. Moreover, synaptic transmission is an inherently stochastic process: a spike arriving at a presynaptic terminal triggers release of a vesicle of neurotransmitter from a release site with a probability that can be much less than one. Changes in release probability represent one of the main mechanisms by which synaptic efficacy is modulated in neural circuits. We propose and investigate a simple model for dynamic stochastic synapses that can easily be integrated into common models for neural computation. We show through computer simulations and rigorous theoretical analysis that this model for a dynamic stochastic synapse increases computational power in a nontrivial way. Our results may have implications for the process(cid:173) ing of time-varying signals by both biological and artificial neural networks.

************************************

Self-similarity Properties of Natural Images
Antonio Turiel, Germán Mato, Néstor Parga, Jean-Pierre Nadal

Scale invariance is a fundamental property of ensembles of nat(cid:173) ural images [1]. Their non Gaussian properties [15, 16] are less well understood, but they indicate the existence of a rich statis(cid:173) tical structure. In this work we present a detailed study of the marginal statistics of a variable related to the edges in the images. A numerical analysis shows that it exhibits extended self-similarity [3, 4, 5]. This is a scaling property stronger than self-similarity: all its moments can be expressed as a power of any given moment. More interesting, all the exponents can be predicted in terms of a multiplicative log-Poisson process. This is the very same model that was used very recently to predict the correct exponents of the structure functions of turbulent flows [6]. These results allow us to study the underlying multifractal singularities. In particular we find that the most singular structures are one-dimensional: the most singular manifold consists of sharp edges.

************************************

Multiple Threshold Neural Logic
Vasken Bohossian, Jehoshua Bruck

We introduce a new Boolean computing element related to the Lin(cid:173) ear Threshold element, which is the Boolean version of the neuron. Instead of the sign function, it computes an arbitrary (with poly(cid:173) nornialy many transitions) Boolean function of the weighted sum of its inputs. We call the new computing element an LT M element, which stands for Linear Threshold with Multiple transitions. The paper consists of the following main contributions related to our study of LTM circuits: (i) the creation of efficient designs of

LTM circuits for the addition of a multiple number of integers and the product of two integers. In particular, we show how to compute the addition of m integers with a single layer of LT M elements. (ii) a proof that the area of the VLSI layout is reduced from O(n2 ) in LT circuits to O(n) in LTM circuits, for n inputs symmetric Boolean functions, and (iii) the characterization of the computing power of LT M relative to LT circuits.

************************************

A General Purpose Image Processing Chip: Orientation Detection
Ralph Etienne-Cummings, Donghui Cai

The generalization ability of a neural network can sometimes be improved dramatically by regularization. To analyze the improve(cid:173) ment one needs more refined results than the asymptotic distri(cid:173) bution of the weight vector. Here we study the simple case of one-dimensional linear regression under quadratic regularization, i.e., ridge regression. We study the random design, misspecified case, where we derive expansions for the optimal regularization pa(cid:173) rameter and the ensuing improvement. It is possible to construct examples where it is best to use no regularization.

************************************

On Efficient Heuristic Ranking of Hypotheses
Steve Chien, Andre Stechert, Darren Mutz

This paper considers the problem of learning the ranking of a set of alternatives based upon incomplete information (e.g., a limited number of observations). We describe two algorithms for hypoth(cid:173) esis ranking and their application for probably approximately cor(cid:173) rect (PAC) and expected loss (EL) learning criteria. Empirical results are provided to demonstrate the effectiveness of these rank(cid:173) ing procedures on both synthetic datasets and real-world data from a spacecraft design optimization problem.

************************************

Learning Continuous Attractors in Recurrent Networks
H. Sebastian Seung

One approach to invariant object recognition employs a recurrent neu(cid:173) ral network as an associative memory. In the standard depiction of the network's state space, memories of objects are stored as attractive fixed points of the dynamics. I argue for a modification of this picture: if an object has a continuous family of instantiations, it should be represented by a continuous attractor. This idea is illustrated with a network that learns to complete patterns. To perform the task of filling in missing in(cid:173) formation, the network develops a continuous attractor that models the manifold from which the patterns are drawn. From a statistical view(cid:173) point, the pattern completion task allows a formulation of unsupervised learning in terms of regression rather than density estimation.

************************************

Boltzmann Machine Learning Using Mean Field Theory and Linear Response Correction
Hilbert Kappen, Francisco de Borja Rodríguez Ortiz

We present a new approximate learning algorithm for Boltzmann Machines, using a systematic expansion of the Gibbs free energy to second order in the weights. The linear response correction to the correlations is given by the Hessian of the Gibbs free energy. The computational complexity of the algorithm is cubic in the number of neurons. We compare the performance of the exact BM learning algorithm with first order (Weiss) mean field theory and second order (TAP) mean field theory. The learning task consists of a fully connected Ising spin glass model on 10 neurons. We conclude that 1) the method works well for paramagnetic problems 2) the TAP correction gives a significant improvement over the Weiss mean field theory, both for paramagnetic and spin glass problems and 3) that the inclusion of diagonal weights improves the Weiss approximation for paramagnetic problems , but not for spin glass problems.

************************************

Incorporating Test Inputs into Learning
Zehra Cataltepe, Malik Magdon-Ismail

In many applications, such as credit default prediction and medical im(cid:173) age recognition, test inputs are available in addition to the labeled train(cid:173) ing examples. We propose a method to incorporate the test inputs into learning. Our method results in solutions having smaller test errors than that of simple training solution, especially for noisy problems or small training sets.

************************************

An Application of Reversible-Jump MCMC to Multivariate Spherical Gaussian Mixtures
Alan Marrs

Applications of Gaussian mixture models occur frequently in the fields of statistics and artificial neural networks. One of the key issues arising from any mixture model application is how to es(cid:173) timate the optimum number of mixture components. This paper extends the Reversible-Jump Markov Chain Monte Carlo (MCMC) algorithm to the case of multivariate spherical Gaussian mixtures using a hierarchical prior model. Using this method the number of mixture components is no longer fixed but becomes a param(cid:173) eter of the model which we shall estimate. The Reversible-Jump MCMC algorithm is capable of moving between parameter sub(cid:173) spaces which correspond to models with different numbers of mix(cid:173) ture components. As a result a sample from the full joint distribu(cid:173) tion of all unknown model parameters is generated. The technique is then demonstrated on a simulated example and a well known vowel dataset.

************************************

A 1, 000-Neuron System with One Million 7-bit Physical Interconnections
Yuzo Hirai

An asynchronous PDM (Pulse-Density-Modulating) digital neural network system has been developed in our laboratory. It consists of one thousand neurons that are physically interconnected via one million 7-bit synapses. It can solve one thousand simultaneous nonlinear first-order differential equations in a fully parallel and continuous fashion. The performance of this system was measured by a winner-take-all network with one thousand neurons. Although the magnitude of the input and network parameters were identi(cid:173) cal for each competing neuron, one of them won in 6 milliseconds. This processing speed amounts to 360 billion connections per sec(cid:173) ond. A broad range of neural networks including spatiotemporal filtering, feedforward, and feedback networks can be run by loading appropriate network parameters from a host system.

************************************

Reinforcement Learning for Continuous Stochastic Control Problems
Rémi Munos, Paul Bourgine

This paper is concerned with the problem of Reinforcement Learn(cid:173) ing (RL) for continuous state space and time stocha.stic control problems. We state the Harnilton-Jacobi-Bellman equation satis(cid:173) fied by the value function and use a Finite-Difference method for designing a convergent approximation scheme. Then we propose a RL algorithm based on this scheme and prove its convergence to the optimal solution.

************************************

Intrusion Detection with Neural Networks
Jake Ryan, Meng-Jang Lin, Risto Miikkulainen

With the rapid expansion of computer networks during the past few years, security has become a crucial issue for modern computer systems. A good way to detect illegitimate use is through monitoring unusual user activity. Methods of intrusion detection based on hand-coded rule sets or predicting commands on-line are laborous to build or not very reliable. This paper proposes a new way of applying neural networks to detect intrusions. We believe that a user leaves a 'print' when using the system; a neural network can be used to learn this print and identify each user much like detectives use thumbprints to place people at crime scenes. If a user's behavior does not match hislher print, the system administrator can be alerted of a possible security breech. A backpropagation neural network called NNID (Neural Network Intrusion Detector) was trained in the identification task and tested experimentally on a system of 10 users. The system was 96% accurate in detecting unusual activity, with 7% false alarm rate. These results suggest that learning user profiles is an effective way for detecting intrusions.

************************************

Incorporating Contextual Information in White Blood Cell Identification
Xubo Song, Yaser Abu-Mostafa, Joseph Sill, Harvey Kasdan

In this paper we propose a technique to incorporate contextual informa(cid:173) tion into object classification. In the real world there are cases where the identity of an object is ambiguous due to the noise in the measurements based on which the classification should be made. It is helpful to re(cid:173) duce the ambiguity by utilizing extra information referred to as context, which in our case is the identities of the accompanying objects. This technique is applied to white blood cell classification. Comparisons are made against "no context" approach, which demonstrates the superior classification performance achieved by using context. In our particular application, it significantly reduces false alarm rate and thus greatly re(cid:173) duces the cost due to expensive clinical tests.

************************************

Function Approximation with the Sweeping Hinge Algorithm
Don Hush, Fernando Lozano, Bill Horne

We present a computationally efficient algorithm for function ap(cid:173) proximation with piecewise linear sigmoidal nodes. A one hidden layer network is constructed one node at a time using the method of fitting the residual. The task of fitting individual nodes is accom(cid:173) plished using a new algorithm that searchs for the best fit by solving a sequence of Quadratic Programming problems. This approach of(cid:173) fers significant advantages over derivative-based search algorithms (e.g. backpropagation and its extensions). Unique characteristics of this algorithm include: finite step convergence, a simple stop(cid:173) ping criterion, a deterministic methodology for seeking "good" local minima, good scaling properties and a robust numerical implemen(cid:173) tation.

************************************

Correlates of Attention in a Model of Dynamic Visual Recognition
Rajesh Rao

Given a set of objects in the visual field, how does the the visual system learn to attend to a particular object of interest while ignoring the rest? How are occlusions and background clutter so effortlessly discounted for when rec(cid:173) ognizing a familiar object? In this paper, we attempt to answer these ques(cid:173) tions in the context of a Kalman filter-based model of visual recognition that has previously proved useful in explaining certain neurophysiological phe(cid:173) nomena such as endstopping and related extra-classical receptive field ef(cid:173) fects in the visual cortex. By using results from the field of robust statistics, we describe an extension of the Kalman filter model that can handle multiple objects in the visual field. The resulting robust Kalman filter model demon(cid:173) strates how certain forms of attention can be viewed as an emergent prop(cid:173) erty of the interaction between top-down expectations and bottom-up sig(cid:173) nals. The model also suggests functional interpretations of certain attention(cid:173) related effects that have been observed in visual cortical neurons. Exper(cid:173) imental results are provided to help demonstrate the ability of the model to perform robust segmentation and recognition of objects and image se(cid:173) quences in the presence of varying degrees of occlusions and clutter.

************************************

Mapping a Manifold of Perceptual Observations
Joshua Tenenbaum

Nonlinear dimensionality reduction is formulated here as the problem of trying to find a Euclidean feature-space embedding of a set of observations that preser

ves as closely as possible their intrinsic metric structure - the distances between points on the observation manifold as measured along geodesic paths. Our isometric feature mapping procedure, or isomap, is able to reliably recover low-dimensional nonlinear structure in realistic perceptual data sets, such as a manifold of face images, where conventional global mapping methods find only local minima. The recovered map provides a canonical set of globally meaningful features, which allows perceptual transformations such as interpolation, extrapolation, and analogy - highly nonlinear transformations in the original observation space - to be computed with simple linear operations in feature space.
************************************
The Rectified Gaussian Distribution
Nicholas Socci, Daniel Lee, H. Sebastian Seung
A simple but powerful modification of the standard Gaussian dis(cid:173) tribution is studied. The variables of the rectified Gaussian are constrained to be nonnegative, enabling the use of nonconvex en(cid:173) ergy functions. Two multimodal examples, the competitive and cooperative distributions, illustrate the representational power of the rectified Gaussian. Since the cooperative distribution can rep(cid:173) resent the translations of a pattern, it demonstrates the potential of the rectified Gaussian for modeling pattern manifolds.
************************************
RCC Cannot Compute Certain FSA, Even with Arbitrary Transfer Functions
Mark Ring
Existing proofs demonstrating the computational limitations of Re(cid:173) current Cascade Correlation and similar networks (Fahlman, 1991; Bachrach, 1988; Mozer, 1988) explicitly limit their results to units having sigmoidal or hard-threshold transfer functions (Giles et aI., 1995; and Kremer, 1996). The proof given here shows that for any finite, discrete transfer function used by the units of an RCC network, there are finite-state automata (FSA) that the network cannot model, no matter how many units are used. The proof also applies to continuous transfer functions with a finite number of fixed-points, such as sigmoid and radial-basis functions.
************************************
Learning Generative Models with the Up Propagation Algorithm
Jong-Hoon Oh, H. Sebastian Seung
Up-propagation is an algorithm for inverting and learning neural network generative models Sensory input is processed by inverting a model that generates patterns from hidden variables using topdown connections The inversion process is iterative utilizing a negative feedback loop that depends on an error signal propagated by bottomup connections The error signal is also used to learn the generative model from examples The algorithm is benchmarked against principal component analysis in experiments on images of handwritten digits.
************************************
Serial Order in Reading Aloud: Connectionist Models and Neighborhood Structure
Jeanne Milostan, Garrison Cottrell
If globally high dimensional data has locally only low dimensional distribu(cid:173) tions, it is advantageous to perform a local dimensionality reduction before further processing the data. In this paper we examine several techniques for local dimensionality reduction in the context of locally weighted linear re(cid:173) gression. As possible candidates, we derive local versions of factor analysis regression, principle component regression, principle component regression on joint distributions, and partial least squares regression. After outlining the statistical bases of these methods, we perform Monte Carlo simulations to evaluate their robustness with respect to violations of their statistical as(cid:173) sumptions. One surprising outcome is that locally weighted partial least squares regression offers the best average results, thus outperforming even factor analysis, the theoretically most appealing of our candidate techniques.

