Computational Differences between Asymmetrical and Symmetrical Networks

Zhaoping Li, Peter Dayan

Symmetrically connected recurrent networks have recently been used as models of a host of neural computations. However, be(cid:173) cause of the separation between excitation and inhibition, biolog(cid:173) ical neural networks are asymmetrical. We study characteristic differences between asymmetrical networks and their symmetri(cid:173) cal counterparts, showing that they have dramatically different dynamical behavior and also how the differences can be exploited for computational ends. We illustrate our results in the case of a network that is a selective amplifier.

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

A Randomized Algorithm for Pairwise Clustering

Yoram Gdalyahu, Daphna Weinshall, Michael Werman

We present a stochastic clustering algorithm based on pairwise sim(cid:173) ilarity of datapoints. Our method extends existing deterministic methods, including agglomerative algorithms, min-cut graph algo(cid:173) rithms, and connected components. Thus it provides a common framework for all these methods. Our graph-based method differs from existing stochastic methods which are based on analogy to physical systems. The stochastic nature of our method makes it more robust against noise, including accidental edges and small spurious clusters. We demonstrate the superiority of our algorithm using an example with 3 spiraling bands and a lot of noise.

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

Risk Sensitive Reinforcement Learning

Ralph Neuneier, Oliver Mihatsch

A directed generative model for binary data using a small number of hidden continuous units is investigated. A clipping nonlinear(cid:173) ity distinguishes the model from conventional principal components analysis. The relationships between the correlations of the underly(cid:173) ing continuous Gaussian variables and the binary output variables are utilized to learn the appropriate weights of the network. The advantages of this approach are illustrated on a translationally in(cid:173) variant binary distribution and on handwritten digit images.

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

Convergence of the Wake-Sleep Algorithm

Shiro Ikeda, Shun-ichi Amari, Hiroyuki Nakahara

The W-S (Wake-Sleep) algorithm is a simple learning rule for the models with hidden variables. It is shown that this algorithm can be applied to a factor analysis model which is a linear version of the Helmholtz ma(cid:173) chine. But even for a factor analysis model, the general convergence is not proved theoretically. In this article, we describe the geometrical un(cid:173) derstanding of the W-S algorithm in contrast with the EM (Expectation(cid:173) Maximization) algorithm and the em algorithm. As the result, we prove the convergence of the W-S algorithm for the factor analysis model. We also show the condition for the convergence in general models.

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

Mean Field Methods for Classification with Gaussian Processes

Manfred Opper, Ole Winther

We discuss the application of TAP mean field methods known from the Statistical Mechanics of disordered systems to Bayesian classifi(cid:173) cation models with Gaussian processes. In contrast to previous ap(cid:173) proaches, no knowledge about the distribution of inputs is needed. Simulation results for the Sonar data set are given.

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

Learning from Dyadic Data

Thomas Hofmann, Jan Puzicha, Michael Jordan

Dyadzc data refers to a domain with two finite sets of objects in which observations are made for dyads , i.e., pairs with one element from either set. This type of data arises naturally in many ap(cid:173) plication ranging from computational linguistics and information retrieval to pr eference analysis and computer vision. In this paper, we present a systematic, domain-independent framework of learn(cid:173) ing from dyadic data by statistical mixture models. Our approach covers different models with fiat and hierarchical latent class struc(cid:173) tures. We propose an annealed version of the standard EM algo(cid:173) rithm for model fitting which is empirically evaluated on a variety of data sets from different domains.

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

Call-Based Fraud Detection in Mobile Communication Networks Using a Hierarchical Regime-Switching Model

Jaakko Hollmén, Volker Tresp

Fraud causes substantial losses to telecommunication carriers. Detec(cid:173) tion systems which automatically detect illegal use of the network can be used to alleviate the problem. Previous approaches worked on features derived from the call patterns of individual users. In this paper we present a call-based detection system based on a hierarchical regime-switching model. The detection problem is formulated as an inference problem on the regime probabilities. Inference is implemented by applying the junc(cid:173) tion tree algorithm to the underlying graphical model. The dynamics are learned from data using the EM algorithm and subsequent discriminative training. The methods are assessed using fraud data from a real mobile communication network.

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

Analyzing and Visualizing Single-Trial Event-Related Potentials

Tzyy-Ping Jung, Scott Makeig, Marissa Westerfield, Jeanne Townsend, Eric Courchesne, Terrence J. Sejnowski

Event-related potentials (ERPs), are portions of electroencephalo(cid:173) graphic (EEG) recordings that are both time- and phase-locked to experimental events. ERPs are usually averaged to increase their signal/noise ratio relative to non-phase locked EEG activ(cid:173) ity, regardless of the fact that response activity in single epochs may vary widely in time course and scalp distribution. This study applies a linear decomposition tool, Independent Component Anal(cid:173) ysis (ICA) [1], to multichannel single-trial EEG records to derive spatial filters that decompose single-trial EEG epochs into a sum of temporally independent and spatially fixed components arising from distinct or overlapping brain or extra-brain networks. Our results on normal and autistic subjects show that ICA can sep(cid:173) arate artifactual, stimulus-locked, response-locked, and. non-event related background EEG activities into separate components, al(cid:173) lowing ( 1) removal of pervasive artifacts of all types from single-trial EEG records, and (2) identification of both stimulus- and response(cid:173) locked EEG components. Second, this study proposes a new visual(cid:173) ization tool, the 'ERP image', for investigating variability in laten(cid:173) cies and amplitudes of event-evoked responses in spontaneous EEG or MEG records. We show that sorting single-trial ERP epochs in order of reaction time and plotting the potentials in 2-D clearly reveals underlying patterns of response variability linked to per(cid:173) formance. These analysis and visualization tools appear broadly applicable to electrophyiological research on both normal and clin(cid:173) ical populations.

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

Learning Nonlinear Dynamical Systems Using an EM Algorithm

Zoubin Ghahramani, Sam Roweis

The Expectation-Maximization (EM) algorithm is an iterative pro(cid:173) cedure for maximum likelihood parameter estimation from data sets with missing or hidden variables [2]. It has been applied to system identification in linear stochastic state-space models, where the state variables are hidden from the observer and both the state and the parameters of the model have to be estimated simulta(cid:173) neously [9]. We present a generalization of the EM algorithm for parameter estimation in nonlinear dynamical systems. The "expec(cid:173) tation" step makes use of Extended Kalman Smoothing to estimate the state, while the "maximization" step re-estimates the parame(

cid:173) ters using these uncertain state estimates. In general, the nonlinear maximization step is difficult because it requires integrating out the uncertainty in the states. However, if Gaussian radial basis func(cid:173) tion (RBF) approximators are used to model the nonlinearities, the integrals become tractable and the maximization step can be solved via systems of linear equations.

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

Replicator Equations, Maximal Cliques, and Graph Isomorphism

Marcello Pelillo

We present a new energy-minimization framework for the graph isomorphism problem which is based on an equivalent maximum clique formulation. The approach is centered around a fundamental result proved by Motzkin and Straus in the mid-1960s, and recently expanded in various ways, which allows us to formulate the maxi(cid:173) mum clique problem in terms of a standard quadratic program. To solve the program we use "replicator" equations, a class of simple continuous- and discrete-time dynamical systems developed in var(cid:173) ious branches of theoretical biology. We show how, despite their inability to escape from local solutions, they nevertheless provide experimental results which are competitive with those obtained us(cid:173) ing more elaborate mean-field annealing heuristics.

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

Orientation, Scale, and Discontinuity as Emergent Properties of Illusory Contour Shape

Karvel Thornber, Lance Williams

A recent neural model of illusory contour formation is based on a distribution of natural shapes traced by particles moving with constant speed in directions given by Brownian motions. The input to that model consists of pairs of position and direction constraints and the output consists of the distribution of contours joining all such pairs. In general, these contours will not be closed and their distribution will not be scale-invariant. In this paper, we show how to compute a scale-invariant distribution of closed contours given position constraints alone and use this result to explain a well known illusory contour effect.

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

Active Noise Canceling Using Analog Neuro-Chip with On-Chip Learning Capability

Jung-Wook Cho, Soo-Young Lee

A modular analogue neuro-chip set with on-chip learning capability is developed for active noise canceling. The analogue neuro-chip set incorporates the error backpropagation learning rule for practical applications, and allows pin-to-pin interconnections for multi-chip boards. The developed neuro-board demonstrated active noise canceling without any digital signal processor. Multi-path fading of acoustic channels, random noise, and nonlinear distortion of the loud speaker are compensated by the adaptive learning circuits of the neuro-chips. Experimental results are reported for cancellation of car noise in real time.

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

Coordinate Transformation Learning of Hand Position Feedback Controller by Using Change of Position Error Norm

Eimei Oyama, Susumu Tachi

In order to grasp an object, we need to solve the inverse kine(cid:173) matics problem, i.e., the coordinate transformation from the visual coordinates to the joint angle vector coordinates of the arm. Al(cid:173) though several models of coordinate transformation learning have been proposed, they suffer from a number of drawbacks. In human motion control, the learning of the hand position error feedback controller in the inverse kinematics solver is important. This paper proposes a novel model of the coordinate transformation learning of the human visual feedback controller that uses the change of the joint angle vector and the corresponding change of the square of the hand position error norm. The feasibility of the proposed model is illustrated using numerical simulations.

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

Multi-Electrode Spike Sorting by Clustering Transfer Functions

Dmitry Rinberg, Hanan Davidowitz, Naftali Tishby

A new paradigm is proposed for sorting spikes in multi-electrode data using ratios of transfer functions between cells and electrodes. It is assumed that for every cell and electrode there is a stable linear relation. These are dictated by the properties of the tissue, the electrodes and their relative geometries. The main advantage of the method is that it is insensitive to variations in the shape and amplitude of a spike. Spike sorting is carried out in two separate steps. First, templates describing the statistics of each spike type are generated by clustering transfer function ratios then spikes are detected in the data using the spike statistics. These techniques were applied to data generated in the escape response system of the cockroach.

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

Fisher Scoring and a Mixture of Modes Approach for Approximate Inference and Learning in Nonlinear State Space Models

Thomas Briegel, Volker Tresp

We present Monte-Carlo generalized EM equations for learning in non(cid:173) linear state space models. The difficulties lie in the Monte-Carlo E-step which consists of sampling from the posterior distribution of the hidden variables given the observations. The new idea presented in this paper is to generate samples from a Gaussian approximation to the true posterior from which it is easy to obtain independent samples. The parameters of the Gaussian approximation are either derived from the extended Kalman filter or the Fisher scoring algorithm. In case the posterior density is mul(cid:173) timodal we propose to approximate the posterior by a sum of Gaussians (mixture of modes approach). We show that sampling from the approxi(cid:173) mate posterior densities obtained by the above algorithms leads to better models than using point estimates for the hidden states. In our exper(cid:173) iment, the Fisher scoring algorithm obtained a better approximation of the posterior mode than the EKF. For a multimodal distribution, the mix(cid:173) ture of modes approach gave superior results.

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

Probabilistic Modeling for Face Orientation Discrimination: Learning from Labeled and Unlabeled Data

Shumeet Baluja

This paper presents probabilistic modeling methods to solve the problem of dis(cid:173) criminating between five facial orientations with very little labeled data. Three models are explored. The first model maintains no inter-pixel dependencies, the second model is capable of modeling a set of arbitrary pair-wise dependencies, and the last model allows dependencies only between neighboring pixels. We show that for all three of these models, the accuracy of the learned models can be greatly improved by augmenting a small number of labeled training images with a large set of unlabeled images using Expectation-Maximization. This is important because it is often difficult to obtain image labels, while many unla(cid:173) beled images are readily available. Through a large set of empirical tests, we examine the benefits of unlabeled data for each of the models. By using only two randomly selected labeled examples per class, we can discriminate between the five facial orientations with an accuracy of 94%; with six labeled examples, we achieve an accuracy of 98%.

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

Direct Optimization of Margins Improves Generalization in Combined Classifiers

Llew Mason, Peter Bartlett, Jonathan Baxter

Cumulative training margin dis(cid:173) tributions for AdaBoost versus our "Direct Optimization Of Margins" (DOOM) algorithm. The dark curve is AdaBoost, the light curve is DOOM. DOOM sacrifices significant training er(cid:173) ror for improved test error (hori(cid:173) zontal marks on margin= 0 line)_

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

## A Neuromorphic Monaural Sound Localizer

John Harris, Chiang-Jung Pu, José Príncipe

We describe the first single microphone sound localization system and its inspiration from theories of human monaural sound localiza(cid:173) tion. Reflections and diffractions caused by the external ear (pinna) allow humans to estimate sound source elevations using only one ear. Our single microphone localization model relies on a specially shaped reflecting structure that serves the role of the pinna. Spe(cid:173) cially designed analog VLSI circuitry uses echo-time processing to localize the sound. A CMOS integrated circuit has been designed, fabricated, and successfully demonstrated on actual sounds.

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

## Boxlets: A Fast Convolution Algorithm for Signal Processing and Neural Networks

Patrice Simard, Léon Bottou, Patrick Haffner, Yann LeCun

Signal processing and pattern recognition algorithms make exten(cid:173) sive use of convolution. In many cases, computational accuracy is not as important as computational speed. In feature extraction, for instance, the features of interest in a signal are usually quite distorted. This form of noise justifies some level of quantization in order to achieve faster feature extraction . Our approach consists of approximating regions of the signal with low degree polynomi(cid:173) als, and then differentiating the resulting signals in order to obtain impulse functions (or derivatives of impulse functions). With this representation, convolution becomes extremely simple and can be implemented quite effectively. The true convolution can be recov(cid:173) ered by integrating the result of the convolution. This method yields substantial speed up in feature extraction and is applicable to convolutional neural networks.

