# 2
# From ANN to Biomimetic Information Processing

Anders Lansner[1,2], Simon Benjaminsson[1], and Christopher Johansson[1]

[1] Royal Institute of Technology (KTH), Dept Computational Biology
[2] Stockholm University, Dept Numerical Analysis and Computer Science,
AlbaNova University Center SE - 106 91 Stockholm
{ala,simonbe,cjo}@csc.kth.se

**Abstract.** Artificial neural networks (ANN) are useful components in today's data analysis toolbox. They were initially inspired by the brain but are today accepted to be quite different from it. ANN typically lack scalability and mostly rely on supervised learning, both of which are biologically implausible features. Here we describe and evaluate a novel cortex-inspired hybrid algorithm. It is found to perform on par with a Support Vector Machine (SVM) in classification of activation patterns from the rat olfactory bulb. On-line unsupervised learning is shown to provide significant tolerance to sensor drift, an important property of algorithms used to analyze chemo-sensor data. Scalability of the approach is illustrated on the MNIST dataset of handwritten digits.

## 2.1 Introduction

Artificial neural networks and related learning based techniques add important functionality to today's signal processing and data analysis toolboxes. In particular, such methods excel in supervised learning and e.g. SVM challenges human performance in specific domains like recognition of isolated handwritten digits. These methods were initially inspired and motivated by analogies with the brain, but today this connection is rarely emphasized. On the contrary, ANN:s are in many aspects different from biology, for instance, by their lack of scalability to brain-sized networks, their focus on deterministic computing, and on supervised learning based on the availability of labelled training examples. All of these features are markedly non-biological.

Current knowledge about the brain suggests that its architecture is highly scalable and run on stochastic computing elements which employ Hebbian type correlation and reinforcement based learning rules rather than supervised ones. In fact, supervised error correction learning techniques are quite suspect from the point of view of neurobiology. Thorpe and Imbert reviewed the arguments some time ago but their remarks are still valid (Thorpe and Imbert 1989). Quinlan suggested that, in fact, the multilayer perceptron is super-competent on many tasks compared to humans, which reduces its plausibility as models of the brain (Quinlan 1991).

Why should we be interested in neurobiology at all? Well, in important respects, our man-made methods and artefacts still lag far behind biological systems. The latter excel in real-time, real world perception and control, handling of input from high dimensional sensor arrays, as well as holistic pattern recognition including

A. Gutiérrez and S. Marco (Eds.): Biologically Inspired Signal Processing, SCI 188, pp. 33–43.
springerlink.com © Springer-Verlag Berlin Heidelberg 2009

figure-ground separation and information fusion. They also demonstrate exceptional compactness, tolerance to hardware faults and low energy consumption. These are attractive properties also from a technological perspective.

With the increasing abundance of sensors and sensor arrays as well as massive amounts of data generated in many different applications of advanced information technology and autonomous systems there is an increasing technical interest in scalable and unsupervised approaches to learning-based data analysis and in robotics. Also, as new molecular scale computing hardware is developed, the interest in robust algorithms for stochastic computing will increase.

A serious complication is that we do not yet fully understand the computational and information processing principles underlying brain function. An increasingly important tool in brain science is quantitative modelling and numerical simulation. In the field of computational neuroscience models at different levels of biophysical detail

