JOURNAL ON

COMMUNICATIONS

VOLUME XLIV. MAY 1993

NEURAL NETWORKS

JOURNAL ON COMMUNICATIONS

A PUBLICATION OF THE SCIENTIFIC SOCIETY FOR TELECOMMUNICATIONS, HUNGARY

SPONSORED BY

Editorial office
Gábor Áron u. 65.
Budapest, P.O.Box 15.
Hungary, H–1525
Phone: (361) 135-1097
(361) 201-7471
Fax: (361) 135-5560

Subscription rates
Hungarian subscribers
1 year, 12 issues 4400 HUF, single copies 540 HUF
Hungarian individual subscribers
1 year, 12 issues 720 HUF, single copies 90 HUF
Foreign subscribers
12 issues 120 USD, 4 English issues 60 USD, single copies 20 USD
Transfer should be made to the Hungarian Foreign Trade Bank, Budapest, H–1821, A/C No. MKKB 203-21411

JOURNAL ON COMMUNICATIONS is published monthly, alternately in English and Hungarian by TYPOTEX Ltd. H–1015 Bp. Batthyány u. 14., phone: (361) 202-1365, fax: (361) 115-4212. Publisher: Zsuzsa Votisky. Type-setting by TYPOTEX Ltd. Printed by HUNGAPRINT, Budapest, Hungary

HUISSN 0866-5583

EDITORIAL

This special issue of the journal on Neural Networks contains four papers. These are samples rather than representatives from a growing number of researchers and engineers working on neural networks in Hungary.

Nowadays, thousands of professionals are working on neural networks, world-wide. About two years ago a breakthrough signalled the start of broad practical applications: the Intel 80170 ETANN (Electronically trainable analog neural network was the first programmable, analog, commercially available VLSI chip). This discrete-time analog chip has a speed of the order of GXPS (billion analog operations or crossing per second). It is a fully connected neural network with 64 neurons. Recently, the cellular neural network (CNN) was invented, in Professor Leon O. Chua's Laboratory in Berkeley. It is a locally connected analog processor array. Due to local interconnections the first CNN chip was able to produce about a trillion operations per second, though with fixed templates. The stored program version of the CNN, the so called CNN Universal Machine and supercomputer, provides a new capability: algorithmic programming (like the microprocessor).

In this issue the first paper deals with a very interesting problem: the use of neural networks for nonlinear dynamic system modelling. Improved convergence and modified structures with increased learning speed mark these solutions.

The next paper uses the CNN Universal Machine and give CNN algorithms for depth detection, in particular, a differential random dot stereogram is generated and used. It is shown how to "see" random dot stereograms, discovered by Béla Julesz.

The third paper considers a "classical problem": the capacity of associative memories. The connections among capacity, efficiency and complexity is a key question. The statistical analysis provides interesting a PC, provides workstation speed for CNN simulation and its software support makes possible to develop complex CNN algorithms with different input/output imaging devices.

The fourth paper, based on a short analysis of slowly converging backpropagation-type learning, introduces new transfer function and error measure to improve learning speed.

The fifth paper is an exhaustive, commercial-minded description of the CNN Workstation, a product of the Dual and Neural Computing Systems Lab. of MTA SzTAKI, which provides a development and experimentation environment in the CNN field. Being furnished with numerous external interfaces, simulators and a hardware accelerator board at that, it is the most cost-effective CNN development system for the time being.

I do hope this special issue will not only convey useful information but some readers will start working on neural networks.

[1] Intel chip & development system
[2] L. O. Chua and Lin Yang: Cellular Neural Networks: Theory and Applications, *IEEE Trans. on CAS*, Vol. 35. No. 10. October 1988.
[3] T. Roska and L. O. Chua: CNN Universal Machine, *CAS*, March 1993.
[4] B. Julesz: Foundations of Cyclopean Perception, *Chicago University Press*, Chicago, 1971.

T. ROSKA

Tamás Roska received the Diploma in electrical engineering from the Technical University of Budapest in 1964 and the Ph.D. and D.Sc. degrees in Hungary in 1973 and 1982, respectively. Since 1964 he has held various research positions. During 1964–1970 he was with the Measuring Instrument Research Institute, Budapest, between 1978 and 1982 with the Research Institute for Telecommunication, Budapest (serving also as the head of department for Circuits, Systems and Computers) and since 1982 he is scientific adviser at the Computer and Automation Institute of the Hungarian Academy of Sciences where he is presently head of the Dual (analogic) and Neural Computing Systems Research Laboratory. Professor Roska has taught several courses at the Technical University of Budapest, presently he is teaching a graduate course on "Electronic operators and neural circuits" (in Budapest and Veszprém). In 1974 and since 1980 in each year, he has been Visiting Scholar at the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory of the University of California at Berkeley. His main research areas in electronic circuits and systems have been: active circuits, computer aided design, nonlinear circuit and systems, and neural circuits, especially cellular neural networks (CNN). He has published several papers, three textbooks, and held several guest seminars at various universities and research institutions in Europe, USA and Japan. Professor Roska is a member of several Hungarian Scientific Societies, a Fellow of the IEEE. Since 1975 he is member of the Technical Committee on Nonlinear Circuits and Systems of the IEEE Circuits and Systems Society. Between 1987–89 he was the founding Secretary and now he serves as Chairman of the Hungary Section of the IEEE. Recently, he has been Associate Editor of the IEEE Transactions on Circuits and Systems, Guest Co-Editor of a special issue on Cellular Neural Networks of the International Journal of Circuit Theory and Applications and the IEEE Transactions on Circuits and Systems. He is a member of the Editorial Board of the international Journal of Circuit Theory and Applications.

APPLICATION OF NEURAL NETWORKS FOR NONLINEAR DYNAMIC SYSTEM MODELLING

R. DUNAY, G. HORVÁTH and B. PATAKI

DEPARTMENT OF MEASUREMENT AND INSTRUMENT ENGINEERING
TECHNICAL UNIVERSITY OF BUDAPEST
BUDAPEST, HUNGARY H-1521

Recently many neural network structures have been developed for modelling nonlinear dynamic operation. Some of these structures use complex systems which are combined from neural networks and linear dynamic systems [1], [2], others apply such neural networks where the basic processing elements show dynamic behaviour. This paper is dealing with some new possibilities of nonlinear system modelling. Two different approaches are presented: in the first case an inherently static network is modified to show dynamic behaviour, in the second case neural network controlled linear filters are used. The suggested networks are extensions of previously developed neural network based systems. The modifications or extensions improve their properties, increase the learning speed and/or reduce the complexity of the whole systems.

1. INTRODUCTION

One of the important application areas of neural networks is the modelling of nonlinear dynamic systems. In these applications the output to be determined depends not only on the current input but on the previous input and/or output values too.

For solving the problem of dynamic system modelling a complex system of neural network(s) and linear dynamic subsystem(s) seems to be appropriate. In this paper two cases are considered:

- In the first case complex systems are used in which the neural subsystem is directly involved in the input-output mapping. This approach is proposed by Narendra and Parthasaraty [1]. In this case the modelling system is comprising the combination of linear dynamic subsystem(s) and neural network(s). The suggested various combined structures are based on error-backpropagation multilayer networks. It can be used for modelling either linear or nonlinear systems, but the uncontrollable behaviour of the neural network can cause stability and convergence problems especially during the training phase. However, the main drawback of this approach is the extremely slow learning capability of the backpropagation neural nets. This slow trainability is the most serious obstacle of using these networks in real time adaptive applications. In the first part of the paper some modified network structures are suggested based on CMAC. CMAC neural network is a real alternative to backpropagated multilayer networks in nonlinear function approximation. It has advantageous properties like learning without local minima, incremental learning capability and much higher learning speed. Further it has a suitable architecture for hardware realization. The suggested modifications extend the CMAC network and form from the inherently static network dynamic neural nets.
- In the second case a different architecture is used where the mapping is done by the linear dynamic subsystem only and the neural subsystem is used to adapt (control) the parameters of the linear dynamics (Sztipánovits [2], Sztipánovits and Pataki [3]). In this case the neural network is not involved directly in the input-output signal mapping, instead a basically linear systems is used, where sophisticated analysis and synthesis techniques are available. The role of the neural network is reduced to control the parameters of the linear filter. In the second part of the paper the characteristics of this model are discussed , and some extensions are suggested. The extended networks improve the performance of the neural network based dynamic models especially during the learning phase, and extend the range of applications.

2. THE EXTENSIONS OF THE CMAC NETWORK

CMAC was originally developed by J. Albus in 1975 [4], but until the latest years it had not aroused much interest. Recently many extensions have been proposed e.g. [5], [6], and an increasing number of applications show that in some fields, especially in the area of learning control it can be applied successfully [7]. Here some possible further extensions are presented which form from the inherently static network a dynamic neuron net. While the advantageous properties of the original CMAC network are preserved, the modified networks are more suitable for modelling of nonlinear dynamic systems.

The CMAC network

CMAC (Cerebellar Model Articulation Controller) network can be considered as an associative memory which performs two subsequent mappings. The first one — referred to as memory addressing scheme — projects an input space point into a set of memory addresses. This set of internal addresses can be considered as a binary association vector with active bits at the addressed locations, zeros elsewhere (the terms *association vector* and *memory address set* will both be used). The association vectors always have C active elements, so every point in the input space selects C weights, stored in the addressed memory. The second mapping calculates the output of the network by summing up the selected C weights. In the association vector notation the output is calculated by a linear combiner with binary inputs.

The first mapping has the following main characteristics: (i) It maps two neighbouring input points into sets of memory locations, where only a few addresses — i.e. few weights — are different. (ii) As the distance between two input points grows, the number of the common addresses in their address sets decreases. The input points far enough from each other should not have any common addresses to avoid undesirable interference. (iii) It implements a nonlinear mapping.

In practical applications the two mappings are implemented by a two-layer network. The first layer is responsible for mapping the input points to the association vectors; this mapping is fixed (can be wired in hardware). The second layer implements the linear combination where the weights can be modified during training. This is the only trainable layer of the CMAC networks; the weights are updated by using the LMS rule.

To implement the first mapping the input variables are divided into overlapped regions where every region is subdivided into quantization levels. This quantization determines the resolution of the network and the shift positions of the overlapping regions. A given value of an input variable activates all the regions where the input is within a quantization interval of the region. To ensure C active bits in the association vectors C quantization intervals have to be used in every region. If the distance between two input points is not more then C quantization intervals then there surely are common bits in their association vectors. As in this case local generalization takes place the parameter C is often called the generalization factor. This local generalization means that the training of a point has not any or has only a weak effect on training of a different point if these points are far enough from each other. The consequence of this property is the incremental learning capability, a feature which cannot be found in backpropagation nets. Global generalization does not exist in this network.

The size of the necessary memory can be very high in case of a multi-input network. Memory size reduction can be achieved by a random compression technique. The basic idea is to project the original address rang into a smaller one. This means that all the elements of the address sets have to be hash-coded [8]. This coding can be realized for example by dividing the addresses with an appropriate polynomial. The applied coding should be uniform and suitably random.

Modified (FIR/IIR) CMAC networks

To form a dynamic net from the CMAC network some time dependence have to be built into it. FIR or IIR filters can be used in the output layer of the modified CMAC network in place of the single weights; the input layer is not changed. The structure of the FIR network is shown in Fig. 1.

Supposing that a single input single output FIR-CMAC network is used the operation of the network is described by:

$$y(k) = \sum_i z_i(k) = \sum_i \mathbf{W}_i^T(k)\mathbf{X}_i(k) \qquad (1)$$

where $\mathbf{X}_i(k) = [x_i(k), x_i(k-1), \ldots x_i(k-N_i)]^T$ is a vector of the delayed binary outputs of the first layer of the CMAC network, $\mathbf{W}_i(k) = [w_i(0), w_i(1), \ldots w_i(N_i)]^T$ is the ith filter coefficient vector and $z_i(k)$ is the output of the ith filter; $y(k)$ is the output of the network and k is the discrete time index.

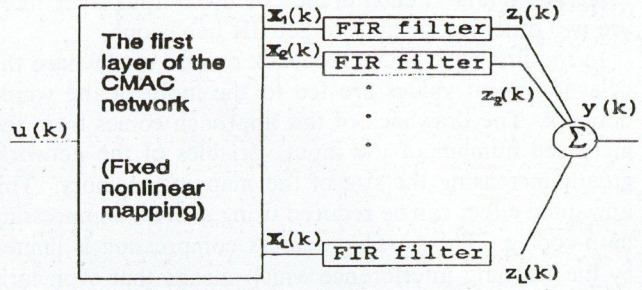


Fig. 1. The FIR-CMAC network

Because of the nonlinear mapping of the first layer, in spite of using linear FIR filters in the output layer, the whole network operates as a piecewise linear filter. Each time instant different bits will be active in the association vector which means that the filters that contribute to the output may be different from step to step. The inputs of the FIR filters are binary vectors consequently the multiplierless structure of the original CMAC networks is maintained. The only price for this time-dependent operation is a moderate increase of the size of the mapping memory.

The FIR learning rule

The learning rule is achieved by applying the LMS algorithm to the modified structure. If a desired response $d(k)$ is given, the output error $E(k)$ and the weight updating equation can be determined as:

$$E(k) = \varepsilon^2(k) = [d(k) - y(k)]^2 = [d(k) - \sum_i z_i(k)]^2 \qquad (2)$$

$$\mathbf{W}_i(k+1) = \mathbf{W}_i(k) - \mu\frac{\partial E(k)}{\partial \mathbf{W}_i(k)} = \mathbf{W}_i(k) + 2\mu\varepsilon(k)\mathbf{X}_i(k) \qquad (3)$$

Similar learning rules were derived for FIR backpropagation networks [9]. However, in that case such simple equation can be obtained only for the last layer, the learning rules for the hidden layers are much more complex. The weight vectors may be updated in each time step or only in each K steps. In the second case the output error has to be accumulated between the updating time instants.

The possibilities of the IIR extensions

When a system output depends on both the previous inputs and the previous outputs IIR structure has to be used for modelling the dynamic behaviour of the system. In these applications FIR-CMAC networks result in rather poor solutions. IIR extension of the CMAC network is also possible, however, in this case many advantageous properties of the CMAC or FIR-CMAC networks will be lost. Applying IIR structures instead of FIR elements the whole network cannot be built without multipliers. What is more, because of the feedback paths at every branch of the output layer, after a few steps of operation, most

of the IIR filters will be working, highly increasing the computational burden. One of the best features of the CMAC network, that — due to the sparse coding of the association vector — only a few weight values have to be summed up to get the output, will also be lost. Instead of using IIR filters in each branch of the output layer there are two other possibilities to get IIR behaviour.

In the first case global feedback can be used where the delayed output values are fed to the input of the whole network. The drawback of this approach comes from the increased number of the input variables of the network, greatly increasing the size of the mapping memory. This unwanted effect can be reduced using address-compressing hash-coding. The possible address compression is limited by the learning interference which means that even farly situated input points will be mapped into overlapping association vectors, making the learning of these points more difficult. The effect of learning interference can partly be overcome by repeated training of the given points.

Secondly, a single IIR filter can be added to the output of the network. This means that a FIR-CMAC network is followed by a filter where the feedback weights $q(m)$ are also trainable (Fig. 2.).

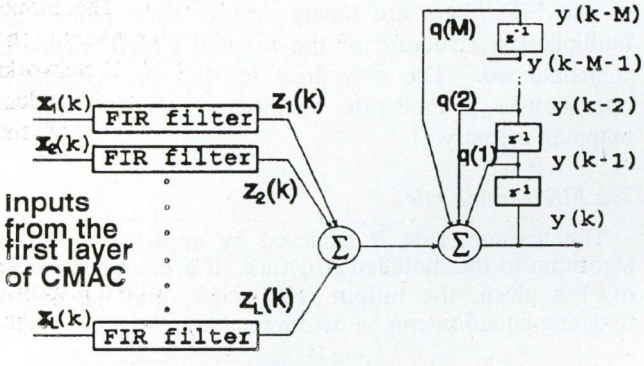


Fig. 2. The IIR extension

The derivation of the learning rule is similar to that of the FIR case. If the operation of the output layer is described by:

$$y(k) = \sum_{l=1}^{L} z_l(k) + \sum_{j=1}^{M} q(j)y(k-j), \qquad (4)$$

where

$$z_l(k) = \mathbf{W}_l^T(k)\mathbf{X}_l(k) = \sum_{n=0}^{N_l} w_l(n)x_l(k-n)$$

the weight updating equations are as follows:

$$w_i^{new}(n) = w_i^{old}(n) - \mu\frac{\partial\varepsilon^2}{\partial w_i(n)};$$

$$q^{new}(m) = q^{old}(m) - \mu\frac{\partial\varepsilon^2}{\partial q(m)} \qquad (5)$$

here:

$$\frac{\partial\varepsilon^2}{\partial w_i(n)} = -2\sum_{k=0}^{K}\varepsilon(k)\frac{\partial y(k)}{\partial w_i(n)};$$

where:

$$\frac{\partial y(k)}{\partial w_i(n)} = x_i(k-n) + \sum_{j=1}^{M} q(j)\frac{\partial y(k-j)}{\partial w_i(n)} \qquad (6)$$

and similarly

$$\frac{\partial\varepsilon^2}{\partial q(m)} = -2\sum_{k=0}^{K}\varepsilon(k)\frac{\partial y(k)}{\partial q(m)};$$

where

$$\frac{\partial y(k)}{\partial q(m)} = y(k-m) + \sum_{j=1}^{M} q(j)\frac{\partial y(k-j)}{\partial q(j)} \qquad (7)$$

The learning rule is valid when the weight values are updated only in every K steps i.e. the gradients of the accumulated error $E = \varepsilon^2 = \sum_{k=0}^{K}\varepsilon^2(k) = \sum_{k=0}^{K}[d(k) - y(k)]^2$ with respect to the weight values have to be determined. The structure and the modelling capability of this network is similar to the cascade arrangement of a static nonlinear system and a dynamic linear system, however, the training of the tow parts are not separated here.

The performance of the modified networks

The modified structures were tested by problems previously solved using dynamic backpropagation networks [1]. In the first problem a system with cascade arrangement of a static nonlinearity $f(u) = 4u(1-4u^2)^{-1}$ and a linear dynamic subsystem with transfer function $W(z) = 0.1z^{-1}+1.0z^{-2}+0.5z^{-3}$ was modelled by a FIR-CMAC network. The simulation result after 10 000 training steps using a uniformly distributed random input signal is shown in Fig. 3(a).