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

## Experimental Results on Learning Stochastic Memoryless Policies for Partially Observable Markov Decision Processes

John Williams, Satinder Singh

Partially Observable Markov Decision Processes (pO "MOPs) constitute an important class of reinforcement learning problems which present unique theoretical and computational difficulties. In the absence of the Markov property, popular reinforcement learning algorithms such as Q-Iearning may no longer be effective, and memory-based methods which remove partial observability via state-estimation are notoriously expensive. An alternative approach is to seek a stochastic memoryless policy which for each observation of the environment prescribes a probability distribution over available actions that maximizes the average reward per timestep. A reinforcement learning algorithm which learns a locally optimal stochastic memoryless policy has been proposed by Jaakkola, Singh and Jordan, but not empirically verified. We present a variation of this algorithm, discuss its implementation, and demonstrate its viability using four test problems.

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

## Attentional Modulation of Human Pattern Discrimination Psychophysics Reproduced by a Quantitative Model

Laurent Itti, Jochen Braun, Dale Lee, Christof Koch

We previously proposed a quantitative model of early visual pro(cid:173) cessing in primates, based on non-linearly interacting visual filters and statistically efficient decision. We now use this model to inter(cid:173) pret the observed modulation of a range of human psychophysical thresholds with and without focal visual attention. Our model - calibrated by an automatic fitting procedure - simultaneously re(cid:173) produces thresholds for four classical pattern discrimination tasks, performed while attention was engaged by another concurrent task. Our model then predicts that the seemingly complex improvements of certain thresholds, which we observed when attention was fully available for the discrimination tasks, can best be explained by a strengthening of competition among early visual filters.

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

## Signal Detection in Noisy Weakly-Active Dendrites

Amit Manwani, Christof Koch

Here we derive measures quantifying the information loss of a synaptic signal due to the presence of neuronal noise sources, as it electrotonically propagates along a weakly-active dendrite. We model the dendrite as an infinite linear cable, with noise sources distributed along its length. The noise sources we consider are thermal noise, channel noise arising from the stochastic nature of voltage-dependent ionic channels (K+ and Na+) and synaptic noise due to spontaneous background activity. We assess the efficacy of information transfer using a signal detection paradigm where the objective is to detect the presence/absence of a presynaptic spike from the post-synaptic membrane voltage. This allows us to analytically assess the role of each of these noise sources in information transfer. For our choice of parameters, we find that the synaptic noise is the dominant noise source which limits the maximum length over which information be reliably transmitted.

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

Kernel PCA and De-Noising in Feature Spaces

Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, Gunnar Rätsch

Kernel PCA as a nonlinear feature extractor has proven powerful as a preprocessing step for classification algorithms. But it can also be con(cid:173) sidered as a natural generalization of linear principal component anal(cid:173) ysis. This gives rise to the question how to use nonlinear features for data compression, reconstruction, and de-noising, applications common in linear PCA. This is a nontrivial task, as the results provided by ker(cid:173) nel PCA live in some high dimensional feature space and need not have pre-images in input space. This work presents ideas for finding approxi(cid:173) mate pre-images, focusing on Gaussian kernels, and shows experimental results using these pre-images in data reconstruction and de-noising on toy examples as well as on real world data.

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

Multiple Paired Forward-Inverse Models for Human Motor Learning and Control

Masahiko Haruno, Daniel M. Wolpert, Mitsuo Kawato

Humans demonstrate a remarkable ability to generate accurate and appropriate motor behavior under many different and oftpn uncprtain environmental conditions. This paper describes a new modular ap(cid:173) proach to human motor learning and control, baspd on multiple pairs of inverse (controller) and forward (prpdictor) models. This architecture simultaneously learns the multiple inverse models necessary for control as well as how to select the inverse models appropriate for a given em'i(cid:173) ronm0nt. Simulations of object manipulation demonstrates the ability to learn mUltiple objects, appropriate generalization to novel objects and the inappropriate activation of motor programs based on visual cues, followed by on-line correction, seen in the "size-weight illusion".

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

SMEM Algorithm for Mixture Models

Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, Geoffrey E. Hinton

We present a split and merge EM (SMEM) algorithm to overcome the local maximum problem in parameter estimation of finite mixture models. In the case of mixture models, non-global maxima often involve having too many components of a mixture model in one part of the space and too few in an(cid:173) other, widely separated part of the space. To escape from such configurations we repeatedly perform simultaneous split and merge operations using a new criterion for efficiently selecting the split and merge candidates. We apply the proposed algorithm to the training of Gaussian mixtures and mixtures of factor analyzers using synthetic and real data and show the effectiveness of using the split and merge operations to improve the likelihood of both the training data and of held-out test data.

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

Learning Lie Groups for Invariant Visual Perception

Rajesh Rao, Daniel Ruderman

One of the most important problems in visual perception is that of visual in(cid:173) variance: how are objects perceived to be the same despite undergoing transfor(cid:173) mations such as translations, rotations or scaling? In this paper, we describe a Bayesian method for learning invariances based on Lie group theory. We show that previous approaches based on first-order Taylor series expansions of inputs can be regarded as special cases of the Lie group approach, the latter being ca(cid:173) pable of handling in principle arbitrarily large transfonnations. Using a matrix(cid:173) exponential based generative model of images, we derive an unsupervised al(cid:173) gorithm for learning Lie group operators from input data containing infinites(cid:173) imal transfonnations. The on-line unsupervised learning algorithm maximizes the posterior probability of generating the training data. We provide experimen(cid:173) tal results suggesting that the proposed method can learn Lie group operators for handling reasonably large I-D translations and 2-D rotations.

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

Tractable Variational Structures for Approximating Graphical Models
David Barber, Wim Wiegerinck

Graphical models provide a broad probabilistic framework with ap(cid:173) plications in speech recognition (Hidden Markov Models), medical diagnosis (Belief networks) and artificial intelligence (Boltzmann Machines). However, the computing time is typically exponential in the number of nodes in the graph. Within the variational frame(cid:173) work for approximating these models, we present two classes of dis(cid:173) tributions, decimatable Boltzmann Machines and Tractable Belief Networks that go beyond the standard factorized approach. We give generalised mean-field equations for both these directed and undirected approximations. Simulation results on a small bench(cid:173) mark problem suggest using these richer approximations compares favorably against others previously reported in the literature.

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

A Micropower CMOS Adaptive Amplitude and Shift Invariant Vector Quantiser
Richard Coggins, Raymond Wang, Marwan Jabri

In this paper we describe the architecture, implementation and experi(cid:173) mental results for an Intracardiac Electrogram (ICEG) classification and compression chip. The chip processes and vector-quantises 30 dimen(cid:173) sional analogue vectors while consuming a maximum of 2.5 J-tW power for a heart rate of 60 beats per minute (1 vector per second) from a 3.3 V supply. This represents a significant advance on previous work which achieved ultra low power supervised morphology classification since the template matching scheme used in this chip enables unsupervised blind classification of abnonnal rhythms and the computational support for low bit rate data compression. The adaptive template matching scheme used is tolerant to amplitude variations, and inter- and intra-sample time shifts.

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

Convergence Rates of Algorithms for Visual Search: Detecting Visual Contours
Alan L. Yuille, James Coughlan

This paper formulates the problem of visual search as Bayesian inference and defines a Bayesian ensemble of problem instances . In particular, we address the problem of the detection of visual contours in noise/clutter by optimizing a global criterion which combines local intensity and geometry information. We analyze the convergence rates of A * search algorithms using results from information theory to bound the probability of rare events within the Bayesian ensemble. This analysis determines characteristics of the domain , which we call order parameters, that determine the convergence rates. In particular, we present a specific admissible A * algorithm with pruning which converges, with high probability, with expected time O(N) in the size of the problem. In addi(cid:173) tion, we briefly summarize extensions of this work which address fundamental limits of target contour detectability (Le. algorithm independent results) and the use of non-admissible heuristics.

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

## Learning to Find Pictures of People

Sergey Ioffe, David Forsyth

Finding articulated objects, like people, in pictures present.s a par(cid:173) ticularly difficult object. recognition problem. We show how t.o find people by finding putative body segments, and then construct.(cid:173) ing assemblies of those segments that are consist.ent with the con(cid:173) straints on the appearance of a person that result from kinematic properties. Since a reasonable model of a person requires at. least nine segments, it is not possible to present every group to a classi(cid:173) fier. Instead, the search can be pruned by using projected versions of a classifier that accepts groups corresponding to people. We describe an efficient projection algorithm for one popular classi(cid:173) fier , and demonstrate that our approach can be used to determine whether images of real scenes contain people.

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

## Visualizing Group Structure

Marcus Held, Jan Puzicha, Joachim Buhmann

Cluster analysis is a fundamental principle in exploratory data analysis, providing the user with a description of the group struc(cid:173) ture of given data. A key problem in this context is the interpreta(cid:173) tion and visualization of clustering solutions in high- dimensional or abstract data spaces. In particular, probabilistic descriptions of the group structure, essential to capture inter-cluster relation(cid:173) ships, are hardly assessable by simple inspection ofthe probabilistic assignment variables. VVe present a novel approach to the visual(cid:173) ization of group structure. It is based on a statistical model of the object assignments which have been observed or estimated by a probabilistic clustering procedure. The objects or data points are embedded in a low dimensional Euclidean space by approximating the observed data statistics with a Gaussian mixture model. The algorithm provides a new approach to the visualization of the inher(cid:173) ent structure for a broad variety of data types, e.g. histogram data, proximity data and co-occurrence data. To demonstrate the power of the approach, histograms of textured images are visualized as an example of a large-scale data mining application.

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

## Modeling Surround Suppression in V1 Neurons with a Statistically Derived Normalization Model

Eero Simoncelli, Odelia Schwartz

We examine the statistics of natural monochromatic images decomposed using a multi-scale wavelet basis. Although the coefficients of this rep(cid:173) resentation are nearly decorrelated, they exhibit important higher-order statistical dependencies that cannot be eliminated with purely linear pro(cid:173) c~ssing. In particular, rectified coefficients corresponding to basis func(cid:173) tions at neighboring spatial positions, orientations and scales are highly correlated. A method of removing these dependencies is to divide each coefficient by a weighted combination of its rectified neighbors. Sev(cid:173) eral successful models of the steady -state behavior of neurons in primary visual cortex are based on such "divisive normalization" computations, and thus our analysis provides a theoretical justification for these models. Perhaps more importantly, the statistical measurements explicitly specify the weights that should be used in computing the normalization signal. We demonstrate that this weighting is qualitatively consistent with re(cid:173) cent physiological experiments that characterize the suppressive effect of stimuli presented outside of the classical receptive field. Our obser(cid:173) vations thus provide evidence for the hypothesis that early visual neural

processing is well matched to these statistical properties of images.
************************************

### Familiarity Discrimination of Radar Pulses

Eric Granger, Stephen Grossberg, Mark Rubin, William Streilein

The ARTMAP-FD neural network performs both identification (placing test patterns in classes encountered during training) and familiarity discrimination (judging whether a test pattern belongs to any of the classes encountered during training). The perfor(cid:173) mance of ARTMAP-FD is tested on radar pulse data obtained in the field, and compared to that of the nearest-neighbor-based NEN algorithm and to a k > 1 extension of NEN.
************************************

### Graphical Models for Recognizing Human Interactions

Nuria Oliver, Barbara Rosario, Alex Pentland

We describe a real-time computer vision and machine learning sys(cid:173) tem for modeling and recognizing human actions and interactions. Two different domains are explored: recognition of two-handed motions in the martial art 'Tai Chi' , and multiple- person interac(cid:173) tions in a visual surveillance task. Our system combines top-down with bottom-up information using a feedback loop, and is formu(cid:173) lated with a Bayesian framework. Two different graphical models (HMMs and Coupled HMMs) are used for modeling both individual actions and multiple-agent interactions, and CHMMs are shown to work more efficiently and accurately for a given amount of train(cid:173) ing. Finally, to overcome the limited amounts of training data, we demonstrate that 'synthetic agents' (Alife-style agents) can be used to develop flexible prior models of the person-to-person inter(cid:173) actions.
************************************

### An Entropic Estimator for Structure Discovery

Matthew Brand

We introduce a novel framework for simultaneous structure and parameter learning in hidden-variable conditional probability models, based on an entropic prior and a solution for its maximum a posteriori (MAP) estimator. The MAP estimate minimizes uncertainty in all respects: cross-entropy between model and data; entropy of the model ; entropy of the data's descriptive statistics. Iterative estimation extinguishes weakly supported parameters, compressing and sparsifying the model. Trimming operators accelerate this process by removing excess parameters and, unlike most pruning schemes, guarantee an increase in posterior probability. Entropic estimation takes a overcomplete random model and simplifies it, inducing the structure of relations between hidden and observed variables. Applied to hidden Markov models (HMMs), it finds a concise finite-state machine representing the hidden structure of a signal. We entropically model music, handwriting, and video time-series, and show that the resulting models are highly concise, structured, predictive, and interpretable: Surviving states tend to be highly correlated with meaningful partitions of the data, while surviving transitions provide a low-perplexity model of the signal dynamics.
************************************

### The Effect of Correlations on the Fisher Information of Population Codes

Hyoungsoo Yoon, Haim Sompolinsky

We study the effect of correlated noise on the accuracy of popu(cid:173) lation coding using a model of a population of neurons that are broadly tuned to an angle in two-dimension. The fluctuations in the neuronal activity is modeled as a Gaussian noise with pairwise correlations which decays exponentially with the difference between the preferred orientations of the pair. By calculating the Fisher in(cid:173) formation of the system, we show that in the biologically relevant regime of parameters positive correlations decrease the estimation capability of the network relative to the uncorrelated population. Moreover strong positive correlations result in information capac(cid:173) ity which saturates to a finite value as th e number of cells in the population grows. In contrast, negative correlations substantially increase the information capacity of the neuronal population.

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

Spike-Based Compared to Rate-Based Hebbian Learning

Richard Kempter, Wulfram Gerstner, J. van Hemmen

A correlation-based learning rule at the spike level is formulated, mathematically analyzed, and compared to learning in a firing-rate description. A differential equation for the learning dynamics is derived under the assumption that the time scales of learning and spiking can be separated. For a linear Poissonian neuron model which receives time-dependent stochastic input we show that spike correlations on a millisecond time scale play indeed a role. Corre(cid:173) lations between input and output spikes tend to stabilize structure formation, provided that the form of the learning window is in accordance with Hebb's principle. Conditions for an intrinsic nor(cid:173) malization of the average synaptic weight are discussed.