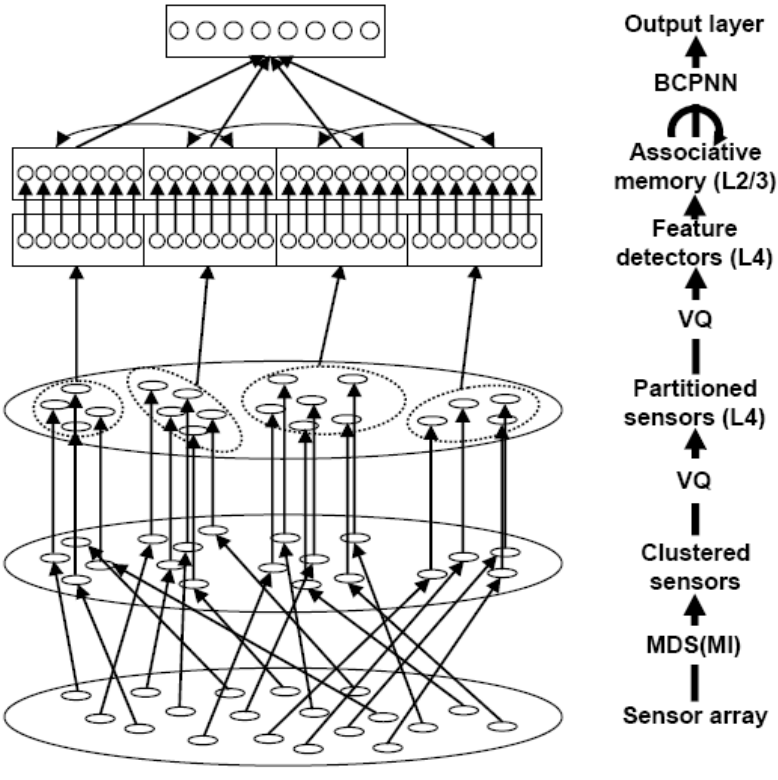



**Fig. 2.1.** Outline of the hybrid algorithm. The unstructured array of sensors is clustered using multi-dimensional scaling (MDS) with a mutual information (MI) based distance measure. Then Vector Quantization (VQ) is used to partition the sensor into correlated groups. Each such group provides input to one module of an associative memory layer. VQ is used again to provide each module unit with a specific receptive field, i.e. to become a feature detector. Finally, classification is done by means of BCPNN.

are developed and investigated to speed up the development of our understanding of how the brain works (De Schutter et al. 2005). In fact, the most reduced models in this field are formulated on a level of abstraction close to that of ANN, so called connectionist models. Such models can serve as the starting point for the design of brain-inspired computational structures, and this is the approach we have taken.

Our overall goal is the development of a generic cortex-inspired computational building block that allows for the design of modular and recursive hierarchical adaptive pattern processing structures useful in technical applications as those mentioned above. This development is in its early stages, and we report here the basic design and evaluation of a novel hybrid algorithm aimed for this purpose.

### 2.1.1 The Underlying Abstract Model of Cortex

We have previously developed and investigated biophysically detailed models of the associative memory function of neocortex based on experimental data (Lundqvist et al. 2006). Based on the knowledge gained we have formulated an abstract network model of cortical layers 2/3 that forms the core of our present approach (Lansner and Holst 1996; Sandberg et al. 2002; Johansson and Lansner 2006a). Layer 5 is also likely to be closely interacting with layers 2/3 and is not represented separately (Hirsch and Martinez 2006).

An important additional operation is the transformation from raw sensor data to the sparse and distributed representations employed in cortical layer 2/3. This transformation is started in the early sub-cortical sensory processing streams but is continued in the forward pathway of cortex that involves its layer 4 as a key component. In our abstract model we represent layer 4 separately as a layer that self-organizes a modular ("hypercolumnar") structure and also decorrelates the input forming specific receptive fields and response properties of units in this layer. The hypercolumnar structure is imposed on layer 2/3 when formed and the layer 4 units drive their companion units in layer 2/3 via specific one-to-one connections. In the simplest case, as in the simulations described in the following, there is a feedforward projection from layer 2/3 to some output layer. In general, this structure can be extended recursively with projections connecting layer 2/3 to a layer 4 in the next level in the hierarchy located in a different cortical area. Long-range recurrent connections may also form between hypercolumns within layer 2/3 at the same level, forming the basis for autoassociation.

## 2.2 Methods

The proposed algorithm for one module works in several stages (Figure 2.1). First a sensor clustering followed by a vector quantization step partitions the input space. Then each group is decorrelated and sparsified in a feature extraction step, again using vector quantization. Finally the data is fed into an associative memory which is used in a feed-forward classification setting. Each step is explained in detail below.