************************************

Adaptation in Speech Motor Control

John Houde, Michael Jordan

Human subjects are known to adapt their motor behavior to a shift of the visual field brought about by wearing prism glasses over their eyes. We have studied the analog of this effect in speech. U sing a device that can feed back transformed speech signals in real time, we exposed subjects to alterations of their own speech feedback. We found that speakers learn to adjust their production of a vowel to compensate for feedback alterations that change the vowel's perceived phonetic identity; moreover, the effect generalizes across consonant contexts and to different vowels.

************************************

An Incremental Nearest Neighbor Algorithm with Queries

Joel Ratsaby

We consider the general problem of learning multi-category classifi(cid:173) cation from labeled examples. We present experimental results for a nearest neighbor algorithm which actively selects samples from different pattern classes according to a querying rule instead of the a priori class probabilities. The amount of improvement of this query-based approach over the passive batch approach depends on the complexity of the Bayes rule. The principle on which this al(cid:173) gorithm is based is general enough to be used in any learning algo(cid:173) rithm which permits a model-selection criterion and for which the error rate of the classifier is calculable in terms of the complexity of the model.

************************************

Competitive On-line Linear Regression

Volodya Vovk

We apply a general algorithm for merging prediction strategies (the Aggregating Algorithm) to the problem of linear regression with the square loss; our main assumption is that the response variable is bounded. It turns out that for this particular problem the Aggre(cid:173) gating Algorithm resembles, but is slightly different from, the well(cid:173) known ridge estimation procedure. From general results about the Aggregating Algorithm we deduce a guaranteed bound on the dif(cid:173) ference between our algorithm's performance and the best, in some sense, linear regression function's performance. We show that the AA attains the optimal constant in our bound, whereas the con(cid:173) stant attained by the ridge regression procedure in general can be 4 times worse.

************************************

Nonlinear Markov Networks for Continuous Variables

Reimar Hofmann, Volker Tresp

We address the problem oflearning structure in nonlinear Markov networks with continuous variables. This can be viewed as non-Gaussian multidi(cid:173) mensional density estimation exploiting certain conditional independencies in the variables. Markov networks are a graphical way of describing con(cid:173) ditional independencies well suited to model relationships which do not ex(cid:173) hibit a natural causal ordering. We use neural network structures to model the quantitative relationships between variables. The main focus in this pa(cid:173) per will be on learning the structure for the purpose of gaining insight into the underlying process. Using two data sets we show that interesting struc(cid:173) tures can be found using our approach. Inference will be briefly addressed.

************************************

On-line Learning from Finite Training Sets in Nonlinear Networks

Peter Sollich, David Barber

Online learning is one of the most common forms of neural net(cid:173) work training. We present an analysis of online learning from finite training sets for non-linear networks (namely, soft-committee ma(cid:173) chines), advancing the theory to more realistic learning scenarios. Dynamical equations are derived for an appropriate set of order parameters; these are ex act in the limiting case of either linear networks or infinite training sets. Preliminary comparisons with simulations suggest that the theory captures some effects of finite training sets, but may not yet account correctly for the presence of local minima.

************************************

Automated Aircraft Recovery via Reinforcement Learning: Initial Experiments

Jeffrey Monaco, David Ward, Andrew Barto

Initial experiments described here were directed toward using reinforce(cid:173) ment learning (RL) to develop an automated recovery system (ARS) for high-agility aircraft. An ARS is an outer-loop flight-control system de(cid:173) signed to bring an aircraft from a range of out-of-control states to straight(cid:173) and-level flight in minimum time while satisfying physical and phys(cid:173) iological constraints. Here we report on results for a simple version of the problem involving only single-axis (pitch) simulated recoveries. Through simulated control experience using a medium-fidelity aircraft simulation, the RL system approximates an optimal policy for pitch-stick inputs to produce minimum-time transitions to straight-and-Ievel flight in unconstrained cases while avoiding ground-strike. The RL system was also able to adhere to a pilot-station acceleration constraint while execut(cid:173) ing simulated recoveries.

************************************

Adaptive Choice of Grid and Time in Reinforcement Learning

Stephan Pareigis

We propose local error estimates together with algorithms for adap(cid:173) tive a-posteriori grid and time refinement in reinforcement learn(cid:173) ing. We consider a deterministic system with continuous state and time with infinite horizon discounted cost functional. For grid re(cid:173) finement we follow the procedure of numerical methods for the Bellman-equation. For time refinement we propose a new criterion, based on consistency estimates of discrete solutions of the Bellman(cid:173) equation. We demonstrate, that an optimal ratio of time to space discretization is crucial for optimal learning rates and accuracy of the approximate optimal value function.

************************************

Graph Matching with Hierarchical Discrete Relaxation

Richard Wilson, Edwin Hancock

Our aim in this paper is to develop a Bayesian framework for match(cid:173) ing hierarchical relational models. The goal is to make discrete la(cid:173) bel assignments so as to optimise a global cost function that draws information concerning the consistency of match from different lev(cid:173) els of the hierarchy. Our Bayesian development naturally distin(cid:173) guishes between intra-level and inter-level constraints. This allows the impact of reassigning a match to be assessed not only at its own (or peer) level ofrepresentation, but also upon its parents and children in the hierarchy.

************************************

Minimax and Hamiltonian Dynamics of Excitatory-Inhibitory Networks

H. Sebastian Seung, Tom Richardson, J. Lagarias, John J. Hopfield

A Lyapunov function for excitatory-inhibitory networks is constructed. The construction assumes symmetric interactions within excitatory and inhibitory populations of neurons, and antisymmetric interactions be(cid:173) tween populations. The Lyapunov function yields sufficient conditions for the global asymptotic stability of fixed points. If these conditions are violated, limit cycles may be stable. The relations of the Lyapunov function to optimization theory and classical mechanics are revealed by minimax and dissipative Hamiltonian forms of the network dynamics.

************************************

Hybrid NN/HMM-Based Speech Recognition with a Discriminant Neural Feature Extraction

Daniel Willett, Gerhard Rigoll

In this paper, we present a novel hybrid architecture for continuous speech recognition systems. It consists of a continuous HMM system extended by an arbitrary neural network that is used as a preprocessor that takes several frames of the feature vector as input to produce more discrimin(cid:173) ative feature vectors with respect to the underlying HMM system. This hybrid system is an extension of a state-of-the-art continuous HMM sys(cid:173) tem, and in fact, it is the first hybrid system that really is capable of outper(cid:173) forming these standard systems with respect to the recognition accuracy. Experimental results show an relative error reduction of about 10% that we achieved on a remarkably good recognition system based on continu(cid:173) ous HMMs for the Resource Management 1 OOO-word continuous speech recognition task.

************************************

2D Observers for Human 3D Object Recognition?

Zili Liu, Daniel Kersten

Converging evidence has shown that human object recognition depends on familiarity with the images of an object. Further, the greater the similarity between objects, the stronger is the dependence on object appearance, and the more important two(cid:173) dimensional (2D) image information becomes. These findings, how(cid:173) ever, do not rule out the use of 3D structural information in recog(cid:173) nition, and the degree to which 3D information is used in visual memory is an important issue. Liu, Knill, & Kersten (1995) showed that any model that is restricted to rotations in the image plane of independent 2D templates could not account for human perfor(cid:173) mance in discriminating novel object views. We now present results from models of generalized radial basis functions (GRBF), 2D near(cid:173) est neighbor matching that allows 2D affine transformations, and a Bayesian statistical estimator that integrates over all possible 2D affine transformations. The performance of the human observers relative to each of the models is better for the novel views than for the familiar template views, suggesting that humans generalize better to novel views from template views. The Bayesian estima(cid:173) tor yields the optimal performance with 2D affine transformations and independent 2D templates. Therefore, models of 2D affine matching operations with independent 2D templates are unlikely to account for human recognition performance.

************************************

A Neural Network Based Head Tracking System

Daniel D. Lee, H. Seung

We have constructed an inexpensive video based motorized tracking system that learns to track a head. It uses real time graphical user inputs or an auxiliary infrared detector as supervisory signals to train a convolutional neural network. The inputs to the neural network consist of normalized luminance and chrominance images and motion information from frame differences. Subsampled images are also used to provide scale invariance. During the online training phases the neural network rapidly adjusts the input weights depending up on the reliability of the different channels in the surrounding environment. This quick adaptation allows the system to robustly track a head even when other objects are moving within a cluttered background.

************************************

Perturbative M-Sequences for Auditory Systems Identification

Mark Kvale, Christoph Schreiner

In this paper we present a new method for studying auditory sys(cid:173) tems based on m-sequences. The method allows us to perturba(cid:173) tively study the linear response of the system in the presence of various other stimuli, such as speech or sinusoidal modulations. This allows one to construct linear kernels (receptive fields) at the same time that other stimuli are being presented. Using the method we calculate the modulation transfer function of single units in the inferior colli cui us of the cat at different operating points and discuss nonlinearities in the response.

************************************

Bach in a Box - Real-Time Harmony

Randall Spangler, Rodney Goodman, Jim Hawkins

We describe a system for learning J. S. Bach's rules of musical har(cid:173) mony. These rules are learned from examples and are expressed as rule-based neural networks. The rules are then applied in real(cid:173) time to generate new accompanying harmony for a live performer. Real-time functionality imposes constraints on the learning and harmonizing processes, including limitations on the types of infor(cid:173) mation the system can use as input and the amount of processing the system can perform. We demonstrate algorithms for gener(cid:173) ating and refining musical rules from examples which meet these constraints. We describe a method for including a priori knowl(cid:173) edge into the rules which yields significant performance gains. We then describe techniques for applying these rules to generate new music in real-time. We conclude the paper with an analysis of experimental results.

************************************

Use of a Multi-Layer Perceptron to Predict Malignancy in Ovarian Tumors

Herman Verrelst, Yves Moreau, Joos Vandewalle, Dirk Timmerman

This paper is concerned with the problem of Reinforcement Learn(cid:173) ing (RL) for continuous state space and time stocha.stic control problems. We state the Harnilton-Jacobi-Bellman equation satis(cid:173) fied by the value function and use a Finite-Difference method for designing a convergent approximation scheme. Then we propose a RL algorithm based on this scheme and prove its convergence to the optimal solution.

************************************

Learning Nonlinear Overcomplete Representations for Efficient Coding

Michael Lewicki, Terrence J. Sejnowski

We derive a learning algorithm for inferring an overcomplete basis by viewing it as probabilistic model of the observed data. Over(cid:173) complete bases allow for better approximation of the underlying statistical density. Using a Laplacian prior on the basis coefficients removes redundancy and leads to representations that are sparse and are a nonlinear function of the data. This can be viewed as a generalization of the technique of independent component anal(cid:173) ysis and provides a method for blind source separation of fewer mixtures than sources. We demonstrate the utility of overcom(cid:173) plete representations on natural speech and show that compared to the traditional Fourier basis the inferred representations poten(cid:173) tially have much greater coding efficiency.

************************************

Receptive Field Formation in Natural Scene Environments: Comparison of Single Cell Learning Rules

Brian Blais, Nathan Intrator, Harel Shouval, Leon Cooper

We study several statistically and biologically motivated learning rules using the same visual environment, one made up of natural scenes, and the same single cell neuronal architecture. This allows us to concentrate on the feature extraction and neuronal coding properties of these rules. Included in these rules are kurtosis and skewness maximization, the quadratic form of the BCM learning rule, and single cell ICA. Using a structure removal method, we demonstrate that receptive fields developed using these rules de(cid:173) pend on a small portion of the distribution. We find that the quadratic form of the BCM rule behaves in a manner similar to a kurtosis maximization rule when the distribution contains kurtotic directions, although the BCM modification equations are compu(cid:173) tationally simpler.

************************************

Reinforcement Learning for Call Admission Control and Routing in Integrated Service Networks

Peter Marbach, Oliver Mihatsch, Miriam Schulte, John Tsitsiklis

We provide a model of the standard watermaze task, and of a more challenging task involving novel platform locations, in which rats exhibit one-trial learning after a few days of training. The model uses hippocampal place cells to support reinforcement learning, and also, in an integrated manner, to build and use allocentric coordinates.