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

Contrast Adaptation in Simple Cells by Changing the Transmitter Release Probability

Péter Adorján, Klaus Obermayer

The contrast response function (CRF) of many neurons in the primary vi(cid:173) sual cortex saturates and shifts towards higher contrast values following prolonged presentation of high contrast visual stimuli. Using a recurrent neural network of excitatory spiking neurons with adapting synapses we show that both effects could be explained by a fast and a slow compo(cid:173) nent in the synaptic adaptation. (i) Fast synaptic depression leads to sat(cid:173) uration of the CRF and phase advance in the cortical response to high contrast stimuli. (ii) Slow adaptation of the synaptic transmitter release probability is derived such that the mutual information between the input and the output of a cortical neuron is maximal. This component-given by infomax learning rule-explains contrast adaptation of the averaged membrane potential (DC component) as well as the surprising experi(cid:173) mental result, that the stimulus modulated component (Fl component) of a cortical cell's membrane potential adapts only weakly. Based on our results, we propose a new experiment to estimate the strength of the ef(cid:173) fective excitatory feedback to a cortical neuron, and we also suggest a relatively simple experimental test to justify our hypothesized synaptic mechanism for contrast adaptation.

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

Optimizing Correlation Algorithms for Hardware-Based Transient Classification

R. Edwards, Gert Cauwenberghs, Fernando Pineda

The perfonnance of dedicated VLSI neural processing hardware depends critically on the design of the implemented algorithms. We have pre(cid:173) viously proposed an algorithm for acoustic transient classification [1]. Having implemented and demonstrated this algorithm in a mixed-mode architecture, we now investigate variants on the algorithm, using time and frequency channel differencing, input and output nonnalization, and schemes to binarize and train the template values, with the goal of achiev(cid:173) ing optimal classification perfonnance for the chosen hardware.

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

Information Maximization in Single Neurons

Martin Stemmler, Christof Koch

Information from the senses must be compressed into the limited range of firing rates generated by spiking nerve cells. Optimal compression uses all firing rates equally often, implying that the nerve cell's response matches the statistics of naturally occurring stimuli. Since changing the voltage-dependent ionic conductances in the cell membrane alters the flow of information, an unsupervised, non-Hebbian, developmental learning rule is derived to adapt the conductances in Hodgkin-Huxley model neurons. By maximizing the rate of information transmission, each firing rate within the model neuron's limited dynamic range is used equally often .

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

Mechanisms of Generalization in Perceptual Learning

Zili Liu, Daphna Weinshall

The learning of many visual perceptual tasks has been shown to be specific to practiced stimuli, while new stimuli require re-Iearning from scratch. Here we demonstrate generalization using a novel paradigm in motion discrimination where learning has been previ(cid:173) ously shown to be specific. We trained subjects to discriminate the directions of moving dots, and verified the previous results that learning does not transfer from the trained direction to a new one. However, by tracking the subjects' performance across time in the new direction, we found that their rate of learning doubled. Therefore, learning generalized in a task previously considered too difficult for generalization. We also replicated, in the second ex(cid:173) periment, transfer following training with "easy" stimuli. The specificity of perceptual learning and the dichotomy between learning of "easy" vs. "difficult" tasks were hypothesized to involve different learning processes, operating at different visual cortical areas. Here we show how to interpret these results in terms of signal detection theory. With the assumption of limited computational resources, we obtain the observed phenomena - direct transfer and change of learning rate - for increasing levels of task 'difficulty. It appears that human generalization concurs with the expected behavior of a generic discrimination system.

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

Probabilistic Image Sensor Fusion

Ravi Sharma, Todd Leen, Misha Pavel

We present a probabilistic method for fusion of images produced by multiple sensors. The approach is based on an image formation model in which the sensor images are noisy, locally linear functions of an underlying, true scene. A Bayesian framework then provides for maximum likelihood or maximum a posteriori estimates of the true scene from the sensor images. Maximum likelihood estimates of the parameters of the image formation model involve (local) second order image statistics, and thus are related to local principal component analysis. We demonstrate the efficacy of the method on images from visible-band and infrared sensors.

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

On-Line Learning with Restricted Training Sets: Exact Solution as Benchmark for General Theories

H. Rae, Peter Sollich, Anthony Coolen

O(ws(s log d+log(dqh/ s))) and O(ws((h/ s) log q) +log(dqh/ s)) are upper bounds for the VC-dimension of a set of neural networks of units with piecewise polynomial activation functions, where s is the depth of the network, h is the number of hidden units, w is the number of adjustable parameters, q is the maximum of the number of polynomial segments of the activation function, and d is the maximum degree of the polynomials; also n(wslog(dqh/s)) is a lower bound for the VC-dimension of such a network set, which are tight for the cases s = 8(h) and s is constant. For the special case q = 1, the VC-dimension is 8(ws log d).

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

Phase Diagram and Storage Capacity of Sequence-Storing Neural Networks

A. Düring, Anthony Coolen, D. Sherrington

We solve the dynamics of Hopfield-type neural networks which store se(cid:173) quences of patterns, close to saturation. The asymmetry of the interaction matrix in such models leads to violation of detailed balance, ruling out an equilibrium statistical mechanical analysis. Using generating functional methods we derive exact closed equations for dynamical order parame(cid:173) ters, viz. the sequence overlap and correlation and response functions. in the limit of an infinite system size. We calculate the time translation invariant solutions of these equations. describing stationary limit-cycles. which leads to a phase diagram. The effective retarded self-interaction usually appearing in symmetric models is here found to vanish, which causes a significa ntly enlarged storage capacity of eYe ~ 0.269. com(cid:173) pared to eYe ~ 0.139 for Hopfield networks s~oring static patterns. Our results are tested against extensive computer simulations and excellent agreement is found.

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

Example-Based Image Synthesis of Articulated Figures
Trevor Darrell

We present a method for learning complex appearance mappings. such as occur with images of articulated objects. Traditional interpolation networks fail on this case since appearance is not necessarily a smooth function nor a linear manifold for articulated objects. We define an ap(cid:173) pearance mapping from examples by constructing a set of independently smooth interpolation networks; these networks can cover overlapping re(cid:173) gions of parameter space. A set growing procedure is used to find ex(cid:173) ample clusters which are well-approximated within their convex hull; interpolation then proceeds only within these sets of examples. With this method physically valid images are produced even in regions of param(cid:173) eter space where nearby examples have different appearances. We show results generating both simulated and real arm images.

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

Shrinking the Tube: A New Support Vector Regression Algorithm
Bernhard Schölkopf, Peter Bartlett, Alex Smola, Robert C. Williamson

A new algorithm for Support Vector regression is described. For a priori chosen 1/, it automatically adjusts a flexible tube of minimal radius to the data such that at most a fraction 1/ of the data points lie outside. More(cid:173) over, it is shown how to use parametric tube shapes with non-constant radius. The algorithm is analysed theoretically and experimentally.

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

Lazy Learning Meets the Recursive Least Squares Algorithm
Mauro Birattari, Gianluca Bontempi, Hugues Bersini

Lazy learning is a memory-based technique that, once a query is re(cid:173) ceived, extracts a prediction interpolating locally the neighboring exam(cid:173) ples of the query which are considered relevant according to a distance measure. In this paper we propose a data-driven method to select on a query-by-query basis the optimal number of neighbors to be considered for each prediction. As an efficient way to identify and validate local models, the recursive least squares algorithm is introduced in the con(cid:173) text of local approximation and lazy learning. Furthermore, beside the winner-takes-all strategy for model selection, a local combination of the most promising models is explored. The method proposed is tested on six different datasets and compared with a state-of-the-art approach.

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

Reinforcement Learning for Trading
John Moody, Matthew Saffell

We propose to train trading systems by optimizing financial objec(cid:173) tive functions via reinforcement learning. The performance func(cid:173) tions that we consider are profit or wealth, the Sharpe ratio and our recently proposed differential Sharpe ratio for online learn(cid:173) ing. In Moody & Wu (1997), we presented empirical results that demonstrate the advantages of reinforcement learning relative to supervised learning. Here we extend our previous work to com(cid:173) pare Q-Learning to our Recurrent Reinforcement Learning (RRL) algorithm. We provide new simulation results that demonstrate the presence of predictability in the monthly S&P 500 Stock Index for the 25 year period 1970 through 1994, as well as a sensitivity analysis that provides economic insight into the trader's structure.

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

Distributional Population Codes and Multiple Motion Models
Richard Zemel, Peter Dayan

Most theoretical and empirical studies of population codes make the assumption that underlying neuronal activities is a unique and unambiguous value of an enc oded quantity. However, population activities can contain additional information about such things as multiple values of or uncertainty about the quantity. We have pre(cid:173) viously suggested a method to recover extra information by treat(cid:173) ing the activities of the population of cells as coding for a com(cid:173) plete distribution over the coded quantity rather than just a single value. We now show how this approach bears on psychophys(cid:173) ical and neurophysiological studies of population codes for mo(cid:173) tion direction in tasks involving transparent motion stimuli. We show that, unlike standard approaches, it is able to recover mul(cid:173) tiple motions from population responses, and also that its output is consistent with both correct and erroneous human performance on psychophysical tasks.

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

Using Collective Intelligence to Route Internet Traffic

David Wolpert, Kagan Tumer, Jeremy Frank

A COllective INtelligence (COIN) is a set of interacting reinforce(cid:173) ment learning (RL) algorithms designed in an automated fashion so that their collective behavior optimizes a global utility function. We summarize the theory of COINs, then present experiments us(cid:173) ing that theory to design COINs to control internet traffic routing. These experiments indicate that COINs outperform all previously investigated RL-based, shortest path routing algorithms.

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

Sparse Code Shrinkage: Denoising by Nonlinear Maximum Likelihood Estimation

Aapo Hyvärinen, Patrik Hoyer, Erkki Oja

Sparse coding is a method for finding a representation of data in which each of the components of the representation is only rarely significantly active. Such a representation is closely related to re(cid:173) dundancy reduction and independent component analysis, and has some neurophysiological plausibility. In this paper, we show how sparse coding can be used for denoising. Using maximum likelihood estimation of nongaussian variables corrupted by gaussian noise, we show how to apply a shrinkage nonlinearity on the components of sparse coding so as to reduce noise. Furthermore, we show how to choose the optimal sparse coding basis for denoising. Our method is closely related to the method of wavelet shrinkage, but has the important benefit over wavelet methods that both the features and the shrinkage parameters are estimated directly from the data.

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

Finite-Dimensional Approximation of Gaussian Processes

Giancarlo Ferrari-Trecate, Christopher Williams, Manfred Opper

Gaussian process (GP) prediction suffers from O(n3) scaling with the data set size n. By using a finite-dimensional basis to approximate the GP predictor, the computational complexity can be reduced. We de(cid:173) rive optimal finite-dimensional predictors under a number of assump(cid:173) tions, and show the superiority of these predictors over the Projected Bayes Regression method (which is asymptotically optimal). We also show how to calculate the minimal model size for a given n. The calculations are backed up by numerical experiments.

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

Controlling the Complexity of HMM Systems by Regularization

Christoph Neukirchen, Gerhard Rigoll

This paper introduces a method for regularization ofHMM systems that avoids parameter overfitting caused by insufficient training data. Regu(cid:173) larization is done by augmenting the EM training method by a penalty term that favors simple and smooth HMM systems. The penalty term is constructed as a mixture model of negative exponential distributions that is assumed to generate the state dependent emission probabilities of the HMMs. This new method is the successful transfer of a well known regularization approach in neural networks to the HMM domain and can be interpreted as a generalization of traditional state-tying for HMM sys(cid:173) tems. The effect of regularization is demonstrated for continuous speech recognition tasks by improving overfitted triphone m odels and by speaker adaptation with limited training data.

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

Scheduling Straight-Line Code Using Reinforcement Learning and Rollouts

Amy McGovern, J. Moss

In 1986, Tanner and Mead [1] implemented an interesting constraint sat(cid:173) isfaction circuit for global motion sensing in a VLSI. We report here a new and improved a VLSI implementation that provides smooth optical flow as well as global motion in a two dimensional visual field. The com(cid:173) putation of optical flow is an ill-posed problem, which expresses itself as the aperture problem. However, the optical flow can be estimated by the use of regularization methods, in which additional constraints are intro(cid:173) duced in terms of a global energy functional that must be minimized . We show how the algorithmic constraints of Hom and Schunck [2] on com(cid:173) puting smooth optical flow can be mapped onto the physical constraints of an equivalent electronic network.

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

Learning to Estimate Scenes from Images

William Freeman, Egon Pasztor

We seek the scene interpretation that best explains image data. For example, we may want to infer the projected velocities (scene) which best explain two consecutive image frames (image). From synthetic data , we model the relationship between image and scene patches, and between a scene patch and neighboring scene patches. Given' a new image, we propagate likelihoods in a Markov network (ignoring the effect of loops) to infer the underlying scene. This yields an efficient method to form low-level scene interpretations. We demonstrate the technique for motion analysis and estimating high resolution images from low-resolution ones.

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

Learning Curves for Gaussian Processes

Peter Sollich

I consider the problem of calculating learning curves (i.e., average generalization performance) of Gaussian processes used for regres(cid:173) sion. A simple expression for the generalization error in terms of the eigenvalue decomposition of the covariance function is derived, and used as the starting point for several approximation schemes. I identify where these become exact, and compare with existing bounds on learning curves; the new approximations, which can be used for any input space dimension, generally get substantially closer to the truth.