### 2.2.1 Partitioning of Input Space

We consider the case of sensors with discrete coded values or value intervals. For sensor $X$ and $Y$, the general dependence is calculated by the mutual information

$$I(X,Y) = \sum_{i \in Y} \sum_{j \in X} p_{ij} \log \frac{p_{ij}}{p_i p_j} \tag{2.1}$$

Here, *i* and *j* are the indexes for the units in each hypercolumn and the probabilities are estimated as

$$p_i = \frac{1}{P} \sum_{\mu=1}^{P} \xi_i^\mu \tag{2.2}$$

$$p_{ij} = \frac{1}{P} \sum_{\mu=1}^{P} \xi_i^\mu \, \xi_j^\mu \tag{2.3}$$

Where $P$ is the number of input patterns and is the unit value at position $i$ for input pattern $\mu$. In case of continuous variables the values in this step needs to be interval coded.

The mutual information is transformed into a distance measure (Kraskov et al. 2005):

$$D(X,Y) = 1 - \frac{I(X,Y)}{J(X,Y)} \tag{2.4}$$

with the joint entropy calculated as

$$J(X,Y) = -\sum_{i \in Y} \sum_{j \in X} p_{ij} \log p_{ij} \tag{2.5}$$

From the full distance measure matrix we can create a multidimensional geometric map fulfilling the distance relations by employing classical multidimensional scaling (Young 1985). The number of dimensions in this map is specified to be as low as possible (without reducing the quality of the map too much) in order to reduce the computational needs in the following step. The number of partitions of the input space is manually specified and sets the number of code vectors in a vector quantization (VQ, see below) of the map produced by the multidimensional scaling. The VQ encoding process on each element in the map decides which group each sensor should belong to. The sensors with high general dependences (as determined by the mutual information) will in this way be grouped together.

### 2.2.2 Decorrelation and Sparsification

For each group, we perform VQ on the input from the subset sensors that belongs to that specific group, resulting in a decorrelated and sparsified code well suited for an associative memory system (Steinert et al. 2006). The VQ is performed by means of Competitive Selective Learning (CSL) (Ueda and Nakano 1994), but another VQ algorithm could have been used. As for standard competitive learning, CSL updates the weight from an input unit *i* to the output unit with highest activity for input pattern $\xi^\mu$ as

$$w_{ij} = w_{ij} + \epsilon\left(\xi_i^\mu - w_{ij}\right) \tag{2.6}$$

$\epsilon$ is the step length of change which decreases during learning. CSL also adds a selection mechanism which avoids local minima by reinitializing weight vectors.

### 2.2.3 Associative Memory

The resulting decorrelated and sparsified code $\xi'^{\mu}$ for input pattern $\mu$ is fed into a BCPNN (Bayesian Confidence Propagating Neural Network) with hypercolumns that uses a supervised correlation based Bayesian learning algorithm (Johansson and Lansner 2006). Here we are only interested in classification, so the input code from the intermediate layer is directly mapped to an output layer having a single hypercolumn, using a feed-forward pass where the weights are learned by the Hebbian-Bayesian learning. If we consider the output code from a group in the intermediate layer as a hypercolumn *Qg*, where each unit corresponds to a code vector from the VQ, and the classes as units in the output hypercolumn, the weight between presynaptic unit and postsynaptic unit is computed as

$$w_{ij} = \begin{cases} 0 & p_i' = 0 \vee p_j' = 0 \\ 1/P & p_{ij}' = 0 \\ \dfrac{p_{ij}'}{p_i' p_j'} & otherwise \end{cases} \tag{2.7}$$

and $p'_i$ and $p'_{ij}$ are probabilities once again estimated according to Eqs. 2.2 and 2.3 above.

For each generated input pattern $\xi'^{\mu}$. Each unit has a bias set to be

$$\beta_j = \begin{cases} \log 1/P^2 & p_j' = 0 \\ \log\ p_j' & otherwise \end{cases} \tag{2.8}$$

When an incoming pattern is processed the activity in postsynaptic unit is calculated as

$$s_j = \log \beta_j + \sum_g \log \sum_{i \in Q_g} w_{ij} o_i \tag{2.9}$$

Here we sum over all groups and all units in each group where is the activation value of unit *i*.