************************************

The Error Coding and Substitution PaCTs
Gareth James, Trevor Hastie

A new class of plug in classification techniques have recently been de(cid:173) veloped in the statistics and machine learning literature. A plug in clas(cid:173) sification technique (PaCT) is a method that takes a standard classifier (such as LDA or TREES) and plugs it into an algorithm to produce a new classifier. The standard classifier is known as the Plug in Classi(cid:173) fier (PiC). These methods often produce large improvements over using a single classifier. In this paper we investigate one of these methods and give some motivation for its success.

************************************

Modeling Complex Cells in an Awake Macaque during Natural Image Viewing
William Vinje, Jack Gallant

We model the responses of cells in visual area VI during natural vision. Our model consists of a classical energy mechanism whose output is divided by nonclassical gain control and texture contrast mechanisms. We apply this model to review movies, a stimulus sequence that replicates the stimulation a cell receives during free viewing of natural images. Data were collected from three cells using five different review movies, and the model was fit separately to the data from each movie. For the energy mechanism alone we find modest but significant correlations (rE = 0.41, 0.43, 0.59, 0.35) between model and data. These correlations are improved somewhat when we allow for suppressive surround effects (rE+G = 0.42, 0.56, 0.60, 0.37). In one case the inclusion of a delayed suppressive surround dramatically improves the fit to the data by modifying the time course of the model's response.

************************************

Generalization in Decision Trees and DNF: Does Size Matter?
Mostefa Golea, Peter Bartlett, Wee Sun Lee, Llew Mason

Recent theoretical results for pattern classification with thresh(cid:173) olded real-valued functions (such as support vector machines, sig(cid:173) moid networks, and boosting) give bounds on misclassification probability that do not depend on the size of the classifier, and hence can be considerably smaller than the bounds that follow from the VC theory. In this paper, we show that these techniques can be more widely applied, by representing other boolean functions as two-layer neural networks (thresholded convex combinations of boolean functions). For example, we show that with high probabil(cid:173) ity any decision tree of depth no more than d that is consistent with m training examples has misclassification probability no more than o ( (~ (Neff VCdim(U) log2 m log d)) 1/2), where U is the class of node decision functions, and Neff ::; N can be thought of as the effective number of leaves (it becomes small as the distribution on the leaves induced by the training data gets far from uniform). This bound is qualitatively different from the VC bound and can be considerably smaller. We use the same technique to give similar results for DNF formulae.

************************************

Local Dimensionality Reduction
Stefan Schaal, Sethu Vijayakumar, Christopher Atkeson

If globally high dimensional data has locally only low dimensional distribu(cid:173) tions, it is advantageous to perform a local dimensionality reduction before further processing the data. In this paper we examine several techniques for local dimensionality reduction in the context of locally weighted linear re(cid:173) gression. As possible candidates, we derive local versions of factor analysis regression, principle component regression, principle component regression on joint distributions, and partial least squares regression. After outlining the statistical bases of these methods, we perform Monte Carlo simulations to evaluate their robustness with respect to violations of their statistical as(cid:173) sumptions. One surprising outcome is that locally weighted partial least squares regression offers the best average results, thus outperforming even factor analysis, the theoretically most appealing of o

ur candidate techniques.
************************************

### A Mathematical Model of Axon Guidance by Diffusible Factors

Geoffrey Goodhill

In the developing nervous system, gradients of target-derived dif(cid:173) fusible factors play an important role in guiding axons to appro(cid:173) priate targets. In this paper, the shape that such a gradient might have is calculated as a function of distance from the target and the time since the start of factor production. Using estimates of the relevant parameter values from the experimental literature, the spatiotemporal domain in which a growth cone could detect such a gradient is derived. For large times, a value for the maximum guidance range of about 1 mm is obtained. This value fits well with experimental data. For smaller times, the analysis predicts that guidance over longer ranges may be possible. This prediction remains to be tested.
************************************

### Asymptotic Theory for Regularization: One-Dimensional Linear Case

Petri Koistinen

The generalization ability of a neural network can sometimes be improved dramatically by regularization. To analyze the improve(cid:173) ment one needs more refined results than the asymptotic distri(cid:173) bution of the weight vector. Here we study the simple case of one-dimensional linear regression under quadratic regularization, i.e., ridge regression. We study the random design, misspecified case, where we derive expansions for the optimal regularization pa(cid:173) rameter and the ensuing improvement. It is possible to construct examples where it is best to use no regularization.
************************************

### Instabilities in Eye Movement Control: A Model of Periodic Alternating Nystagmus

Ernst Dow, Thomas Anastasio

Nystagmus is a pattern of eye movement characterized by smooth rota(cid:173) tions of the eye in one direction and rapid rotations in the opposite di(cid:173) rection that reset eye position. Periodic alternating nystagmus (PAN) is a form of uncontrollable nystagmus that has been described as an un(cid:173) stable but amplitude-limited oscillation. PAN has been observed previ(cid:173) ously only in subjects with vestibulo-cerebellar damage. We describe results in which PAN can be produced in normal subjects by prolonged rotation in darkness. We propose a new model in which the neural cir(cid:173) cuits that control eye movement are inherently unstable, but this insta(cid:173) bility is kept in check under normal circumstances by the cerebellum. Circumstances which alter this cerebellar restraint, such as vestibulo(cid:173) cerebellar damage or plasticity due to rotation in darkness, can lead to PAN.
************************************

### Neural Basis of Object-Centered Representations

Sophie Denève, Alexandre Pouget

We present a neural model that can perform eye movements to a particular side of an object regardless of the position and orienta(cid:173) tion of the object in space, a generalization of a task which has been recently used by Olson and Gettner [4] to investigate the neu(cid:173) ral structure of object-centered representations. Our model uses an intermediate representation in which units have oculocentric recep(cid:173) tive fields- just like collicular neurons- whose gain is modulated by the side of the object to which the movement is directed, as well as the orientation of the object. We show that these gain modulations are consistent with Olson and Gettner's single cell recordings in the supplementary eye field. This demonstrates that it is possible to perform an object-centered task without a representation involv(cid:173) ing an object-centered map, viz., without neurons whose receptive fields are defined in object-centered coordinates. We also show that the same approach can account for object-centered neglect, a situ(cid:173) ation in which patients with a right parietal lesion neglect the left side of objects regardless of the orientation of the objects.

************************************

Modelling Seasonality and Trends in Daily Rainfall Data

Peter Williams

This paper presents a new approach to the problem of modelling daily rainfall using neural networks. We first model the conditional distribu(cid:173) tions of rainfall amounts, in such a way that the model itself determines the order of the process, and the time-dependent shape and scale of the conditional distributions. After integrating over particular weather pat(cid:173) terns, we are able to extract seasonal variations and long-term trends.

************************************

The Storage Capacity of a Fully-Connected Committee Machine

Yuansheng Xiong, Chulan Kwon, Jong-Hoon Oh

We study the storage capacity of a fully-connected committee ma(cid:173) chine with a large number K of hidden nodes. The storage capac(cid:173) ity is obtained by analyzing the geometrical structure of the weight space related to the internal representation . By examining the as(cid:173) ymptotic behavior of order parameters in the limit of large K, the storage capacity Q c is found to be proportional to ]{ Jln ]{ up to the leading order. This result satisfies the mathematical bound given by Mitchison and Durbin , whereas the replica-symmetric solution in a conventional Gardner's approach violates this bound.

************************************

Structural Risk Minimization for Nonparametric Time Series Prediction

Ron Meir

The problem of time series prediction is studied within the uniform con(cid:173) vergence framework of Vapnik and Chervonenkis. The dependence in(cid:173) herent in the temporal structure is incorporated into the analysis, thereby generalizing the available theory for memoryless processes. Finite sam(cid:173) ple bounds are calculated in terms of covering numbers of the approxi(cid:173) mating class, and the tradeoff between approximation and estimation is discussed. A complexity regularization approach is outlined, based on Vapnik's method of Structural Risk Minimization, and shown to be ap(cid:173) plicable in the context of mixing stochastic processes.

************************************

Selecting Weighting Factors in Logarithmic Opinion Pools

Tom Heskes

A simple linear averaging of the outputs of several networks as e.g. in bagging [3], seems to follow naturally from a bias/variance decomposition of the sum-squared error. The sum-squared error of the average model is a quadratic function of the weighting factors assigned to the networks in the ensemble [7], suggesting a quadratic programming algorithm for finding the "optimal" weighting factors. If we interpret the output of a network as a probability statement, the sum-squared error corresponds to minus the loglikelihood or the Kullback-Leibler divergence, and linear averaging of the out(cid:173) puts to logarithmic averaging of the probability statements: the logarithmic opinion pool. The crux of this paper is that this whole story about model aver(cid:173) aging, bias/variance decompositions, and quadratic programming to find the optimal weighting factors, is not specific for the sum(cid:173) squared error, but applies to the combination of probability state(cid:173) ments of any kind in a logarithmic opinion pool, as long as the Kullback-Leibler divergence plays the role of the error measure. As examples we treat model averaging for classification models under a cross-entropy error measure and models for estimating variances.

************************************

Reinforcement Learning with Hierarchies of Machines

Ronald Parr, Stuart Russell

We present a new approach to reinforcement learning in which the poli(cid:173) cies considered by the learning process are constrained by hierarchies of partially specified machines. This allows for the use of prior knowledge to reduce the search space and provides a framework in which knowledge can be transferred across problems and in which component solutions can be recombined to solve larger and more complicated problems. Our approach can be seen as providing a link between reinforcement learn(cid:173) ing and "behavior-based" or "teleo-reactive" approaches to control. We present provably convergent algorithms for problem-solving and learn(cid:173) ing with hierarchical machines and demonstrate their effectiveness on a problem with several thousand states.

************************************

### Multiplicative Updating Rule for Blind Separation Derived from the Method of Scoring

Howard Yang

For blind source separation, when the Fisher information matrix is used as the Riemannian metric tensor for the parameter space, the steepest descent algorithm to maximize the likelihood function in this Riemannian parameter space becomes the serial updating rule with equivariant property. This algorithm can be further simplified by using the asymptotic form of the Fisher information matrix around the equilibrium.

************************************

### On the Separation of Signals from Neighboring Cells in Tetrode Recordings

Maneesh Sahani, John Pezaris, Richard Andersen

We discuss a solution to the problem of separating waveforms pro(cid:173) duced by multiple cells in an extracellular neural recording. We take an explicitly probabilistic approach, using latent-variable mod(cid:173) els of varying sophistication to describe the distribution of wave(cid:173) forms produced by a single cell. The models range from a single Gaussian distribution of waveforms for each cell to a mixture of hidden Markov models. We stress the overall statistical structure of the approach, allowing the details of the generative model chosen to depend on the specific neural preparation.

************************************

### Task and Spatial Frequency Effects on Face Specialization

Matthew Dailey, Garrison Cottrell

There is strong evidence that face processing is localized in the brain. The double dissociation between prosopagnosia, a face recognition deficit occurring after brain damage, and visual object agnosia, difficulty recognizing otber kinds of complex objects, indicates tbat face and non(cid:173) face object recognition may be served by partially independent mecha(cid:173) nisms in the brain. Is neural specialization innate or learned? We sug(cid:173) gest that this specialization could be tbe result of a competitive learn(cid:173) ing mechanism that, during development, devotes neural resources to the tasks they are best at performing. Furtber, we suggest that the specializa(cid:173) tion arises as an interaction between task requirements and developmen(cid:173) tal constraints. In this paper, we present a feed-forward computational model of visual processing, in which two modules compete to classify input stimuli. When one module receives low spatial frequency infor(cid:173) mation and the other receives high spatial frequency information, and the task is to identify the faces while simply classifying the objects, the low frequency network shows a strong specialization for faces. No otber combination of tasks and inputs shows this strong specialization. We take these results as support for the idea that an innately-specified face processing module is unnecessary.

************************************

### Extended ICA Removes Artifacts from Electroencephalographic Recordings

Tzyy-Ping Jung, Colin Humphries, Te-Won Lee, Scott Makeig, Martin McKeown, Vicente Iragui, Terrence J. Sejnowski

Severe contamination of electroencephalographic (EEG) activity by eye movements, blinks, muscle, heart and line noise is a serious problem for EEG interpretation and analysis. Rejecting contami(cid:173) nated EEG segments results in a considerable loss of information and may be impractical for clinical data. Many methods have been proposed to remove eye movement and blink artifacts from EEG recordings. Often regression in the time or frequency domain is performed on simultaneous EEG and electrooculographic (EOG) recordings to derive parameters characterizing the appearance and spread of EOG artifacts in the EEG channels. However, EOG records also contain brain signals [1, 2], so regressing out EOG ac(cid:173) tivity inevitably involves subtracting a portion of the relevant EEG signal from each recording as well. Regression cannot be used to remove muscle noise or line noise, since these have no reference channels. Here , we propose a new and generally applicable method for removing a wide variety of artifacts from EEG records. The method is based on an extended version of a previous Indepen(cid:173) dent Component Analysis (lCA) algorithm [3, 4] for performing blind source separation on linear mixtures of independent source signals with either sub-Gaussian or super-Gaussian distributions. Our results show that ICA can effectively detect, separate and re(cid:173) move activity in EEG records from a wide variety of artifactual sources, with results comparing favorably to those obtained using regression-based methods.