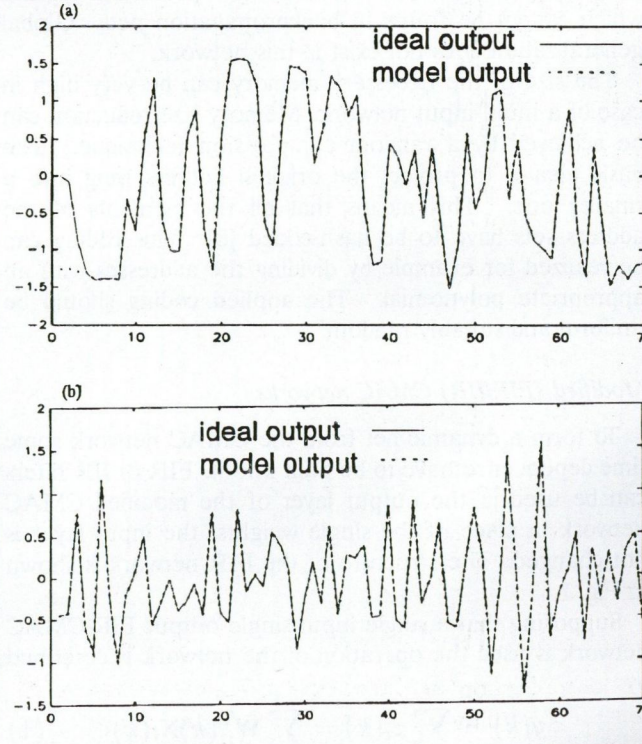


Fig. 3. The outputs of the ideal system and the modified CMAC network in the first (a) and in the second test case (b)

In the second test case a similar cascade structure was modelled by an IIR-CMAC network (Fig. 2). The transfer function of the linear part was: $W(z) = (1 - 0.3z^{-1} - 0.6z^{-2})^{-1}$, with static nonlinearity $f(u) = 0.6\sin(\pi u) + 0.3\sin(3\pi u) + 0.1\sin(5\pi u)$ [1]. Fig. 3(b) shows the results after 10 000 training steps; here again uniformly distributed random signal was used as input. In both tests the solutions were practically identical or very close to the backpropagation based solutions, but the simulation time necessary to get the results were very different. Not only the time of a learning step but the number of learning cycles were reduced. When a simulation program written in C is running on a 40 MHz 386 PC the CMAC-based solutions are approximately ten times faster than the backprop-based nets. Similar speed ratio can be obtained in other tests. The speed up factor would be increased further if parallel hardware solutions of similar complexity had been applied. A further example is shown in Fig. 4.

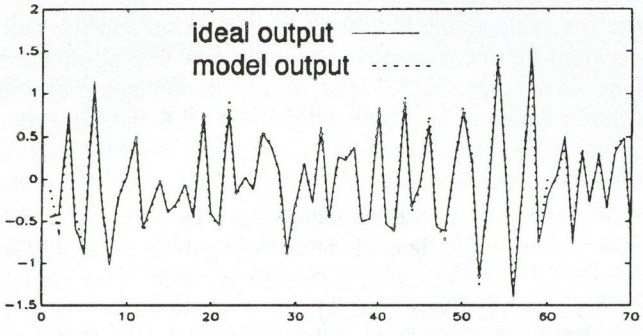


Fig. 4. The result of the second test case using the network of Fig. 5.

It is the solution of the second test problem using a feedback FIR-CMAC network containing only a single feedback path (Fig. 5). The simulation shows that IIR behaviour can be obtained using this simple FIR-CMAC structure, without the need for a longer tapped delay line in the feedback path. It is obvious that using CMAC structure where all necessary delayed output values are fed back these tasks can be solved, but because of the increasing number of the input variables we will meet the difficulties mentioned before.

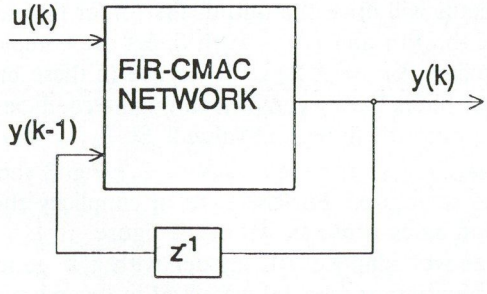


Fig. 5. FIR-CMAC with global feedback

3. NEURAL NETWORK CONTROLLED LINEAR FILTERS

The application of the backpropagation network controlled digital filter promises some helpful features. An interesting structure proposed by Sztipánovics [2] applies special linear filter component: the resonator based digital filter (RBDF) developed by Péceli [10]. The RBDF has some advantageous features; it is structurally passive, provides minimum roundoff noise, can suppress zero-input limit cycles etc. From the implementational point of view it is a highly parallel structure which provides the adaptive system with substantial advantages. On the other hand both components have drawbacks as well: (i) the RBDF structure is suitable for both FIR and IIR filtering problems, but its application in an adaptive IIR context is not straight forward because of stability problems, (ii) the neural network can exhibit poor convergence properties during training.

The filter section of the model

The proposed time-varying filter is a resonator based digital filter structure. The filter structure is based on some concepts of the observer theory. The key part of the filter is a conceptual state variable model of the input signal, where the state variables are the components of a discrete transformation (Fig. 6).

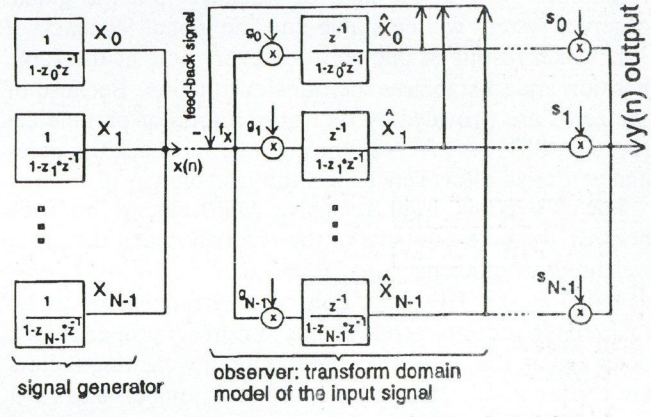


Fig. 6. Concept and structure of resonator based digital filter

Arbitrary discrete orthogonal transformation can be used but in most of the cases we prefer to use Fourier transformation, and that will be used throughout this paper. Discrete Fourier transformation is performed if we use N parallel first order complex resonators, where: (i) the poles of the resonators $(z_i = \exp(2\pi i/N)i = 0, 1, 2, \ldots, N-1)$ are the Nth roots of 1, (ii) the input weighting factors are proportional to the corresponding poles $(g_i = z_i/N \; i = 0, 1, 2, \ldots, N-1)$. In that case the outputs of the observer loop $(\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_{N-1})$ give the recursive discrete Fourier transform (RDFT) of the last N samples of the input time series $(x(n-1), x(n-2), \ldots, x(n-N))$, [10]. The sum of the outputs of the observer loop $(\hat{X}_0 + \hat{X}_1 + \ldots + \hat{X}_{N-1})$ gives a one-step prediction of the input signal $(\hat{x}(n))$ based on the last N samples. The outputs of the transformation loop always give a correct DFT of the last N input samples, but the one step prediction is not always errorless. If the inputs signal is periodic, having time period of N (i.e. having harmonic components of the resonator pole frequencies of the loop) the signal $x(n)$ can be composed of the transformation components without error. In that case there is no difference between the input $x(n)$ and the feedback signal $f_b(n)$, therefore

the feedback difference signal $f_x(n) = x(n) - f_b(n)$ is zero. The resonators work with zero input and their outputs will provide sinusoidal signals of the free running frequencies. This will be referred to as the case of correct signal model. If the input signal has different harmonics which are not embodied in the resonators, the input signal $(x(n))$ cannot be predicted from the N-point time series, therefore cannot be reconstructed from the N point DFT of this record. In that case the feedback signal can follow the input with some error only, the feedback difference signal is not zero, it will tune the resonators to give frequencies different from the free-running ones. It will be referred to as the case of not correct signal model. On the basis of the observer in Fig. 6 the realization of both FIR and IIR filters is possible. For FIR filters, the transformation loop is fixed, only the output weighting factors have to be set for the proper FIR filter transfer characteristics. In that case the observer is dead-beat in N steps, where N denotes the transformation size. Because the resonator pole positions are on the unit circle, the resonators work at the limit of stability. But the global observer system will be stable due the global feedback of -1, which results in one zero for every pole at the same position and so stabilizes the transfer function. Because of the zeros are provided by the same resonator parameters $(z_i;\ i = 0, 1, \ldots, N-1)$ through the feedback the poles and zeros will effectively cancel the instability.

For IIR filters fixed recursive DFT cannot be used because the pole positions of the resonators and the input weighting coefficients $(g_i;\ i = 0, 1 \ldots, N-1)$ will depend on the IIR filter pole zero arrangement to be realized. In order to achieve good sensitivity properties the poles are on the unit circle and the input weighting factors are chosen $g_i = r_i z_i$, where r_i are real numbers and $r_0 + r_1 + \ldots r_{N-1} \leq 1$ [10]. The neural network controlled adaptive filter structure proposed by Sztipánovics [2] is shown in Fig. 7. Only the output weighting factors of a fixed RDFT loop are adapted therefore only FIR filters can be implemented by this model. Two architectures were investigated with and without feedback $G(y)$ and it was shown that this type of feedback transformation helps to improve the noise rejection properties of the filter [2].

Extending the structure to IIR filters

The resonators of the RBDF structure work at the limit of instability. The negative feedback of -1 stabilizes the loop. But for IIR filters the parameters of the loop especially the resonator pole positions should be changed as well. Because the adaptation algorithm of the pole positions includes the same pole positions, and it is out of the negative feedback loop, unfortunately the learning process performed by the neural network will be instable. One should modify the structure such that any IIR filter use the same stable loop for transformation and the pole positions be unchanged. In that case new free parameters have to be introduced because the denominator of the transfer function has to be defined. (An IIR filter of order N has $2N-1$ free parameters, compared to a FIR filter of the same order which has N free parameters only. On the other hand the transformation is dead beat so it provides only a finite memory.) A possible solution can be found for that problem if we use the same recursive Discrete Fourier Transformation loop as in Fig. 6, but we modify the output weighting by using first order IIR filters (in form of $b_1/(1 + a_2 z^{-1})$) instead of the complex numbers $(s_i; i = 0, 1, \ldots, N-1)$. In this case the transfer function of the filter can be written in the following form:

$$H(z) = \frac{\sum_{i=0}^{N-1} [b_{1i} g_i z^{-1} \prod_{k=0, k\neq i}^{N-1} (1 + a_{2i} z^{-1})(1 - z_i z^{-1})}{\prod_{k=0}^{N-1} (1 + a_{2k} z^{-1})(1 - z_k z^{-1}) + \sum_{i=0}^{N-1} [g_i z^{-1} (1 + a_{2i} z^{-1}) \prod_{k=0, k\neq i}^{N-1} (1 + a_{2k})(1 - z_k z^{-1})]} \quad (8)$$

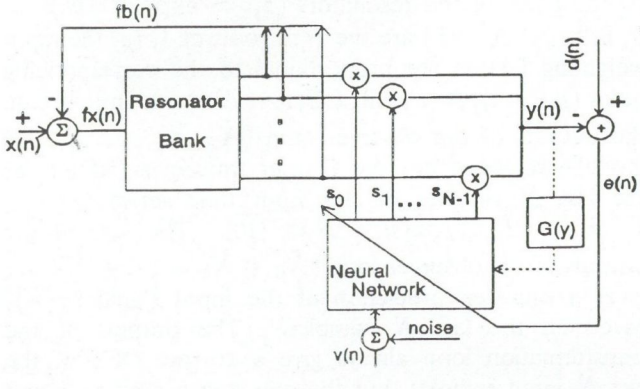


Fig. 7. Neural network controlled FIR filter structure

The training algorithm can set the parameters of the output weighting filters with some constraints to have real-valued output function for $y(n)$. (Because the loop is similar to the FIR case, Discrete Fourier Transformation components will drive the output first order filters, so the input of the kth and $(N-k)$th filters are complex conjugate pairs: $X_k = X^*_{N-k}$. If we use in these branches first order filters having complex conjugate coefficients, the resulting output will be real valued. So $b_{1i} = b_{1(N-i)^*}$; $a_{2i} = a_{2(N-i)^*}$; $i = 0, \ldots, N-1$.) Fig. 8 shows the extended structure. For the sake of simplicity the error evaluation block is not shown in the figure.

The above adaptive IIR model with the generalized error evaluation process [3] was used in several computer simulations. In Fig. 9 the ideal and the actual output signals are shown. The input signal $(x(n))$ was composed of some sinusoidal signals (some of them are in resonator positions others are between resonator positions) and a stochastic error term. The ideal output $(d(n))$ was a filtered version of the input signal by using two second order IIR low-pass filters. The filter parameters providing the ideal signal were changed periodically. The model

had 16 parallel resonators (therefore 16th order IIR filters could be formed using that filter) and it was trained to follow these changes.

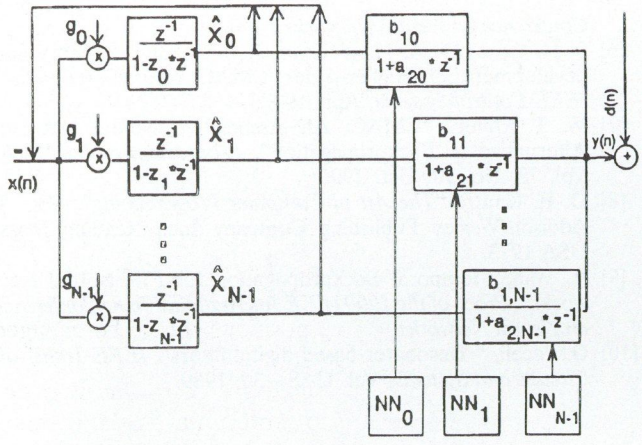


Fig. 8. Neural network controlled IIR filter

In Fig. 9 a part of the training is shown where the time window is: $9300 <= n <= 9400$. The filter characteristics providing the ideal output are changed at $n = 9325$. It is clear that the adaptive filter could follow the change after a short transient. In Fig. 10 the output error $(d(n) - y(n)$: solid line) and the feedback error of the transformation of the ideal signal $(f_d(n)$: dotted line) are shown in the same time window. One can see the transient after changing the filter providing the ideal output signal. In the period $9300 <= n <= 9325$ the low-pass filter having higher cut-off frequency is used, the output error is of lower level than the transformation error [3], because the high frequency noise could not be properly modelled by the transformation. In this part the output error is used dominantly for training. In the second part $(9325 < n < 9400)$ when lower cut-off frequency filter is used, the output error is dominant over the error of transformation. In this part dominantly the transform domain component errors are used for training.

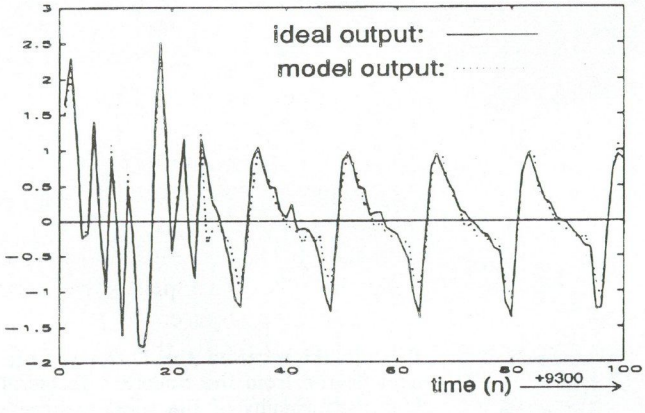


Fig. 9. Output of a time varying filter structure and the output of the model

The performance of the neural network controlled FIR structure and the controlled IIR structure was compared in several simulations. The results shown in Fig. 11 were obtained from simulations in which the models were taught using the input and output signals of a second-order IIR lowpass filter and a fourth-order IIR bandpass filter. Both the FIR and IIR models had 16 parallel branches therefore they were able to provide 16th order FIR and IIR filters respectively. (Because the FIR model had a maximum order higher than the IIR filters used as ideal systems, it was supposed to be able to approximate the transfer characteristics.) The average output error versus the training steps performed is shown in Fig. 11. The neural network controlled IIR structure had a better performance, the difference would have been even larger if the order of the ideal system was closer to the maximum possible order of the models.

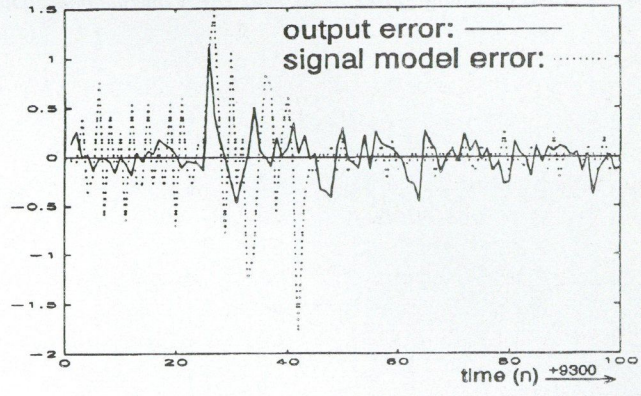


Fig. 10. Output error and feedback error signals

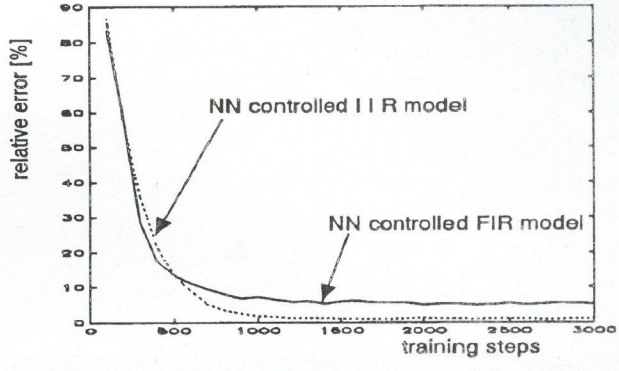


Fig. 11. Modelling IIR filters by NN controlled FIR and IIR systems

4. CONCLUSIONS

In this paper some new methods for improving the convergence of neural networks for dynamic system modelling were investigated. In the first part the CMAC network was extended, in the second part a structural modification of the resonator based filter was suggested to extend the applicability of the structure to adaptive IIR filters. Further possibility is when in the neural network controlled filter structures the backpropagation networks will be replaced by CMAC or modified CMAC networks. This will be studied in the near future.

REFERENCES

[1] S. K. Narendra, K. Pathasarathy, "Identification and Control of Dynamical Systems Using Neural Networks", *IEEE Trans. Neural Networks*, Vol. 1. 1990.

[2] J.Sztipánovics, "Dynamic Backpropagation Algorithm for Neural Network Controlled Resonator-Bank Architecture", *IEEE Trans. Circuits and Systems-II*, Vol. 39., No. 2, Feb. 1992.

[3] J. Sztipánovics, B. Pataki, "Training Algorithm for Neural Network Controlled Resonator-based Digital Filters", *Híradástechnika*, to appear (In Hungarian).

[4] J. S. Albus, "A New Approach to Manipulator Control: The Cerebellar Model Articulation Controller (CMAC)", *Journal of Dynamic Systems, Measurement and Control* Sept. 1975.

[5] Kim, Chun-Shin Lin, "Use of Adaptive Resolution for Better CMAC Learning", *Proc. of the 1992 IEEE International Joint Conference on Neural Networks*.

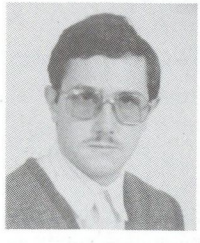
[6] S. H. Lane, D. A. Handelman, J. J. Gelfand, "Theory and development of Higher-order CMAC Neural Networks", *IEEE Control Systems*, Apr. 1992.

[7] W. T. Miller, "CMAC: An Associative Neural Network Alternative to Backpropagation", *Proceedings of the IEEE*, Vol. 78. No. 10. Oct. 1990.

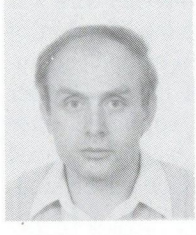
[8] D. E. Knuth, "*The Art of Computer Programming*", Vol. 3. Addison-Wesley Publishing Company Inc., Reading Mass. USA 1973.

[9] A. Wan, "Temporal Backpropagation for FIR Neural Networks", *Proc. of the 1990 IEEE International Joint Conference on Neural Networks*.