The final output is calculated by a softmax function, controlled by the gain parameter *G*, over all the units in the output layer:

$$o_j = \frac{e^{Gs_j}}{\sum_{k \in Q_o} e^{Gs_k}} \tag{2.10}$$

In a classification task, the unit with the highest output is taken as the classification result.

### 2.2.4 Data Sets

In this study we used two different datasets, one of activation patterns from rat olfactory bulb and one of isolated handwritten digits.

***Rat Olfactory Bulb Activation Patterns***
The olfactory bulb activation data of Leon and Johnson was used as one of the evaluation data sets (Leon and Johnson). We used a subset comprising 2-deoxyglucose (2-DG) imaged activation patterns from 94 different odour stimuli. These spatial activation patterns were clustered in 60 different local spatial clusters. The mean activity within each such cluster was transformed to the range [0,1], whereby 94 patterns with 60 components were obtained (Marco et al. 2006).

The classification task was to separate these compounds into their chemical classes, i.e. acids (24), aldehydes (19), alcohols (16), ketones (17), esters (6), hydrocarbons (8), and misc (4). A random subset of 75% of these patterns was used for training and the rest comprised the validation set.

The robustness to sensor drift of the method under study was evaluated using a simple synthetic drift model. A gain for each of the 60 sensors was initiated to 1 after which the gain factor was subject for over 100 random-walk steps taken from a Gaussian distribution with = 0.01. In the on-line learning condition while testing drift robustness, the last unsupervised vector quantization step was run continuously.

***MNIST Data***
The MNIST data set consists of handwritten images, 28x28 pixels large with 256 gray levels (Figure 2.2). It has a training set of 60,000 samples and a test set of 10,000 samples. Specialized classifiers based on SVM have been reported to be more than 99% correct on the test set while a standard single layered network typically achieves 88% with no preprocessing (LeCun et al. 1998).

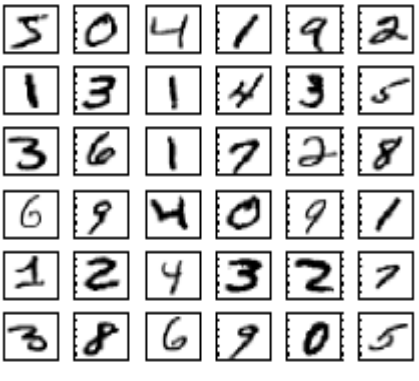
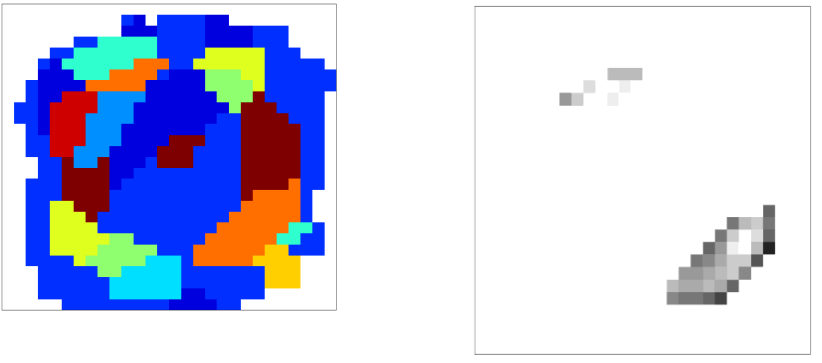


**Fig. 2.2.** Samples from the MNIST data set



(a) (b)

**Fig. 2.3.** Patches generated from the MNIST data by the MI + MDS + VQ + VQ steps of the hybrid algorithm. (a) The 12 different patches are colour coded. Note that some patches comprise more than one subfield. (b) Example of the specific receptive field of one of the 10 units in the patch marked with orange (with two subfields).