************************************

Radial Basis Functions: A Bayesian Treatment

David Barber, Bernhard Schottky

Bayesian methods have been successfully applied to regression and classification problems in multi-layer perceptrons. We present a novel application of Bayesian techniques to Radial Basis Function networks by developing a Gaussian approximation to the posterior distribution which, for fixed basis function widths, is analytic in the parameters. The setting of regularization constants by cross(cid:173) validation is wasteful as only a single optimal parameter estimate is retained. We treat this issue by assigning prior distributions to these constants, which are then adapted in light of the data under a simple re-estimation formula.

************************************

Blind Separation of Radio Signals in Fading Channels

Kari Torkkola

We apply information maximization / maximum likelihood blind source separation [2, 6) to complex valued signals mixed with com(cid:173) plex valued nonstationary matrices. This case arises in radio com(cid:173) munications with baseband signals. We incorporate known source signal distributions in the adaptation, thus making the algorithms less "blind". This results in drastic reduction of the amount of data needed for successful convergence. Adaptation to rapidly changing signal mixing conditions, such as to fading in mobile communica(cid:173) tions, becomes now feasible as demonstrated by simulations.

************************************

Computing with Action Potentials

John J. Hopfield, Carlos Brody, Sam Roweis

Most computational engineering based loosely on biology uses contin(cid:173) uous variables to represent neural activity. Yet most neurons communi(cid:173) cate with action potentials. The engineering view is equivalent to using a rate-code for representing information and for computing. An increas(cid:173) ing number of examples are being discovered in which biology may not be using rate codes. Information can be represented using the timing of action potentials, and efficiently computed with in this representation. The "analog match" problem of odour identification is a simple problem which can be efficiently solved using action potential timing and an un(cid:173) derlying rhythm. By using adapting units to effect a fundamental change of representation of a problem, we map the recognition of words (hav(cid:173) ing uniform time-warp) in connected speech into the same analog match problem. We describe the architecture and preliminary results of such a recognition system. Using the fast events of biology in conjunction with an underlying rhythm is one way to overcome the limits of an event(cid:173) driven view of computation. When the intrinsic hardware is much faster than the time scale of change of inputs, this approach can greatly increase the effective computation per unit time on a given quantity of har dware.

************************************

New Approximations of Differential Entropy for Independent Component Analysis and Projection Pursuit

Aapo Hyvärinen

We derive a first-order approximation of the density of maximum entropy for a continuous 1-D random variable, given a number of simple constraints. This results in a density expansion which is somewhat similar to the classical polynomial density expansions by Gram-Charlier and Edgeworth. Using this approximation of density, an approximation of 1-D differential entropy is derived. The approximation of entropy is both more exact and more ro(cid:173) bust against outliers than the classical approximation based on the polynomial density expansions, without being computationally more expensive. The approximation has applications, for example, in independent component analysis and projection pursuit.

************************************

Independent Component Analysis for Identification of Artifacts in Magnetoencephalographic Recordings

Ricardo Vigário, Veikko Jousmäki, Matti Hämäläinen, Riitta Hari, Erkki Oja

We have studied the application of an independent component analysis (ICA) approach to the identification and possible removal of artifacts from a magnetoencephalographic (MEG) recording. This statistical tech(cid:173) nique separates components according to the kurtosis of their amplitude distributions over time, thus distinguishing between strictly periodical signals, and regularly and irregularly occurring signals. Many artifacts belong to the last category. In order to assess the effectiveness of the method, controlled artifacts were produced, which included saccadic eye movements and blinks, increased muscular tension due to biting and the presence of a digital watch inside the magnetically shielded room. The results demonstrate the capability of the method to identify and clearly isolate the produced artifacts.

************************************

Classification by Pairwise Coupling

Trevor Hastie, Robert Tibshirani

We discuss a strategy for polychotomous classification that involves estimating class probabilities for each pair of classes, and then cou(cid:173) pling the estimates together. The coupling model is similar to the Bradley-Terry method for paired comparisons. We study the na(cid:173) ture of the class probability estimates that arise, and examine the performance of the procedure in simulated datasets. The classifiers used include linear discriminants and nearest neighbors: applica(cid:173) tion to support vector machines is also briefly described.

************************************

Hybrid Reinforcement Learning and Its Application to Biped Robot Control

Satoshi Yamada, Akira Watanabe, Michio Nakashima

A learning system composed of linear control modules, reinforce(cid:173) ment learning modules and selection modules (a hybrid reinforce(cid:173) ment learning system) is proposed for the fast learning of real-world control problems. The selection modules choose one appropriate control module dependent on the state. This hybrid learning sys(cid:173) tem was applied to the control of a stilt-type biped robot. It learned the control on a sloped floor more quickly than the usual reinforce(cid:173) ment learning because it did not need to learn the control on a flat floor, where the linear control module can control the robot. When it was trained by a 2-step learning (during the first learning step, the selection module was trained by a training procedure con(cid:173) trolled only by the linear controller), it learned the control more quickly. The average number of trials (about 50) is so small that the learning system is applicable to real robot control.

************************************

Synaptic Transmission: An Information-Theoretic Perspective

Amit Manwani, Christof Koch

Here we analyze synaptic transmission from an infonnation-theoretic perspective. We derive c1osed-fonn expressions for the lower-bounds on the capacity of a simple model of a cortical synapse under two explicit coding paradigms. Under the "signal estimation" paradigm, we assume the signal to be encoded in the mean firing rate of a Poisson neuron. The perfonnance of an optimal linear estimator of the signal then provides a lower bound on the capacity for signal estimation. Under the "signal detection" paradigm, the presence or absence of the signal has to be de(cid:173) tected. Perfonnance of the optimal spike detector allows us to compute a lower bound on the capacity for signal detection. We find that single synapses (for empirically measured parameter values) transmit infonna(cid:173) tion poorly but significant improvement can be achieved with a small amount of redundancy.

************************************

Just One View: Invariances in Inferotemporal Cell Tuning

Maximilian Riesenhuber, Tomaso Poggio

In macaque inferotemporal cortex (IT), neurons have been found to re(cid:173) spond selectively to complex shapes while showing broad tuning ("in(cid:173) variance") with respect to stimulus transformations such as translation and scale changes and a limited tuning to rotation in depth. Training monkeys with novel, paperclip-like objects, Logothetis et al. 9 could in(cid:173) vestigate whether these invariance properties are due to experience with exhaustively many transformed instances of an object or if there are mech(cid:173) anisms that allow the cells to show response invariance also to previously unseen instances of that object. They found object-selective cells in an(cid:173) terior IT which exhibited limited invariance to various transformations after training with single object views. While previous models accounted for the tuning of the cells for rotations in depth and for their selectiv(cid:173) ity to a specific object relative to a population of distractor objects,14,1 the model described here attempts to explain in a biologically plausible way the additional properties of translation and size invariance. Using the same stimuli as in the experiment, we find that model IT neurons exhibit invariance properties which closely parallel those of real neurons. Simulations show that the model is capable of unsupervised learning of view-tuned neurons.

************************************

Agnostic Classification of Markovian Sequences

Ran El-Yaniv, Shai Fine, Naftali Tishby

Classification of finite sequences without explicit knowledge of their statistical nature is a fundamental problem with many important applications. We propose a new information theoretic approach to this problem which is based on the following ingredients: (i) se(cid:173) quences are similar when they are likely to be generated by the same source; (ii) cross entropies can be estimated via "universal compres(cid:173) sion"; (iii) Markovian sequences can be asymptotically-optimally merged. With these ingredients we design a method for the classification of discrete sequences whenever they can be compressed. We introduce the method and illustrate its application for hierarchical clustering of languages and for estimating similarities of protein sequences.

************************************

Visual Navigation in a Robot Using Zig-Zag Behavior

M. Lewis

We implement a model of obstacle avoidance in flying insects on a small, monocular robot. The result is a system that is capable of rapid navigation through a dense obstacle field. The key to the system is the use of zigzag behavior to articulate the body during movement. It is shown that this behavior compensates for a parallax blind spot surrounding the focus of expansion nor(cid:173) mally found in systems without parallax behavior. The system models the coop(cid:173) eration of several behaviors: halteres-ocular response (similar to VOR), optomotor response, and the parallax field computation and mapping to motor system. The resulting system is neurally plausible, very simple, and should be easily h osted on a VLSI hardware.

************************************

Generalized Prioritized Sweeping

David Andre, Nir Friedman, Ronald Parr

Prioritized sweeping is a model-based reinforcement learning method that attempts to focus an agent's limited computational resources to achieve a good estimate of the value of environment states. To choose ef(cid:173) fectively where to spend a costly planning step, classic prioritized sweep(cid:173) ing uses a simple heuristic to focus computation on the states that are likely to have the largest errors. In this paper, we introduce generalized prioritized sweeping, a principled method for generating such estimates in a representation-specific manner. This allows us to extend prioritized sweeping beyond an explicit, state-based representation to deal with com(cid:173) pact representations that are necessary for dealing with large state spaces. We apply this method for generalized model approximators (such as Bayesian networks), and describe preliminary experiments that compare our approach with classical prioritized sweeping.

************************************

Multiresolution Tangent Distance for Affine-invariant Classification

Nuno Vasconcelos, Andrew Lippman

The ability to rely on similarity metrics invariant to image transforma(cid:173) tions is an important issue for image classification tasks such as face or character recognition. We analyze an invariant metric that has performed well for the latter - the tangent distance - and study its limitations when applied to regular images, showing that the most significant among these (convergence to local minima) can be drastically reduced by computing the distance in a multiresolution setting. This leads to the multi resolution tangent distance, which exhibits significantly higher invariance to im(cid:173) age transformations, and can be easily combined with robust estimation procedures.

************************************

Effects of Spike Timing Underlying Binocular Integration and Rivalry in a Neural Model of Early Visual Cortex

Erik Lumer

In normal vision, the inputs from the two eyes are inte(cid:173) grated into a single percept. When dissimilar images are presented to the two eyes, however, perceptual integra(cid:173) tion gives way to alternation between monocular inputs, a phenomenon called binocular rivalry. Although recent evidence indicates that binocular rivalry involves a mod(cid:173) ulation of neuronal responses in extrastriate cortex, the basic mechanisms responsible for differential processing of con:6.icting and congruent stimuli remain unclear. Using a neural network that models the mammalian early visual system, I demonstrate here that the desynchronized fir(cid:173) ing of cortical-like neurons that first receive inputs from the two eyes results in rivalrous activity patterns at later stages in the visual pathway. By contrast, synchronization of firing among these cells prevents such competition. The temporal coordination of cortical activity and its effects on neural competition emerge naturally from the network connectivity and from its dynamics. These results suggest that input-related differences in relative spike timing at an early stage of visual processing may give rise to the phenomena both of perceptual integration and rivalry in binocular vision.

************************************

Detection of First and Second Order Motion

Alexander Grunewald, Heiko Neumann

A model of motion detection is presented. The model contains three stages. The first stage is unoriented and is selective for con(cid:173) trast polarities. The next two stages work in parallel. A phase insensitive stage pools across different contrast polarities through a spatiotemporal filter and thus can detect first and second order motion. A phase sensitive stage keeps contrast polarities separate, each of which is filtered through a spatiotemporal filter, and thus only first order motion can be detected. Differential phase sensitiv(cid:173) ity can therefore account for the detectio n of first and second order motion. Phase insensitive detectors correspond to cortical complex cells, and phase sensitive detectors to simple cells.

************************************

A Framework for Multiple-Instance Learning

Oded Maron, Tomás Lozano-Pérez

Multiple-instance learning is a variation on supervised learning, where the task is to learn a concept given positive and negative bags of instances. Each bag may contain many instances, but a bag is labeled positive even if only one of the instances in it falls within the concept. A bag is labeled negative only if all the instances in it are negative. We describe a new general framework, called Diverse Density, for solving multiple-instance learning problems. We apply this framework to learn a simple description of a person from a series of images (bags) containing that person, to a stock selection problem, and to the drug activity prediction problem.

************************************

Monotonic Networks

Joseph Sill

Monotonicity is a constraint which arises in many application do(cid:173) mains. We present a machine learning model, the monotonic net(cid:173) work, for which monotonicity can be enforced exactly, i.e., by virtue offunctional form . A straightforward method for implementing and training a monotonic network is described. Monotonic networks are proven to be universal approximators of continuous, differen(cid:173) tiable monotonic functions. We apply monotonic networks to a real-world task in corporate bond rating prediction and compare them to other approaches.

************************************

Relative Loss Bounds for Multidimensional Regression Problems

Jyrki Kivinen, Manfred K. K. Warmuth

We study on-line generalized linear regression with multidimensional outputs, i.e., neural networks with multiple output nodes but no hidden nodes. We allow at the final layer transfer functions such as the soft(cid:173) max function that need to consider the linear activations to all the output neurons. We use distance functions of a certain kind in two completely independent roles in deriving and analyzing on-line learning algorithms for such tasks. We use one distance function to define a matching loss function for the (possibly multidimensional) transfer function, which al(cid:173) lows us to generalize earlier results from one-dimensional to multidimen(cid:173) sional outputs. We use another distance function as a tool for measuring progress made by the on-line updates. This shows how previously stud(cid:173) ied algorithms such as gradient descent and exponentiated gradient fit into a common framework. We evaluate the performance of the algo(cid:173) rithms using relative loss bounds that compare the loss of the on-line algoritm to the best off-line predictor from the relevant model class, thus completely eliminating probabilistic assumptions about the data.