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

Facial Memory Is Kernel Density Estimation (Almost)

Matthew Dailey, Garrison Cottrell, Thomas Busey

We compare the ability of three exemplar-based memory models, each using three different face stimulus representations, to account for the probability a human subject responded "old" in an old/new facial mem(cid:173) ory experiment. The models are 1) the Generalized Context Model, 2) SimSample, a probabilistic sampling model, and 3) MMOM, a novel model related to kernel density estimation that explicitly encodes stim(cid:173) ulus distinctiveness. The representations are 1) positions of stimuli in MDS "face space," 2) projections of test faces onto the "eigenfaces" of the study set, and 3) a representation based on response to a grid of Gabor filter jets. Of the 9 model/representation combinations, only the distinc(cid:173) tiveness model in MDS space predicts the observed "morph familiarity inversion" effect, in which the subjects' false alarm rate for morphs be(cid:173) tween similar faces is higher than their hit rate for many of the studied faces. This evidence is consistent with the hypothesis that human mem(cid:173) ory for faces is a kernel density estimation task, with the caveat that dis(cid:173) tinctive faces require larger kernels than do typical faces.

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

VLSI Implementation of Motion Centroid Localization for Autonomous Navigation

Ralph Etienne-Cummings, Viktor Gruev, Mohammed Ghani

A circuit for fast, compact and low-power focal-plane motion centroid localization is presented. This chip, which uses mixed signal CMOS components to implement photodetection, edge detection, ON-set detection and centroid localization, models the retina and superior colliculus. The centroid localization circuit uses time-windowed asynchronously triggered row and column address events and two linear resistive grids to provide the analog coordinates of the motion centroid. This VLSI chip is used to realize fast lightweight autonavigating vehicles. The obstacle avoiding line-following algorithm is discussed.

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

Discontinuous Recall Transitions Induced by Competition Between Short- and Long-Range Interactions in Recurrent Networks
N. Skantzos, C. Beckmann, Anthony Coolen

We present exact analytical equilibrium solutions for a class of recur(cid:173) rent neural network models, with both sequential and parallel neuronal dynamics, in which there is a tunable competition between nearest(cid:173) neighbour and long-range synaptic interactions. This competition is found to induce novel coexistence phenomena as well as discontinuous transitions between pattern recall states, 2-cycles and non-recall states.

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

An Integrated Vision Sensor for the Computation of Optical Flow Singular Points
Charles Higgins, Christof Koch

A robust, integrative algorithm is presented for computing the position of the focus of expansion or axis of rotation (the singular point) in optical flow fields such as those generated by self-motion. Measurements are shown of a fully parallel CMOS analog VLSI motion sensor array which computes the direction of local motion (sign of optical flow) at each pixel and can directly implement this algorithm. The flow field singular point is computed in real time with a power consumption of less than 2 m W. Computation of the singular point for more general flow fields requires measures of field expansion and rotation, which it is shown can also be computed in real-time hardware, again using only the sign of the optical flow field. These measures, along with the location of the singular point, provide robust real-time self-motion information for the visual guidance of a moving platform such as a robot.

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

Robust, Efficient, Globally-Optimized Reinforcement Learning with the Parti-Game Algorithm
Mohammad Al-Ansari, Ronald Williams

Parti-game (Moore 1994a; Moore 1994b; Moore and Atkeson 1995) is a reinforcement learning (RL) algorithm that has a lot of promise in over(cid:173) coming the curse of dimensionality that can plague RL algorithms when applied to high-dimensional problems. In this paper we introduce mod(cid:173) ifications to the algorithm that further improve its performance and ro(cid:173) bustness. In addition, while parti-game solutions can be improved locally by standard local path-improvement techniques, we introduce an add-on algorithm in the same spirit as parti-game that instead tries to improve solutions in a non-local manner.

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

Learning a Hierarchical Belief Network of Independent Factor Analyzers
Hagai Attias

Many belief networks have been proposed that are composed of binary units. However, for tasks such as object and speech recog(cid:173) nition which produce real-valued data, binary network models are usually inadequate. Independent component analysis (ICA) learns a model from real data, but the descriptive power of this model is severly limited. We begin by describing the independent factor analysis (IFA) technique, which overcomes some of the limitations of ICA. We then create a multilayer network by cascading single(cid:173) layer IFA models. At each level, the IFA network extracts real(cid:173) valued latent variables that are non-linear functions of the input data with a highly adaptive functional form

, resulting in a hier(cid:173) archical distributed representation of these data. Whereas exact maximum-likelihood learning of the network is intractable, we de(cid:173) rive an algorithm that maximizes a lower bound on the likelihood, based on a variational approach.

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

A Reinforcement Learning Algorithm in Partially Observable Environments Using Short-Term Memory

Nobuo Suematsu, Akira Hayashi

We describe a Reinforcement Learning algorithm for partially observ(cid:173) able environments using short-term memory, which we call BLHT. Since BLHT learns a stochastic model based on Bayesian Learning, the over(cid:173) fitting problem is reasonably solved. Moreover, BLHT has an efficient implementation. This paper shows that the model learned by BLHT con(cid:173) verges to one which provides the most accurate predictions of percepts and rewards, given short-term memory.

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

Maximum-Likelihood Continuity Mapping (MALCOM): An Alternative to HMMs

David Nix, John Hogden

We describe Maximum-Likelihood Continuity Mapping (MALCOM), an alternative to hidden Markov models (HMMs) for processing sequence data such as speech. While HMMs have a discrete "hidden" space con(cid:173) strained by a fixed finite-automaton architecture, MALCOM has a con(cid:173) tinuous hidden space-a continuity map-that is constrained only by a smoothness requirement on paths through the space. MALCOM fits into the same probabilistic framework for speech recognition as HMMs, but it represents a more realistic model of the speech production process. To evaluate the extent to which MALCOM captures speech production information, we generated continuous speech continuity maps for three speakers and used the paths through them to predict measured speech articulator data. The median correlation between the MALCOM paths obtained from only the speech acoustics and articulator measurements was 0.77 on an independent test set not used to train MALCOM or the predictor. This unsupervised model achieved correlations over speak(cid:173) ers and articulators only 0.02 to 0.15 lower than those obtained using an analogous supervised method which used articulatory measurements as well as acoustics ..

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

Semiparametric Support Vector and Linear Programming Machines

Alex Smola, Thilo-Thomas Frieß, Bernhard Schölkopf

Semiparametric models are useful tools in the case where domain knowledge exists about the function to be estimated or emphasis is put onto understandability of the model. We extend two learning algorithms - Support Vector machines and Linear Programming machines to this case and give experimental results for SV ma(cid:173) chines.

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

Regularizing AdaBoost

Gunnar Rätsch, Takashi Onoda, Klaus R. Müller

Boosting methods maximize a hard classification margin and are known as powerful techniques that do not exhibit overfitting for low noise cases. Also for noisy data boosting will try to enforce a hard margin and thereby give too much weight to outliers, which then leads to the dilemma of non-smooth fits and overfitting. Therefore we propose three algorithms to allow for soft margin classification by introducing regularization with slack variables into the boosting concept: (1) AdaBoostreg and regularized versions of (2) linear and (3) quadratic programming AdaBoost. Experiments show the usefulness of the proposed algorithms in comparison to another soft margin classifier: the support vector machine.

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

Graph Matching for Shape Retrieval

Benoit Huet, Andrew Cross, Edwin Hancock

We propose a new in-sample cross validation based method (randomized GACV) for choosing smoothing or bandwidth parameters that govern the bias-variance or fi

t-complexity tradeoff in 'soft' classification. Soft clas(cid:173) sification refers to a learning procedure which estimates the probability that an example with a given attribute vector is in class 1 vs class O. The target for optimizing the the tradeoff is the Kullback-Liebler distance between the estimated probability distribution and the 'true' probabil(cid:173) ity distribution, representing knowledge of an infinite population. The method uses a randomized estimate of the trace of a Hessian and mimics cross validation at the cost of a single relearning with perturbed outcome data.
************************************
DTs: Dynamic Trees
Christopher Williams, Nicholas Adams
In this paper we introduce a new class of image models, which we call dynamic trees or DTs. A dynamic tree model specifies a prior over a large number of trees, each one of which is a tree-structured belief net (TSBN). Experiments show that DTs are capable of generating images that are less blocky, and the models have better translation invariance properties than a fixed, "balanced" TSBN. We also show that Simulated Annealing is effective at finding trees which have high posterior probability.
************************************
Adding Constrained Discontinuities to Gaussian Process Models of Wind Fields
Dan Cornford, Ian Nabney, Christopher Williams
Gaussian Processes provide good prior models for spatial data, but can be too smooth. In many physical situations there are discontinuities along bounding surfaces, for example fronts in near-surface wind fields. We describe a modelling method for such a constrained discontinuity and demonstrate how to infer the model parameters in wind fields with MCMC sampling.
************************************
The Belief in TAP
Yoshiyuki Kabashima, David Saad
We show the similarity between belief propagation and TAP, for decoding corrupted messages encoded by Sourlas's method. The latter is a special case of the Gallager error-correcting code, where the code word comprises products of J{ bits selected randomly from the original message. We examine the efficacy of solutions obtained by the two methods for various values of J{ and show that solutions for J{ 2': 3 may be sensitive to the choice of initial conditions in the case of unbiased patterns. Good approximations are obtained generally for J{ = 2 and for biased patterns in the case of J{ 2': 3, especially when Nishimori's temperature is being used.
************************************
Synergy and Redundancy among Brain Cells of Behaving Monkeys
Itay Gat, Naftali Tishby
Determining the relationship between the activity of a single nerve cell to that of an entire population is a fundamental question that bears on the basic neural computation paradigms. In this paper we apply an information theoretic approach to quantify the level of cooperative activity among cells in a behavioral context. It is possible to discriminate between synergetic activity of the cells vs . redundant activity, depending on the difference between the infor(cid:173) mation they provide when measured jointly and the information they provide independently. We define a synergy value that is pos(cid:173) itive in the first case and negative in the second and show that the synergy value can be measured by detecting the behavioral mode of the animal from simultaneously recorded activity of the cells. We observe that among cortical cells positive synergy can be found, while cells from the basal ganglia, active during the same task, do not exhibit similar synergetic activity.
************************************
Classification on Pairwise Proximity Data
Thore Graepel, Ralf Herbrich, Peter Bollmann-Sdorra, Klaus Obermayer
We investigate the problem of learning a classification task on data represented in terms of their pairwise proximities. This representa(cid:173) tion does

not refer to an explicit feature representation of the data items and is thus more general than the standard approach of us(cid:173) ing Euclidean feature vectors, from which pairwise proximities can always be calculated. Our first approach is based on a combined linear embedding and classification procedure resulting in an ex(cid:173) tension of the Optimal Hyperplane algorithm to pseudo-Euclidean data. As an alternative we present another approach based on a linear threshold model in the proximity values themselves, which is optimized using Structural Risk Minimization. We show that prior knowledge about the problem can be incorporated by the choice of distance measures and examine different metrics W.r.t. their gener(cid:173) alization. Finally, the algorithms are successfully applied to protein structure data and to data from the cat's cerebral cortex. They show better performance than K-nearest-neighbor classification.

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

Using Analytic QP and Sparseness to Speed Training of Support Vector Machines
John Platt

Training a Support Vector Machine (SVM) requires the solution of a very large quadratic programming (QP) problem. This paper proposes an al(cid:173) gorithm for training SVMs: Sequential Minimal Optimization, or SMO. SMO breaks the large QP problem into a series of smallest possible QP problems which are analytically solvable. Thus, SMO does not require a numerical QP library. SMO's computation time is dominated by eval(cid:173) uation of the kernel, hence kernel optimizations substantially quicken SMO. For the MNIST database, SMO is 1.7 times as fast as PCG chunk(cid:173) ing; while for the UCI Adult database and linear SVMs, SMO can be 1500 times faster than the PCG chunking algorithm.

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

Dynamically Adapting Kernels in Support Vector Machines
Nello Cristianini, Colin Campbell, John Shawe-Taylor

The kernel-parameter is one of the few tunable parameters in Sup(cid:173) port Vector machines, controlling the complexity of the resulting hypothesis . Its choice amounts to model selection and its value is usually found by means of a validation set. We present an algo(cid:173) rithm which can automatically perform model selection with little additional computational cost and with no need of a validation set . In this procedure model selection and learning are not separate, but kernels are dynamically adjusted during the learning process to find the kernel parameter which provides the best possible upper bound on the generalisation error. Theoretical results motivating the approach and experimental results confirming its validity are presented.

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

Temporally Asymmetric Hebbian Learning, Spike liming and Neural Response Variability
L. Abbott, Sen Song

Recent experimental data indicate that the strengthening or weakening of synaptic connections between neurons depends on the relative timing of pre- and postsynaptic action potentials. A Hebbian synaptic modification rule based on these data leads to a stable state in which the excitatory and inhibitory inputs to a neuron are balanced, producing an irregular pattern of firing. It has been proposed that neurons in vivo operate in such a mode.

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

Learning Mixture Hierarchies
Nuno Vasconcelos, Andrew Lippman

The hierarchical representation of data has various applications in do(cid:173) mains such as data mining, machine vision, or information retrieval. In this paper we introduce an extension of the Expectation-Maximization (EM) algorithm that learns mixture hierarchies in a computationally ef(cid:173) ficient manner. Efficiency is achieved by progressing in a bottom-up fashion, i.e. by clustering the mixture components of a given level in the hierarchy to obtain those of the level above. This cl ustering requires onl y knowledge of the mixture paramet ers, there being no need to resort to intermediate samples. In addition to practical applications, the algorithm allows a new interpretation of EM that makes clear the relationship with non-parametric kernel-based estimation methods, provides explicit con(cid:173) trol over the trade-off between the bias and variance of EM estimates, and offers new insights about the behavior of deterministic annealing methods commonly used with EM to escape local minima of the likelihood.

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

On the Optimality of Incremental Neural Network Algorithms
Ron Meir, Vitaly Maiorov

We study the approximation of functions by two-layer feedforward neu(cid:173) ral networks, focusing on incremental algorithms which greedily add units, estimating single unit parameters at each stage. As opposed to standard algorithms for fixed architectures, the optimization at each stage is performed over a small number of parameters, mitigating many of the difficult numerical problems inherent in high-dimensional non-linear op(cid:173) timization. We establish upper bounds on the error incurred by the al(cid:173) gorithm, when approximating functions from the Sobolev class, thereby extending previous results which only provided rates of convergence for functions in certain convex hulls of functional spaces. By comparing our results to recently derived lower bounds, we show that the greedy algo(cid:173) rithms are nearly optimal. Combined with estimation error results for greedy algorithms, a strong case can be made for this type of approach.