We can illustrate each step in the previous section by applying the proposed method to a real classification task.

In our example, the bit depth of all MNIST images is lowered by reducing the number of gray levels to eight. One input hypercolumn, corresponding to one image pixel, then can take on eight different values.

The general dependences between the image pixels are calculated by the mutual information. After multidimensional scaling the resulting matrix is grouped into $P$ partitions by performing vector quantization. The result is a 28x28 map which shows how the pixels should be grouped, see Figure 2.3a for the case $P = 12$. Note that this is an entirely data driven approach that is independent of sensor modality. In the case of images, this step replaces the commonly used square tiling of the image. However, such tiling can only be applied when the correlation structure of the data is known beforehand to be two-dimensional.

We again perform vector quantization on each subset of sensors and form $Q$ code vectors for each group. This gives us a total of $P \cdot Q$ units in the intermediate layer between the input and associative layer. Each code vector corresponds to a receptive field, an example of which is seen in Figure 2.3b, where we have backtracked the connections between a single code vector and the input sensors in a setting where $Q = 10$.

### 2.2.5 MLP/BP and SVM Software

The MLP/BP code used here to process the olfactory bulb activation patterns was from MATLAB® 7.3.0 NN-toolbox, using the scaled conjugate gradient learning rule with weight regularization. The SVM code used the osu-svm toolbox for MATLAB® (Ma et al. 2006). Parameters were in both cases selected to obtain best average performance on the validation set. Average and SEM of classification performance were calculated.

## 2.3 Results

This result section has three main parts, the first showing a straight-forward comparison of our novel hybrid algorithm with other methods, the second demonstrating the drift-tolerance of this algorithm relative to other methods, and the third demonstrating its scaling performance.

### 2.3.1 Evaluation on Olfactory Bulb Activation Patterns

We compared the results on the classification of the olfactory bulb activation pattern data set using different methods. The MLP/BP, one-layer, and SVM networks were

**Table 2.1.** Classification performance on validation set

| METHOD | %correct (validation) |
|---|---|
| Onelayer | 45 % |
| MLP/BP w reg | 64% |
| SVM (Poly) | 66% |
| SVM (RBF) | 70% |
| VQ-BCPNN (1) | 69% |
| VQ-BCPNN (7) | 60% |

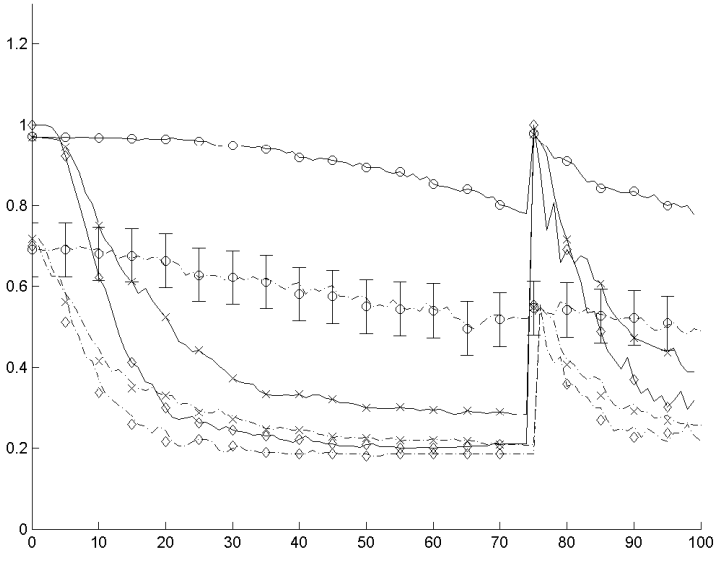



**Fig. 2.4.** Drift robustness of SVM, new method and new on-line learning method. Solid and dash-dotted lines represent performance on training and test sets respectively. Diamond, cross and circle refers to SVM, new method, and new on-line method respectively. Error bars are given only for performance of new on-line method on test data. At step 75 a complete recalibration is performed.

run as described in the methods section. The hybrid algorithm was run using two setups, one with just a single sensor partition and the other with the 60 sensors partitioned into seven groups. The total number of units in the BCPNN input layer was 70 in both cases. The results of this comparison are given in Table 2.1. The hybrid algorithm performs on par with SVM when only a single partition is used.