************************************

Gradients for Retinotectal Mapping

Geoffrey Goodhill

The initial activity-independent formation of a topographic map in the retinotectal system has long been thought to rely on the matching of molecular cues expressed in gradients in the retina and the tectum. However, direct experimental evidence for the existence of such gradients has only emerged since 1995. The new data has provoked the discussion of a new set of models in the ex(cid:173) perimentalliterature. Here, the capabilities of these models are an(cid:173) alyzed, and the gradient shapes they predict in vivo are derived.

************************************

Multi-modular Associative Memory

Nir Levy, David Horn, Eytan Ruppin

Motivated by the findings of modular structure in the association cortex, we study a multi-modular model of associative memory that can successfully store memory patterns with different levels of ac(cid:173) tivity. We show th at the segregation of synaptic conductances into intra-modular linear and inter-modular nonlinear ones considerably enhances the network's memory retrieval performance. Compared with the conventional, single-module associative memory network, the multi-modular network has two main advantages: It is less sus(cid:173) ceptible to damage to columnar input, and its response is consistent with the cognitive data pertaining to category specific impairment.

************************************

Analysis of Drifting Dynamics with Neural Network Hidden Markov Models

Jens Kohlmorgen, Klaus-Robert Müller, Klaus Pawelzik

We present a method for the analysis of nonstationary time se(cid:173) ries with multiple operating modes. In particular, it is possible to detect and to model both a switching of the dynamics and a less abrupt, time consuming drift from one mode to another. This is achieved in two steps. First, an unsupervised training method pro(cid:173) vides prediction experts for the inherent dynamical modes. Then, the trained experts are used in a hidden Markov model that allows to model drifts. An application to physiological wake/sleep data demonstrates that analysis and modeling of real-world time series can be improved when the drift paradigm is taken into account.

************************************

The Observer-Observation Dilemma in Neuro-Forecasting

Hans-Georg Zimmermann, Ralph Neuneier

We explain how the training data can be separated into clean informa(cid:173) tion and unexplainable noise. Analogous to the data, the neural network is separated into a time invariant structure used for forecasting, and a noisy part. We propose a unified theory connecting the optimization al(cid:173) gorithms for cleaning and learning together with algorithms that control the data noise and the parameter noise. The combined algorithm allows a data-driven local control of the liability of the network parameters and therefore an improvement in generalization. The approach is proven to be very useful at the task of forecasting the German bond market.

************************************

A Model of Early Visual Processing

Laurent Itti, Jochen Braun, Dale Lee, Christof Koch

We propose a model for early visual processing in primates. The model consists of a population of linear spatial filters which inter(cid:173) act through non-linear excitatory and inhibitory pooling. Statisti(cid:173) cal estimation theory is then used to derive human psychophysical thresholds from the responses of the entire population of units. The model is able to reproduce human thresholds for contrast and ori(cid:173) entation discrimination tasks, and to predict contrast thresholds in the presence of masks of varying orientation and spatial frequency.

************************************

A Simple and Fast Neural Network Approach to Stereovision

Rolf Henkel

A neural network approach to stereovision is presented based on aliasing effects of simple disparity estimators and a fast coherence(cid:173) detection scheme. Within a single network structure, a dense dis(cid:173) parity map with an associated validation map and, additionally, the fused cyclopean view of the scene are available. The network operations are based on simple, biological plausible circuitry; the algorithm is fully parallel and non-iterative.

************************************

Data-Dependent Structural Risk Minimization for Perceptron Decision Trees

John Shawe-Taylor, Nello Cristianini

A novel neural network model of pre-attention processing in visual(cid:173) search tasks is presented. Using displays of line orientations taken from Wolfe's experiments [1992], we study the hypothesis that the distinction between parallel versus serial processes arises from the availability of global information in the internal representations of the visual scene. The model operates in two phases. First, the visual displays are compressed via principal-component-analysis. Second, the compressed data is processed by a target detector mod(cid:173) ule in order to identify the existence of a target in the display. Our main finding is that targets in displays which were found exper(cid:173) imentally to be processed in parallel can be detected by the sys(cid:173) tem, while targets in experimentally-serial displays cannot . This fundamental difference is explained via variance analysis of the compressed representations, providing a numerical criterion distin(cid:173) guishing parallel from serial displays. Our model yields a mapping of response-time slopes that is similar to Duncan and Humphreys's "search surface" [1989], providing an explicit formulation of their intuitive notion of feature similarity. It presents a neural realiza(cid:173) tion of the processing that may underlie the classical metaphorical explanations of visual search.

************************************

Using Expectation to Guide Processing: A Study of Three Real-World Applications
Shumeet Baluja

In many real world tasks, only a small fraction of the available inputs are important at any particular time. This paper presents a method for ascertaining the relevance of inputs by exploiting temporal coherence and predictability . The method pro(cid:173) posed in this paper dynamically allocates relevance to inputs by using expectations of their future values. As a model of the task is learned, the model is simulta(cid:173) neously extended to create task-specific predictions of the future values of inputs. Inputs which are either not relevant, and therefore not accounted for in the model, or those which contain noise, will not be predicted accurately. These inputs can be de-emphasized, and, in turn, a new, improved, model of the task created. The tech(cid:173) niques presented in this paper have yielded significant improvements for the vision-based autonomous control of a land vehicle, vision-based hand tracking in cluttered scenes, and the detection of faults in the etching of semiconductor wafers.

************************************

Features as Sufficient Statistics
Davi Geiger, Archisman Rudra, Laurance Maloney

An image is often represented by a set of detected features. We get an enormous compression by representing images in this way. Fur(cid:173) thermore, we get a representation which is little affected by small amounts of noise in the image. However, features are typically chosen in an ad hoc manner. tures can be obtained using sufficient statistics. The idea of sparse data representation naturally arises. We treat the I-dimensional and 2-dimensional signal reconstruction problem to make our ideas concrete.

************************************

Factorizing Multivariate Function Classes
Juan Lin

The mathematical framework for factorizing equivalence classes of multivariate functions is formulated in this paper. Independent component analysis is shown to be a special case of this decompo(cid:173) sition. Using only the local geometric structure of a class repre(cid:173) sentative, we derive an analytic solution for the factorization. We demonstrate the factorization solution with numerical experiments and present a preliminary tie to decorrelation.

************************************

An Annealed Self-Organizing Map for Source Channel Coding
Matthias Burger, Thore Graepel, Klaus Obermayer

We derive and analyse robust optimization schemes for noisy vector quantization on the basis of deterministic annealing. Starting from a cost function for central clustering that incorporates distortions from channel noise we develop a soft topographic vector quantization al(cid:173) gorithm (STVQ) which is based on the maximum entropy principle and which performs a maximum-likelihood estimate in an expectation(cid:173) maximization (EM) fashion. Annealing in the temperature parameter f3 leads to phase tr ansitions in the existing code vector representation dur(cid:173) ing the cooling process for which we calculate critical temperatures and modes as a function of eigenvectors and eigenvalues of the covariance matrix of the data and the transition matrix of the channel noise. A whole family of vector quantization algorithms is derived from STVQ, among them a deterministic annealing scheme for Kohonen's self-organizing map (SOM). This algorithm, which we call SSOM, is then applied to vector quantization of image data to be sent via a noisy binary symmetric channel. The algorithm's performance is compared to those of LBG and STVQ. While it is naturally superior to LBG, which does not take into account channel noise, its results compare very well to those of STVQ, which is computationally much more demanding.

************************************

How to Dynamically Merge Markov Decision Processes

Satinder Singh, David Cohn

We are frequently called upon to perform multiple tasks that com(cid:173) pete for our attention and resource. Often we know the optimal solution to each task in isolation; in this paper, we describe how this knowledge can be exploited to efficiently find good solutions for doing the tasks in parallel. We formulate this problem as that of dynamically merging multiple Markov decision processes (MDPs) into a composite MDP, and present a new theoretically-sound dy(cid:173) namic programming algorithm for finding an optimal policy for the composite MDP. We analyze various aspects of our algorithm and illustrate its use on a simple merging problem.

************************************

Refractoriness and Neural Precision

Michael Berry, Markus Meister

The relationship between a neuron's refractory period and the precision of its response to identical stimuli was investigated. We constructed a model of a spiking neuron that combines probabilistic firing with a refractory period. For realistic refractoriness, the model closely reproduced both the average firing rate and the response precision of a retinal ganglion cell. The model is based on a "free" firing rate, which exists in the absence of refractoriness. This function may be a better description of a spiking neuron's response than the peri-stimulus time histogram.

************************************

Active Data Clustering

Thomas Hofmann, Joachim Buhmann

Active data clustering is a novel technique for clustering of proxim(cid:173) ) ity data which utilizes principles from sequential experiment design in order to interleave data generation and data analysis. The pro(cid:173) posed active data sampling strategy is based on the expected value of information, a concept rooting in statistical decision theory. This is considered to be an important step towards the analysis of large(cid:173) scale data sets, because it offers a way to overcome the inherent data sparseness of proximity data. '''Ie present applications to unsu(cid:173) pervised texture segmentation in computer vision and information retrieval in document databases.

************************************

Coding of Naturalistic Stimuli by Auditory Midbrain Neurons

Hagai Attias, Christoph Schreiner

It is known that humans can make finer discriminations between familiar sounds (e.g. syllables) than between unfamiliar ones (e.g. different noise segments). Here we show that a corresponding en(cid:173) hancement is present in early auditory processing stages. Based on previous work which demonstrated that natural sounds had robust statistical properties that could be quantified, we hypothesize that the auditory system exploits those properties to construct efficient neural codes. To test this hypothesis, we measure the informa(cid:173) tion rate carried by auditory spike trains on narrow-band stimuli whose amplitude modulation has naturalistic characteristics, and compare it to the information rate on stimuli with non-naturalistic modulation. We find that naturalistic inputs significantly enhance the rate of transmitte d information, indicating that auditiory neu(cid:173) ral responses are matched to characteristics of natural auditory scenes.

************************************

### Enhancing Q-Learning for Optimal Asset Allocation

Ralph Neuneier

This paper enhances the Q-Iearning algorithm for optimal asset alloca(cid:173) tion proposed in (Neuneier, 1996 [6]). The new formulation simplifies the approach by using only one value-function for many assets and al(cid:173) lows model-free policy-iteration. After testing the new algorithm on real data, the possibility of risk management within the framework of Markov decision problems is analyzed. The proposed methods allows the construction of a multi-period portfolio management system which takes into account transaction costs, the risk preferences of the investor, and several constraints on the allocation.

************************************

### Nonparametric Model-Based Reinforcement Learning

Christopher Atkeson

This paper describes some of the interactions of model learning algorithms and planning algorithms we have found in exploring model-based reinforcement learning. The paper focuses on how lo(cid:173) cal trajectory optimizers can be used effectively with learned non(cid:173) parametric models. We find that trajectory planners that are fully consistent with the learned model often have difficulty finding rea(cid:173) sonable plans in the early stages of learning. Trajectory planners that balance obeying the learned model with minimizing cost (or maximizing reward) often do better, even if the plan is not fully consistent with the learned model.

************************************

### On Parallel versus Serial Processing: A Computational Study of Visual Search

Eyal Cohen, Eytan Ruppin

A novel neural network model of pre-attention processing in visual(cid:173) search tasks is presented. Using displays of line orientations taken from Wolfe's experiments [1992], we study the hypothesis that the distinction between parallel versus serial processes arises from the availability of global information in the internal representations of the visual scene. The model operates in two phases. First, the visual displays are compressed via principal-component-analysis. Second, the compressed data is processed by a target detector mod(cid:173) ule in order to identify the existence of a target in the display. Our main finding is that targets in displays which were found exper(cid:173) imentally to be processed in parallel can be detected by the sys(cid:173) tem, while targets in experimentally-serial displays cannot . This fundamental difference is explained via variance analysis of the compressed representations, providing a numerical criterion distin(cid:173) guishing parallel from serial displays. Our model yields a mapping of response-time slopes that is similar to Duncan and Humphreys's "search surface" [1989], providing an explicit formulation of their intuitive notion of feature similarity. It presents a neural realiza(cid:173) tion of the processing that may underlie the classical metaphorical explanations of visual search.

************************************

### A Superadditive-Impairment Theory of Optic Aphasia

Michael C. Mozer, Mark Sitton, Martha Farah

Accounts of neurological disorders often posit damage to a specific functional pathway of the brain. Farah (1990) has proposed an alterna(cid:173) tive class of explanations involving partial damage to multiple path(cid:173) ways. We explore this explanation for optic aphasia, a disorder in which severe perfonnance deficits are observed when patients are asked to name visually presented objects, but surprisingly, performance is rela(cid:173) tively nonnal on naming objects from auditory cues and on gesturing the appropriate use of visually presented objects. We model this highly specific deficit through partial damage to two pathways-one that maps visual input to semantics, and the other that maps semantics to naming responses. The effect of this damage is superadditive, meaning that tasks which require one pathway or the other show little or no perfor(cid:173) mance deficit, but the damage is manifested when a task requires both pathways (i.e., naming visually presented objects). Our model explains other phenomena associated with optic aphasia, and makes testable experimental predictions.