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

Optimizing Admission Control while Ensuring Quality of Service in Multimedia Networks via Reinforcement Learning
Timothy Brown, Hui Tong, Satinder Singh

This paper examines the application of reinforcement learning to a telecommunications networking problem . The problem requires that rev(cid:173) enue be maximized while simultaneously meeting a quality of service constraint that forbids entry into certain states. We present a general solution to this multi-criteria problem that is able to earn significantly higher revenues than alternatives.

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

Coding Time-Varying Signals Using Sparse, Shift-Invariant Representations
Michael Lewicki, Terrence J. Sejnowski

A common way to represent a time series is to divide it into short(cid:173) duration blocks, each of which is then represented by a set of basis functions. A limitation of this approach, however, is that the tem(cid:173) poral alignment of the basis functions with the underlying structure in the time series is arbitrary. We present an algorithm for encoding a time series that does not require blocking the data. The algorithm finds an efficient representation by inferring the best temporal po(cid:173) sitions for functions in a kernel basis. These can have arbitrary temporal extent and are not constrained to be orthogonal. This allows the model to capture structure in the signal that may occur at arbitrary temporal positions and preserves the relative temporal structure of underlying events. The model is shown to be equivalent to a very sparse and highly over complete basis. Under this model, the mapping from the data to the representation is nonlinear, but can be computed efficiently. This form also allows the use of ex(cid:173) isting methods for adapting the basis itself to data. This approach is applied to speech data and results in a shift invariant, spike-like representation that resembles coding in the cochlear nerve.

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

Evidence for a Forward Dynamics Model in Human Adaptive Motor Control
Nikhil Bhushan, Reza Shadmehr

Based on computational principles, the concept of an internal model for adaptive control has been divided into a forward and an inverse model. However, there is as yet little evidence that learning control by the eNS is through adaptation of one or the other. Here we examine two adaptive control a rchitectures, one based only on the inverse model and other based on a combination of forward and inverse models. We then show that for reaching movements of the hand in novel force fields, only the learning of the forward model results in key characteristics of performance that match the kine(cid:173) matics of human subjects. In contrast, the adaptive control system that relies only on the inverse model fails to produce the kinematic patterns observed in the subjects, despite the fact that it is more stable. Our results provide evidence that learning control of novel dynamics is via formation of a forward model.

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

Restructuring Sparse High Dimensional Data for Effective Retrieval

Charles Isbell, Paul Viola

The task in text retrieval is to find the subset of a collection of documents relevant to a user's information request, usually expressed as a set of words. Classically, documents and queries are represented as vectors of word counts. In its simplest form, relevance is defined to be the dot product between a document and a query vector-a measure of the number of common terms. A central difficulty in text retrieval is that the presence or absence of a word is not sufficient to determine relevance to a query. Linear dimensionality reduction has been proposed as a tech(cid:173) nique for extracting underlying structure from the document collection. In some domains (such as vision) dimensionality reduction reduces computational com(cid:173) plexity. In text retrieval it is more often used to improve retrieval performance. We propose an alternative and novel technique that produces sparse represen(cid:173) tations constructed from sets of highly-related words. Documents and queries are represented by their distance to these sets, and relevance is measured by the number of common clusters. This technique significantly improves retrieval per(cid:173) formance, is efficient to compute and shares properties with the optimal linear projection operator and the independent components of documents.

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

Classification in Non-Metric Spaces

Daphna Weinshall, David Jacobs, Yoram Gdalyahu

A key question in vision is how to represent our knowledge of previously encountered objects to classify new ones. The answer depends on how we determine the similarity of two objects. Similarity tells us how relevant each previously seen object is in determining the category to which a new object belongs. Here a dichotomy emerges. Complex notions of similar(cid:173) ity appear necessary for cognitive models and applications, while simple notions of similarity form a tractable basis for current computational ap(cid:173) proaches to classification. We explore the nature of this dichotomy and why it calls for new approaches to well-studied problems in learning. We begin this process by demonstrating new computational methods for supervised learning that can handle complex notions of similarity. (1) We discuss how to implement parametric met.hods that represent a class by its mean when using non-metric similarity functions; and (2) We review non-parametric methods that we have developed using near(cid:173) est neighbor classification in non-metric spaces. Point (2) , and some of the background of our work have been described in more detail in [8].

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

Approximate Learning of Dynamic Models

Xavier Boyen, Daphne Koller

Inference is a key component in learning probabilistic models from par(cid:173) tially observable data. When learning temporal models, each of the many inference phases requires a traversal over an entire long data se(cid:173) quence; furthermore, the data structures manipulated are exponentially large, making this process computationally expensive. In [2], we describe an approximate inference algorithm for monitoring stochastic processes, and prove bounds on its approximation error. In this paper, we apply this algorithm as an approximate forward propagation step in an EM algorithm for learning temporal Bay esian networks. We provide a related approxi(cid:173) mation for the backward step, and prove error bounds for the combined algorithm. We show empirically that, for a real-life domain, EM using our inference algorithm is much faster than EM using exact inference, with almost no degradation in quality of the learned model. We extend our analysis to the online learning task, showing a bound on the error resulting from restricting attention to a small window of observations. We present an online EM learning algorithm for dynamic systems, and show that it learns much faster than standard offline EM.

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

Independent Component Analysis of Intracellular Calcium Spike Data

Klaus Prank, Julia Börger, Alexander von zur Mühlen, Georg Brabant, Christof Schöfl

Calcium (Ca2+)is an ubiquitous intracellular messenger which reg(cid:173) ulates cellular processes, such as secretion, contraction, and cell proliferation. A number of different cell types respond to hormonal stimuli with periodic oscillations of the intracellular free calcium concentration ([Ca2+]i). These Ca2+ signals are often organized in complex temporal and spatial patterns even under conditions of sustained stimulation. Here we study the spatio-temporal as(cid:173) pects of intracellular calcium ([Ca 2+]i) oscillations in clonal J3-cells (hamster insulin secreting cells, HIT) under pharmacological stim(cid:173) ulation (Schofi et al., 1996). We use a novel fast fixed-point al(cid:173) gorithm (Hyvarinen and Oja, 1997) for Independent Component Analysis (ICA) to blind source separation of the spatio-temporal dynamics of [Ca2+]i in a HIT-cell. Using this approach we find two significant independent components out of five differently mixed in(cid:173) put signals: one [Ca2+]i signal with a mean oscillatory period of 68s and a high frequency signal with a broadband power spectrum with considerable spectral density. This results is in good agree(cid:173) ment with a study on high-frequency [Ca2+]j oscillations (Palus et al., 1998) Further theoretical and experimental studies have to be performed to resolve the question on the functional impact of intracellular signaling of these independent [Ca2+]i signals.

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

A Model for Associative Multiplication

G. Christianson, Suzanna Becker

Despite the fact that mental arithmetic is based on only a few hun(cid:173) dred basic facts and some simple algorithms, humans have a diffi(cid:173) cult time mastering the subject, and even experienced individuals make mistakes. Associative multiplication, the process of doing multiplication by memory without the use of rules or algorithms, is especially problematic. Humans exhibit certain characteristic phenomena in performing associative multiplications, both in the type of error and in the error frequency. We propose a model for the process of associative multiplication, and compare its perfor(cid:173) mance in both these phenomena with data from normal humans and from the model proposed by Anderson et al (1994).

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

Batch and On-Line Parameter Estimation of Gaussian Mixtures Based on the Joint Entropy

Yoram Singer, Manfred K. K. Warmuth

We describe a new iterative method for parameter estimation of Gaus(cid:173) sian mixtures. The new method is based on a framework developed by Kivinen and Warmuth for supervised on-line learning. In contrast to gra(cid:173) dient descent and EM, which estimate the mixture's covariance matrices, the proposed method estimates the inverses of the covariance matrices. Furthennore, the new parameter estimation procedure can be applied in both on-line and batch settings. We show experimentally that it is typi(cid:173) cally faster than EM, and usually requires about half as many iterations as EM.

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

Learning Macro-Actions in Reinforcement Learning

Jette Randlov

We present a method for automatically constructing macro-actions from scratch f
rom primitive actions during the reinforcement learning process. The overall
idea is to reinforce the tendency to perform action b after action a if s
uch a pattern of actions has been rewarded. We test the method on a bic
ycle task, the car-on-the-hill task, the race-track task and some grid-world ta
sks. For the bicycle and race-track tasks the use of macro-actions approxim
ately halves the learning time, while for one of the grid-world tasks the lear
ning time is reduced by a factor of 5. The method did not work for the car-o
n-the-hill task for reasons we discuss in the conclusion.
************************************

Fast Neural Network Emulation of Dynamical Systems for Computer Animation
Radek Grzeszczuk, Demetri Terzopoulos, Geoffrey E. Hinton
Computer animation through the numerical simulation of physics-based graphics
models offers unsurpassed realism, but it can be computation(cid:173) ally de
manding. This paper demonstrates the possibility of replacing the numerical sim
ulation of nontrivial dynamic models with a dramatically more efficient "N
euroAnimator" that exploits neural networks. Neu(cid:173) roAnimators are
automatically trained off-line to emulate physical dy(cid:173) namics thr
ough the observation of physics-based models in action. De(cid:173) pending
on the model, its neural network emulator can yield physically realistic a
nimation one or two orders of magnitude faster than conven(cid:173) tiona
l numerical simulation. We demonstrate NeuroAnimators for a va(cid:173) riety
of physics-based models.
************************************

Neural Networks for Density Estimation
Malik Magdon-Ismail, Amir Atiya
We introduce two new techniques for density estimation. Our ap(cid:173) pro
ach poses the problem as a supervised learning task which can be performed
using Neural Networks. We introduce a stochas(cid:173) tic method for le
arning the cumulative distribution and an analo(cid:173) gous deterministic t
echnique. We demonstrate convergence of our methods both theoretically and e
xperimentally, and provide com(cid:173) parisons with the Parzen estimate. Our
theoretical results demon(cid:173) strate better convergence properties than the
Parzen estimate.
************************************

Improved Switching among Temporally Abstract Actions
Richard S. Sutton, Satinder Singh, Doina Precup, Balaraman Ravindran
In robotics and other control applications it is commonplace to have a pre(cid:1
73) existing set of controllers for solving subtasks, perhaps hand-crafte
d or previously learned or planned, and still face a difficult problem of how
to choose and switch among the controllers to solve an overall task as well as
possible. In this paper we present a framework based on Markov decision proc
esses and semi-Markov decision processes for phrasing this problem, a basic the
orem regarding the improvement in performance that can be ob(cid:173) tained by
switching flexibly between given controllers, and example appli(cid:173) cations
of the theorem. In particular, we show how an agent can plan with these high
-level controllers and then use the results of such planning to find an even be
tter plan, by modifying the existing controllers, with negligible additional co
st and no re-planning. In one of our examples, the complexity of the problem is
reduced from 24 billion state-action pairs to less than a million state-con
troller pairs.
************************************

Dynamics of Supervised Learning with Restricted Training Sets
Anthony Coolen, David Saad
We study the dynamics of supervised learning in layered neural net(cid:
173) works, in the regime where the size p of the training set is proportiona
l to the number N of inputs. Here the local fields are no longer described
by Gaussian distributions. We use dynamical replica theory to predict t
he evolution of macroscopic observables, including the relevant error me
asures, incorporating the old formalism in the limit piN --t 00.

```
************************************
Efficient Bayesian Parameter Estimation in Large Discrete Domains
Nir Friedman, Yoram Singer
Requests for name changes in the electronic proceedings will be accepted with no
 questions asked. However name changes may cause bibliographic tracking issues.
  Authors are asked to consider this carefully and discuss it with their co-auth
ors prior to requesting a name change in the electronic proceedings.
************************************
A Polygonal Line Algorithm for Constructing Principal Curves
Balázs Kégl, Adam Krzyzak, Tamás Linder, Kenneth Zeger
Principal curves have been defined as "self consistent" smooth curves whic
h pass through the "middle" of a d-dimensional probability distri(cid:173) butio
n or data cloud. Recently, we [1] have offered a new approach by defining pr
incipal curves as continuous curves of a given length which minimize the expec
ted squared distance between the curve and points of the space randomly chosen
according to a given distribution. The new definition made it possible to car
ry out a theoretical analysis of learning principal curves from training data.
 In this paper we propose a practical construction based on the new definition.
  Simulation results demonstrate that the new algorithm compares favorably with
 previous methods both in terms of performance and computational complexity.
************************************
Neuronal Regulation Implements Efficient Synaptic Pruning
Gal Chechik, Isaac Meilijson, Eytan Ruppin
Human and animal studies show that mammalian brain undergoes massive synaptic
pruning during childhood , removing about half of the synapses until puberty
. We have previously shown that main(cid:173) taining network memory perfo
rmance while synapses are deleted, requires that synapses are properly
 modified and pruned, remov(cid:173) ing the weaker synapses. We now show t
hat neuronal regulation , a mechanism recently observed to maintain the avera
ge neuronal in(cid:173) put field , results in weight-dependent synaptic modific
ation . Under the correct range of the degradation dimension and synapti
c up(cid:173) per bound, neuronal regulation removes the weaker synapses
  and judiciously modifies the remaining synapses . It implements near o
ptimal synaptic modification, and maintains the memory perfor(cid:173) mance
of a network undergoing massive synaptic pruning. Thus , this paper show
s that in addition to the known effects of Hebbian changes, neuronal reg
ulation may play an important role in the self-organization of brain net
works during development.
************************************
Finite-Sample Convergence Rates for Q-Learning and Indirect Algorithms
Michael Kearns, Satinder Singh
In this paper, we address two issues of long-standing interest in the
 re(cid:173) inforcement learning literature. First, what kinds of perform
ance guar(cid:173) antees can be made for Q-learning after only a fini
te number of actions? Second, what quantitative comparisons can be made
  between Q-learning and model-based (indirect) approaches, which use ex
perience to estimate next-state distributions for off-line value iteratio
n? We first show that both Q-learning and the indirect approach enjoy
  rather rapid convergence to the optimal policy as a function of the
num(cid:173) ber of state transitions observed. In particular, on the o
rder of only (Nlog(1/c)/c2 )(log(N) + loglog(l/c)) transitions are suffici
ent for both algorithms to come within c of the optimal policy, in a
n idealized model that assumes the observed transitions are "well-mixed
" throughout an N-state MDP. Thus, the two approaches have roughly th
e same sample complexity. Perhaps surprisingly, this sample complexity
is far less than what is required for the model-based approach to actually co
nstruct a good approximation to the next-state distribution. The result
also shows that the amount of memory required by the model-based approac
h is closer to N than to N 2 • For either approach, to remove the a
ssumption that the observed tran(cid:173) sitions are well-mixed, we con
```