Drift tolerance was tested according to the description above using this data set and results are shown in Figure 2.4. As can be seen, the new algorithm with on-line learning has a much superior drift tolerance under these conditions.

### 2.3.2 Classification of MNIST Data

The algorithm was run on the entire MNIST data set. Of the 10,000 images in the test set, 95% were correctly classified (Johansson 2006). When only a feedforward configuration of BCPNN was used, with no intermediate layer generated by the hybrid algorithm, 84% of the images were correctly classified. Note that the learning in this case is not gradient descent but one-shot and correlation based.

Scaling performance of the algorithm and its dependence on the number of units in each hypercolumn is illustrated in Figure 2.5. As can be seen, the performance levels off at about 95% when there are more than one hundred units in each hypercolumn. Since there are eight hypercolumns, the total number of units in the internal layer is in this case up to one thousand.

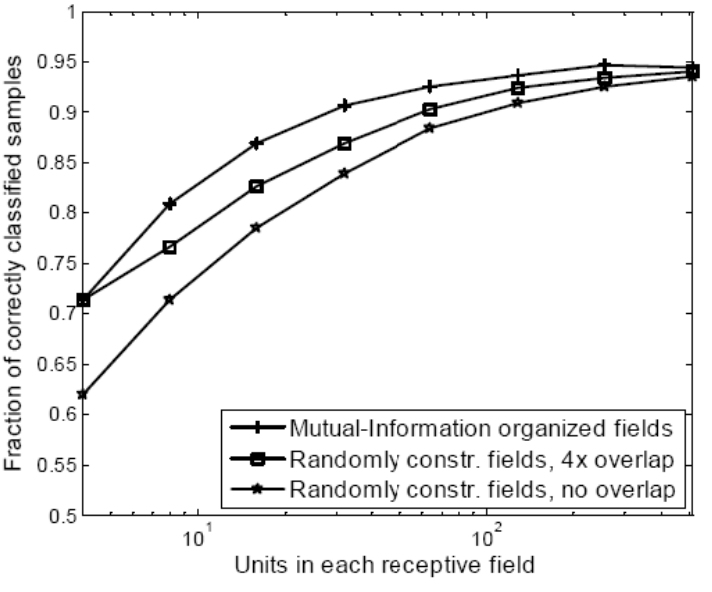



**Fig. 2.5.** Scaling performance of the new hybrid algorithm. Dependence of classification performance on the number of units in each hypercolumn (Johansson 2006).

## 2.4 Discussion and Conclusions

In this study of the performance of a novel hybrid algorithm for pattern processing we have proposed and described the different steps of the algorithm and evaluated its performance with regard to classification, drift robustness, and scaling. The algorithm is entirely data driven and does not make any assumptions on correlation structure, e.g. when processing image data.

When compared to an rbf-SVM approach on a small dataset of olfactory bulb activation patterns the new algorithm performed on par with SVM when no subdivision of the input space was done. With subdivision of the input space in seven disjoint groups, performance dropped significantly, from 70% to 60%. This suggests that the algorithm did in this case not find a set of seven independent groups of sensors. No method was able to reach beyond 70% which suggests that the problem is not separable, even non-linearily so. Comparison using more standard pattern classification benchmark datasets is ongoing.

In the test of robustness to sensor drift it was shown that when the unsupervised part of the algorithm was allowed to run in on-line training mode drift robustness much superior to SVM and the new algorithm with no on-line learning was demonstrated. This is a promising result, but further characterization of this property is required. Additional evaluation is currently ongoing on a real chemosensor dataset.

On the MNIST dataset the algorithm was able to reach 95% performance on the test set. This is not as good as a carefully designed SVM that reaches beyond 99%. On the other hand, our aim here is to develop a cortex-inspired algorithm with similar learning performance as a human being. It is not known how well humans do on the full MNIST dataset but it is not unlikely to be close to 95% (e.g. worse than SVM) given that many handwritten digits in this dataset are truly ambiguous.