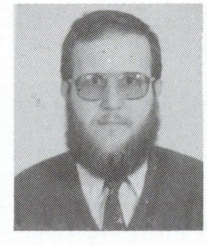
[10] G. Péceli, " Resonator-based digital filters", *IEEE Trans. on Circuits and Systems*, Vol. CAS–36, 1989.



Rezső Dunay graduated from the Budapest Technical University in 1992 as an electrical engineer. Currently, he is a first-year postgraduate student in the Department of Measurement and Instrument Engineering of Budapest Technical University. His research interest are in signal processing systems including neural networks, applications of learning structures and neural networks in complex measuring systems.



Gábor Horváth received the Diploma Engineer degree from the Budapest Technical University Faculty of Electrical Engineering in 1970. Since then he has been at the Department of Measurement and Instrument Engineering of Budapest Technical University where held various teaching positions, currently he is an Associate Professor. He received the Candidate of Engineering Science degree in digital signal processing from the Hungarian Academy of Sciences in 1988. His educational and research activity is related to microprocessor system design, digital signal processing and neural networks.



Béla Pataki received the Diploma Engineer degree from the Budapest Technical University Faculty of Electrical Engineering in 1978. Until 1982 he worked in the factory Works for Electronic Measuring Gear (EMG), he took part in development of complex test systems. Since then he has been at the Department of Measurement and Instrument Engineering of Budapest Technical University, currently he is an Assistant Lecturer. His educational and research activity is related to digital signal processing and neural network based dynamic models.

USING CNN TO "SEE" RANDOM-DOT STEREOGRAMS - DUAL CNN MODELS OF STEREO VISION

A. G. RADVÁNYI

DUAL AND NEURAL COMPUTING SYSTEMS LABORATORY
COMPUTER & AUTOMATION INSTITUTE, HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST P.O.B. 63, H-1518

The random-dot stereogram coding 3D information in its internal correlation is for probing human stereopsis. We report dual CNN algorithms that can reveal 3D surfaces coded in stereograms. The concept of difference stereogram is introduced and used for coding smooth surfaces. Its importance is due to the fact that difference stereograms of real objects can be created in an optical environment using projector and camera.

1. INTRODUCTION

An interesting question about the human visual system is whether we identify an object before we put it in its proper perspective, or whether there exists a mechanism - the Cyclopean eye - in the visual cortex, that can perceive "pure" depth based merely on the correlation and parallax of the left and right retinal images. Using random-dot stereograms [4] almost anyone can try and testify to oneself the existence of the Cyclopean mechanism.

The random-dot stereogram (RDS) is a two-dimensional, seemingly random pattern consisting of pairwise horizontally correlating internal segments. Shifting a RDS horizontally upon itself, correlating (identical) areas will overlap again and again. Viewing properly an RDS devoid of any (monocularly) observable cue, the Cyclopean eye will find the correlating areas and "see" them in different perspective depths, depending on their horizontal distance in the stereogram.

In Section 2, a brief summary of the binocular stereo vision and the Cyclopean mechanism is given. The main goal of the Section is to demonstrate that human brain is capable of finding correlation — if exists — between the two retinal images and — based on it — building up a depth pattern.

In Section 3, methods creating different types of RDSs — stereo pairs, auto-stereograms and difference stereograms — are discussed, laying stress upon the internal structure of stereograms.

In Sections 4 and 5, dual CNN algorithms that can find the perspective hidden in stereograms, are dealt with.

In Section 6, an optical set-up with projector and camera, to produce difference stereograms of real objects is outlined. Potentially, if connected to a dual CNN hardware it can detect 3D depth variations in its scope.

2. FUNDAMENTALS OF HUMAN STEREO-VISION

It is a well-known experience, that two points lying in the surrounding world at different depths are separated by different distances in the left and right retinal images, due to the different positions of the two eyes. In other words, objects lying at different depths are seen at different angles by the two eyes (Fig. 1.). The difference of angles called relative binocular disparity has fundamental role in perceiving stereodepth. In the horizontal plane of the two eyes, the points having zero relative disparity with respect to a specific point A, are seen at the same depth as A.

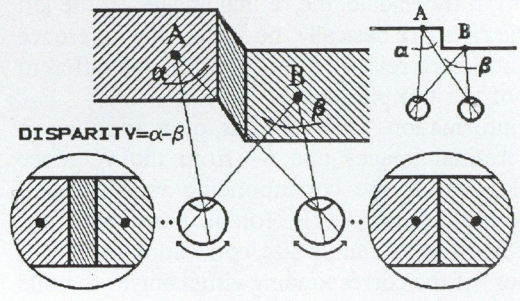


Fig. 1. Retinal images of a step on wall

When we fixate a point A, i.e. its image falls on the fovea — the point of highest visual acuity — in both eyes, every other points of zero relative disparity have their images in "corresponding points" of the two retinas. We see objects with images in corresponding points of the retinas as single. The points and objects of non-zero relative disparity, being significantly behind or before the fixation depth thus have their retinal images at a distance from the corresponding points, are seen as double. In between, there is the area for binocular single vision and for stereopsis. The retinal images of an object in the single vision area, falling close to corresponding points, are fused by the brain, giving a single percept of the object.

It was the astronomer Johannes Kepler who first suggested that the binocular disparity is the stimulus that produces depth sensation. In the Keplerian model, the brain — when deciding the relative spatial depth of two objects — first selects the corresponding images in the left and right retinas, and then calculates their relative disparity. In case of two identical objects, however, the four retinal images are identical too, that might render the selection of corresponding images ambiguous. The outcome of a mistaken correspondence would be a perception of "ghost" images where no real object exists. But, in fact, we never see such ghost images.

This obvious ambiguity of the Keplerian model raises the question, that whether we really need to identify objects first, before placing them in proper spatial depth. Random-dot stereograms (RDSs) of Julesz [4] give strong evidence that it is not so. A correlation mechanism

— that after Julesz we call Cyclopean eye — seems working instead, simply to find maximum correlating image segments in the two retinas. The mechanism itself, in spite of its numerous models in computational neuroscience [9], is still an enigma.

In the next Section, an algorithmic treatise will follow on RDSs that have proved to be excellent means to enlighten several vision-related issues in psychology. At first, however, to get familiar with them, an intuitive derivation of RDSs comes.

2.1. Correlating retinal images

Consider a vertical wall with a little step on it as shown in Fig. 1. Let wallpapers of different patterns cover the three wall planes. Fig. 1. shows the draft retinal images too, for the case when we take an observation position just opposite to the step and fixate it. Except for a slight horizontal shift in the middle, the retinal images on the left side and on the right are basically the same. The difference is due to the non-zero relative disparities, i.e. the different spatial depth of the two planes.

The only information brain obtains of a situation is the pairs of retinal images and — from the vergence movement of eyes — the corresponding visual angles. Scanning through fixation levels, for all pairs of retinal images it can find the maximum size correlating areas and "keep a register" of the corresponding visual angles to build three dimensional stereoscopic vision.

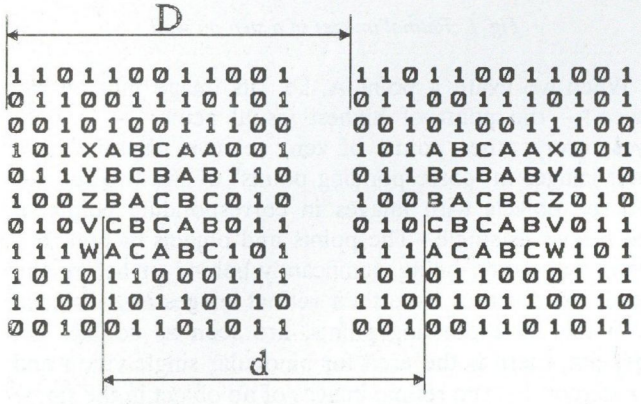


Fig. 2. Correlating areas in RDS

If we accept for fact the existence of Cyclopean eye, meaning that brain is capable of finding maximum size identical — or at least similar, correlating — areas in the retinal images, then we can reverse the above reasoning to derive the concept of RDS and create images solely for the Cyclopean eye. Fig. 2. shows a pair of patterns, both composed of two segments made of letters and numerals, respectively. The segments of letters and that of numerals as well, are identical in the two patterns, but they are located at different relative positions. Let us project the two patterns respectively onto the left and right retinas. The brain will correlate the retinal images and find those two settings for eyes when the pairs of identical segments fall on or close to corresponding points. The correlating segments will fuse and form a three dimensional object as we cannot distinguish this view from that of real objects (Fig. 3.). Such figures that give rise to stereoscopic illusion are called stereograms, and stereopairs when — like this case — they are in couples.

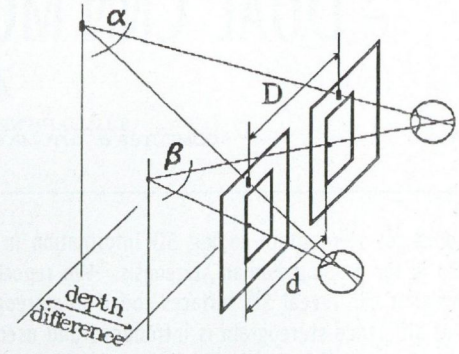


Fig. 3. Perceiving objects in stereo depth

We do not necessarily need to separate correlating figures for the two retinas. A single picture in itself can embody suitable autocorrelation to convey stereoscopic sensation. For the simplest case, if we consider horizontal stripes of periodic patterns as in Fig. 4/a, projected onto both retinas in such a way, that adjoining periods fall close to corresponding points, than — just as above — the correlating segments will be recognized, and as "surface" segments put in proper visual depth (Fig. 4/b) according to the different period length of stripes, resulting in perceiving different non-zero relative disparities. Since depth sensation is directly related to the local horizontal distance of correlating image fragments, with a suitable modulation of period length, that will be detailed in the next Section, more complex surfaces can also be coded in autocorrelating patterns called autostereograms.

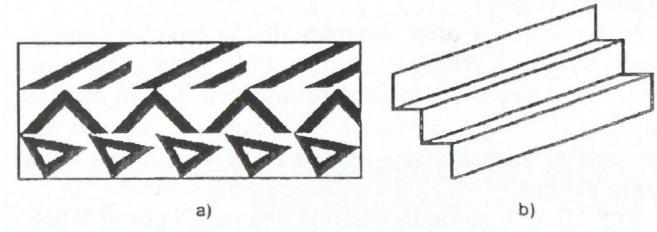


Fig. 4. Periodic pattern and the corresponding surface

As the actual local patterns of stereograms bear no importance whatsoever in surface coding, and even any monocular regularity may interfere with binocular depth sensation, using random patterns as generic ones to create stereograms is clearly advantageous.

(Projecting stereograms properly onto the retinas we can do simply by fixating either before or behind their plane. These cases are called "crossed" and "uncrossed" disparities, respectively. Even without experience, those fixations can easily be achieved. For the crossed case, one has to fixate a pointed object held between the plane of stereogram and the observer. In the uncrossed case, the ones who cannot simply look far away, behind the plane of

figures can use a transparent copy of them, held at some distance from a white background with a fixation mark, a black dot on it. Having the proper position found, the three dimensional percept would gradually emerge in some seconds.)

3. TYPES OF RANDOM-DOT STEREOGRAMS

The RDS is a visually perceptible rectangular pattern conveying 3D depth information coded in its internal correlation. Its mathematical representation, which for simplicity we also call RDS, is a two-variable function $g(x, y)$ coding a depth pattern represented by the $s(x, y)$ surface function, where the independent variables are horizontal and vertical coordinates. In visualization the $g(x, y)$ values are converted into different visual attributes as colour, brightness or texture.

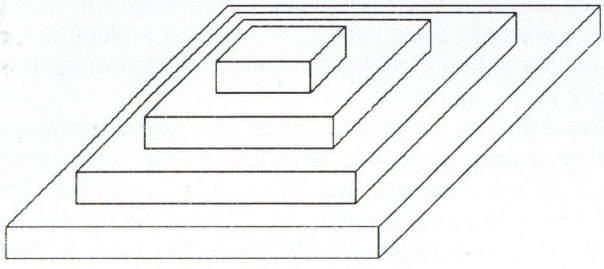


Fig. 5. Step pyramid

In RDSs, depth is transformed into correlation, which is defined as follows:

g at the points (x_c, y) and $(x_c + d, y)$ is correlated, if $g(x, y) = g(x + d, y)$ for $x_c \leq x \leq x_c + l$, where l is a suitable correlation length.

On its either — say left — side, the RDS contains a rectangular area with no correlation inside. Almost all the rest of RDS is composed of patches that are copies of some other areas lying to the left. The copy mechanism ensures internal correlation in the above sense. The position of areas to be copied is determined point by point in accordance with the depth pattern to be coded.

Let x_D and y_D respectively be the horizontal and vertical dimensions of the RDS, and $r(x, y)$ an internally uncorrelated pattern for $0 \leq x \leq x_D$ and $0 \leq y \leq y_D$; and let x_P be the width of the left uncorrelated area. We introduce an $X(s, x)$ copy-source-selector function to be defined later, accordingly to the different types of RDSs. In the row-wise generation of RDS, based on the surface to be coded, it will be its role to single out a point (X, y) from the left for copying.

Then for $0 \leq y \leq y_D$

$$
\begin{array}{ll}
g(x, y) = r(x, y) & \text{if } x < x_P, \\
g(x, y) = g[X(s(x, y), x), y] & \text{if } x \leq x_P \\
 & \text{and } (X, y) \text{ location has} \\
 & \text{not been copied yet,} \\
g(x, y) = r(x, y) & \text{otherwise}
\end{array}
$$

The width x_P sets constraints upon the dynamics of depth variations for the sake of a meaningful resulting stereogram. In other words, to generate RDS for surfaces of large variation in depth needs sufficiently large x_P.

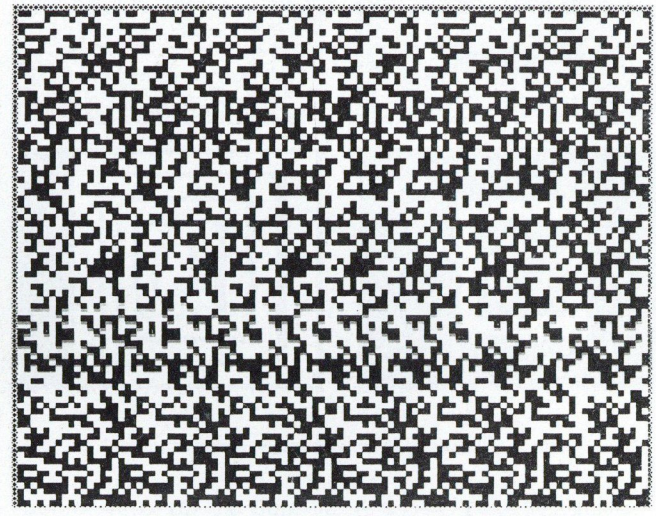


Fig. 6. Autostereogram of a step pyramid

If x_D is some multiple of x_P, we obtain an **auto-stereogram** [5], that can be inspected simply fixating behind or before its plane. The two fixation positions give reverse depth sensation. Fig. 6. shows an auto-stereogram of the step-pyramid seen in Fig. 5.

If $x_D = 2 * x_P$, we obtain a random-dot **stereopair**, with either half uncorrelated in itself (Fig. 7). For visual inspection, its halves can be separated and projected onto the left and right retinas. It is obvious, that the stereopair is a special type auto-stereogram, containing only the first two periods of the otherwise longer quasi-periodic structure.

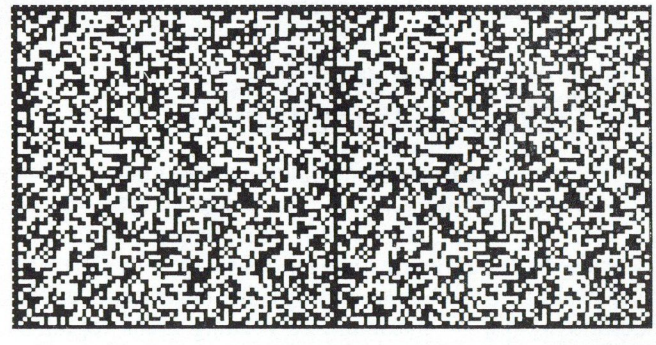


Fig. 7. Stereopair of a step pyramid

Both for stereopairs and auto-stereograms the copy-source-selector function is the same and is as follows: $X(s, x) = x - x_p + s$. For constant surface depth *at that point*, this copy-source-selector function obviously produces a strictly periodic pattern of period length $x_P - s$. For varying surface it shows how the period length is modulated locally by the changing depth values. According to the generation rules above, the copy mechanism is to work for flat and left to right descending surface segments only. Points for the ascending segments are taken from an uncorrelated pattern.

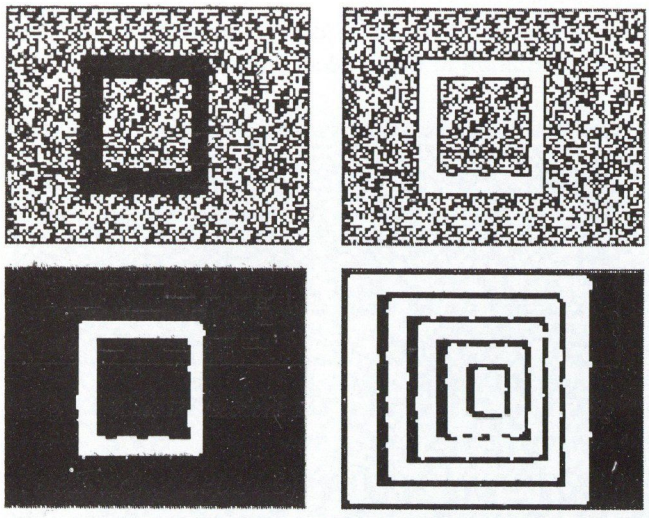


Fig. 8. Phases of iterative surface reconstruction

Using $X(s,x) = (x+s) \bmod x_P$ as copy-source-selector function leads to **difference stereograms**. The difference stereogram is for coding the changes in surface depth, instead of depth itself. Each flat segment will generate a periodic pattern of the same periodicity irrespective of its spatial depth. Local depth changes leave their marks on the stereogram only within a horizontal neighbourhood x_P. Difference stereograms have two distinct advantages from the aspect of depth coding. As was mentioned, in the other cases x_P sets constraints upon the dynamics of depth variations and vice versa. Now x_P has to be sufficiently large for coding local variations only, requiring in general a reduced width for x_P since local changes of depth on surfaces are usually smaller than the absolute depth difference of extreme points. The other advantage is that difference stereograms can also be produced by simple optical means, as will be seen later.

4. DUAL CNN ALGORITHM TO EXTRACT DEPTH FROM RDS

When perceiving depth, the binocular fusion mechanism of the human visual system — the Cyclopean eye — picks out fields correlating in the two retinal images. Correlation is tested by overlapping the retinal images at different relative shifts. Based on an entropy-like measure — called neurontropy [6] — maximum correlating areas together with their relative shifts that are directly proportional to the depth of corresponding areas on the surface are extracted.

The function of this neural mechanism can be broken into steps realizable in the framework of the Dual CNN paradigm [2], incorporating logic operations in addition to the analog CNN ones. Discrete binary (say black on white) RDSs of all the three types, built of black and white pixels, can be processed in the dual CNN framework. When producing RDS, the coded surface is pixelized, i.e. its base is discretized along both dimensions, and the pixels are to store discrete values for the third dimension.