sider  a  model  in  which  the  transitions  are  determined by a fixed,  arbit
rary exploration policy.  Bounds on the number  of transitions  required  in  or
der  to  achieve  a  desired  level  of performance  are  then  related  to  the
  stationary  distribution  and  mixing  time  of  this  policy.
************************************
Global Optimisation of Neural Network Models via Sequential Sampling
João de Freitas, Mahesan Niranjan, Arnaud Doucet, Andrew Gee
We  propose a novel strategy for  training neural networks using se(cid:173) que
ntial sampling-importance resampling algorithms.  This global  optimisation stra
tegy allows  us  to learn the  probability distribu(cid:173) tion of the network
  weights  in  a  sequential framework.  It is  well  suited to applications inv
olving on-line, nonlinear, non-Gaussian or  non-stationary signal processing.
************************************
Non-Linear PI Control Inspired by Biological Control Systems
Lyndon Brown, Gregory Gonye, James Schwaber
A  non-linear  modification  to  PI  control  is  motivated  by  a  model  of a
signal transduction pathway active in mammalian blood pres(cid:173) sure regulat
ion.  This control algorithm,  labeled  PII  (proportional  with  intermittent i
ntegral),  is  appropriate for  plants requiring ex(cid:173) act set-point match
ing and disturbance attenuation in the presence  of infrequent  step  changes  i
n  load  disturbances or  set-point.  The  proportional  aspect  of the  control
ler is  independently designed  to  be  a  disturbance  attenuator  and  set-poi
nt  matching  is  achieved  by  intermittently invoking an integral controller.
 The mechanisms  observed in the Angiotensin 11/ AT1  signaling pathway are used
 to  control the switching of the integral control.  Improved performance  over
 PI control is  shown  on  a  model  of cyc1opentenol  production.  A sign chang
e in  plant gain at the desirable operating point causes  traditional  PI  contr
ol  to  result  in  an  unstable  system.  Applica(cid:173) tion  of this  new
approach  to  this  problem  results  in  stable exact  set-point matching for
achievable set-points.
************************************
Linear Hinge Loss and Average Margin
Claudio Gentile, Manfred K. K. Warmuth
We  describe a unifying method for proving relative loss  bounds for on(cid:173)
 line linear threshold classification algorithms, such as the Perceptron and  th
e Winnow algorithms.  For classification problems the discrete loss  is  used,
i.e., the total  number of prediction mistakes.  We  introduce a con(cid:173) ti
nuous loss function, called the "linear hinge loss", that can be employed  to de
rive the updates of the algorithms.  We  first  prove bounds w.r.t.  the  linear
 hinge loss  and  then  convert them  to  the  discrete  loss.  We  intro(cid:17
3) duce a notion of "average margin" of a set of examples .  We  show how  relat
ive loss  bounds based on the  linear hinge  loss can  be converted to  relative
 loss bounds i.t.o.  the discrete loss using the average margin.
************************************
Learning Instance-Independent Value Functions to Enhance Local Search
Robert Moll, Andrew Barto, Theodore Perkins, Richard S. Sutton
Reinforcement learning methods can be used to improve the performance  of local
 search  algorithms  for  combinatorial  optimization  by  learning  an  evaluat
ion  function  that  predicts  the  outcome  of search.  The  eval(cid:173) uati
on  function  is  therefore  able  to  guide  search  to  low-cost  solutions  b
etter than  can  the original  cost function.  We  describe a  reinforcement  le
arning method for enhancing local search that combines aspects of pre(cid:173) v
ious work by Zhang and Dietterich (1995) and Boyan and Moore (1997,  Boyan  1998
).  In  an  off-line  learning  phase,  a  value function  is  learned  that is
 useful for guiding search for multiple problem sizes and instances.  We illustr
ate our technique by  developing several  such functions for the  Dial-A-Ride Pr
oblem.  Our learning-enhanced local  search algorithm ex(cid:173) hibits  an  im
provement of more  then  30%  over  a  standard  local  search  algorithm.
************************************
Support Vector Machines Applied to Face Recognition

P. Phillips

Face recognition is a K class problem. where K is the number of known individuals; and support vector machines (SVMs) are a binary classi(cid:173) fication method. By reformulating the face recognition problem and re(cid:173) interpreting the output of the SVM classifier. we developed a SVM -based face recognition algorithm. The face recognition problem is formulated as a problem in difference space. which models dissimilarities between two facial images. In difference space we formulate face recognition as a two class problem. The classes are: dissimilarities between faces of the same person. and dissimilarities between faces of different people. By modifying the interpretation of the decision surface generated by SVM. we generated a similarity metric between faces that is learned from ex(cid:173) amples of differences between faces. The SVM-based algorithm is com(cid:173) pared with a principal component analysis (PeA) based algorithm on a difficult set of images from the FEREf database. Performance was mea(cid:173) sured for both verification and identification scenarios. The identification performance for SVM is 77-78% versus 54% for PCA. For verification. the equal error rate is 7% for SVM and 13 % for PCA.

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

A Theory of Mean Field Approximation

Toshiyuki Tanaka

I present a theory of mean field approximation based on information ge(cid:173) ometry. This theory includes in a consistent way the naive mean field approximation, as well as the TAP approach and the linear response the(cid:173) orem in statistical physics, giving clear information-theoretic interpreta(cid:173) tions to them.

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

Unsupervised and Supervised Clustering: The Mutual Information between Parameters and Observations

Didier Herschkowitz, Jean-Pierre Nadal

Recent works in parameter estimation and neural coding have demonstrated that optimal performance are related to the mutual information between parameters and data. We consider the mutual information in the case where the dependency in the parameter (a vector 8) of the conditional p.d.f. of each observation (a vector 0, is through the scalar product 8.~ only. We derive bounds and asymptotic behaviour for the mutual information and compare with results obtained on the same model with the" replica technique" .

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

Stationarity and Stability of Autoregressive Neural Network Processes

Friedrich Leisch, Adrian Trapletti, Kurt Hornik

We analyze the asymptotic behavior of autoregressive neural net(cid:173) work (AR-NN) processes using techniques from Markov chains and non-linear time series analysis. It is shown that standard AR-NNs without shortcut connections are asymptotically stationary. If lin(cid:173) ear shortcut connections are allowed, only the shortcut weights determine whether the overall system is stationary, hence standard conditions for linear AR processes can be used.

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

A Principle for Unsupervised Hierarchical Decomposition of Visual Scenes

Michael C. Mozer

Structure in a visual scene can be described at many levels of granular(cid:173) ity. At a coarse level, the scene is composed of objects; at a finer level, each object is made up of parts, and the parts of subparts. In this work, I propose a simple principle by which such hierarchical structure can be extracted from visual scenes: Regularity in the relations among different parts of an object is weaker than in the internal structure of a part. This principle can be applied recursively to define part-whole relationships among elements in a scene. The principle does not make use of object models, categories, or other sorts of higher-level knowledge; rather, part-whole relationships can be established based on the statistics of a set of sample visual scenes. I illustrate with a model that performs unsu(cid:173) pervised decomposition of simple scenes. The model can account for the results from a human learning experiment on the ontogeny of part(cid:173) whole relationships.

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

Gradient Descent for General Reinforcement Learning
Leemon Baird, Andrew Moore

A simple learning rule is derived, the VAPS algorithm, which can be instantiated to generate a wide range of new reinforcement(cid:173) learning algorithms. These algorithms solve a number of open problems, define several new approaches to reinforcement learning, and unify different approaches to reinforcement learning under a single theory. These algorithms all have guaranteed convergence, and include modifications of several existing algorithms that were known to fail to converge on simple MOPs. These include Q(cid:173) In addition to these learning, SARSA, and advantage learning. it also generates pure policy-search value-based algorithms reinforcement-learning algorithms, which learn optimal policies without learning a value function. search and value-based algorithms to be combined, thus unifying two very different approaches to reinforcement learning into a single Value and Policy Search (V APS) algorithm. And these algorithms converge for POMDPs without requiring a proper belief state . Simulations results are given, and several areas for future research are discussed.

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

Vertex Identification in High Energy Physics Experiments
Gideon Dror, Halina Abramowicz, David Horn

In High Energy Physics experiments one has to sort through a high flux of events, at a rate of tens of MHz, and select the few that are of interest. One of the key factors in making this decision is the location of the vertex where the interaction , that led to the event , took place. Here we present a novel solution to the problem of finding the location of the vertex, based on two feedforward neu(cid:173) ral networks with fixed architectures, whose parameters are chosen so as to obtain a high accuracy. The system is tested on simu(cid:173) lated data sets , and is shown to perform better than conventional algorithms.

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

Probabilistic Visualisation of High-Dimensional Binary Data
Michael Tipping

We present a probabilistic latent-variable framework for data visu(cid:173) alisation, a key feature of which is its applicability to binary and categorical data types for which few established methods exist. A variational approximation to the likelihood is exploited to derive a fast algorithm for determining the model parameters. Illustrations of application to real and synthetic binary data sets are given.

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

Computation of Smooth Optical Flow in a Feedback Connected Analog Network
Alan A. Stocker, Rodney Douglas

In 1986, Tanner and Mead [1] implemented an interesting constraint sat(cid:173) isfaction circuit for global motion sensing in a VLSI. We report here a new and improved a VLSI implementation that provides smooth optical flow as well as global motion in a two dimensional visual field. The com(cid:173) putation of optical flow is an ill-posed problem, which expresses itself as the aperture problem. However, the optical flow can be estimated by the use of regularization methods, in which additional constraints are intro(cid:173) duced in terms of a global energy functional that must be minimized . We show how the algorithmic constraints of Hom and Schunck [2] on com(cid:173) puting smoot h optical flow can be mapped onto the physical constraints of an equivalent electronic network.

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

Applications of Multi-Resolution Neural Networks to Mammography
Clay Spence, Paul Sajda

We have previously presented a coarse-to-fine hierarchical pyra(cid:173) mid/neural network (HPNN) architecture which combines multi(cid:173) scale image processing techniques with neural networks. In this paper we present applications of this general architecture to two problems in mammographic Computer-Aided Diagnosis (CAD). The first application is the detection of microcalcifications. The <:oarse-to-fine HPNN was designed to learn large-scale context in(cid:173) formation for detecting small objects like microcalcifications. Re(cid:173) ceiver operating characteristic (ROC) analysis suggests that the hierarchical architecture improves detection performance of a well established CAD system by roughly 50 %. The second application is to detect mammographic masses directly. Since masses are large, extended objects, the coarse-to-fine HPNN architecture is not suit(cid:173) able for this problem. Instead we construct a fine-to-coarse HPNN architecture which is designed to learn small-scale detail structure associated with the extended objects. Our initial results applying the fine-to-coarse HPNN to mass detection are encouraging, with detection performance improvements of about 36 %. We conclude that the ability of the HPNN architecture to integrate information across scales, both coarse-to-fine and fine-to-coarse, makes it well suited for detecting objects which may have contextual clues or detail structure occurring at scales other than the natural scale of the object.

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

Divisive Normalization, Line Attractor Networks and Ideal Observers
Sophie Denève, Alexandre Pouget, Peter Latham

Gain control by divisive inhibition, a.k.a. divisive normalization, has been proposed to be a general mechanism throughout the vi(cid:173) sual cortex. We explore in this study the statistical properties of this normalization in the presence of noise. Using simulations, we show that divisive normalization is a close approximation to a maximum likelihood estimator, which, in the context of population coding, is the same as an ideal observer. We also demonstrate ana(cid:173) lytically that this is a general property of a large class of nonlinear recurrent networks with line attractors. Our work suggests that divisive normalization plays a critical role in noise filtering, and that every cortical layer may be an ideal observer of the activity in the preceding layer.

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

Robot Docking Using Mixtures of Gaussians
Matthew Williamson, Roderick Murray-Smith, Volker Hansen

This paper applies the Mixture of Gaussians probabilistic model, com(cid:173) bined with Expectation Maximization optimization to the task of sum(cid:173) marizing three dimensional range data for a mobile robot. This provides a flexible way of dealing with uncertainties in sensor information, and al(cid:173) lows the introduction of prior knowledge into low-level perception mod(cid:173) ules. Problems with the basic approach were solved in several ways: the mixture of Gaussians was reparameterized to reflect the types of objects expected in the scene, and priors on model parameters were included in the optimization process. Both approaches force the optimization to find 'interesting' objects, given the sensor and object characteristics. A higher level classifier was used to interpret the results provided by the model, and to reject spurious solutions.

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

Semi-Supervised Support Vector Machines
Kristin Bennett, Ayhan Demiriz

We introduce a semi-supervised support vector machine (S3yM) method. Given a training set of labeled data and a working set of unlabeled dat a, S3YM constructs a support vector machine us(cid:173) ing both the training and working sets. We use S3 YM to solve the transduction problem using overall risk minimization (ORM) posed by Yapnik. The transduction problem is to estimate the value of a classification function at the given points in the working set. This contrasts with the standard inductive learning problem of estimating the classification function at all possible values and then using the fixed function to deduce the classes of the working set data. We propose a general S3YM model that minimizes both the misclassification error and the function capacity based on all the available data. We show how the S3YM model for I-norm lin(cid:173) ear support vector machines can be converted to a mixed-integer program and then solved exactly using integer programming. Re(cid:173) sults of S3YM and the standard I-norm support vector machine approach are compared on ten data sets. Our computational re(cid:173) sults support the statistical learning theory results showing that incorporating working data improves generalization when insuffi(cid:173) cient training information is available. In every case, S3YM either improved or showed no significant difference in generalization com(cid:173) pared to the traditional approach.