************************************

S-Map: A Network with a Simple Self-Organization Algorithm for Generative Topographic Mappings

Kimmo Kiviluoto, Erkki Oja

The S-Map is a network with a simple learning algorithm that com(cid:173) bines the self-organization capability of the Self-Organizing Map (SOM) and the probabilistic interpretability of the Generative To(cid:173) pographic Mapping (GTM). The simulations suggest that the S(cid:173) Map algorithm has a stronger tendency to self-organize from ran(cid:173) dom initial configuration than the GTM. The S-Map algorithm can be further simplified to employ pure Hebbian learning, with(cid:173) out changing the qualitative behaviour of the network.

************************************

Bayesian Model of Surface Perception

William Freeman, Paul Viola

Image intensity variations can result from several different object surface effects, including shading from 3-dimensional relief of the object, or paint on the surface itself. An essential problem in vision , which people solve naturally, is to attribute the proper physical cause, e.g. surface relief or paint, to an observed image. We ad(cid:173) dressed this problem with an approach combining psychophysical and Bayesian computational methods. We assessed human performance on a set of test images, and found that people made fairly consistent judgements of surface properties. Our computational model assigned simple prior probabilities to different relief or paint explanations for an image, and solved for the most probable interpretation in a Bayesian framework. The ratings of the test images by our algorithm compared surprisingly well with the mean ratings of our subjects.

************************************

Training Methods for Adaptive Boosting of Neural Networks

Holger Schwenk, Yoshua Bengio

"Boosting" is a general method for improving the performance of any learning algorithm that consistently generates classifiers which need to perform only slightly better than random guessing. A recently proposed and very promising boosting algorithm is AdaBoost [5]. It has been ap(cid:173) plied with great success to several benchmark machine learning problems using rather simple learning algorithms [4], and decision trees [1, 2, 6]. In this paper we use AdaBoost to improve the performances of neural networks. We compare training methods based on sampling the training set and weighting the cost function. Our system achieves about 1.4% error on a data base of online handwritten digits from more than 200 writers. Adaptive boosting of a multi-layer network achieved 1.5% error on the UCI Letters and 8.1 % error on the UCI satellite data set.

************************************

An Analog VLSI Model of the Fly Elementary Motion Detector

Reid Harrison, Christof Koch

Flies are capable of rapidly detecting and integrating visual motion in(cid:173) formation in behaviorly-relevant ways. The first stage of visual motion processing in flies is a retinotopic array of functional units known as el(cid:173) ementary motion detectors (EMDs). Several decades ago, Reichardt and colleagues developed a correlation-based model of motion detection that described the behavior of these neural circuits. We have implemented a variant of this model in a 2.0-JLm analog CMOS VLSI process. The re(cid:173) sult is a low-power, continuous-time analog circuit with integrated pho(cid:173) toreceptors that responds to motion in real time. The responses of the circuit to drifting sinusoidal gratings qualitatively resemble the temporal frequency response, spatial frequency response, and direction selectivity of motion-sensitive neurons observed in insects. In addition to its pos(cid:173) sible engineering applications, the circuit could potentially be used as a building block for constructing hardware models of higher-level insect motion integration.

************************************

Experiences with Bayesian Learning in a Real World Application

Peter Sykacek, Georg Dorffner, Peter Rappelsberger, Josef Zeitlhofer

This paper reports about an application of Bayes' inferred neu(cid:173) ral network classifiers in the field of automatic sleep staging. The reason for using Bayesian learning for this task is two-fold. First, Bayesian inference is known to embody regularization automati(cid:173) cally. Second, a side effect of Bayesian learning leads to larger variance of network outputs in regions without training data. This results in well known moderation effects, which can be used to detect outliers. In a 5 fold cross-validation experiment the full Bayesian solution found with R. Neals hybrid Monte Carlo algo(cid:173) rithm, was not better than a single maximum a-posteriori (MAP) solution found with D.J. MacKay's evidence approximation. In a second experiment we studied the properties of both solutions in rejecting classification of movement artefacts.

************************************

Analytical Study of the Interplay between Architecture and Predictability

Avner Priel, Ido Kanter, David Kessler

We study model feed forward networks as time series predictors in the stationary limit. The focus is on complex, yet non-chaotic, behavior. The main question we address is whether the asymptotic behavior is governed by the architecture, regardless the details of the weights . We find hierarchies among classes of architectures with respect to the attract or dimension of the long term sequence they are capable of generating; larger number of hidden units can generate higher dimensional attractors. In the case of a perceptron, we develop the stationary solution for general weights, and show that the flow is typically one dimensional. The relaxation time from an arbitrary initial condition to the stationary solution is found to scale linearly with the size of the network. In multilayer networks, the number of hidden units gives bounds on the number and dimension of the possible attractors. We conclude that long term prediction (in the non-chaotic regime) with such models is governed by attractor dynamics related to the architecture.

************************************

A Hippocampal Model of Recognition Memory

Randall O'Reilly, Kenneth Norman, James McClelland

A rich body of data exists showing that recollection of specific infor(cid:173) mation makes an important contribution to recognition memory, which is distinct from the contribution of familiarity, and is not adequately cap(cid:173) tured by existing unitary memory models. Furthennore, neuropsycholog(cid:173) ical evidence indicates that recollection is sub served by the hippocampus. We present a model, based largely on known features of hippocampal anatomy and physiology, that accounts for the following key character(cid:173) istics of recollection: 1) false recollection is rare (i.e., participants rarely claim to recollect having studied nonstudied items), and 2) increasing in(cid:173) terference leads to less recollection but apparently does not compromise the quality of recollection (i.e., the extent to which recollected infonna(cid:173) tion veridically reflects events that occurred at study).

************************************

Wavelet Models for Video Time-Series

Sheng Ma, Chuanyi Ji

In this work, we tackle the problem of time-series modeling of video traffic. Different from the existing methods which model the time(cid:173) series in the time domain, we model the wavelet coefficients in the wavelet domain. The strength of the wavelet model includes (1) a unified approach to model both the long-range and the short-range dependence in the video traffic simultaneously, (2) a computation(cid:173) ally efficient method on developing the model and generating high quality video traffic, and (3) feasibility of performance analysis us(cid:173) ing the model.

************************************

Multi-time Models for Temporally Abstract Planning

Doina Precup, Richard S. Sutton

Planning

************************************

Linear Concepts and Hidden Variables: An Empirical Study

Adam Grove, Dan Roth

Some learning techniques for classification tasks work indirectly, by first trying to fit a full probabilistic model to the observed data. Whether this is a good idea or not depends on the robustness with respect to deviations from the postulated model. We study this question experimentally in a restricted, yet non-trivial and interesting case: we consider a conditionally independent attribute (CIA) model which postulates a single binary-valued hidden variable z on which all other attributes (i.e., the target and the observables) depend. In this model, finding the most likely value of anyone variable (given known values for the others) reduces to testing a linear function of the observed values.

************************************

Comparison of Human and Machine Word Recognition

Markus Schenkel, Cyril Latimer, Marwan Jabri

We present a study which is concerned with word recognition rates for heavily degraded documents. We compare human with machine read(cid:173) ing capabilities in a series of experiments, which explores the interaction of word/non-word recognition, word frequency and legality of non-words with degradation level. We also study the influence of character segmen(cid:173) tation, and compare human performance with that of our artificial neural network model for reading. We found that the proposed computer model uses word context as efficiently as humans, but performs slightly worse on the pure character recognition task.

************************************

Characterizing Neurons in the Primary Auditory Cortex of the Awake Primate Using Reverse Correlation

R. DeCharms, Michael Merzenich

While the understanding of the functional role of different classes of neurons in the awake primary visual cortex has been extensively studied since the time of Hubel and Wiesel (Hubel and Wiesel, 1962), our understanding of the feature selectivity and functional role of neurons in the primary auditory cortex is much farther from com(cid:173) plete. Moving bars have long been recognized as an optimal stimulus for many visual cortical neurons, and this finding has recently been confirmed and extended in detail using reverse correlation methods (Jones and Palmer, 1987; Reid and Alonso, 1995; Reid et al., 1991; llingach et al., 1997). In this study, we recorded from neurons in the primary auditory cortex of the awake primate, and used a novel re(cid:173) verse correlation technique to compute receptive fields (or preferred stimuli), encompassing both multiple frequency components and on(cid:173) going time. These spectrotemporal receptive fields make clear that neurons in the primary auditory cortex, as in the primary visual cor(cid:173) tex, typically show considerable structure in their feature processing properties, often including multiple excitatory and inhibitory regions in their receptive fields. These neurons can be sensitive to stimulus edges in frequency composition or in time, and sensitive to stimulus transitions such as changes in frequency. These neurons also show strong responses and selectivity to continuous frequency modulated stimuli analogous to visual drifting gratings.

************************************

Phase Transitions and the Perceptual Organization of Video Sequences

Yair Weiss

Estimating motion in scenes containing multiple moving objects remains a difficult problem in computer vision. A promising ap(cid:173) proach to this problem involves using mixture models, where the motion of each object is a component in the mixture. However, ex(cid:173) isting methods typically require specifying in advance the number of components in the mixture , i.e. the number of objects in the scene.

************************************

Regression with Input-dependent Noise: A Gaussian Process Treatment

Paul Goldberg, Christopher Williams, Christopher Bishop

Gaussian processes provide natural non-parametric prior distribu(cid:173) tions over regression functions. In this paper we consider regression problems where there is noise on the output, and the variance of the noise depends on the inputs. If we assume that the noise is a smooth function of the inputs, then it is natural to model the noise variance using a second Gaussian process, in addition to the Gaussian process governing the noise-free output value. We show that prior uncertainty about the parameters controlling both pro(cid:173) cesses can be handled and that the posterior distribution of the noise rate can be sampled from using Markov chain Monte Carlo methods. Our results on a synthetic data set give a posterior noise variance that well-approximates the true variance.

************************************

A Generic Approach for Identification of Event Related Brain Potentials via a Competitive Neural Network Structure

Daniel Lange, Hava Siegelmann, Hillel Pratt, Gideon Inbar

We present a novel generic approach to the problem of Event Related Potential identification and classification, based on a competitive N eu(cid:173) ral Net architecture. The network weights converge to the embedded signal patterns, resulting in the formation of a matched filter bank. The network performance is analyzed via a simulation study, exploring identification robustness under low SNR conditions and compared to the expected performance from an information theoretic perspective. The classifier is applied to real event-related potential data recorded during a classic odd-ball type paradigm; for the first time, within(cid:173) session variable signal patterns are automatically identified, dismiss(cid:173) ing the strong and limiting requirement of a-priori stimulus-related selective grouping of the recorded data.

************************************

On the Infeasibility of Training Neural Networks with Small Squared Errors

Van H. Vu

We demonstrate that the problem of training neural networks with small (average ) squared error is computationally intractable. Con(cid:173) sider a data set of M points (Xi, Yi), i = 1,2, ... , M, where Xi are input vectors from Rd, Yi are real outputs (Yi E R). For a net- work 10 in some class F of neural networks, (11M) L~l (fO(Xi)(cid:173) Yi)2)1/2 - inlfEF(l/ M) "2:f!1 (f(Xi) - YJ2)1/2 is the (avarage) rel(cid:173) ative error occurs when one tries to fit the data set by 10. We will prove for several classes F of neural networks that achieving a rela(cid:173) tive error smaller than some fixed positive threshold (independent from the size of the data set) is NP-hard.

************************************

Structure Driven Image Database Retrieval

Jeremy De Bonet, Paul Viola

A new algorithm is presented which approximates the perceived visual similarity between images. The images are initially trans(cid:173) formed into a feature space which captures visual structure, tex(cid:173) ture and color using a tree of filters. Similarity is the inverse of the distance in this perceptual feature space. Using this algorithm we have constructed an image database system which can perform example based retrieval on large image databases. Using carefully constructed target sets, which limit variation to only a single visual characteristic, retrieval rates are quantitatively compared to those of standard methods.

************************************

Bayesian Robustification for Audio Visual Fusion

Javier Movellan, Paul Mineiro

We discuss the problem of catastrophic fusion in multimodal recog(cid:173) nition systems. This problem arises in systems that need to fuse different channels in non-stationary environments. Practice shows that when recognition modules within each modality are tested in contexts inconsistent with their assumptions, their influence on the fused product tends to increase, with catastrophic results. We ex(cid:173) plore a principled solution to this problem based upon Bayesian ideas of competitive models and inference robustification: each sensory channel is provided with simple white-noise context mod(cid:173) els, and the perceptual hypothesis and context are jointly esti(cid:173) mated. Consequently, context deviations are interpreted as changes in white noise contamination strength, automatically adjusting the influence of the module. The approach is tested on a fixed lexicon automatic audiovisual speech recognition problem with very good results.

************************************

A Neural Network Model of Naive Preference and Filial Imprinting in the Domestic Chick

Lucy Hadden

Filial imprinting in domestic chicks is of interest in psychology, biology, and computational modeling because it exemplifies simple, rapid, in(cid:173) nately programmed learning which is biased toward learning about some objects. Hom et al. have recently discovered a naive visual preference for heads and necks which develops over the course of the first three days of life. The neurological basis of this predisposition is almost en(cid:173) tirely unknown; that of imprinting-related learning is fairly clear. This project is the first model of the predisposition consistent with what is known about learning in imprinting. The model develops the predisposi(cid:173) tion appropriately, learns to "approach" a training object, and replicates one interaction between the two processes. Future work will replicate more interactions between imprinting and the predisposition in chicks, and analyze why the system works.