An iterative dual CNN algorithm can find the surface layers in depth, one after the other, in successive phases. Each iteration corresponds to a specific discrete depth and results in the image of surface segments, fragments of the whole surface that lie in that depth layer just investigated. Since in the dual CNN framework [3] an external control is supposed to exist, the value of depth just investigated is always known. Combining the results of successive phases yields also the level line — or contour — structure of the whole surface.

In the following outline of the algorithm, the RDS is split into left and right images. Both images and the results as well, are $x_D - x_P$ wide, the left image being the left part of the RDS and the right one its right part of that width. For CNN terminology, one is considered as input picture, the other as initial state.

The N-th iteration of the algorithm consists of the following steps:

Step 1.: We shift the right image by one pixel to the left and replace it with the result.

Step 2.: The task to reveal sufficiently large correlating areas having the same pattern, at identical position in the left and the shifted right image, involves the following steps (Fig. 8.).

Step 2/a.: We perform a logical *equivalence* operation that by making pixel level correlations, will immediately reveal to an onlooker the areas being searched for. The image obtained is contiguously black on correlating areas and shows random black and white noise pattern elsewhere.

Step 2/b.: To demarcate correlating areas from noise by CNN, we introduce the next simple criterion: the black pixels having no white neighbours (most probably) belong to correlating areas. Using the *edge* [7] template, that inverts the internal pixels in solid black areas, we can find just those pixels that fulfil the criterion.

Step 2/c.: The final step in finding correlating areas is the removal of noisy pattern. Combining by logical *or* operation, the result of *edge* (Step 2/b) and the inverted result of logical *equivalence* (Step 2/a) will yield the white image of the N-th layer of the coded surface, on a black background.

The logical *inversion* and logical *or* operations can be performed one after the other by the logic part of dual CNN. Equivalently, a composite *inverse-or* operation can be defined and performed using the analog template. ($A_{22} = 1, B_{22} = -1, I = 1$)

The success of this noise removal step is highly dependent on the local randomness of RDS, or — conversely — on the adequacy of the demarcation criterion above. According to experience, small error patches may remain in the results. Applying a small positive value in the feedback off-center positions of the *inverse-or* template can improve noise removal by filtering out isolated errors or even one pixel wide stubs, depending on the value applied.

Step 3.: The surface layers found in successions can be aggregated by logical *and* operation to produce the level line structure of the coded surface. Essentially, the succession of logical *and* cuts and pastes those edges that were found in the demarcation step (2/b). Moreover, it has a noise filtering side-effect in the sense that any error that lies entirely inside areas on other depth layers (before or behind in depth) will disappear. On the other hand, errors

that cross surface edges will locally corrupt the level line result. Fig. 8. also shows the level lines of the step-pyramid seen in Fig. 5.

5. THE DIFFERENCE RDS AND ITS USE WITH REGULAR DOT PATTERN

RDSs usually code surface depth into correlating patterns. As follows from the generation rules, the difference RDS of a surface is the auto-stereogram of the "difference" surface. The difference surface can be produced simply by reducing the depth of each pixel of the original surface by the depth of that pixel laying in distance x_P to the left. Using a small enough x_P, relative to the minimum horizontal distance between unit depth changes, yields a three-level $(-1, 0, +1)$ difference surface. Such difference surface having only 2 bit dynamics can be coded with small x_P, that allows fine spatial resolution. Fig. 9/a and 9/b show the difference surfaces of the 76 pixel wide step-pyramid, for $x_P = 2$ and $x_P = 12$, respectively.

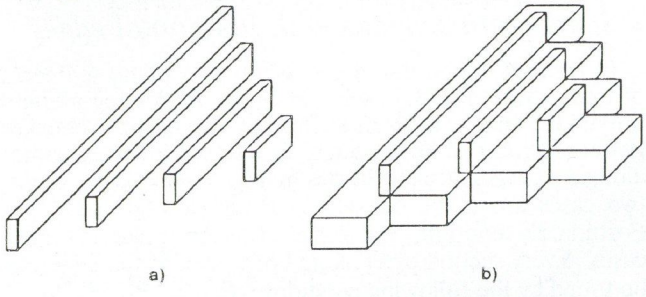


Fig. 9. 3D image of left (positive) parts of difference surfaces of the step pyramid

From the aspect of processing RDSs with a CNN, the difference RDS, due to its narrow correlation band, is fundamentally different from the other types of RDSs. The global correlation property of those is transformed into a local one of difference RDSs, that can be detected in a simple CNN window of neighbourhood 2 or 3. By that way, both flat areas and ascending or descending slopes of the coded surface, can be extracted in single CNN steps directly from the RDS.

The difference RDS plays especially important role when working with "smooth" discretized surfaces. A surface with a given discretization is smooth if the depth difference of neighbouring pixels is at most 1, and none of the 2 by 2 pixel areas form a saddle. At a saddle, all pixel values along either diagonal are higher than any pixel value along the other one.

1	2
2	1

A saddle

In the case of narrow band difference RDSs, the randomness of pattern looses its importance. Just the knowledge of pattern structure may contribute to the reconstruction of the coded surface. (Hence, we modify the copy procedure by allowing copying of a pixel more than ones.) In the following, we will elaborate a dual CNN algorithm to make out a smooth surface from its narrow band difference stereogram using regular checker pattern. For the first step, we analyse the stereograms of horizontal and vertical depth steps of unit size, the building blocks of stereograms of more complex smooth surfaces, to obtain criteria for finding them. Next, we convert the criteria into simple CNN templates of neighbourhood 1 for extracting edges, slopes, valleys and ridges, locally extreme areas, etc. To be able to uniquely identify edges in the resulting pictures, we require that they be separated everywhere by at least one pixel wide flat rims. We will apply and illustrate the method for the surface in Fig. 10., where different shadings are to show different depths.

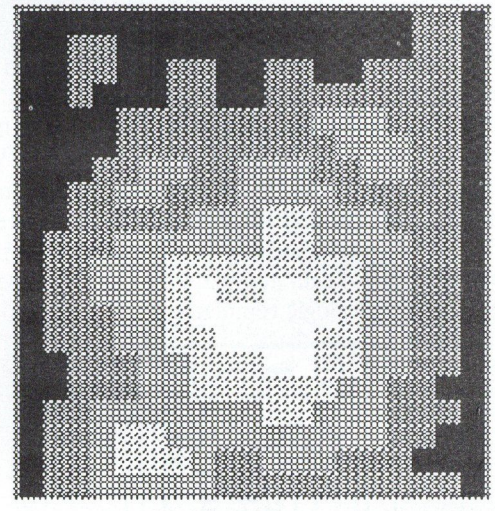


Fig. 10. Smooth test-surface

We consider a narrow band difference stereogram with an $x_P = 4$ pixel wide left band showing a checker pattern of $2 * 2$ pixel sized squares (Fig. 11). A part of such stereogram of a plane together with the pixel grid-lines is shown in Fig. 12/a. A vertical edge (i.e. a depth-unit step on smooth surface) generates either of the local patterns shown in Figs. 12/b and 12/c, depending on its being an ascending or descending one in horizontal direction. Similarly, Figs. 12/d and 12/e show traces that horizontal edges leave on the stereogram, depending on the exact location of the edge relative to the nearest checker grid line. Note, however, that due to the horizontally oriented nature of the generation algorithm, these traces of horizontal edges are always sort of after-effects of an adjoining vertical edge to the left. That is why, on the contrary to the vertical case, for horizontal edges the vertically ascending or descending nature of the surface there, never can be determined locally.

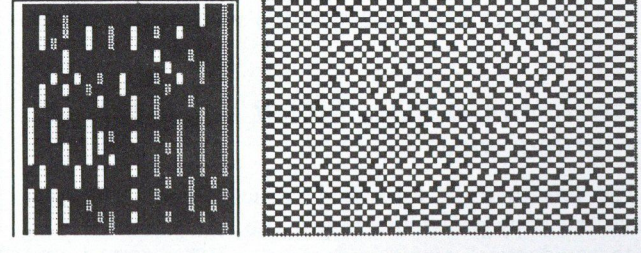


Fig. 11. Difference surface and stereogram

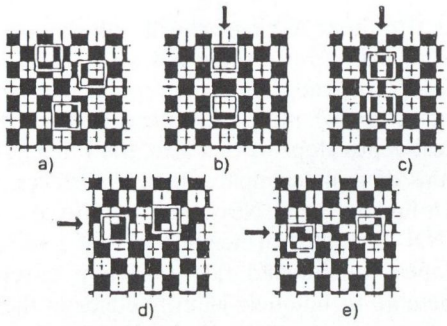


Fig. 12. Typical checker patterns

• *Stereogram sections for flat surface segments*

Those surface segments void of any change in depth leave the original checker pattern unchanged (Fig. 12/a). Considering its properties in a $3 * 3$ window, sufficient condition is $P_{11} = P_{33} \bigcap P_{13} = P_{31} \bigcap P_{11} \neq P_{13} \bigcap P_{12} \neq P_{32} \bigcap P_{21} \neq P_{23}$, i.e., in words, both the neighbouring corners and the opposite edges have pairwise different values. Checking this criterion may require an appropriate non-linear template or means consecutive application of simple logic templates.

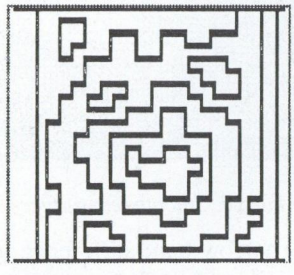


Fig. 13. Plane plateaus

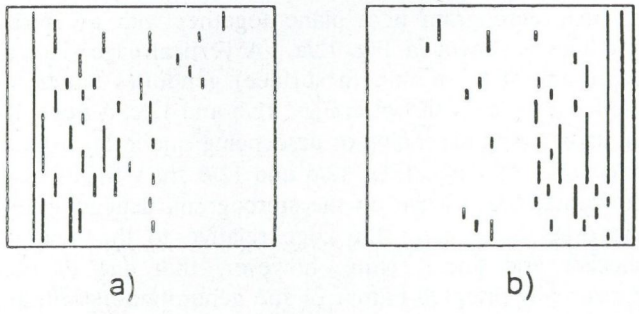


Fig. 14. Left to right ascending (a) and descending (b) edges

However, this criterion is of little practical value, as it is too complex and no further properties of edges found can be extracted. The following set of simpler criteria for finding other details will give the surface plateaus too (Fig. 13.).

• *Stereogram sections for left to right ascending vertical edges*

The left to right ascending vertical edges, in effect, make the copy procedure repeat the last pixel, resulting in three identical pixels in succession (Fig. 12/b). The corresponding criterion is $P_{21} = P_{22} = P_{23}$, that can easily be checked by two linear CNN templates, one for the white pixels, the other for the black ones, as follows: $(A_{22} = 1, I_0 = -0.5, B_{21} = B_{22} = B_{23} = \pm 1/3)$, being the two templates different only in the sign of B values. To obtain edges, the results of the two templates are to be logically *or*ed (Fig. 14/a).

• *Stereogram sections for left to right descending vertical edges*

The left to right descending vertical edges, in effect, make the copy procedure leave out one pixel, resulting in three alternating pixels in succession (Fig. 12/c). The corresponding criterion is $P_{21} = P_{22} = P_{23}$, that can easily be checked by two linear CNN templates, one for the white-black-white case, the other for the opposite one, as follows: $(A_{22} = 1, I_0 = -5, B_{21} = B_{22} = B_{23} = \pm 1/3)$, being the two templates different only in the sign of B values. To obtain edges, the results of the two templates are to be logically *or*ed (Fig. 14/b).

• *Stereogram sections with horizontal edges*

In case of horizontal edges, when two plane segments of exactly one unit difference in depth are joining along a horizontal line, a horizontal shift of the original checker pattern there, can be detected. Since that line runs either along a checker grid or in the middle of checker squares, two cases are to be considered (Figures 12/d and 12/e). Fortunately enough, the same criterion holds for both cases. Every second pixel along both sides of the edge can be found by the following condition: $P_{11} = P_{31} \neq P_{13} = = P_{33}$, leading to templates

$$(A_{22} = 4,\ I_0 = -1,\ B_{11} = B_{31} = \pm 1,\ B_{13} = B_{33} = \mp 1)$$

To obtain edges (Fig. 15), the results of the two templates are to be logically *or*ed and the one pixel wide gaps filled by template:

$$A_{22} = 1,\ I_0 = 1,\ B_{21} = B_{23} = 0.5,\ B_{22} = 1.5).$$

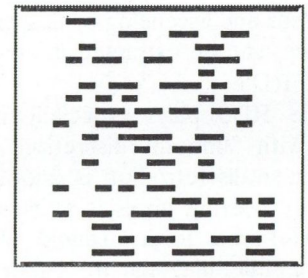


Fig. 15. Horizontal edges

• *Determining the orientation of horizontal edges found*

If we consider the constraints for smoothness and the location and orientation of adjoining vertical edges we can determine the orientation of horizontal edges too. Both ends of horizontal edges join to either end of a vertical edge. To determine orientation, it is enough to look only at one of the ends. The four possible joints and conclusions about the ascending or descending nature of depth change at horizontal edges are summarized in the next table.

Features of left connecting vertical edge	goes upwards	goes downwards
left to right descending	downward ascending horizontal edge	downward descending horizontal edge
left to right ascending	downward descending horizontal edge	downward ascending horizontal edge

We can separate horizontal edges according to their orientation, in four steps:

1 – Using three simple templates, we single out the left end of downward ascending edges by first selecting all the ends common with vertical edges that are either descending and going upwards or ascending and going downwards, and then discard ends on the right-hand side.

2 – To obtain downward ascending edges at full length (Fig. 16/a), we erase with the next template propagating from left to right, all the horizontal edges that are not masked by the ends found in the previous step:

$$(A_{21} = 1.5,\ A_{22} = 3,\ I_0 = -1.5,\ B_{22} = 1.5)$$

3 – Applying the difference template ($A_{22} = 2$, $I_0 = -1$, $B_{22} = -2$) for all horizontal edges and the downward ascending ones, gives the downward descending edges shown in Fig. 16/b.

4 – The horizontal edges found as bordering pixels of adjoining planes, are two pixel wide. The two thinning templates

$(A_{22} = 2, I_0 = -3, B_{22} = 4,\ B_{32} = 4)$ and
$(A_{22} = 2, I_0 = -3, B_{22} = 4,\ B_{12} = 4)$,

respectively for edges, descending and ascending downward, yield one pixel wide lines retaining pixels on the higher level planes only (Fig. 16).

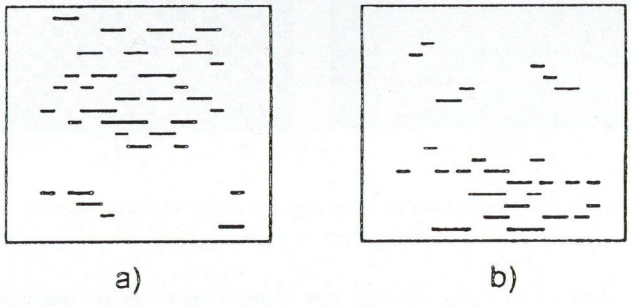


Fig. 16. Downward ascending (a) and (b) descending edges

• *Finding slopes, ridges, valleys, hollows and peaks*

We define a descending/ascending slope on a pixelized surface as a straight line of some direction along which each pixel value is lower/greater than or equal to any of the previous ones. Starting from the former results all horizontal and vertical slopes can easily be found using propagating templates. By self-explaining definition, ridges and valleys on a surface are places common for ascending and descending slopes of the same direction. Consequently, the intersections of horizontal/vertical slopes, that can be produced by logically *and*ing them, yield the vertical/horizontal ridges and valleys of the surface. Those intersections of horizontal and vertical ridges that entirely fill in a closed contour line are the local maxima, the peaks of the surface. Similarly, the intersections of valleys are the local minima, the hollows.

We generate slopes as shadows of one type of edges, stretching until the first opposite type edge encountered. Producing ascending slopes calls for propagating the ascending edges and halting propagation by the descending ones, and vice versa. The appropriate template is as follows, where propagation comes from the direction of the non-zero off-center feedback entry A_{ij} if the set of edges to be propagated is the initial state and the halting set is the input picture of the CNN:

$$(A_{ij} = 1.5\ \ A_{22} = 1.8,\ B_{22} = -1.2)$$

Some of the horizontal and vertical slopes are shown in Fig. 17. Horizontal and vertical valleys obtained as intersections of slopes are shown in Fig. 18.

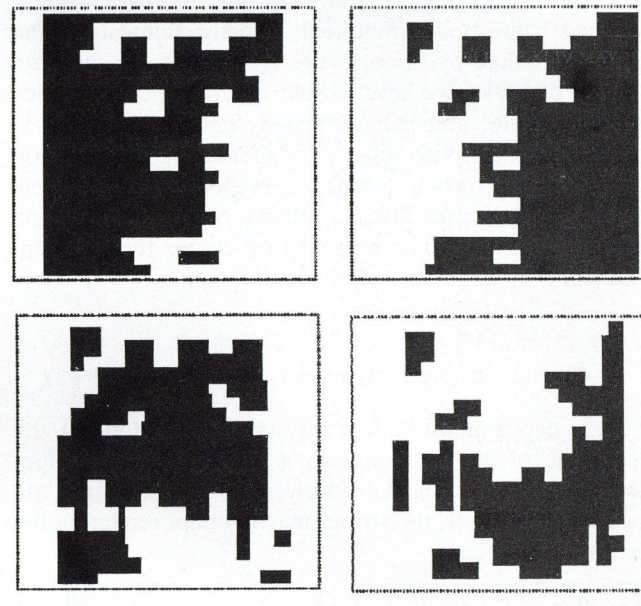


Fig. 17. Horizontal and vertical slopes

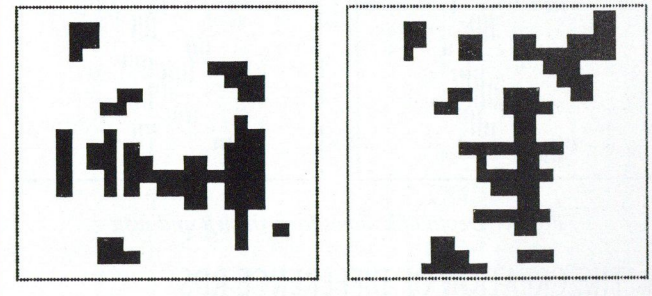


Fig. 18. Horizontal and vertical valleys

To find surface maxima/minima first we fatten the intersecting segments of ridges/valleys by one pixel, using template ($I_0 = 8$, $B_{ij} = 1$). From the result taken as initial state for the following template, we erase those areas that do not totally fit into one of the contour line loops, (Fig. 19) the input picture for the template:

$(A_{12} = A_{21} = A_{23} = A_{32} = 1,\ A_{22} = 5,$
$I_0 = -5.25,\ B_{22} = 2).$

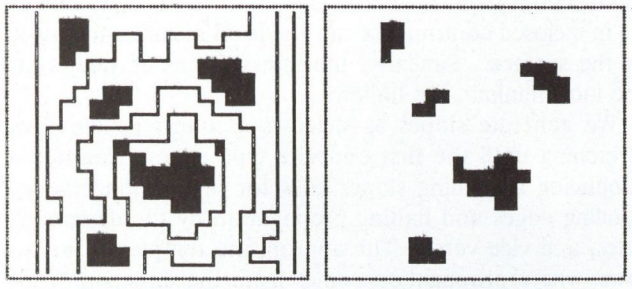


Fig. 19. Valley intersections among level lines and the resulting hollows

• *Measuring depth of local extreme areas*

We define the local depth of extreme areas — peaks and hollows — as the number of surface steps — found as edges — of the same type in one direction before an opposite type step is encountered. In this respect the local depth is a direction dependent measure. Similarly to the generation of slopes, we shift one type of edges using a modified connected component detector template. The shifting is halted by the opposite type edges. To measure "height" of peaks, we have to shift ascending edges loaded for initial state, and halt shifting by descending edges in input picture, and vice versa when measuring depth of hollows. The appropriate templates, respectively for vertical and horizontal cases, are as follows, where propagation goes in the direction of non-zero off-center feedback entries from the negative to the positive sign:

$$(A_{12} = \pm 1,\ A_{22} = 2,\ A_{32} = \mp 1,\ I_0 = -3,\ B_{22} = -3)$$
$$(A_{21} = \pm 1,\ A_{22} = 2,\ A_{23} = \mp 1, I_0 = -3,\ B_{22} = -3)$$

Some depth measurement results are shown in Fig. 20. The value of depth of hollows is the number of stripes inside the hollow area. Obviously, if deep hollows are too close to each other, the stripes may overlap, rendering the results meaningless.