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

A Phase Space Approach to Minimax Entropy Learning and the Minutemax Approximations

James Coughlan, Alan L. Yuille

There has been much recent work on measuring image statistics and on learning probability distributions on images. We observe that the mapping from images to statistics is many-to-one and show it can be quantified by a phase space factor. This phase space approach throws light on the Minimax Entropy technique for learning Gibbs distributions on images with potentials derived from image statistics and elucidates the ambiguities that are inherent to determining the potentials. In addition, it shows that if the phase factor can be approximated by an analytic distribution then this approximation yields a swift "Minutemax" algorithm that vastly reduces the computation time for Minimax entropy learning. An illustration of this concept, using a Gaussian to approximate the phase factor, gives a good approximation to the results of Zhu and Mumford (1997) in just seconds of CPU time. The phase space approach also gives insight into the multi-scale potentials found by Zhu and Mumford (1997) and suggests that the forms of the potentials are influenced greatly by phase space considerations. Finally, we prove that probability distributions learned in feature space alone are equivalent to Minimax Entropy learning with a multinomial approximation of the phase factor.

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

Almost Linear VC Dimension Bounds for Piecewise Polynomial Networks

Peter Bartlett, Vitaly Maiorov, Ron Meir

We compute upper and lower bounds on the VC dimension of feedforward networks of units with piecewise polynomial activa(cid:173) tion functions. We show that if the number of layers is fixed, then the VC dimension grows as W log W, where W is the number of parameters in the network. This result stands in opposition to the case where the number of layers is unbounded, in which case the VC dimension grows as W 2 •

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

Bayesian Modeling of Facial Similarity

Baback Moghaddam, Tony Jebara, Alex Pentland

In previous work [6, 9, 10], we advanced a new technique for direct visual matching of images for the purposes of face recognition and image retrieval , using a probabilistic measure of similarity based primarily on a Bayesian (MAP) analysis of image differ(cid:173) ences, leading to a "dual" basis similar to eigenfaces [13]. The performance advantage of this probabilistic matching technique over standard Euclidean nearest-neighbor eigenface matching was recently demonstrated using results from DARPA's 1996 "FERET" face recognition competition, in which this probabilis

tic matching algorithm was found to be the top performer. We have further developed a simple method of replacing the costly com put ion of nonlinear (online) Bayesian similarity measures by the relatively inexpensive computation of linear (offline) subspace projections and simple (online) Euclidean norms, thus resulting in a significant computational speed-up for implementation with very large image databases as typically encountered in real-world applications.
***********************************

The Bias-Variance Tradeoff and the Randomized GACV
Grace Wahba, Xiwu Lin, Fangyu Gao, Dong Xiang, Ronald Klein, Barbara Klein

We propose a new in-sample cross validation based method (randomized GACV) for choosing smoothing or bandwidth parameters that govern the bias-variance or fit-complexity tradeoff in 'soft' classification. Soft clas(cid:173) sification refers to a learning procedure which estimates the probability that an example with a given attribute vector is in class 1 vs class 0. The target for optimizing the the tradeoff is the Kullback-Liebler distance between the estimated probability distribution and the 'true' probabil(cid:173) ity distribution, representing knowledge of an infinite population. The method uses a randomized estimate of the trace of a Hessian and mimics cross validation at the cost of a single relearning with perturbed outcome data.
***********************************

The Role of Lateral Cortical Competition in Ocular Dominance Development
Christian Piepenbrock, Klaus Obermayer

Lateral competition within a layer of neurons sharpens and localizes the response to an input stimulus. Here, we investigate a model for the ac(cid:173) tivity dependent development of ocular dominance maps which allows to vary the degree of lateral competition. For weak competition, it re(cid:173) sembles a correlation-based learning model and for strong competition, it becomes a self-organizing map. Thus, in the regime of weak compe(cid:173) tition the receptive fields are shaped by the second order statistics of the input patterns, whereas in the regime of strong competition, the higher moments and "features" of the individual patterns become important. When correlated localized stimuli from two eyes drive the cortical de(cid:173) velopment we find (i) that a topographic map and binocular, localized receptive fields emerge when the degree of competition exceeds a critical value and (ii) that receptive fields exhibit eye dominance beyond a sec(cid:173) ond critical value. For anti-correlated activity between the eyes, the sec(cid:173) ond order statistics drive the system to develop ocular dominance even for weak competition, but no topography emerges. Topography is estab(cid:173) lished only beyond a critical degree of competition.
***********************************

General Bounds on Bayes Errors for Regression with Gaussian Processes
Manfred Opper, Francesco Vivarelli

Based on a simple convexity lemma, we develop bounds for differ(cid:173) ent types of Bayesian prediction errors for regression with Gaussian processes. The basic bounds are formulated for a fixed training set. Simpler expressions are obtained for sampling from an input distri(cid:173) bution which equals the weight function of the covariance kernel, yielding asymptotically tight results. The results are compared with numerical experiments.
***********************************

Bayesian PCA
Christopher Bishop

The technique of principal component analysis (PCA) has recently been expressed as the maximum likelihood solution for a generative latent variable model. In this paper we use this probabilistic reformulation as the basis for a Bayesian treatment of PCA. Our key result is that ef(cid:173) fective dimensionality of the latent space (equivalent to the number of retained principal components) can be determined automatically as part of the Bayesian inference procedure. An important application of this framework is to mixtures of probabilistic PCA models, in which each component can determine its own effective complexity.

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

A Precise Characterization of the Class of Languages Recognized by Neural Nets under Gaussian and Other Common Noise Distributions

Wolfgang Maass, Eduardo Sontag

We consider recurrent analog neural nets where each gate is subject to Gaussian noise, or any other common noise distribution whose probabil(cid:173) ity density function is nonzero on a large set. We show that many regular languages cannot be recognized by networks of this type, for example the language {w E {O, I} * I w begins with O}, and we give a precise characterization of those languages which can be recognized. This result implies severe constraints on possibilities for constructing recurrent ana(cid:173) log neural nets that are robust against realistic types of analog noise. On the other hand we present a method for constructing feed forward analog neural nets that are robust with regard to analog noise of this type.

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

General-Purpose Localization of Textured Image Regions

Ruth Rosenholtz

We suggest a working definition of texture: Texture is stuff that is more compactly represented by its statistics than by specifying the configuration of its parts. This definition suggests that to fmd texture we look for outliers to the local statistics, and label as texture the regions with no outliers. We present a method, based upon this idea, for labeling points in natural scenes as belonging to texture regions, while simultaneously allowing us to label low(cid:173) level, bottom-up cues for visual attention. This method is based upon recent psychophysics results on processing of texture and popout.

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

Making Templates Rotationally Invariant. An Application to Rotated Digit Recognition

Shumeet Baluja

This paper describes a simple and efficient method to make template-based object classification invariant to in-plane rotations. The task is divided into two parts: orientation discrimination and classification. The key idea is to perform the orientation discrimination before the classification. This can be accom(cid:173) plished by hypothesizing, in turn, that the input image belongs to each class of interest. The image can then be rotated to maximize its similarity to the train(cid:173) ing images in each class (these contain the prototype object in an upright orien(cid:173) tation). This process yields a set of images, at least one of which will have the object in an upright position. The resulting images can then be classified by models which have been trained with only upright examples. This approach has been successfully applied to two real-world vision-based tasks: rotated handwritten digit recognition and rotated face detection in cluttered scenes.

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

Outcomes of the Equivalence of Adaptive Ridge with Least Absolute Shrinkage

Yves Grandvalet, Stéphane Canu

Adaptive Ridge is a special form of Ridge regression, balancing the quadratic penalization on each parameter of the model. It was shown to be equivalent to Lasso (least absolute shrinkage and selection operator), in the sense that both procedures produce the same estimate. Lasso can thus be viewed as a particular quadratic penalizer. From this observation, we derive a fixed point algorithm to compute the Lasso solution. The analogy provides also a new hyper-parameter for tun(cid:173) ing effectively the model complexity. We finally present a series ofpossi(cid:173) ble extensions oflasso performing sparse regression in kernel smoothing, additive modeling and neural net training.

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

Bayesian Modeling of Human Concept Learning

Joshua Tenenbaum

I consider the problem of learning concepts from small numbers of pos(cid:173) itive examples, a feat which humans perform routinely but which com(cid:173) pu ters are rarely capable of. Bridging machine learning and cognitive science perspectives, I present both theoretical analysis and an empirical study with human subjects for the simple task oflearning concepts corre(cid:173) sponding to axis-aligned rectangles in a multidimensional feature space. Existing learning models, when applied to this task, cannot explain how subjects generalize from only a few examples of the concept. I propose a principled Bayesian model based on the assumption that the examples are a random sample from the concept to be learned. The model gives precise fits to human behavior on this simple task and provides qualitati ve insights into more complex, realistic cases of concept learning.

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

Maximum Conditional Likelihood via Bound Maximization and the CEM Algorithm
Tony Jebara, Alex Pentland

We present the CEM (Conditional Expectation Maximi::ation) al(cid:173) gorithm as an extension of the EM (Expectation M aximi::ation) algorithm to conditional density estimation under missing data. A bounding and maximization process is given to specifically optimize conditional likelihood instead of the usual joint likelihood. We ap(cid:173) ply the method to conditioned mixture models and use bounding techniques to derive the model's update rules . Monotonic conver(cid:173) gence, computational efficiency and regression results superior to EM are demonstrated.

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

Source Separation as a By-Product of Regularization
Sepp Hochreiter, Jürgen Schmidhuber

This paper reveals a previously ignored connection between two important fields: regularization and independent component anal(cid:173) ysis (ICA). We show that at least one representative of a broad class of algorithms (regularizers that reduce network complexity) extracts independent features as a by-product. This algorithm is Flat Minimum Search (FMS), a recent general method for finding low-complexity networks with high generalization capability. FMS works by minimizing both training error and required weight pre(cid:173) cision. According to our theoretical analysis the hidden layer of an FMS-trained autoassociator attempts at coding each input by a sparse code with as few simple features as possible. In experi(cid:173) ments the method extracts optimal codes for difficult versions of the "noisy bars" benchmark problem by separating the underlying sources, whereas ICA and PCA fail. Real world images are coded with fewer bits per pixel than by ICA or PCA.

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

Analog VLSI Cellular Implementation of the Boundary Contour System
Gert Cauwenberghs, James Waskiewicz

We present an analog VLSI cellular architecture implementing a simpli(cid:173) . fied version of the Boundary Contour System (BCS) for real-time image processing. Inspired by neuromorphic models across several layers of visual cortex, the design integrates in each pixel the functions of sim(cid:173) ple cells, complex cells, hyper-complex cells, and bipole cells, in three orientations interconnected on a hexagonal grid. Analog current-mode CMOS circuits are used throughout to perform edge detection, local inhi(cid:173) bition, directionally selective long-range diffusive kernels, and renormal(cid:173) izing global gain control. Experimental results from a fabricated 12 x 10 pixel prototype in 1.2 J-tm CMOS technology demonstrate the robustness of the architecture in selecting image contours in a cluttered and noisy background.

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

Unsupervised Classification with Non-Gaussian Mixture Models Using ICA
Te-Won Lee, Michael Lewicki, Terrence J. Sejnowski

We present an unsupervised classification algorithm based on an ICA mixture model. The ICA mixture model assumes that the observed data can be categorized into several mutually exclusive data classes in which the components in each class are generated by a linear mixture of independent sources. The algorithm finds the independent sources, the mixing matri x for each class and also computes the class membership probability for each data point. This approach extends the Gaussian mixture model so that the classes can have non-Gaussian structure. We demonstrate that this method can learn efficient codes to represent images of natural scenes and text. The learned classes of basis functions yield a better approximation of the underlying distributions of the data, and thus can provide greater coding efficiency. We believe that this method is well suited to modeling structure in high-dimensional data and has many potential applications.

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

Barycentric Interpolators for Continuous Space and Time Reinforcement Learning

Rémi Munos, Andrew Moore

In order to find the optimal control of continuous state-space and time reinforcement learning (RL) problems, we approximate the value function (VF) with a particular class of functions called the barycentric interpolators. We establish sufficient conditions under which a RL algorithm converges to the optimal VF, even when we use approximate models of the state dynamics and the reinforce(cid:173) ment functions .

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

Exploiting Generative Models in Discriminative Classifiers

Tommi Jaakkola, David Haussler

Generative probability models such as hidden ~larkov models pro(cid:173) vide a principled way of treating missing information and dealing with variable length sequences. On the other hand , discriminative methods such as support vector machines enable us to construct flexible decision boundaries and often result in classification per(cid:173) formance superior to that of the model based approaches. An ideal classifier should combine these two complementary approaches. In this paper, we develop a natural way of achieving this combina(cid:173) tion by deriving kernel functions for use in discriminative methods such as support vector machines from generative probability mod(cid:173) els. We provide a theoretical justification for this combination as well as demonstrate a substantial improvement in the classification performance in the context of D~A and protein sequence analysis.

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

Learning a Continuous Hidden Variable Model for Binary Data

Daniel Lee, Haim Sompolinsky

A directed generative model for binary data using a small number of hidden continuous units is investigated. A clipping nonlinear(cid:173) ity distinguishes the model from conventional principal components analysis. The relationships between the correlations of the underly(cid:173) ing continuous Gaussian variables and the binary output variables are utilized to learn the appropriate weights of the network. The advantages of this approach are illustrated on a translationally in(cid:173) variant binary distribution and on handwritten digit images.