************************************

Shared Context Probabilistic Transducers

Yoshua Bengio, Samy Bengio, Jean-Franc Isabelle, Yoram Singer

Recently, a model for supervised learning of probabilistic transduc(cid:173) ers represented by suffix trees was introduced. However, this algo(cid:173) rithm tends to build very large trees, requiring very large amounts of computer memory. In this paper, we propose anew, more com(cid:173) pact, transducer model in which one shares the parameters of distri(cid:173) butions associated to contexts yielding similar conditional output distributions . We illustrate the advantages of the proposed algo(cid:173) rithm with comparative experiments on inducing a noun phrase recogmzer.

************************************

Learning to Schedule Straight-Line Code

J. Moss, Paul Utgoff, John Cavazos, Doina Precup, Darko Stefanovic, Carla Brodley, David Scheeff

Program execution speed on modem computers is sensitive, by a factor of two or more, to the order in which instructions are presented to the proces(cid:173) sor. To realize potential execution efficiency, an optimizing compiler must employ a heuristic algorithm for instruction scheduling. Such algorithms are painstakingly hand-crafted, which is expensive and time-consuming. We show how to cast the instruction scheduling problem as a learning task, ob(cid:173) taining the heuristic scheduling algorithm automatically. Our focus is the narrower problem of scheduling straight-line code (also called basic blocks of instructions). Our empirical results show that just a few features are ad (cid:173) equate for quite good performance at this task for a real modem processor, and that any of several supervised learning methods perform nearly opti(cid:173) mally with respect to the features used.

************************************

Approximating Posterior Distributions in Belief Networks Using Mixtures

Christopher Bishop, Neil Lawrence, Tommi Jaakkola, Michael Jordan

Exact inference in densely connected Bayesian networks is computation(cid:173) ally intractable, and so there is considerable interest in developing effec(cid:173) tive approximation schemes. One approach which has been adopted is to bound the log likelihood using a mean-field approximating distribution. While this leads to a tractable algorithm, the mean field distribution is as(cid:173) sumed to be factorial and hence unimodal. In this paper we demonstrate the feasibility of using a richer class of approximating distributions based on mixtures of mean field distributions. We derive an efficient algorithm for updating the mixture parameters and apply it to the problem of learn(cid:173) ing in sigmoid belief networks. Our results demonstrate a systematic improvement over simple mean field theory as the number of mixture components is increased.

************************************

The Canonical Distortion Measure in Feature Space and 1-NN Classification

Jonathan Baxter, Peter Bartlett

We prove that the Canonical Distortion Measure (CDM) [2, 3] is the optimal distance measure to use for I nearest-neighbour (l-NN) classifi(cid:173) cation, and show that it reduces to squared Euclidean distance in feature space for function classes that can be expressed as linear combinations of a fixed set of features. PAC-like bounds are given on the sample(cid:173) complexity required to learn the CDM. An experiment is presented in which a neural network CDM was learnt for a Japanese OCR environ(cid:173) ment and then used to do I-NN classification.

************************************

A Non-Parametric Multi-Scale Statistical Model for Natural Images

Jeremy De Bonet, Paul Viola

The observed distribution of natural images is far from uniform. On the contrary, real images have complex and important struc(cid:173) ture that can be exploited for image processing, recognition and analysis. There have been many proposed approaches to the prin(cid:173) cipled statistical modeling of images, but each has been limited in either the complexity of the models or the complexity of the im(cid:173) ages. We present a non-parametric multi-scale statistical model for images that can be used for recognition, image de-noising, and in a "generative mode" to synthesize high quality textures.

************************************

A Solution for Missing Data in Recurrent Neural Networks with an Application to Blood Glucose Prediction

Volker Tresp, Thomas Briegel

We consider neural network models for stochastic nonlinear dynamical systems where measurements of the variable of interest are only avail(cid:173) able at irregular intervals i.e. most realizations are missing. Difficulties arise since the solutions for prediction and maximum likelihood learn(cid:173) ing with missing data lead to complex integrals, which even for simple cases cannot be solved analytically. In this paper we propose a spe(cid:173) cific combination of a nonlinear recurrent neural predictive model and a linear error model which leads to tractable prediction and maximum likelihood adaptation rules. In particular, the recurrent neural network can be trained using the real-time recurrent learning rule and the linear error model can be trained by an EM adaptation rule, implemented us(cid:173) ing forward-backward Kalman filter equations. The model is applied to predict the glucose/insulin metabolism of a diabetic patient where blood glucose measurements are only available a few times a day at irregular intervals. The new model shows considerable improvement with respect to both recurrent neural networks trained with teacher forcing or in a free running mode and various linear models.

************************************

Regularisation in Sequential Learning Algorithms
João de Freitas, Mahesan Niranjan, Andrew Gee

In this paper, we discuss regularisation in online/sequential learn(cid:173) ing algorithms. In environments where data arrives sequentially, techniques such as cross-validation to achieve regularisation or model selection are not possible. Further, bootstrapping to de(cid:173) termine a confidence level is not practical. To surmount these problems, a minimum variance estimation approach that makes use of the extended Kalman algorithm for training multi-layer percep(cid:173) trons is employed. The novel contribution of this paper is to show the theoretical links between extended Kalman filtering, Sutton's variable learning rate algorithms and Mackay's Bayesian estima(cid:173) tion framework. In doing so, we propose algorithms to overcome the need for heuristic choices of the initial conditions and noise covariance matrices in the Kalman approach.

************************************

An Improved Policy Iteration Algorithm for Partially Observable MDPs
Eric Hansen

A new policy iteration algorithm for partially observable Markov decision processes is presented that is simpler and more efficient than an earlier policy iteration algorithm of Sondik (1971,1978). The key simplification is representation of a policy as a finite-state controller. This representation makes policy evaluation straightforward. The pa(cid:173) per's contribution is to show that the dynamic-programming update used in the policy improvement step can be interpreted as the trans(cid:173) formation of a finite-state controller into an improved finite-state con(cid:173) troller. The new algorithm consistently outperforms value iteration as an approach to solving infinite-horizon problems.

************************************

The Asymptotic Convergence-Rate of Q-learning
Csaba Szepesvári

In this paper we show that for discounted MDPs with discount factor, > 1/2 the asymptotic rate of convergence of Q-Iearning if R(1 - ,) < 1/2 and O( Jlog log tit) otherwise is O(1/tR (1-1') provided that the state-action pairs are sampled from a fixed prob(cid:173) ability distribution. Here R = Pmin/Pmax is the ratio of the min(cid:173) imum and maximum state-action occupation frequencies. The re(cid:173) sults extend to convergent on-line learning provided that Pmin > 0, where Pmin and Pmax now become the minimum and maximum state-action occupation frequencies corresponding to the station(cid:173) ary distribution.

************************************

Learning Human-like Knowledge by Singular Value Decomposition: A Progress Report
Thomas Landauer, Darrell Laham, Peter Foltz

Singular value decomposition (SVD) can be viewed as a method for unsupervised training of a network that associates two classes of events reciprocally by linear connections through a single hidden layer. SVD was used to learn and represent relations among very large numbers of words (20k-60k) and very large numbers of natural text passages (lk- 70k) in which they occurred. The result was 100-350 dimensional "semantic spaces" in which any trained or newly aibl word or passage could be represented as a vector, and similarities were measured by the cosine of the contained angle between vectors. Good accmacy in simulating human judgments and behaviors has been demonstrated by performance on multiple-choice vocabulary and domain knowledge tests, emulation of expert essay evaluations, and in several other ways. Examples are also given of how the kind of knowledge extracted by this method can be applied.

************************************

Toward a Single-Cell Account for Binocular Disparity Tuning: An Energy Model May Be Hiding in Your Dendrites
Bartlett Mel, Daniel Ruderman, Kevin Archie

Converging evidence has shown that human object recognition depends on familiarity with the images of an object. Further, the greater the similarity between objects, the stronger is the dependence on object appearance, and the more important two(cid:173) dimensional (2D) image information becomes. These findings, how(cid:173) ever, do not rule out the use of 3D structural information in recog(cid:173) nition, and the degree to which 3D information is used in visual memory is an important issue. Liu, Knill, & Kersten (1995) showed that any model that is restricted to rotations in the image plane of independent 2D templates could not account for human perfor(cid:173) mance in discriminating novel object views. We now present results from models of generalized radial basis functions (GRBF), 2D near(cid:173) est neighbor matching that allows 2D affine transformations, and a Bayesian statistical estimator that integrates over all possible 2D affine transformations. The performance of the human observers relative to each of the models is better for the novel views than for the familiar template views, suggesting that humans generalize better to novel views from template views. The Bayesian estima(cid:173) tor yields the optimal performance with 2D affine transformations and independent 2D templates. Therefore, models of 2D affine matching operations with independent 2D templates are unlikely to account for human recognition performance.

************************************

Recurrent Neural Networks Can Learn to Implement Symbol-Sensitive Counting
Paul Rodriguez, Janet Wiles

Recently researchers have derived formal complexity analysis of analog computation in the setting of discrete-time dynamical systems. As an empirical constrast, training recurrent neural networks (RNNs) produces self -organized systems that are realizations of analog mechanisms. Pre(cid:173) vious work showed that a RNN can learn to process a simple context-free language (CFL) by counting. Herein, we extend that work to show that a RNN can learn a harder CFL, a simple palindrome, by organizing its re(cid:173) sources into a symbol-sensitive counting solution, and we provide a dy(cid:173) namical systems analysis which demonstrates how the network: can not only count, but also copy and store counting infonnation.

************************************

Modeling Acoustic Correlations by Factor Analysis
Lawrence Saul, Mazin Rahim

Hidden Markov models (HMMs) for automatic speech recognition rely on high dimensional feature vectors to summarize the short(cid:173) time properties of speech. Correlations between features can arise when the speech signal is non-stationary or corrupted by noise. We investigate how to model these correlations using factor analysis, a statistical method for dimensionality reduction . Factor analysis uses a small number of parameters to model the covariance struc(cid:173) ture of high dimensional data. These parameters are estimated by an Expectation-Maximization (EM) algorithm that can be em(cid:173) bedded in the training procedures for HMMs. We evaluate the combined use of mixture densities and factor analysis in HMMs that recognize alphanumeric strings. Holding the total number of parameters fixed, we find that these methods, properly combined, yield better models than either method on its own.

************************************

MELONET I: Neural Nets for Inventing Baroque-Style Chorale Variations
Dominik Hörnel

MELONET I is a multi-scale neural network system producing baroque-style melodic variations. Given a melody, the system in(cid:173) vents a four-part chorale harmonization and a variation of any chorale voice, after being trained on music pieces of composers like J. S. Bach and J . Pachelbel. Unlike earlier approaches to the learn(cid:173) ing of melodic structure, the system is able to learn and reproduce high-order structure like harmonic, motif and phrase structure in melodic sequences. This is achieved by using mutually interacting feedforward networks operating at different time scales, in combi(cid:173) nation with Kohon en networks to classify and recognize musical structure. The results are chorale partitas in the style of J. Pachel(cid:173) bel. Their quality has been judged by experts to be comparable to improvisations invented by an experienced human organist.

************************************

Hippocampal Model of Rat Spatial Abilities Using Temporal Difference Learning
David Foster, Richard Morris, Peter Dayan

We provide a model of the standard watermaze task, and of a more challenging task involving novel platform locations, in which rats exhibit one-trial learning after a few days of training. The model uses hippocampal place cells to support reinforcement learning, and also, in an integrated manner, to build and use allocentric coordinates.

************************************

EM Algorithms for PCA and SPCA
Sam Roweis

I present an expectation-maximization (EM) algorithm for principal component analysis (PCA). The algorithm allows a few eigenvectors and eigenvalues to be extracted from large collections of high dimensional data. It is computationally very efficient in space and time. It also natu(cid:173) rally accommodates missing infonnation. I also introduce a new variant of PC A called sensible principal component analysis (SPCA) which de(cid:173) fines a proper density model in the data space. Learning for SPCA is also done with an EM algorithm. I report results on synthetic and real data showing that these EM algorithms correctly and efficiently find the lead(cid:173) ing eigenvectors of the covariance of datasets in a few iterations using up to hundreds of thousands of datapoints in thousands of dimensions.

************************************

Hierarchical Non-linear Factor Analysis and Topographic Maps
Zoubin Ghahramani, Geoffrey E. Hinton

We first describe a hierarchical, generative model that can be viewed as a non-linear generalisation of factor analysis and can be implemented in a neural network. The model performs per(cid:173) ceptual inference in a probabilistically consistent manner by using top-down, bottom-up and lateral connections. These connections can be learned using simple rules that require only locally avail(cid:173) able information. We then show how to incorporate lateral con(cid:173) nections into the generative model. The model extracts a sparse, distributed, hierarchical representation of depth from simplified random-dot stereograms and the localised disparity detectors in the first hidden layer form a topographic map. When presented with image patches from natural scenes, the model develops topo(cid:173) graphically organised local feature detectors.