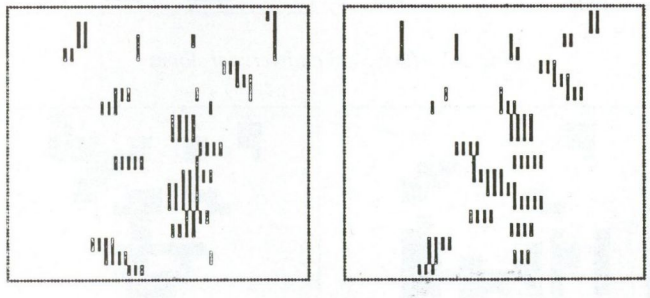


Fig. 20. Depth of hollows from the left and right

6. THE CREATION OF DIFFERENCE RDS WITH OPTICAL MEANS

Apart from its inevitable importance in vision research, up until this point to produce and probe RDS may have seemed an enjoyable and aesthetic play, devoid of any significance from the engineering point of view. However, in principle, difference RDSs of real objects can be produced directly by optical means. Moreover, research of optical CNN devices is reportedly under way [8], that together with the possibility of making continuous, real time difference RDSs of objects may lead to a "stereoscopic" robot eye.

Let us place a projector and a camera side by side at the same height in front of a white wall (Fig. 21) and project perpendicularly onto the wall a horizontally periodic random-dot (or checker) pattern, i.e. the RDS of a single plane, from such a distance that the parallax over the projection area be negligible. The camera — from beside the projector — is to see the whole area at a specific angle, that depends on its distance from the projector, and is the same for the whole area. In fact, the camera will see the RDS, or what is the same in this case, the difference RDS of a single plane.

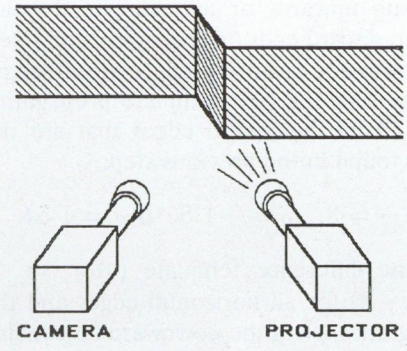


Fig. 21. Creating difference RDS with optical devices

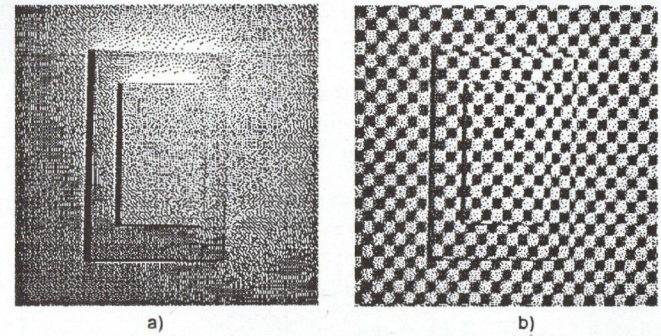


Fig. 22. Camera image of an angular object with and without checker pattern projected onto it

Let us place some angular objects in front of the wall in such a way, that their surfaces be either parallel with or perpendicular to the wall. From the projector no change in the projected pattern can be noticed. However, from the camera definite horizontal shifts of pattern segments that now fall on the inserted objects, can be detected. The resulting view from the camera is the difference RDS of the whole arrangement, except for the images of those surfaces perpendicular to the wall, which go "empty" in the optical method and would be filled with an uncorrelated pattern in the computer algorithm. A camera image of objects in Fig. 22/a is shown in Fig. 22/b. The amount of horizontal shift is determined by the change in depth and the camera angle. We consider a change in depth as depth resolution, if it causes one pixel horizontal shift in the

camera image. For a given pixel size the depth resolution can be adjusted by the camera angle.

Presumably, using electronically controlled devices, the surface in front of the above set-up, can be scanned row-wise by a flying window and processed in real time by a tiny CNN array. Moreover, in a feedback loop, the resolution and the size of the flying window can be adjusted adaptively to the extracted local features of the surface.

6. CONCLUSIONS

From its very beginning, research and development in neural networks has often been initiated by functions of live organism, aiming either at modelling life functions to enhance understanding, or at mimicking life performances in practical task solving. In this paper we have followed both ways, by giving first an iterative functional model for stereo vision, and second a - for the time being hypothetical - method to build stereoscopic eye. The biological relevance of the model has not yet been investigated, nevertheless, experience proves that building stereoscopic percept by human brain is also an iterative sort of process. As far as the stereoscopic eye is concerned, although its algorithm seems self contained, much work is ahead in its optical realization.

REFERENCES

[1a] L. O. Chua and L. Yang, "Cellular neural networks:Theory", *IEEE Transactions on Circuits and Systems*, Vol.35, 1988.

[1b] L. O. Chua and L. Yang, "Cellular neural networks: Applications", ibid.,

[2] T. Roska and L. O. Chua, "Dual CNN analog software", *Report DNS-1-92*, Dual and Neural Computing Systems Res. Lab., Comp. Aut. Inst., Hung. Acad. Sci., 1992.

[3a] K. Halonen, V. Porra, T. Roska and L. O. Chua, "VLSI Implementation of a reconfigurable CNN containing local logic", *Proc. CNNA-90*, 1990.

[3b] A. Radványi, K. Halonen and T. Roska, "The CNNL Simulator and some time-varying CNN templates", *Report DNS-9-91*, Dual and Neural Computing Systems Res. Lab., Comp. Aut. Inst., Hung. Acad. Sci., 1991.

[4a] B. Julesz, "Binocular Depth Perception of Computer-Generated Patterns", *Bell Syst. Tech. J.* 39, 1960.

[4b] B. Julesz, "Foundations of Cyclopean Perception", *Chicago University Press*, Chicago, 1971.

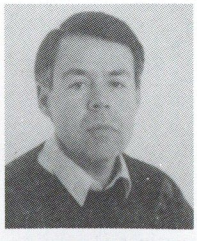
[5] C. W. Tyler, "Sensory processing of binocular disparity", in *C.M.Schor & K.J.Cuiffreda (Eds.),"Vergence Eye Movements"*, Boston; Butterworth, 1983.

[6] B. Julesz, C. W. Tyler, "Neurontropy, an Entropy-Like Measure of Neural Correlation, in Binocular Fusion and Rivalry", *Biological Cybernetics* 23, 1976.

[7] T. Roska, A. Radványi, T. Kozek and T. Boros, "Dual CNN software library", *Report DNS-7-1991*, Dual and Neural Computing Systems Res.Lab., Comp. Aut. Inst., Hung. Acad. Sci., 1991.

[8] N. Frühauf and E. Lüder, "Realization of CNNs by optical parallel processing with spatial light valves", *Proc. IEEE CNNA-90*, 1990.

[9] R. Blake and H. R. Wilson, "Neural Models of Stereoscopic Vision", *TINS*, Vol. 14, No. 10, 1991., Elsevier Science Publishers Ltd



András G. Radványi graduated at the Technical University of Budapest, Faculty for Electrical Engineering in 1969. He received the Ph. D. degree in Technical Sciences from the Hungarian Academy of Sciences in 1981. From 1969 to 1982 he was with the Research Institute for Telecommunications where he worked on computer aided simulation of electronic circuits. Since 1982 he has been a senior researcher in the Computer and Automation Institute of the Hungarian Academy of Sciences, dealing with theory and application of cellular neural networks.

LIMIT ON THE EFFICIENCY OF SPARSELY ENCODED ASSOCIATIVE MEMORIES

J. LEVENDOVSZKY
Department of Telecommunications
Technical University of Budapest,
Stoczek u.2., H-1111, Budapest, Hungary

W. MOMMAERTS
Department of Mathematics
Katholieke Universiteit Leuven
Celestijnenlaan 200B, B-3001 Heverlee, Belgium

E.C. VAN DER MEULEN
Department of Mathematics
Katholieke Universiteit Leuven
Celestijnenlaan 200B, B-3001 Heverlee, Belgium

Information, retrieval from a distorted or fragmented version of a signal, has played a central role in numerous areas of applications like signal processing, pattern recognition and error correcting decoding. Therefore, structures and algorithms, capable of performing the above-mentioned task, have long since been under detailed investigation and often referred to as associative memories. The primary parameter of an associative memory is the amount of information (capacity), which can be retrieved with a certain efficiency, where the efficiency defines the average probability of the correct recollection. On the other hand, the complexity of the corresponding algorithm or structure used for association, has to be seriously taken into account, since the increasing capacity usually involves rapidly growing complexity. Therefore, the aim of the theoretical analysis performed on different types of associative mappings, is to reveal the fundamental relations between capacity, efficiency and complexity, in order to find a good compromise for the engineering design. The paper is concerned with the statistical analysis of classical feedforward structures (one layered perceptron [1,2], attentive associative memory [3]). The feedback structures are neglected from the paper on the ground of stabilty reasons, however, several investigations were performed to assess their information theoretical capacity and applicability in different domains [4,5,6]. Expressions for the design of a relatively new structure, often called as sparsely encoded associative memory, are also derived.

1. DEFINITIONS AND BACKGROUND THEORY

1.1. Hetero-Associative Mapping (HAM)

The mapping $y = F(\overline{w}, \overline{x})$ is called Hetero-Associative Mapping $HAM(S, Q)$ if it satisfies the following equations for any given set denoted by S and Q.

$$\overline{s}^\alpha \in S,\ \overline{q}^\alpha \in Q:\ \alpha = 1, \ldots, M$$

$$F(\overline{w}, \overline{s}^\alpha) = \overline{q}^\alpha \quad \forall \overline{s}^\alpha \in S, \overline{q}^\alpha \in Q \qquad (1)$$

where $s_i^\alpha, q_i^\alpha \in \{-1, 1\}$ and $\overline{w}$ is the weight vector (the freedom of F) which can be adapted to achieve the required mapping from S to Q.

1.2. Capacity of an associative mapping

The capacity of F is determined by the number of elements in the sets S and Q providing the input-output pairs.

$$\mathrm{Cap}F = M.$$

Sometimes the relative capacity is used

$$\mathrm{RelCap}F = M/N,$$

where N denotes the dimension of $\overline{q}$.

1.3. Efficiency of an associative mapping

In general, it cannot be guaranteed that the equations

$$F(\overline{w}, \overline{s}^\alpha) = \overline{q}^\alpha, \quad \forall \overline{s}^\alpha \in S, \overline{q}^\alpha \in Q$$

hold with probability one. In that case there exists a sequence of probabilities

$$P^{\alpha\alpha} := P(\overline{y} = \overline{q}^\alpha | \overline{x} = \overline{s}^\alpha)$$

which forms the efficiency in the following way:

$$\mathrm{Eff} := (1/M) \sum_{\alpha=1}^{M} P^{\alpha\alpha}.$$

It is trivial that Eff $\leq$ 1 and Eff= 1 if equations (1) are satisfied.

1.4. The cost of an associative mapping

If the binary mapping is implemented by digital processing elements then one of the most important features is the number of bits needed to be stored in the weights. Assuming that each weight has only discrete values in the interval $(-L, \ldots, -1, 0, 1, \ldots, L)$, the cost is defined as:

$$\mathrm{Cost} := \dim \overline{w} \log(2L + 1),$$

where dim $\overline{w}$ denotes the dimension of $\overline{w}$.

2. ONE LAYERED PERCEPTRON

The classical mapping, suggested to carry out HAM(S,Q), is summarized in (2).

$$y = \mathrm{sgn}\{\sum_{j=1}^{N} W_{ij} x_j\} \quad i = 1, \ldots, N \qquad (2)$$

where $W_{ij} := (1/N) \sum_{\alpha=1}^{M} q_i^\alpha s_j^\alpha$, if $(\overline{s}^\alpha, \overline{q}^\alpha)$ $\alpha = 1, \ldots, M$ binary pairs have to be mapped into each other. The corresponding network is depicted in Fig. 1.

Assuming that the actual input is $\overline{s}^\beta \in S$ and rewriting (2) we obtain

$$y_i = \mathrm{sgn}\{\sum_{j=1}^{N} W_{ij} x_j\} = \mathrm{sgn}\{\sum_{j=1}^{N} (\frac{1}{N} \sum_{\alpha=1}^{M} q_i^\alpha s_j^\alpha s_j^\beta\} =$$

$$\mathrm{sgn}\{\sum_{\alpha=1}^{M} (\frac{1}{N} \sum_{j=1}^{N} q_i^\alpha s_j^\alpha s_j^\beta\} =$$

$$\mathrm{sgn}\{q_i^\beta (\frac{1}{N}) \sum_{j=1}^{N} (s_j^\beta)^2 + \sum_{\alpha=1, \alpha\neq\beta}^{M} q_i^\alpha (\frac{1}{N} \sum_{j=1}^{N} s_j^\alpha s_j^\beta\}.$$

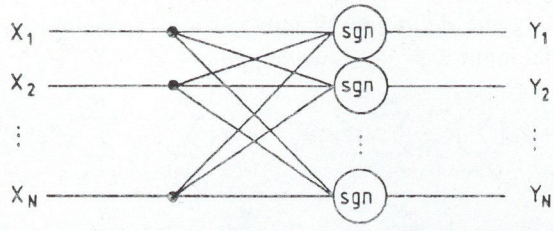


Fig. 1. One layered perceptron

The second term in the last summation represents the noise which can cause false recollection. Therefore, the following conditions must be satisfied for the correct retrieval:

$$\mu_i := \sum_{\alpha=1,\alpha\neq\beta}^{M} q_i^\alpha (\frac{1}{N}) \sum_{j=1}^{N} s_j^\alpha s_j^\beta < 1 \quad \text{if } q_i^\beta = -1$$

$$\mu_i := \sum_{\alpha=1,\alpha\neq\beta}^{M} q_i^\alpha (\frac{1}{N}) \sum_{j=1}^{N} s_j^\alpha s_j^\beta > -1 \quad \text{if } q_i^\beta = 1 \qquad (3)$$

taking into account that $(1/N)\sum_{j=1}^{N}(s_j^\beta)^2 = 1$.

Unfortunately, the number of $\overline{s}^\alpha, \overline{q}^\alpha$ vectors which satisfy condition (3) is severely limited. The trivial solution to fulfill condition (3) is to choose $\overline{s}^\alpha$, $\alpha = 1,\ldots,M$ as an orthogonal vector set, which restricts the capacity to N (ie. maximum N orthogonal vectors can be chosen in an N dimensional space), though. Therefore,

$$\text{RelCap}_{\text{one layered perceptron}} = 1$$

which is rather small, being a severe drawback in applications.

The cost function of this structure can be easily calculated taking into account that $W_{ij} := (\frac{1}{N})\sum_{\alpha=1}^{M} q_i^\alpha s_j^\alpha$ has maximum $M+1$ possible discrete values (due to the binary vector components) in the $(-M/N,\ldots,M/N)$ interval, resulting

$$\text{Cost}_{\text{one layered perceptron}} = N^2 \log_2(M+1).$$

If condition (3) holds then the efficiency is equal to one, since the mapping is correct with probability one, regardless which input vector is selected from S (presently, we do not deal with input vectors which are not elements of S, however, a rather limited capacity has been obtained. Thus, violating condition (3) one can increase the capacity, which involves a decreasing efficiency. One of the most familiar methods to increase M is to take randomly chosen binary vectors as $\overline{s}^\alpha, \overline{q}^\alpha = 1,\ldots,M > N$ with equally probability. In this case, condition (3) is satisfied only with a certain probability, determined by the distribution of $\overline{\mu}$ noise vector. Assuming that $\overline{\mu}$ is subject to Gaussian distribution (based on the central limit theorem if N is large) the mean vector and covariance matrix can be easily calculated as

$$\text{E}\mu_i = 0 \qquad \text{E}\mu_i\mu_j = \delta_{ij}\frac{(M-1)}{N}$$

which provides the following efficiency:

$$P(\overline{y} = \overline{q}^\alpha | \overline{x} = \overline{s}^\alpha) = \prod_{i=1}^{N} P(y_i = q_i^\alpha | \overline{x} = \overline{s}^\alpha) =$$

$$\prod_{i=1}^{N} P(\mu_i^{>1}_{<-1} \,|^{q_i^\alpha=-1}_{q_i^\alpha=1}) = \Phi^N(\sqrt{\frac{N}{M-1}})$$

$$\text{Eff}_{\text{one layered perceptron}} = \frac{1}{M}\sum_{\alpha=1}^{M} \Phi^N(\frac{N}{M-1}) =$$

$$\Phi^N(\sqrt{\frac{N}{M-1}})$$

where

$$\Phi(x) := \int_{-\infty}^{x} (1/\sqrt{2\pi}) e^{-x^2/2} dx\,.$$

Investigating the efficiency with respect to M it is clear that increasing M (growing capacity) results in very bad efficiency. Therefore, other structures must be seeked to achieve higher capacities with reasonable efficiency.

3. ONE LAYERED PERCEPTRON WITH DISTRIBUTED NONLINEARITIES

As it was seen earlier, the capacity is bounded due to the condition imposed on the noise, as expressed in condition (3). In this expression the inner products play a central role since if one is able to enhance the first term $\frac{1}{N}\sum_{j=1}^{N}(s_j^\beta)^2$ and supress the inner products $\frac{1}{N}\sum_{j=1}^{N} s_j^\alpha s_j^\beta$ $\alpha \neq \beta$ which belong to the noise term, the capacity can be increased. This can be done by introducing proper nonlinear transformations on the inner products. In this way, the one layered perceptron with distributed nonlinearities is defined by mapping (4) (see Fig. 2)

$$y_i = \text{sgn}\{\sum_{\alpha=1}^{M} q_i^\alpha \psi(\frac{1}{N}\sum_{j=1}^{N} s_j^\alpha x_j]\}. \qquad (4)$$

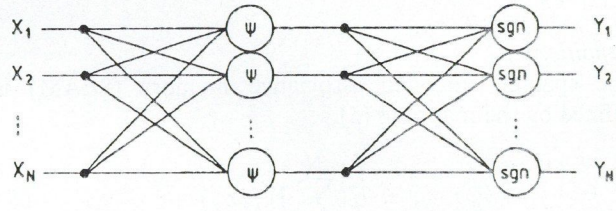


Fig. 2. One layered perceptron with distributed nonlinearities

Assuming that the actual input is $\overline{s}^\beta \in S$ and separating the terms in the summation again, we obtain

$$y_i = \text{sgn}\{q_i^\beta \psi[(\frac{1}{N})\sum_{j=1}^{N}(s_j^\beta)^2] + \sum_{\alpha=1,\alpha\neq\beta}^{M} q_i^\alpha \psi[(\frac{1}{N}\sum_{j=1}^{N} s_j^\alpha s_j^\beta]\}$$

which makes clear that the nonlinearities can be used to decrease the noise. It can be proven that any arbitrary capacity can be achieved by this structure with proper nonlinearities [7]. The trade off for the increased capacity, however, is the growing complexity which is implied by the increased number of nodes. The cost can be easily calculated as

$$\text{Cost}_{\text{one layered perceptron with DN}} = 2MN >>$$

$$N^2 \log_2(M+1).$$

Since the increase in efficiency was accomplished by a growing cost function, being propotional to the capacity, instead of a logarithmic dependence the economic realization of HAM can not be achieved with the structure discussed above. Hence, we aim to solve the following problem:

How is it possible to find a new $F(\overline{w}, \overline{x}) = \overline{y}$ associative mapping, which is cost invariant (Cost $= O(\log_2(M))$) and at the same time maximizes the efficiency for a given capacity?