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

Perceiving without Learning: From Spirals to Inside/Outside Relations

Ke Chen, DeLiang Wang

As a benchmark task, the spiral problem is well known in neural net(cid:173) works. Unlike previous work that emphasizes learning, we approach the problem from a generic perspective that does not involve learning. We point out that the spiral problem is intrinsically connected to the in(cid:173) side/outside problem. A generic solution to both problems is proposed based on oscillatory correlation using a time delay network. Our simu(cid:173) lation results are qualitatively consistent with human performance, and we interpret human limitations in terms of synchrony and time delays, both biologically plausible. As a special case, our network without time delays can always distinguish these figures regardless of shape, position, size, and orientation.

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

Optimizing Classifers for Imbalanced Training Sets

Grigoris Karakoulas, John Shawe-Taylor

Following recent results [9, 8] showing the importance of the fat(cid:173) shattering dimension in explaining the beneficial effect of a large margin on generalization performance, the current paper investi(cid:173) gates the implications of these results for the case of imbalanced datasets and develops two approaches to setting the threshold. The approaches are incorporated into ThetaBoost, a boosting al(cid:173) gorithm for dealing with unequal loss functions. The performance of ThetaBoost and the two approaches are tested experimentally.

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

Blind Separation of Filtered Sources Using State-Space Approach
Liqing Zhang, Andrzej Cichocki

In this paper we present a novel approach to multichannel blind separation/generalized deconvolution, assuming that both mixing and demixing models are described by stable linear state-space sys(cid:173) tems. We decompose the blind separation problem into two pro(cid:173) cess: separation and state estimation. Based on the minimization of Kullback-Leibler Divergence, we develop a novel learning algo(cid:173) rithm to train the matrices in the output equation. To estimate the state of the demixing model, we introduce a new concept, called hidden innovation, to numerically implement the Kalman filter. Computer simulations are given to show the validity and high ef(cid:173) fectiveness of the state-space approach.

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

Recurrent Cortical Amplification Produces Complex Cell Responses
Frances Chance, Sacha Nelson, L. Abbott

Cortical amplification has been proposed as a mechanism for enhancing the selectivity of neurons in the primary visual cortex. Less appreciated is the fact that the same form of amplification can also be used to de-tune or broaden selectivity. Using a network model with recurrent cortical circuitry, we propose that the spatial phase invariance of complex cell responses arises through recurrent amplification of feedforward input. Neurons in the network respond like simple cells at low gain and com(cid:173) plex ceUs at high gain. Similar recurrent mechanisms may playa role in generating invariant representations of feedforward input elsewhere in the visual processing pathway.

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

Viewing Classifier Systems as Model Free Learning in POMDPs
Akira Hayashi, Nobuo Suematsu

Classifier systems are now viewed disappointing because of their prob(cid:173) lems such as the rule strength vs rule set performance problem and the credit assignment problem. In order to solve the problems, we have de(cid:173) veloped a hybrid classifier system: GLS (Generalization Learning Sys(cid:173) tem). In designing GLS, we view CSs as model free learning in POMDPs and take a hybrid approach to finding the best generalization, given the total number of rules. GLS uses the policy improvement procedure by Jaakkola et al. for an locally optimal stochastic policy when a set of rule conditions is given. GLS uses GA to search for the best set of rule conditions.

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

Reinforcement Learning Based on On-Line EM Algorithm
Masa-aki Sato, Shin Ishii

In this article, we propose a new reinforcement learning (RL) method based on an actor-critic architecture. The actor and the critic are approximated by Normalized Gaussian Networks (NGnet), which are networks of local linear regression units. The NGnet is trained by the on-line EM algorithm proposed in our pre(cid:173) vious paper. We apply our RL method to the task of swinging-up and stabilizing a single pendulum and the task of balancing a dou(cid:173) ble pendulum near the upright position. The experimental results show that our RL method can be applied to optimal control prob(cid:173) lems having continuous state/action spaces and that the method achieves good control with a small number of trial-and-errors.

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

Learning Multi-Class Dynamics
Andrew Blake, Ben North, Michael Isard

Standard techniques (eg. Yule-Walker) are available for learning Auto-Regressive process models of simple, directly observable, dy(cid:173) namical processes. When sensor noise means that dynamics are observed only approximately, learning can still been achieved via Expectation-Maximisation (EM) together with Kalman Filtering. However, this does not handle more complex dynamics, involving multiple classes of motion. For that problem, we show here how EM can be combined with the CONDENSATION algorithm, which is based on propagation of random sample-sets. Experiments have been performed with visually observed juggling, and plausible dy(cid:173) namical models are found to emerge from the learning process.

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

Markov Processes on Curves for Automatic Speech Recognition
Lawrence Saul, Mazin Rahim

We investigate a probabilistic framework for automatic speech recognition based on the intrinsic geometric properties of curves. In particular, we analyze the setting in which two variables-one continuous (~), one discrete (s )-evolve jointly in time. We sup(cid:173) pose that the vector ~ traces out a smooth multidimensional curve and that the variable s evolves stochastically as a function of the arc length traversed along this curve. Since arc length does not depend on the rate at which a curve is traversed, this gives rise to a family of Markov processes whose predictions, Pr[sl~]' are invariant to nonlinear warpings of time. We describe the use of such models, known as Markov processes on curves (MPCs), for automatic speech recognition, where ~ are acoustic feature trajec(cid:173) tories and s are phonetic transcriptions. On two tasks-recognizing New Jersey town names and connected alpha-digits- we find that MPCs yield lower word error rates than comparably trained hidden Markov models.

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

Discovering Hidden Features with Gaussian Processes Regression
Francesco Vivarelli, Christopher Williams

We study the dynamics of supervised learning in layered neural net(cid:173) works, in the regime where the size p of the training set is proportional to the number N of inputs. Here the local fields are no longer described by Gaussian distributions. We use dynamical replica theory to predict the evolution of macroscopic observables, including the relevant error measures, incorporating the old formalism in the limit piN --t 00.

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

Utilizing lime: Asynchronous Binding
Bradley C. Love

Historically, connectionist systems have not excelled at represent(cid:173) ing and manipulating complex structures. How can a system com(cid:173) posed of simple neuron-like computing elements encode complex relations? Recently, researchers have begun to appreciate that rep(cid:173) resentations can extend in both time and space. Many researchers have proposed that the synchronous firing of units can encode com(cid:173) plex representations. I identify the limitations of this approach and present an asynchronous model of binding that effectively rep(cid:173) resents complex structures. The asynchronous model extends the synchronous approach. I argue that our cognitive architecture uti(cid:173) lizes a similar mechanism.

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

Very Fast EM-Based Mixture Model Clustering Using Multiresolution Kd-Trees
Andrew Moore

Clust ering is impor ta nt in m any fields including m anufac tlll'ing , biol og~', fin ance , a nd astronomy. l\Iixturp models arp a popula r ap(cid:173) proach due to their st.atist.ical found a t.ions, and EM is a very pop(cid:173) ular l1wthocl for fillding mixture models. EM, however , requires lllany accesses of the dat a , a nd thus h as been dismissed a s imprac(cid:173) t ical (e.g. [9]) for d ata mining of enormous dataset.s.

We present a nt' \· algorit.hm, baspd on thp l1ultiresolution ~.'Cl-trees of [5] , which dramatically reelucps the cost of EtlI-baspd clusteriug , wit.h savings rising linearl:; wit.h the number of datapoints. Although prespnt.pd lwre for maximum likplihoocl estimation of Gaussian mixt.ure mod(cid:173) f'ls , it. is also applicable to non-(~aussian models (provided class densit.ies are monotonic in Mahalanobis dist.ance), mixed categori(cid:173) cal/ nUllwric clusters. anel Bayesian nwthocls such as Antoclass [1].

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

Tight Bounds for the VC-Dimension of Piecewise Polynomial Networks
Akito Sakurai
O(ws(s log d+log(dqh/ s))) and O(ws((h/ s) log q) +log(dqh/ s)) are upper bounds for the VC-dimension of a set of neural networks of units with piecewise polynomial activation functions, where s is the depth of the network, h is the number of hidden units, w is the number of adjustable parameters, q is the maximum of the number of polynomial segments of the activation function, and d is the maximum degree of the polynomials; also n(wslog(dqh/s)) is a lower bound for the VC-dimension of such a network set, which are tight for the cases s = 8(h) and s is constant. For the special case q = 1, the VC-dimension is 8(ws log d).

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

The Effect of Eligibility Traces on Finding Optimal Memoryless Policies in Partially Observable Markov Decision Processes
John Loch
Agents acting in the real world are confronted with the problem of making good decisions with limited knowledge of the environment. Partially observable Markov decision processes (POMDPs) model decision problems in which an agent tries to maximize its reward in the face of limited sensor feedback. Recent work has shown empirically that a reinforcement learning (RL) algorithm called Sarsa(A) can efficiently find optimal memoryless policies, which map current observations to actions, for POMDP problems (Loch and Singh 1998). The Sarsa(A) algorithm uses a form of short-term memory called an eligibility trace, which distributes temporally delayed rewards to observation-action pairs which lead up to the reward. This paper explores the effect of eligibility traces on the ability of the Sarsa(A) algorithm to find optimal memoryless policies. A variant of Sarsa(A) called k-step truncated Sarsa(A) is applied to four test problems taken from the recent work of Littman, Littman, Cassandra and Kaelbling, Parr and Russell, and Chrisman. The empirical results show that eligibility traces can be significantly truncated without affecting the ability of Sarsa(A) to find optimal memoryless policies for POMDPs.

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

Exploratory Data Analysis Using Radial Basis Function Latent Variable Models
Alan Marrs, Andrew Webb
Two developments of nonlinear latent variable models based on radial basis functions are discussed: in the first, the use of priors or constraints on allowable models is considered as a means of preserving data structure in low-dimensional representations for visualisation purposes. Also, a resampling approach is introduced which makes more effective use of the latent samples in evaluating the likelihood.

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

A V1 Model of Pop Out and Asymmetty in Visual Search
Zhaoping Li
Visual search is the task of finding a target in an image against a background of distractors. Unique features of targets enable them to pop out against the background, while targets defined by lacks of features or conjunctions of features are more difficult to spot. It is known that the ease of target detection can change when the roles of figure and ground are switched. The mechanisms underlying the ease of pop out and asymmetry in visual search have been elusive. This paper shows that a model of segmentation in VI ba

sed on intracortical interactions can explain many of the qualitative aspects of visual search.
************************************

### A High Performance k-NN Classifier Using a Binary Correlation Matrix Memory

Ping Zhou, Jim Austin, John Kennedy

This paper presents a novel and fast k-NN classifier that is based on a binary CMM (Correlation Matrix Memory) neural network. A robust encoding method is developed to meet CMM input requirements . A hardware implementation of the CMM is described, which gives over 200 times the speed of a current mid-range workstation, and is scaleable to very large problems. When tested on several benchmarks and compared with a simple k-NN method, the CMM classifier gave less than I % lower accuracy and over 4 and 12 times speed-up in software and hardware respectively.
************************************

### Exploring Unknown Environments with Real-Time Search or Reinforcement Learning

Sven Koenig

Learning Real-Time A* (LRTA*) is a popular control method that interleaves plan(cid:173) ning and plan execution and has been shown to solve search problems in known environments efficiently. In this paper, we apply LRTA * to the problem of getting to a given goal location in an initially unknown environment. Uninformed LRTA * with maximal lookahead always moves on a shortest path to the closest unvisited state, that is, to the closest potential goal state. This was believed to be a good exploration heuristic, but we show that it does not minimize the worst-case plan-execution time compared to other uninformed exploration methods. This result is also of interest to reinforcement-learning researchers since many reinforcement learning methods use asynchronous dynamic programming, interleave planning and plan execution, and exhibit optimism in the face of uncertainty, just like LRTA *.
************************************

### Inference in Multilayer Networks via Large Deviation Bounds

Michael Kearns, Lawrence Saul

We study probabilistic inference in large, layered Bayesian net(cid:173) works represented as directed acyclic graphs. We show that the intractability of exact inference in such networks does not preclude their effective use. We give algorithms for approximate probabilis(cid:173) tic inference that exploit averaging phenomena occurring at nodes with large numbers of parents. We show that these algorithms compute rigorous lower and upper bounds on marginal probabili(cid:173) ties of interest, prove that these bounds become exact in the limit of large networks, and provide rates of convergence.
************************************

### Basis Selection for Wavelet Regression

Kevin Wheeler, Atam Dhawan

A wavelet basis selection procedure is presented for wavelet re(cid:173) gression. Both the basis and threshold are selected using cross(cid:173) validation. The method includes the capability of incorporating prior knowledge on the smoothness (or shape of the basis functions) into the basis selection procedure. The results of the method are demonstrated using widely published sampled functions. The re(cid:173) sults of the method are contrasted with other basis function based methods.
************************************

### Where Does the Population Vector of Motor Cortical Cells Point during Reaching Movements?

Pierre Baraduc, Emmanuel Guigon, Yves Burnod

Visually-guided arm reaching movements are produced by distributed neural networks within parietal and frontal regions of the cerebral cortex. Experimental data indicate that (I) single neurons in these regions are broadly tuned to parameters of movement; (2) appropriate commands are elaborated by populations of neurons; (3) the coordinated action of neu(cid:173) rons can be visualized using a neuronal population vector (NPV). How(cid:173) ever, the N

PV provides only a rough estimate of movement parameters (direction, velocity) and may even fail to reflect the parameters of move(cid:173) ment when arm posture is changed. We designed a model of the cortical motor command to investigate the relation between the desired direction of the movement, the actual direction of movement and the direction of the NPV in motor cortex. The model is a two-layer self-organizing neural network which combines broadly-tuned (muscular) proprioceptive and (cartesian) visual information to calculate (angular) motor commands for the initial part of the movement of a two-link arm. The network was trained by motor babbling in 5 positions. Simulations showed that (1) the network produced appropriate movement direction over a large part of the workspace; (2) small deviations of the actual trajectory from the desired trajectory existed at the extremities of the workspace; (3) these deviations were accompanied by large deviations of the NPV from both trajectories. These results suggest the NPV does not give a faithful image of cortical processing during arm reaching movements.
************************************