************************************

Learning Path Distributions Using Nonequilibrium Diffusion Networks
Paul Mineiro, Javier Movellan, Ruth Williams

We propose diffusion networks, a type of recurrent neural network with probabilistic dynamics, as models for learning natural signals that are continuous in time and space. We give a formula for the gradient of the log-likelihood of a path with respect to the drift parameters for a diffusion network. This gradient can be used to optimize diffusion networks in the nonequilibrium regime for a wide variety of problems paralleling techniques which have succeeded in engineering fields such as system identification, state estimation and signal filtering. An aspect of this work which is of particu(cid:173) lar interest to computational neuroscience and hardware design is that with a suitable choice of activation function, e.g., quasi-linear sigmoidal, the gradient formula is local in space and time.

************************************

Unsupervised On-line Learning of Decision Trees for Hierarchical Data Analysis
Marcus Held, Joachim Buhmann

An adaptive on-line algorithm is proposed to estimate hierarchical data structures for non-stationary data sources. The approach is based on the principle of minimum cross entropy to derive a decision tree for data clustering and it employs a metalearning idea (learning to learn) to adapt to changes in data characteristics. Its efficiency is demonstrated by grouping non-stationary artifical data and by hierarchical segmentation of LANDSAT images.

************************************

Recovering Perspective Pose with a Dual Step EM Algorithm

Andrew Cross, Edwin Hancock

This paper describes a new approach to extracting 3D perspective structure from 2D point-sets. The novel feature is to unify the tasks of estimating transformation geometry and identifying point(cid:173) correspondence matches. Unification is realised by constructing a mixture model over the bi-partite graph representing the correspon(cid:173) dence match and by effecting optimisation using the EM algorithm. According to our EM framework the probabilities of structural cor(cid:173) respondence gate contributions to the expected likelihood function used to estimate maximum likelihood perspective pose parameters. This provides a means of rejecting structural outliers.

************************************

Learning to Order Things

William W. Cohen, Robert E. Schapire, Yoram Singer

There are many applications in which it is desirable to order rather than classify instances. Here we consider the problem of learning how to order, given feedback in the form of preference judgments, i.e., statements to the effect that one instance should be ranked ahead of another. We outline a two-stage approach in which one first learns by conventional means a preference Junction, of the form PREF( u, v), which indicates whether it is advisable to rank u before v. New instances are then ordered so as to maximize agreements with the learned preference func(cid:173) tion. We show that the problem of finding the ordering that agrees best with a preference function is NP-complete, even under very restrictive assumptions. Nevertheless, we describe a simple greedy algorithm that is guaranteed to find a good approximation. We then discuss an on-line learning algorithm, based on the "Hedge" algorithm, for finding a good linear combination of ranking "experts." We use the ordering algorithm combined with the on-line learning algorithm to find a combination of "search experts," each of which is a domain-specific query expansion strategy for a WWW search engine, and present experimental results that demonstrate the merits of our approach.

************************************

Analog VLSI Model of Intersegmental Coordination with Nearest-Neighbor Coupling

Girish Patel, Jeremy Holleman, Stephen DeWeerth

We have a developed an analog VLSI system that models the coordina(cid:173) tion of neurobiological segmental oscillators. We have implemented and tested a system that consists of a chain of eleven pattern generating cir(cid:173) cuits that are synaptically coupled to their nearest neighbors. Each pat(cid:173) tern generating circuit is implemented with two silicon Morris-Lecar neurons that are connected in a reciprocally inhibitory network. We dis(cid:173) cuss the mechanisms of oscillations in the two-cell network and explore system behavior based on isotropic and anisotropic coupling, and fre(cid:173) quency gradients along the chain of oscillators.

************************************

Globally Optimal On-line Learning Rules

Magnus Rattray, David Saad

We present a method for determining the globally optimal on-line learning rule for a soft committee machine under a statistical me(cid:173) chanics framework. This work complements previous results on locally optimal rules, where only the rate of change in general(cid:173) ization error was considered. We maximize the total reduction in generalization error over the whole learning process and show how the resulting rule can significantly outperform the locally optimal rule.

************************************

Ensemble Learning for Multi-Layer Networks

David Barber, Christopher Bishop

Bayesian treatments of learning in neural networks are typically based either on local Gaussian approximations to a mode of the posterior weight distribution, or on Markov chain Monte Carlo simulations. A third approach, called ensemble learning, was in(cid:173) troduced by Hinton and van Camp (1993). It aims to approximate the posterior distribution by minimizing the Kullback-Leibler di(cid:173) vergence between the true posterior and a parametric approximat(cid:173) ing distribution. However, the derivation of a deterministic algo(cid:173) rithm relied on the use of a Gaussian approximating distribution with a diagonal covariance matrix and so was unable to capture the posterior correlations between parameters. In this paper, we show how the ensemble learning approach can be extended to full(cid:173) covariance Gaussian distributions while remaining computationally tractable. We also extend the framework to deal with hyperparam(cid:173) eters, leading to a simple re-estimation procedure. Initial results from a standard benchmark problem are encouraging.

************************************

Estimating Dependency Structure as a Hidden Variable

Marina Meila, Michael Jordan

This paper introduces a probability model, the mixture of trees that can account for sparse, dynamically changing dependence relationships. We present a family of efficient algorithms that use EM and the Minimum Spanning Tree algorithm to find the ML and MAP mixture of trees for a variety of priors, including the Dirichlet and the MDL priors.

************************************

Ensemble and Modular Approaches for Face Detection: A Comparison

Raphaël Feraud, Olivier Bernier

A new learning model based on autoassociative neural networks is developped and applied to face detection. To extend the de(cid:173) tection ability in orientation and to decrease the number of false alarms, different combinations of networks are tested: ensemble, conditional ensemble and conditional mixture of networks. The use of a conditional mixture of networks allows to obtain state of the art results on different benchmark face databases.

************************************

Stacked Density Estimation

Padhraic Smyth, David Wolpert

In this paper, the technique of stacking, previously only used for supervised learning, is applied to unsupervised learning. Specifi(cid:173) cally, it is used for non-parametric multivariate density estimation, to combine finite mixture model and kernel density estimators. Ex(cid:173) perimental results on both simulated data and real world data sets clearly demonstrate that stacked density estimation outperforms other strategies such as choosing the single best model based on cross-validation, combining with uniform weights, and even the sin(cid:173) gle best model chosen by "cheating" by looking at the data used for independent testing.

************************************

An Analog VLSI Neural Network for Phase-based Machine Vision

Bertram Shi, Kwok Hui

We describe the design, fabrication and test results of an analog CMOS VLSI neural network prototype chip intended for phase-based machine vision algorithms. The chip implements an image filtering operation similar to Gabor-filtering. Because a Gabor filter's output is complex valued, it can be used to define a phase at every pixel in an image. This phase can be used in robust algorithms for disparity estimation and bin(cid:173) ocular stereo vergence control in stereo vision and for image motion analysis. The chip reported here takes an input image and generates two outputs at every pixel corresponding to the real and imaginary parts of the output.

************************************

Combining Classifiers Using Correspondence Analysis

Christopher Merz

Several effective methods for improving the performance of a sin(cid:173) gle learning algorithm have been developed recently. The general approach is to create a set of learned models by repeatedly apply(cid:173) ing the algorithm to different versions of the training data, and then combine the learned models' predictions according to a pre(cid:173) scribed voting scheme. Little work has been done in combining the predictions of a collection of models generated by many learning algorithms having different representation and/or search strategies. This paper describes a method which uses the strategies of stack(cid:173) ing and correspondence analysis to model the relationship between the learning examples and the way in which they are classified by a collection of learned models. A nearest neighbor method is then applied within the resulting representation to classify previously unseen examples. The new algorithm consistently performs as well or better than other combining techniques on a suite of data sets.

************************************

### Synchronized Auditory and Cognitive 40 Hz Attentional Streams, and the Impact of Rhythmic Expectation on Auditory Scene Analysis

Bill Baird

We have developed a neural network architecture that implements a the(cid:173) ory of attention, learning, and trans-cortical communication based on adaptive synchronization of 5-15 Hz and 30-80 Hz oscillations between cortical areas. Here we present a specific higher order cortical model of attentional networks, rhythmic expectancy, and the interaction of hi~her- order and primar¥, cortical levels of processing. It accounts for the' mis(cid:173) match negativity' of the auditory ERP and the results of psychological experiments of Jones showing that auditory stream segregation depends on the rhythmic structure of inputs. The timing mechanisms of the model allow us to explain how relative timing information such as the relative order of events between streams is lost when streams are formed. The model suggests how the theories of auditory perception and attention of Jones and Bregman may be reconciled.

************************************

### From Regularization Operators to Support Vector Kernels

Alex Smola, Bernhard Schölkopf

We derive the correspondence between regularization operators used in Regularization Networks and Hilbert Schmidt Kernels appearing in Sup(cid:173) port Vector Machines. More specifica1ly, we prove that the Green's Func(cid:173) tions associated with regularization operators are suitable Support Vector Kernels with equivalent regularization properties. As a by-product we show that a large number of Radial Basis Functions namely condition(cid:173) ally positive definite functions may be used as Support Vector kernels.

************************************

### Two Approaches to Optimal Annealing

Todd Leen, Bernhard Schottky, David Saad

We employ both master equation and order parameter approaches to analyze the asymptotic dynamics of on-line learning with dif(cid:173) ferent learning rate annealing schedules. We examine the relations between the results obtained by the two approaches and obtain new results on the optimal decay coefficients and their dependence on the number of hidden nodes in a two layer architecture.

************************************

### Bidirectional Retrieval from Associative Memory

Friedrich Sommer, Günther Palm

Similarity based fault tolerant retrieval in neural associative mem(cid:173) ories (N AM) has not lead to wiedespread applications. A draw(cid:173) back of the efficient Willshaw model for sparse patterns [Ste61, WBLH69], is that the high asymptotic information capacity is of little practical use because of high cross talk noise arising in the retrieval for finite sizes. Here a new bidirectional iterative retrieval method for the Willshaw model is presented, called crosswise bidi(cid:173) rectional (CB) retrieval, providing enhanced performance. We dis(cid:173) cuss its asymptotic capaci ty limit, analyze the first step, and com(cid:173) pare it in experiments with the Willshaw model. Applying the very efficient CB memory model either in information retrieval systems or as a functional model for reciprocal cortico-cortical pathways requires more than robustness against random noise in the input: Our experiments show also the segmentation ability of CB-retrieval with addresses containing the superposition of pattens, provided even at high memory load.

************************************

Using Helmholtz Machines to Analyze Multi-channel Neuronal Recordings
Virginia de, R. DeCharms, Michael Merzenich

One of the current challenges to understanding neural information processing in biological systems is to decipher the "code" carried by large populations of neurons acting in parallel. We present an algorithm for automated discovery of stochastic firing patterns in large ensembles of neurons. The algorithm, from the "Helmholtz Machine" family, attempts to predict the observed spike patterns in the data. The model consists of an observable layer which is directly activated by the input spike patterns, and hidden units that are ac(cid:173) tivated through ascending connections from the input layer. The hidden unit activity can be propagated down to the observable layer to create a prediction of the data pattern that produced it. Hidden units are added incrementally and their weights are adjusted to im(cid:173) prove the fit between the predictions and data, that is, to increase a bound on the probability of the data given the model. This greedy strategy is not globally optimal but is computationally tractable for large populations of neurons. We show benchmark data on artifi(cid:173) cially constructed spike trains and promising early results on neuro(cid:173) physiological data collected from our chronic multi-electrode cortical implant.

************************************

Silicon Retina with Adaptive Filtering Properties
Shih-Chii Liu

This paper describes a small, compact circuit that captures the temporal and adaptation properties both of the photoreceptor and of the laminar layers of the fly. This circuit uses only six transis(cid:173) tors and two capacitors. It is operated in the subthreshold domain. The circuit maintains a high transient gain by using adaptation to the background intensity as a form of gain control. The adapta(cid:173) tion time constant of the circuit can be controlled via an external bias. Its temporal filtering properties change with the background intensity or signal-to-noise conditions. The frequency response of the circuit shows that in the frequency range of 1 to 100 Hz, the circuit response goes from highpass filtering under high light levels to lowpass filtering under low light levels (Le., when the signal-to(cid:173) noise ratio is low). A chip with 20x20 pixels has been fabricated in 1.2J.Lm ORBIT CMOS nwell technology.

************************************

Statistical Models of Conditioning
Peter Dayan, Theresa Long

Conditioning experiments probe the ways that animals make pre(cid:173) dictions about rewards and punishments and use those predic(cid:173) tions to control their behavior. One standard model of condition(cid:173) ing paradigms which involve many conditioned stimuli suggests that individual predictions should be added together. Various key results show that this model fails in some circumstances, and mo(cid:173) tivate an alternative model, in which there is attentional selection between different available stimuli. The new model is a form of mixture of experts, has a close relationship with some other exist(cid:173) ing psychological suggestions, and is statistically well-founded.

************************************

The Efficiency and the Robustness of Natural Gradient Descent Learning Rule
Howard Yang, Shun-ichi Amari

The inverse of the Fisher information matrix is used in the natu(cid:173) ral gradient descent algorithm to train single-layer and multi-layer perceptrons. We have discovered a new scheme to represent the Fisher information matrix of a st ochastic multi-layer perceptron. Based on this scheme, we have designed an algorithm to compute the natural gradient. When the input dimension n is much larger than the number of hidden neurons, the complexity of this algo(cid:173) rithm is of order O(n). It is confirmed by simulations that the natural gradient descent learning rule is not only efficient but also robust.

************************************