4. SPARSELY ENCODED ASSOCIATIVE MEMORY

In order to increase the capacity or efficiency, we have so far manipulated the inner products only, resulting in a higher cost. However, investigating the noise term again

$$\mu_i := \sum_{\alpha=1,\alpha\neq\beta}^{M} q_i^\alpha (\frac{1}{N}) \sum_{j=1}^{N} s_j^\alpha s_j^\beta < 1 \text{if} q_i^\beta = -1$$

manipulated involving higher cost

It can be manipulated without increasing the cost

it can be seen that if q_i^α is mostly zero with respect to α then the effect of high inner products can be easily eliminated. Therefore, we have to select the random variable q_i^α to be 0 with high probability and to be 1 with low probability, changing the components of the stored binary vectors from $\{-,1\}$ to $\{0,1\}$. This involves the idea of fixing the number of 1-s in each stored memory word as ρ fragment of N components. In this case only ρN components can be 1, thus the probability of $q_i^\alpha = 0$ is $1-\rho$. Choosing ρ appropiately small the probablity of $|\mu_i| < 1$ can be reasonable high, so the efficiency can be improved.

Definition:
The sparsely encoded associative memory (SEAM) is defined by the mapping (5).

$$y_i = \mathcal{H}(\sum_{j=1}^{N} W_{ij} x_j) \tag{5}$$

where $\mathcal{H}(x) := \frac{1}{2}(1+\text{sgn}(x))$ $W_{ij} := \frac{1}{N}\sum_{\alpha=1}^{M} q_i^\alpha s_j^\alpha$ and the possible values of s_j^α and q_i^α are zero or one and the number of ones in each $\overline{s}^\alpha$ and $\overline{q}^\alpha$ are fixed: ρN.

Theorem:
Applying the SEAM,

$$\frac{N!}{(\rho N)!((1-\rho)N)!}$$

capacity and

$$\epsilon = \{\rho + [(1-\rho) + \rho\frac{\binom{N-\rho N}{\rho N}}{\binom{N}{\rho N}}]^{\binom{N}{\rho N}-1}\}^N \tag{6}$$

efficiency can be achieved with invariant $N\log(M+1)$ cost.

Proof:
Putting the definition W into (5) and assuming that the actual input is $\overline{s}^\beta \in S$, we obtain

$$y_i = \mathcal{H}\{\sum_{j=1}^{N}(\frac{1}{N}\sum_{\alpha=1}^{M} q_i^\alpha s_j^\alpha s_j^\beta\} = \mathcal{H}\{\sum_{\alpha=1}^{M}(\frac{1}{N}\sum_{j=1}^{N} q_i^\alpha s_j^\alpha s_j^\beta\} =$$

$$\mathcal{H}\{q_i^\beta(\frac{1}{N})\sum_{j=1}^{N}(s_j^\beta)^2 + \sum_{\alpha=1,\alpha\neq\beta}^{M} q_i^\alpha(\frac{1}{N}\sum_{j=1}^{N} s_j^\alpha s_j^\beta\}.$$

The first term in the argument is

$$q_i^\beta(\frac{1}{N})\sum_{j=1}^{N}(s_j^\beta)^2 = 0 \quad \text{if } q_i^\beta = 0 \text{ or}$$

$$q_i^\beta(\frac{1}{N})\sum_{j=1}^{N}(s_j^\beta)^2 = \rho \quad \text{if } q_i^\beta = 1.$$

If the signal term is 1 then the noise can take any value (due to the possible binary values, it is always positive) without disturbing the correct recollection. If the signal term is zero then the noise has to be zero, also, otherwise false recollection is obtained. The corresponding probability can be calculated in the following way:

$$P(\text{noise}_i = 0) =$$

$$P(q_i^\alpha = 0 \cup [q_i^\alpha = 1 \cap \sum_{j=1}^{N} s_j^\alpha s_j^\beta = 0 \text{ for all } \alpha,\beta) =$$

$$\{P(q_i^\alpha = 0) + P(\sum_{j=1}^{N} s_j^\alpha s_j^\beta = 0)P(q_i^\alpha = 1)\}^{M-1} =$$

$$(1-\rho+P_I\rho)^{M-1}$$

where

$$P_I := P(\sum_{j=1}^{N} s_j^\alpha s_j^\beta = 0|\beta) =$$

$$\frac{\binom{N-\rho N}{\rho N}}{\binom{N}{\rho N}} = \frac{(N-\rho N)!^2}{(N-2\rho N)!N!}.$$

As a consequence, the conditional probabilities which form the efficiency can be calculated as follows:

$$P(\overline{y} = \overline{q}^\alpha|\overline{x} = \overline{s}^\alpha) = \prod_{i=1}^{N} P(y_i = q_i^\alpha|\overline{x} = \overline{s}^\alpha) =$$

$$\prod_{i=1}^{N}\{P(q_i^\beta = 1) + (q_i^\beta = 0)P(\text{noise}_i = 0)\} =$$

$$\prod_{i=1}^{N}\{\rho + [1-\rho+\rho\frac{(N-\rho N)!^2}{(N-2\rho N)!N!}]^{\frac{N!}{(N-\rho N)!(\rho N)!}-1}\} =$$

$$\{\rho + [1-\rho+\rho\frac{(N-\rho N)!^2}{(N-2\rho N)!N!}]^{\frac{N!}{(N-\rho N)!(\rho N)!}-1}\}^N.$$

Since this expression does not depend on β, it results the efficiency. □

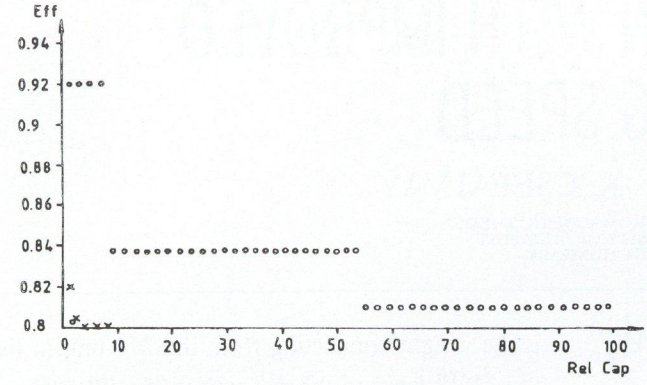


Fig. 3. Comparison of efficiencies

5. NUMERICAL RESULTS

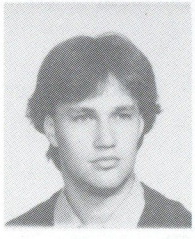
The improvement of the previously detailed method can be clearly seen in Fig. 3., where the capacity (horizontal axis) versus efficiency (vertical axis) is depicted, after numerically evaluating (6). The circles show the higher efficiencies guaranteed by the sparsely encoded structure compared with the efficiencies given by the one layered perceptron (asterisks).

6. CONCLUSION

The possible realizations of the optimal associative mapping were investigated with respect to a certain cost and efficiency. It was pointed out that by using sparsely encoded structure the efficiency or the capacity (depending on the actual strategy) can be increased without increasing the cost. This analysis is also useful for the engineering design.

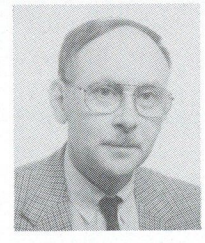
REFERENCES

[1] H. Nielsen, *Learning Machines.* McGraw Hill, 1968.
[2] M. Minsky, S. Papert, *Perceptrons.* MIT Press, Cambridge, 1969.
[3] D. Psaltis, J. Hong, *Shift invariant optical associative meories.* Optical Engineering, 26 No.1., 1987.
[4] J.J. Hopfield, D.W. Tank, *Neural computation in the decisions of optimization problems.* Biological Cybern. 52 (1985)
[5] J. Levendovszky. *Convergence properties of the Hopfield net in the case of generalized nonlinearities.* Proceedings Neuro-Nimes'90, Neural Networks and their Applications, pp.341-350, Nimes, France, 1990.
[6] J. Levendovszky, W. Mommaerts, E.C. van der Meulen, *Hysteretic neural networks for global optimization of quadratic forms.* Neural Network World, Vol.2., No.5. pp. 475-496, 1992.
[7] J. Levendovszky, *Distributed nonlinearities and multilayered perceptrons.* Technical Report at Oxford University, 1990.



János Levendovszky received his degree in Electrical Engineering from the Technical University of Budapest in 1986. In the same year he was granted a three year scholarship by the Hungarian Academy of Sciences in order to research for a Ph.D. He obtained his Ph.D. in 1989 on the topic of adaptive algorithms and equalization in communication systems. In 1989 he was allotted a Soros Scholarship at the University of Oxford, U.K., where he was involved in research and teaching on the subject of neural networks until 1990. From 1990 to 1993 he was invited to the Katholieke Universiteit Leuven, Belgium, as a free researcher at the Department of Mathematics, where his research concerned the mathematical foundations of neural networks. Presently, he is with the Department of Telecommunication, Technical University of Budapest. Dr. Levendovszky is the author of several papers devoted the statistical problems of neural computation.

Walter Mommaerts received his degree in Mathematics from the Katholieke Universiteit Leuven, Belgium. Afterwards, he was granted a research assistantship by the NFWO to do researh in the area principal component analysis, maximum entropy methods and neural networks. Presently he is a candidate for the Ph.D. His interest and research activity have been focused on the statistical problems of signal processing and neural networks in which fields he published several papers.



Edward C. van der Meulen received the B.S. degree in mathematics and physics and M.S. degree in mathematics from the University of Leiden, Netherlands, and the Ph.D. degree in statistics from the University of California Berkeley, in 1958, 1962, and 1968, respectively. From 1968 to 1975, he served on the faculty of the Department of Statistics at the University of Rochester, Rochester, N.Y. From September 1972 to January 1974, ha was on leave at the Mathematical Center in Amsterdam. He has been a Professor of Mathematics at the Katholieke Universiteit Leuven, Belgium since 1975. During 1975-1991 he was for short periods Visiting Professor or Visiting Scientist at Stanford University, Ohio State University, Syracuse University, Technical University of Budapest, and South China Institute of Technology. His research interest include multiuser information theory and information-theoretic statistical inference. In 1983, he organized the Fourth Benelux Symposium on Information Theory, Haasrode, Belgium. Dr. van der Meulen is a member of ASA, IMS, ISI, MAA, INNS, the Dutch Society for Statistics, and Sigma Xi. He is also a Member of the Editorial Board of the American Journal of Mathematical and Management Sciences. Currently he is Chairman of the Information Theory Chapter in the Benelux Section of IEEE.

BACKPROPAGATION WITH IMPROVED LEARNING SPEED

N. ELHADI and K. CSÉFALVAY

DEPARTMENT OF ELECTROMAGNETIC THEORY
TECHNICAL UNIVERSITY OF BUDAPEST
H-1521 BUDAPEST, HUNGARY

In this paper a new algorithm is introduced that increases the learning speed by a factor of two and reduces the occurrence of the local minima. Advantages of the new algorithm are demonstrated by computer experiments.

1. INTRODUCTION

The feedforward network which will be considered here is of the type described in [2], [3], [4]. The backpropagation algorithm can be used for adjusting the weights. The architecture of such networks is formed from several layers of processing elements, which are the input layer, the output layer, and one or more hidden layers in between. An example of a three layer network is shown in Fig. 1. The processing elements in the input layer accept the components x_i of the input pattern vector $\mathbf{X}$ of dimension n. The fully connected network outputs form an m-dimensional output pattern vector $\mathbf{Y}$. It has been shown in [3] that Three layers are enough to perform any classification task, we shall therefore deal only with three layer networks.

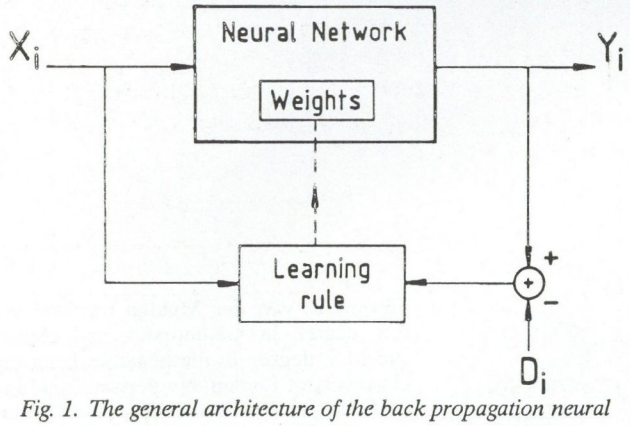


Fig. 1. The general architecture of the back propagation neural network

Each processing element is connected to all processing elements in the previous layer. The net input of a processing element is simply the linear weighted sum of all its inputs. The output of the j^{th} unit in the i^{th} layer is given by

$$y_j^{(i)} = f(net_j^{(i)}) \tag{1}$$

where

$$net_j^{(i)} = \sum_{k=1}^{N^{(i-1)}} (W_{kj}^{i} \cdot y_k^{(i-1)} + \theta_j^{(i)})$$

$$f(x) = \frac{1}{1+e^{-x}}$$

$W_{kj}^{(i)}$: is the weight connecting from the k^{th} unit in the $(i-1)^{th}$ layer to the j^{th} unit in the i^{th} layer,

$\theta^{(i)}$: is the threshold,

$N^{(i-1)}$: is the number of processing elements in the $(i-1)^{th}$ layer.

In a similar manner, we can determine the output of each processing element in other layers.

The network performs a mapping $F : (A \subset R^n) \rightarrow (B \subset R^m)$. In order to determine the appropriate weights to implement the mapping F the network should pass a training phase where a set of training examples $(\mathbf{X}_1, \mathbf{D}_1), (\mathbf{X}_2, \mathbf{D}_2), \ldots, (\mathbf{X}_q, \mathbf{D}_q)$ are presented to the network. $\mathbf{X}_i$ is the input vector and $\mathbf{D}_i$ is the corresponding desired response. The scheme of the backpropagation algorithm is shown in Fig. 2. The network training involves two phases. During the first phase an input pattern vector $\mathbf{X}$ is presented to the network, and the output pattern vector $\mathbf{Y}$ is computed (using arbitrary weight factors at the start).

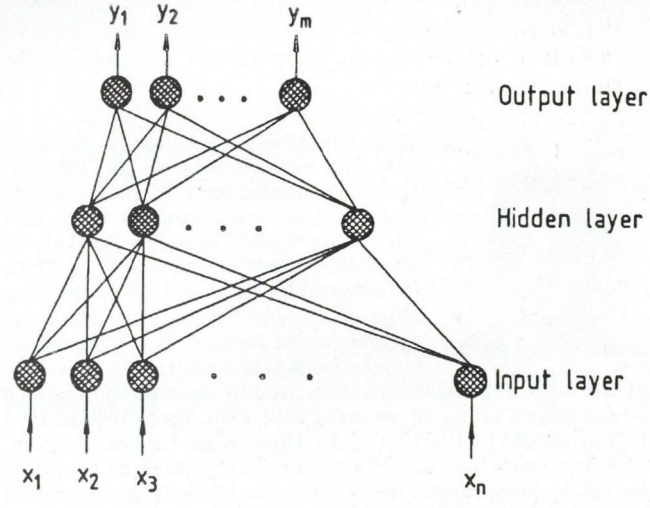


Fig. 2. The basic backpropagation model

In the second phase an error δ is computed for the output layer:

$$\delta_j^{(3)} = (d_j - y_j^{(3)})(1 - y_j^{(3)})y_j^{(3)} \tag{2}$$

where d_j and $y_j^{(3)}$ are the j^{th} component of the desired and actual output vectors.

The error for hidden units is determined in terms of the error of the output units:

$$\delta_j^{(2)} = y_j^{(2)}(1 - y_j^{(2)}) \sum_{i=1}^{N^{(3)}} \delta_i^{(3)} W_{ji}^{(3)} \qquad (3)$$

The weights are updated according to the following equation:

$$\Delta W_{ji}^{(k)}(n+1) = \qquad (4)$$

$$\eta \delta_j^{(k-1)} y^{(k-1)} + \alpha(W_{ji}^{(k)}(n) - W_{ji}^{(k)}(n-1))$$

where n is the time index, η is the learning rate, and α is the momentum term.

2. A NEW ALGORITHM

As pointed out in [1], [4], [5] a well known difficulty with the backpropagation algorithm is that its convergence is slow. Another problem is the occurrence of local minima, when the algorithm converges to some value other than the global minimum. If we examine the error and the update equations given by (2), (3), and (4) we can make the following general remarks.

- From Equation (4) we see that, if an input to a processing element is set to zero, then there will be no weight modification at that connection. This property is more important for the weights connecting the input layer to the hidden layer since the input vector may contain many zero components.
- From Equations (2) and (3) we see that the error will reach its maximum when the outputs from a layer is nearly half way between the two output values. Since the amount of change of the weights is proportional to this error, the weight modification will be faster at the start when the outputs are approximately equal to 0.5. But as we proceed with the iteration process the outputs will all approach either one or zero. Therefore the error values will be smaller which results in a slowdown of the learning rate.
- During the iteration process not all the input patterns require the same number of iteration to be correctly classified. In spite of this fact, for the training we use all the training pairs even if an input pattern is classified correctly.

To overcome these difficulties and achieve better performance the following modifications are proposed:

- First, we modify the transfer function for the processing units to be symmetrical around zero rather than being in the range from zero to one. This is done by modifying the transfer function for each processing unit in the hidden and the output layers to be the hyperbolic tangent as given by Equation (5). Fig. 3. shows the traditional activation function and the proposed one. The transfer function of the input layer is also modified as given by Equation (6). By doing so, the weight modification will be carried out regardless of the value of the input patterns close to zero.

$$f(x) = \frac{1}{2} tanh(\frac{x}{2}) \qquad (5)$$

$$g(x) = x - \frac{1}{2} \qquad (6)$$

- The error has to be zero if and only if the desired response matches the actual response. As mentioned before this is not the case when using the original algorithm. To eliminate this undesirable situation, a new parameter ρ will be introduced, and by this we can guarantee that the error signal will be zero only if the output matches the desired response.
- To move faster along the error surface, the definition of the error on input/output patterns is redefined as

$$E_p = \frac{1}{\beta} \sum_{j=1}^{N^{(3)}} |d_j - y_j^{(3)}|^\beta \qquad (7)$$

This new definition will result in larger weight modification steps and therefore it will increase the convergence speed.