Since associative memory implementing attractor dynamics, reinforcement learning and boosting approaches are all quite feasible from a biological learning perspective our ambition is to extend and evaluate this novel approach in such tasks and to focus on scalable parallel implementation to allow processing of data from arrays of millions of sensors.

## Acknowledgements

This work was partly supported by grants from the Swedish Science Council (Vetenskapsrådet, VR-621-2004-3807) and from the European Union (GOSPEL NoE FP6-2004-IST-507610 and FACETS project, FP6-2004-IST-FETPI-015879), and the Swedish Foundation for Strategic Research (via Stockholm Brain Institute).

## References

De Schutter, E., Ekeberg, Ö., Hellgren Kotaleski, J., Achard, P., Lansner, A.: Biophysically detailed modelling of microcircuits and beyond. Trends Neurosci. 28, 562–569 (2005)

Hirsch, J.A., Martinez, L.M.: Laminar processing in the visual cortical column. Curr. Opin. Neurobiol. 16, 377–384 (2006)

Johansson, C.: An attractor memory model of neocortex. School of Computer Science and Communication. Stockholm, Royal Institute of Technology, Sweden. PhD (2006)

Johansson, C., Lansner, A.: Attractor memory with self-organizing input. In: Ijspeert, A.J., Masuzawa, T., Kusumoto, S. (eds.) BioADIT 2006. LNCS, vol. 3853, pp. 265–280. Springer, Heidelberg (2006)

Johansson, C., Lansner, A.: A hierarchical brain-inspired computing system. In: International Symposium on Nonlinear Theory and its applications (NOLTA), Bologna, Italy, pp. 599–603 (2006b)

Kraskov, A., Stögbauer, H., Andrzejak, R.G., Grassberger, P.: Hierarchical clustering using mutual information. Europhys. Lett. 70(2), 278 (2005)

Lansner, A., Holst, A.: A higher order Bayesian neural network with spiking units. Int. J. Neural Systems 7(2), 115–128 (1996)

LeCun, Y.L., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE 86(11), 2278–2324 (1998)

Leon, M., Johnson, B.A.: Glomerular activity response archive (2006), `http://leonserver.bio.uci.edu`

Lundqvist, M., Rehn, M., Djurfeldt, M., Lansner, A.: Attractor dynamics in a modular network model of the neocortex. Network: Computation in Neural Systems 17, 253–276 (2006)

Ma, J., Zhao, Y., Ahalt, S., Eads, D.: OSU-SVM for Matlab (2006), `http://svm.sourceforge.net`

Marco, S., Lansner, A., Gutierrez Galvez, A.: Exploratory analysis of the rat olfactory bulb activity. Abstract. ECRO 2006, Granada, Spain (2006)

Quinlan, P.: Connectionism and psychology. A psychological perspective on connectionist research. New York, Harvester, Whaetsheaf (1991)

Sandberg, A., Lansner, A., Petersson, K.-M., Ekeberg, Ö.: Bayesian attractor networks with incremental learning. Network: Computation in Neural Systems 13(2), 179–194 (2002)

Steinert, R., Rehn, M., Lansner, A.: Recognition of handwritten digits using sparse codes generated by local feature extraction methods. In: 14th European Symposium on Artificial Neural Networks (ESANN) 2006, Brugge, Belgium, pp. 161–166 (2006)

Thorpe, S., Imbert, M.: Biological constraints on connectionist modelling. In: Pfeiffer, R., Berlin, E. (eds.) Connectionism in Perspective. Springer, Berlin (1989)

Ueda, N., Nakano, R.: A new competitive learning approach based on an equidistortion principle for designing optimal vector quantizers. Neural Networks 7(8), 1211–1227 (1994)

Young, F.W.: Multidimensional scaling. Encyclopedia of Statistical Sciences 5, 649–659 (1985)