- During the course of training not all the output pattern vectors converge in the same number of iterations. In order to increase the learning rate, we have to stop adapting weights by using the patterns which are already convergent. This means we have to check at each iteration wether the output vector matches the desired response. If there is a match, then the actual training pair is removed from the training set. The checking is carried out each sweep of the training pairs by comparing the maximum absolute node error (the error of node (i) in the output layer is $d_i - y_i$) to some specified threshold value γ. If the absolute value of the error is greater than the threshold, then we update the weights using that pattern, otherwise we skip that pattern. By this we ensure that if there is some disturbance for some of the correctly classified patterns then these patterns will be used again to modify the weights.

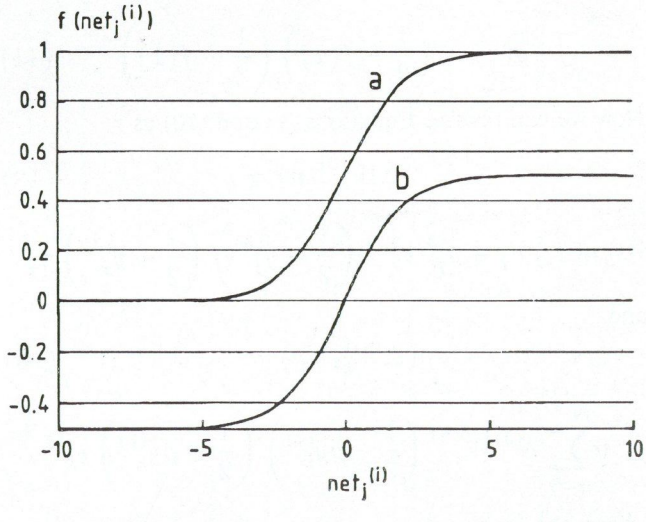


Fig. 3. Activation functions: (a) traditional activation function, (b) the modified function

3. THE UPDATE EQUATIONS FOR THE NEW ALGORITHM

Now we shall derive the error using the above approach. A gradient based learning rule can be expressed as,

$$\Delta W(n) = -\eta \frac{\partial E_p}{\partial W} \tag{8}$$

We use the new definition of E_p given by Equation (7) to get faster convergence. The changes in the output layer weights $W_{ij}^{(3)}$ (the weight connecting the i^{th} unit in the hidden layer to the j^{th} unit the output layer) are given by

$$\begin{aligned}\Delta W_y^{(3)}(n) &= -\eta \frac{\partial}{W_{ij}^{(3)}} \left(\frac{1}{\beta} \sum_{j=1}^{m} |d_j - y_j^{(3)}|^\beta \right) \\ &= \eta \; sgn(d_j)|d_j - y_j^{(3)}|^{\beta-1} g(net_j^{(3)}) y_i^{(2)} \\ &= \eta \; \delta_j^{(3)} y_j^{(3)} \end{aligned} \tag{9}$$

Where $sgn()$ is the sign function and $g()$ is the derivative of the activation function.

For the hidden layer the changes of the weights $W_{ki}^{(2)}$ are given by

$$\begin{aligned}\Delta W_{ki}^{(2)} n &= -\eta \frac{\partial}{\partial W_{ki}^{(2)}} \left(\frac{1}{\beta} \sum_{j=1}^{m} |d_j - y_j^{(3)}|^\beta \right) \\ &= \eta \sum_{j=1}^{N^{(3)}} sgn(d_j)|d_j - o_j^{(3)}|^{\beta-1} g(net_j^{(3)}) W_{ij}^{(3)} g(net_i^{(2)}) y_k^{(1)} \\ &= \sum_{j=1}^{N^{(3)}} \delta_j^{(3)} w_{ij}^{(3)} g(net_i^{(2)}) y_k^{(1)} \\ &= \eta \; \delta_i^{(2)} y_k \end{aligned} \tag{10}$$

the derivative of the activation function is given by

$$g(x) = \left(\frac{1}{2} - f(x) \right) \left(\frac{1}{2} - f(x) \right) \tag{11}$$

Now we can rewrite Equations (9) and (10) as

$$\Delta W_{ij}^{(3)}(n) = \tag{12}$$

$$\eta sgn(d_j)|d_j - y_j^{(3)}|^{\beta-1} \left(\frac{1}{2} - y_j^{(3)} \right) \left(\frac{1}{2} + y_j^{(3)} \right) \gamma_j^{(2)}$$

and

$$\Delta W_{ki}^{(2)}(n) = \tag{13}$$

$$\eta \sum_{j=1}^{N^{(3)}} \delta^{(3)} W_j^{(3)} \left(\frac{1}{2} - \rho y_k^{(2)} \right) \left(\frac{1}{2} + \rho y_k^{(2)} \right) y_k^{(1)}$$

where the parameter ρ is a small constant wich was introduced to ensure that the weight modification will only stop when the desired response matches the actual output.

From Equation (12) we see that the error of the third layer is:

$$\delta_j^{(3)} = sgn(d_j)|d_j - y_j^{(3)}|^{\beta-1} g(net_j^{(3)}) \tag{14}$$

the above equation can be approximated to simplify the calculation:

$$\delta_j^{(3)} = sgn(d_j)|d_j|^{\beta-1} \left(1 - (\beta - 1)\frac{y_j^{(3)}}{d_j} \right) g(net_j^{(3)}) \tag{15}$$

substituting the value of $g(net^{(3)})$ and after simplification we get:

$$\delta_j^{(3)} = (k_1 - k_2 y_j^{(3)})(0.25 - (y_j^{(3)})^2) \tag{16}$$

where $k_1 = sgn(d_j)2^{(1-\beta)}$ and $k_2 = (\beta - 1)2^{(2-\beta)}$. For the hidden units the error is given by

$$\delta_j^{(2)} = (0.25 - k_3(y_j^{(2)}))^2 \sum_{i=1}^{N^{(3)}} \delta_i^{(3)} W_{ji}^{(3)} \tag{17}$$

where $k_3 = \rho^2$.

In the algorithm we introduced a threshold γ. Each time we make an iteration we check for the following condition:

$$\max_j |d_j - y_j^{(3)}| > \gamma \tag{18}$$

If the above inequality holds, we modify the weights, otherwise we skip the weight update and continue with the next patterns. As proposed in [4], a momentum term is added to the update equations, this term will resist erratic weight changes caused by high spatial frequenties in the error surface.

An outline of the new algorithm can be given by the following:

STEP 1 Initialization. Set all weights and threshold values to small random numbers with zero mean. Set the parameters $(\beta, \alpha, \eta, \rho, \gamma)$

STEP 2 Present an input vector **X** and compute the corresponding actual output, and

$$\varepsilon = \max_j (d_j - y_j) \qquad 1 \le j \le N^{(3)}$$

STEP 3 If the value of ε is greater than the threshold γ, then we go to step (4), otherwiese we go to step (2).

STEP 4 Update the weights by

$$W_{ij}^{(3)}(n+1) = W_{ij}^{(3)}(n) + \eta \delta_i^{(3)} y_i^{(2)} + \alpha(W_{ij}^{(3)}(n) - W_{ij}^{(3)}(n-1)).$$

$$W_{ij}^{(2)}(n+1) = W_{ij}^{(2)}(n) + \eta \delta_i^{(2)} x_i + \alpha(W_{ij}^{(2)}(n) - W_{ij}^{(2)}(n-1)).$$

where n is the time index and $\delta_i^{(3)}$ and $\delta_i^{(2)}$ are given by Equations (16) and (17).

STEP 5 If we reach convergence, then we stop; otherwise we go to STEP 2.

4. COMPARISON OF PERFORMANCE

In order to have an empirical validation, extensive performance comparisons between the new rule and the standard backpropagation was performed. We have investigated many network structures, and various problems. We have seen that the new rule performed consistently better that the old one. In every case the speed of convergence increased by more that two times. To compare the two

rules we use the average mean square error per pattern given by

$$Error = \frac{1}{2P_t} \sum_{p=1}^{P_t} \sum_{i=1}^{N^{(3)}} (d_{pi} - y_{pi}^{(3)})^2 \qquad (19)$$

where

P_t: is the total number of the training patterns,

d_{pi}: is the i^{th} component of the p^{th} desired pattern,

$y_{pi}^{(3)}$: is the i^{th} component of the p^{th} output pattern.

From Equations (2), (3), (16) and (17) we can see that the error calculation for one iteration requires $(3N^{(3)} + 3N^{(2)} + N^{(3)}N^{(2)})$ multiplications for the new rule. The old rule requires $(2N^{(3)} + 2N^{(2)} + N^{(3)}N^{(2)})$ multiplications. The number of additions is equal for both rules and is equal to $(2N^{(3)} + N^{(3)}N^{(2)})$. From this it is obvious that the two algorithms have the same computational complexity.

We have used a neural network with seven inputs, five outputs, and six hidden units. Fig. 4 compares the performance of the two algorithms on the same classification problem. We can see that the new algorithm converged in 3800 learning cycles which required 239,400 multiplications and 152,000 additions. The other algorithms converged in 10,000 cycles which required about 520,000 multiplications and 152,000 additions. The speed of convergence is approximately two and a half times faster.

Secondly, test a set of classification problems was run on both types of algorithms. Table 1. shows the iterations per pattern required to converge when we use different number of pattern vectors to be learned. In every case the speed of convergence has been improved more than two times.

Table 1.

Iterations/pattern required for convergence with different number of patterns. For the new rule $\eta = 0.9$, $\beta = 1.7$, $\rho = 0.02$, $\gamma = 0.1$, for the old rule $\eta = 0.9$, $\alpha = 0.2$

Number of patterns	Iteration/pattern	
	New	Old
40	35	82
50	40	96
60	45	110
70	53	108
80	48	96
90	45	104
100	56	127
110	50	114
120	52	109

Thirdly, effects of topology on performance were tested. Table 2. shows the results for a number of different problems and network structures. The network structures are given by the set of numbers $\{N^{(1)}, N^{(2)}, N^{(3)}\}$, where the numbers give the of processing elements in the input, hidden and output layers, respectively.

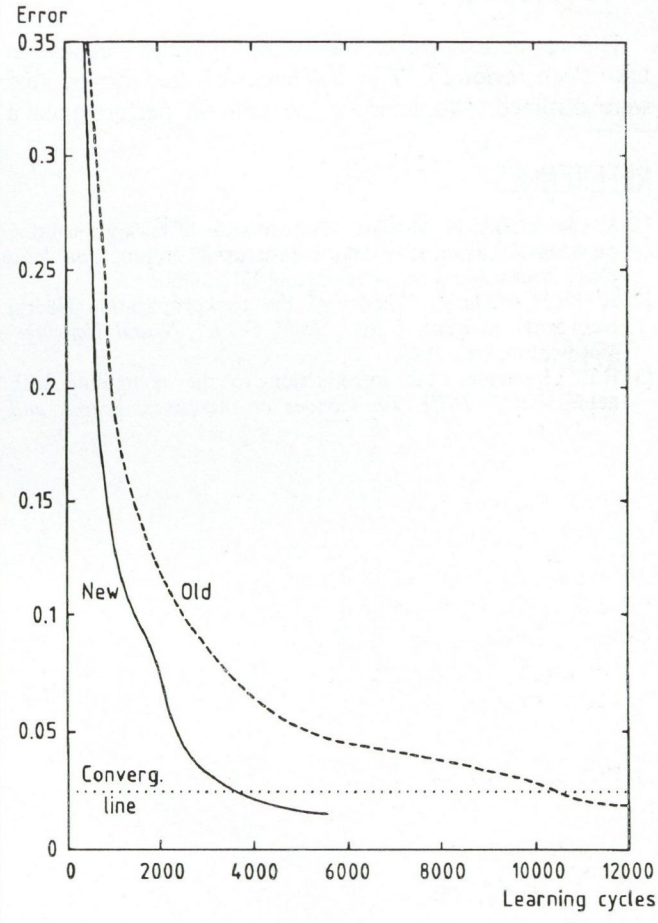


Fig. 4. Learning curves for the two algorithms: the new algorithm converges in 3800 learning cycles where the other converges in 10000 cycles

Table 2.

Simulation results. The first column presents the network structure and the problem. In the results for the new and old rules, 'Tc' stands for the number of convergent trials, 'T' is the total number of trials, 'I/P' is the average iterations/pattern required to converge. For the new rule $\eta = 0.9$, $\alpha = 0.1$, $\beta = 1.7$, $\rho = 0.01$, $\gamma = 0.1$, for the old rule $\eta = 0.9$, $\alpha = 0.1$

Network structure and Problem	New rule		Old rule	
	Tc/T	Ave I/P	Tc/T	Ave I/P
2,2.1 XOR	9/10	42	8/10	96
16,16,16 Random association	9/10	14	8/10	32
60,30,40 Random associations	6/6	20	6/6	38
70,30,30 Pattern classifier	6/6	24	6/6	47

5. CONCLUSION

The major aspects of the backpropagation algorithm have been reviewed. The drawbacks of the learning rule were outlined. To improve the general performance a new learning rule was proposed. Tests show that the convergence speed is increased by a factor of two in various classification problems while the computational complexity of the original algorithm is preserved.

REFERENCES

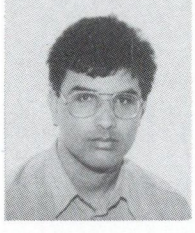
[1] V. Cherkassky, N. Vassilas, "Performance of backpropagation networks for associative database retrieval", in proc. *Int. Joint Conf. Neural Networks*, Washington, DC, 1989.

[2] R. Hecht-Nielson, "Theory of the backpropagation Neural Networks" in proc. *Int. Joint Conf. Neural Networks*, Washington, DC, 1989.

[3] R.P. Lippmann, "An introduction to the computing with neural nets," *IEEE Transactions on Acoustics, Speech and Signal processing*, 1987.

[4] D.E. Rumelhart, G.E. Hinton, R.J. Williams, "Learning internal representations by error propagation," *PDP: Vol.1, Chap.8,* Eds. D. E. Rumelhart, 1986.

[5] B. Widrow, M. Lehr, "30 years of adaptive Neural Networks: peceptron, Madaline and Backpropagation," *Proceedings of the IEEE*, Vol.78, No.9, 1990.



Nouri Elhadi was born in Tripoli in 1958. He received the B.Sc degree in Electrical Engineering from Fateh University, Tripoli in 1981, and the M.Sc degree from the Technical University of Budapest in 1988. He is currently a Ph.D student, his research interests include neural networks and pattern recognition.

Klára Cséfalvay graduated from the Department of Electromagnetic Theory of the Technical University of Budapest in 1966, received the dr. techn. degree in 1986. Since 1966 she has been with the Department of Electromagnetic Theory of the TUB. Her research subjects in addition to network analysis include digital signal processing, adaptive theory and neural computing.

THE CNN WORKSTATION

The CNN Workstation is a development system for various forms of Cellular Neural Networks. Three inputs are provided for images: scanner, camera and interactive graphics. The images, the cloning templates ("instructions"), and the set of instructions ("subroutines") all have their standard interface formats. A CNN software library is attached to the system. Three simulators are available: a multi-layer software simulator CNNM, a multi-processor digital hardware accelerator CNNHAC, and a dual CNN simulator with local logic CNNL.

1. INTRODUCTION

The 3-dimensional regular, analog, nonlinear, dynamic, locally connected processor array, called cellular neural network (CNN), has been invented in 1988 in Berkeley [1]. The CNN is composed of identical, relatively simple, analog, nonlinear dynamic units (processing units), which are placed on a regular 3D geometric grid (several 2D layers), and the analog interactions between the units are local (within a finite neighborhood). The interactions between the units are simple interconnections, although, they may be nonlinear and delay-type as well [2]. Here, the term "neural" is just characterizing the few types of uniform units and the large number of interactions.

The CNN can be considered as the 3D analog alternative of the 2D cellular logic automation invented by John von Neumann.

Extending the CNN paradigm with local (global) logic (without any A/D or D/A converters), the so-called dual CNN [7,6] is combining the strength of array dynamics and logic.

The CNN Workstation, is a development system for various forms of Cellular Neural Networks. Three inputs are provided for images: scanner, camera and interactive graphics. The images (or, equivalently, any analog 2D signal arrays), the cloning templates ("instructions"), and the set of instructions ("subroutines") all have their standard interface formats. A CNN software library is attached to the system. Three simulators are available: a multi-layer software simulator CNNM, a multi-processor digital hardware accelerator CNNHAC (an optional PC "add-on-board"), and a dual CNN simulator CNNL with local logic.

The CNN Workstation is based on standard "PC AT" personal computers, running under DOS and written in C.

2. CONFIGURATION

The CNN Workstation framework is for experimenting in image processing applications using CNN tools. As to capture images, one may have various sources and devices producing different storage format. The CNN Workstation can accept some standard graphic file formats and provides for conversions among them. Typical image sources are computer graphics, optical scanners and video devices (camera, VCR). To show input, intermediate and resulting images CNN Workstation uses high resolution VDU screen and laser printer for hardcopy. From its Control Panel, CNN Workstation components can be selected with mouse.

Computer

The CNN Workstation is running on IBM PC AT compatibles with 640 kbyte memory and DOS 3.xx or higher.

Video adapter and display

CNN Workstation components require standard EGA 640x350 or VGA 640x480 16 colour graphics display. In the latter case 16 grey level mode is also supported in controlling the display, ANSI enhanced control sequences are applied.

Mouse

For selecting components CNN Workstation uses Microsoft Mouse Menu software and a two-button Microsoft Mouse compatible.

Image sources

- *Scanners*: The lavish selection on scanner market comes with a wide variety of control software, typically using elaborate window environment, hence it has not seem feasible and viable to incorporate any or some of them in CNN Workstation directly. Rather, the PCX file format is defined as scanner input format, that each scanner (and computer graphics) software can produce.
- *Video signal sources*: To capture images from VCRs or cameras, a frame grabber add-on board is to be installed in the computer. On the contrary to the scanners, frame grabbers are worth interfacing on a lower lever, partly for the sake of speed and flexibility, partly for the lack of satisfactory software interface shipped with some of them. The CNN Workstation integrates a Visionetics Frame Grabber VFG512/8 board, providing interactive software environment to capture still images and sample motion picture segments with it. The output image of frame grabber as well as its control menus are appearing on the analog monitor or TV set connected to its video output.

Printer

The images processed can be printed out directly on on-line printer or sent into a disk file for later off-line printing by binary DOS copy. There are two printing modes available; for small black-and-white or coarse grey images, character graphics can be used on any type of printers having IBM-PC or PC-8 symbol set. The character graphics resolution in black-and-white mode is 1 pixel/character in the horizontal and 2 pixels/character in the vertical direction. In the coarse grey mode having three grey shades between the black-and-white levels the resolution is the half of that in both directions. The printable image dimensions are limited by the sheet size. For printing large, high resolution pictures of fine grey

shades, Hewlett-Packard LaserJet Series II compatible printers of 300 dot/inch resolution can be used. For half-toning 8*8 dot size fields are applied resulting in 37 pixel/inch picture resolution on printouts.

CNN Workstation files and directories

For the sake of simplicity the CNN Workstation related files are divided into two groups: the tools and the data files. All the programs, utilities, menus etc. are classified as tools and all the user generated files i.e. images of any formats, templates, control files etc. are data files. Two directories called CNNTOOLS and PIC are set up for the tools and the data files, respectively. That two directories are to be installed as subdirectories in any directory (called CNN Workstation root directory) using the SETUP procedure on the distribution disk, which also copies all the CNN Workstation files into proper place. (Alternatively, the distribution disk may contain the CNN Workstation files archived and compressed, keeping the proper path information, as well.)

The CNN Workstation services can also be started directly from DOS prompt as regular programs. All the programs when started without any parameter, or with insufficient or erroneous parameters give a list on the screen of all of its obligatory and optional parameters using decriptive names, sometimes together with brief further explanation.

Configuration summary

Computer: IBM PC AT compatible, 640 kbyte,
DOS 3.xx or higher
Video adapter and display: EGA 640x350 or
VGA 640x480 16 colour
Mouse: Microsoft Mouse compatible
Image sources:

- Scanners: PCX file format
- Video signal sources: Visionetics Frame Grabber VFG512/8

Image display for frame grabber: analog monitor,
TV set
Printer: Hewlett-Packard LaserJet Series II compatible
Installation: CNNSETUP or UnZip
Directories: CNNTOOLS, PIC

The CNN Workstation hardware environment is shown in Fig. 1.

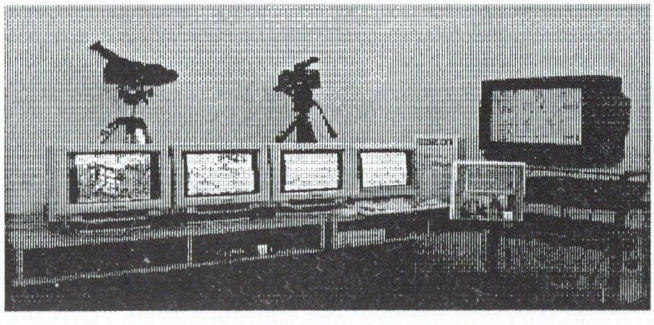


Fig. 1. CNN Workstation Hardware Environment

3. THE CNN WORKSTATION FRAMEWORK

The CNN Workstation can be started from CNN Workstation root directory, and its services can be activated through the Control Panel (also referred as CP). Except for the numerals to be used for referencing later in this material, it looks on the screen as shown in Fig. 2.

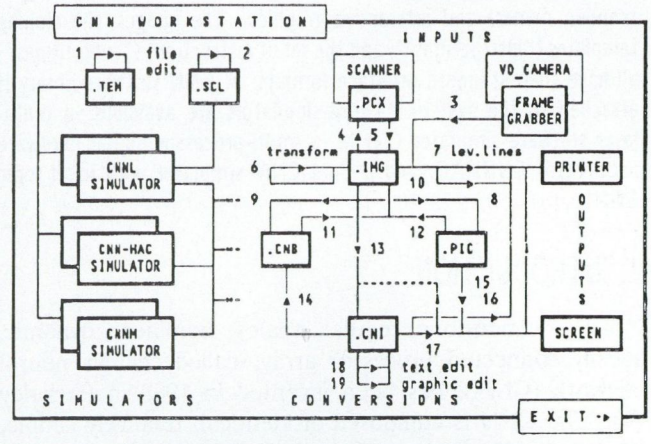


Fig. 2. CNN Workstation Control Panel

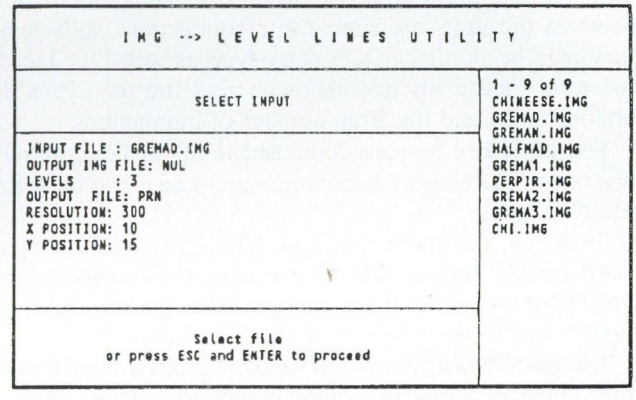


Fig. 3. A Query Session screen

Besides serving as main control means, the Control Panel is to depict the essential modules of CNN Workstation as well as the logic relations and interconnections among them. The CNN Workstation modules can be grouped as follows:

- Image sources, input devices
- Image output devices
- Databases
- Conversions among image databases
- Utilities and editors
- CNN simulator packages

To activate a CNN Workstation service it is to be selected by pointing at it and pressing either mouse button. When selecting a simulator the cursor is to be moved into the corresponding rectangle. All the other CNN Workstation services are represented by arrows along logic interconnection lines. To activate either service, the corresponding arrow should be selected. The CNN Workstation services are performed by a couple of programs. Each program has its command line parameters, some of which are

optional and may have predefined default values. When starting these programs through the CNN Workstation Control Panel, a query session ensues, asking for the pertinent command line parameters and offering default values whenever relevant, as shown in Fig. 3. In addition to command line parameters the CNN Workstation service programs may have further parameters to be given during their run according to the corresponding Users' Instructions.

4. CNN WORKSTATION DATABASES

The CNN Workstation databases are represented on the Control Panel by blocks marked with three-letter file extension codes: .TEM, .SCL, .PCX, .IMG, .CNB, .CNI., .PIC. They are to contain either images of different encoding formats or CNN algorithms of different complexity. A simple CNN algorithm is the CNN template (.TEM), which may be single- or multi-layer, linear, nonlinear and delay-type. A collection of various templates can be found in Dual CNN Software Library [10]. The elaborate control sequences for CNNL Simulator, written in Simulator Command Language (.SCL) are examples for complex CNN algorithms, CNN analog software. Detailed descriptions of TEM and SCL storage formats can be found in the Users' Manuals of CNNM and CNNL Simulators, respectively.

In order to match image sources and CNN Workstation services, CNN Workstation accepts five image encoding formats and provides for conversions among them. Considering at most 256 distinct luminescence levels on the grey scale no conversion or sequence of conversions would loose accuracy. (The term "luminescence" referring usually to that perceptible feature of the original picture, is used interchangeably with term "grey level".) The image dimension limitations are different for different CNN Workstation services, usually the 512-pixel frame size cannot be exceeded. However, utilities are at hand to cut images or to modify image dimensions — magnify or reduce — up to 3000-pixel frame size.

The five image databases are as follow:

- PCX: images compressed by run length encoding from scanners and computer graphics packages, having 128 bytes long standard header containing, among other less important items, the image dimensions and the pixel luminescence encoding parameters: the number of bits per pixel (bpp) and the number of colour planes (ncp). In CNN Workstation the acceptible (bpp:ncp) pairs are (1:1) = black-and-white, (1:4) = 16 grey levels (EGA) and (8:1) = 256 grey levels (VGA).
- IMG: uncompressed image storage format of VISIONETICS frame grabber software, having 18 bytes long header followed row-wise by the image contents in 256 grey levels (1 byte/pixel) encoding.
- PIC: uncompressed image storage format with 2*4 ASCII bytes (characters) for horizontal and vertical dimensions followed row-wise by the image contents in 256 grey levels (1 byte/pixel) encoding.
- CNI: text-editable (ASCII character) image storage format beginning with the horizontal and vertical dimensions followed row-wise by real number values of pixel grey levels normalized between −1 and +1 standing for white and black, respectively. In CNI database sampled motion pictures can also be stored by sequencing blocks of the above format, separated by lines beginning with #.
- CNB: uncompressed binary storage format beginning with 2*2 bytes for the horizontal and vertical dimensions followed row-wise by pixel grey levels normalized as in CNI, encoded as 7.9 fixed point binary number, 2 bytes/pixel. In CNB database, similarly to the CNI case, sampled motion pictures can also be stored by sequencing blocks of the above format. (CNB is the core-image format of CNN-HAC hardware accelerator on-board software written for Texas TMS320C2x fixed point digital signal processors. Except for direct downloding onto CNN-HAC board, no CNN Workstation service can make use of it.)

5. FORMAT CONVERTERS AMONG IMAGE DATABASES

The image format conversions available are shown on the Control Panel by arrows. The original luminescence fidelity of image sources is always preserved; converting, however, real valued CNI images the continuous shading cannot be maintained, each pixel luminescence value is rounded according to the target database.

For the black-and-white case the picture-background convention is not the same for the different input and output devices. On paper-type input and output devices, as scanners and printers, usually white is considered as background and black as picture, which convention has been adopted in CNI and CNB databases, meaning that the higher the numerical value of a pixel luminescence is, the darker the corresponding pixel is considered to be. On the other hand, the convention in computer graphics and on video image sources is just the opposite: the black parts are considered as background and the white ones as image, which is reflected in the encoding method applied in PCX, IMG and PIC databases, in assigning higher numerical values to brighter pixels. In case of computer displays, depending on the level of application, the background-foreground convention is usually reduced to setup parameter or palette selection, neither of them providing a sound base for image encoding methods. Therefore most of the CNN Workstation image format converters can optionally invert images when transferring from a database to another one.

The IMG image database is in central position in the sense that images stored in it can directly be converted into any other database. Conversion from IMG to CNI database offers several additional options and services as follows:

- displaying images in black-and-white, pseudo colours, or grey shades
- cutting out rectangular image segments either by giving corner coordinates, or by using markers on the image frame
- reduction of image size, with adjustable threshold for black-and-white images
- calculation and adjustment of the average grey level
- outputting sampled image sequences from the sampled

tableau produced by the FRAME GRABBER service CP 3.

6. CNN WORKSTATION IMAGE INPUTS AND OUTPUTS

The practical applicability and usefulness of CNN Workstation is highly depending on the ease of acquiring images from various sources. As most common image sources integrated in PC environments, computer graphics packages, scanners and video signal frame grabbers are widespread, the CNN Workstation accepts images from those sources.

6.1. Images input from scanner

The wide variety on the market of image scanners and elaborate software tools shipped with them renders the idea to define and develop a generally befitting scanner interface subsystem unfeasible, if not impracticable. Rather, in CNN Workstation environment the compressed PCX image file format, which all scanner software can produce, is defined as scanner interface, see Control Panel SCANNER.

6.2. Capturing images from video signal sources

The video camcoders, video tape recorders and the television broadcasting provide image processing experimentation with pictures of most diverse sort, including still ones as well as sampled sequences of motion pictures. The conversion of pictures from video signal format of some standard (PAL, NTSC, SECAM) to a computationally comprehensible one; capturing, sampling, discretizing and finally storing it as a rectangular array of integers standing for the average luminescence of tiny picture fragments to be associated with the discrete image pixels from now on, is done by a frame grabber board installed in the PC the CNN Workstation is running on. As the image storage format of frame grabbers are not yet standardized and sampling of motion pictures needs direct control over the frame grabber, the CNN Workstation is integrating a proprietary software package developed to control a VISIONETICS frame grabber board. The FRAME GRABBER service can be started at CP 3.

6.3. Image output devices and services in CNN Workstation

The simulators, utilities and conversion routines of CNN Workstation produce images of wide range in dimensions, resolution and gradation. The image output devices of CNN Workstation are screen and printer. To comply with image parameters and typical PC configurations trade-off in appearance has been made.

Images on screen

Most of CNN Workstation services require EGA or VGA type display to be used in 640*350 or 640*480 high resolution mode. Grey tone images are shown using 16 pseudo colours on EGA and 16 grey levels on VGA screen. Large, high resolution images are truncated, medium size images are magnified to fill the screen and small ones usually displayed in character mode, as necessary, depending also on the service in use.

Depending on the video mode applied and the adjustment of display unit, the visible horizontal and vertical resolution may be different resulting in rectangular distortion of square-like shapes. To compensate distortion some CNN Workstation services provide for manual or automatic means to adjust the aspect ratio of resolutions.

Being the screen the natural device for user communication almost all the CNN Workstation services use it as output device. There are only two simple services having no other purpose than displaying images. One is at CP 15 to show PIC, the other is CP17 to show CNI type images, as shown in Fig. 4.

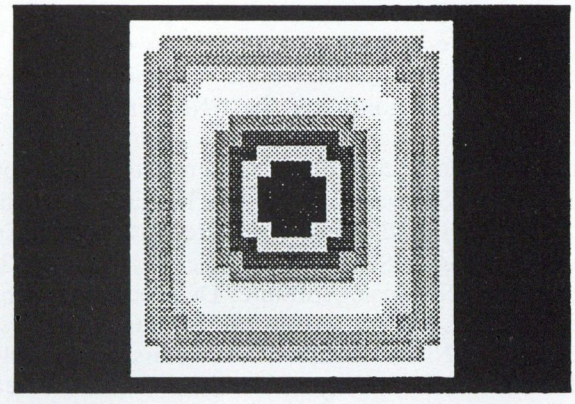


Fig. 4. Grey-scale CNI type image on screen

Printing images

Similarly to displaying images, large, high resolution images are to be printed on graphics printer in high resolution mode, while small size images can be printed on any printer in character mode using IBM-PC or PC-8 symbol set.

From CP 16, CNI type images can be printed out in black-and-white and/or coarse grey character mode. Optionally, the numerical values of pixel grey levels can be printed out as well.

In high resolution graphics mode the CNN Workstation printing services generate Hewlett-Packard LaserJet II control sequences. In black-and-white mode the printer resolution can be 75, 100, 150 or 300 dot/inch. In grey tone mode the half-toning algorithm is using 8*8 dot size patterns of 300 dot/inch resolution for grey pixels providing 64 grey levels and 37.5 pixel/inch image resolution.

From CP 8, IMG type grey scale images can be printed out in high resolution. Optionally the output can be redirected to disk file for later off-line printing or can be disabled when using the routine only to show images on the screen. The routine when running offers several options to choose from, concerning image size, magnification, positioning, inverting, cutting etc.

7. UTILITIES AND EDITORS

7.1. Editing images and data

The CNN templates stored in database TEM control sequences stored in database SCL and also the images

encoded in database CNI, all are having editable character (ASCII) format. On Control Panel selection of CP 1, CP 2 and CP 18 respectively gives access to text editor environment in the three databases mentioned above. The CNN Workstation distribution disk contains a text-editor, but any other editor can be activated from CNN Workstation simply by modifying the calling procedure accordingly.

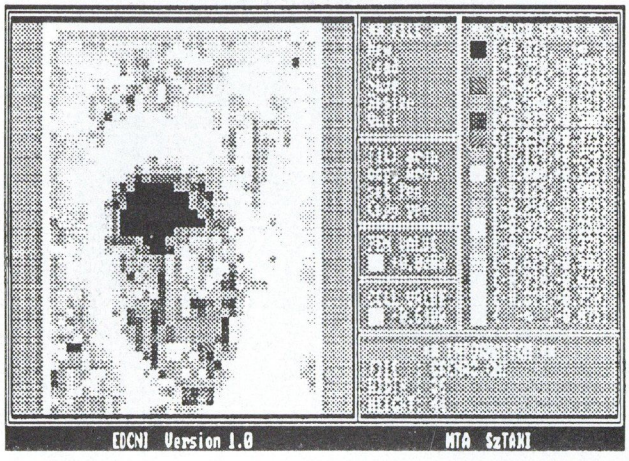


Fig. 5. Screen Editor for CNI type images

The images in CNI database can be inspected and modified on screen graphically as shown in Fig. 5 by selecting CP 19. Details of usage can be found in its Users' Manual [9].

7.2. Image transformations

Using CNN Workstation service CP 6 separate, horizontal and vertical magnification/reduction can be performed with high luminescence fidelity, on images in database IMG. The magnification/reduction rate are given as real numbers.

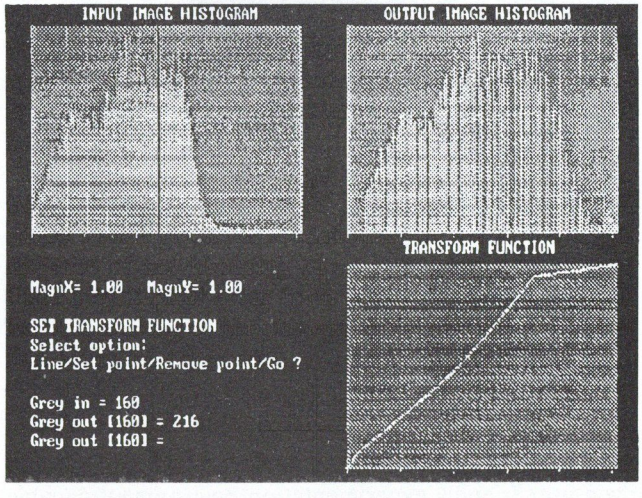


Fig. 6. Grey-scale image histogram transformation

Before performing the transformation, the luminescence histogram of image can be inspected on screen and a piece-wise linear function to be used during transformation as luminescence look-up table can be set up by giving the coordinates of corner points. When defining that function, its actual shape and also the resulting luminescence histogram are shown on the screen (see Fig. 6). In accordance with IMG encoding format 256 distinct grey levels (0-255) can be used.

7.3. Drawing constant grey level lines

Similarly to geographical maps of mountains representing height by constant level lines, selecting CP 7 a drawing of constant grey level lines of images in database IMG can be produced, as shown in Fig. 7 (CP 7 is realizing an uneven coarse grey transformation and a CNN algorithm of nonlinear control /B/ template.) The drawing shown on screen can be positioned and printed out with selectable resolution on Hewlett-Packard LaserJet II compatibles and stored in database IMG as well. The number of grey levels initially to be distributed evenly, is to be prescribed in the beginning, the eventual "height" of levels can be adjusted interactively.

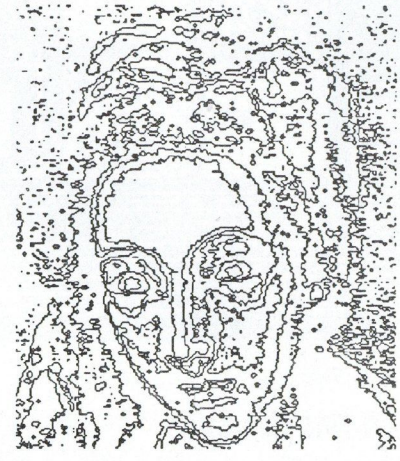


Fig. 7. Constant levels on a grey-scale image

8. CNN WORKSTATION SIMULATORS

The purpose of CNN Workstation Simulators is to support both theoretical research and practical experimentation aiming at designing, testing and controlling CNN chips of VLSI realization.

8.1. The Dual CNNL Simulator

The CNNL Simulator is a menu driven software tool for experimenting with a simulated CNNL [5] chip either on hardware or on functional level in the time domain. It can be used as a general Cellular Neural Network Simulator having various interactive control and flexible output features, as well.

The CNNL architecture and chip have been conceived for dual - i.e. composite neural and digital - computations by programmably sequencing several different templates, combining through logic function the consecutive neural results and using in-cell feedback mechanism for cascade, iterative and recursive algorithms.

The CNNL chip accommodates an array of composite cells, all containing an analog CNN part and a digital shift register with some logic. Each cell is connected to the

nearest horizontal and vertical neighbours. Voltage values are communicated to and from the chip through write- and read-lines, respectively. Row-select-lines are for selecting one at a time row for reading or writing.

The Simulator computes the (analog) voltage transient values and digital register levels in a CNNL taking into account its previous state as well as the values on the external control and signal lines. The Simulator has two independent user interfaces. It can be controlled either through a user-friendly on-screen interactive window menu system or by user edited control files containing a series of commands according to the Simulator Command Language, stored in database SCL.

Detailed description of how to use the CNNL Simulator and its menu system, together with the SCL syntax can be found in the CNNL Simulator Users' and Reference Guide [6]. A typical screen is shown in Fig. 8.

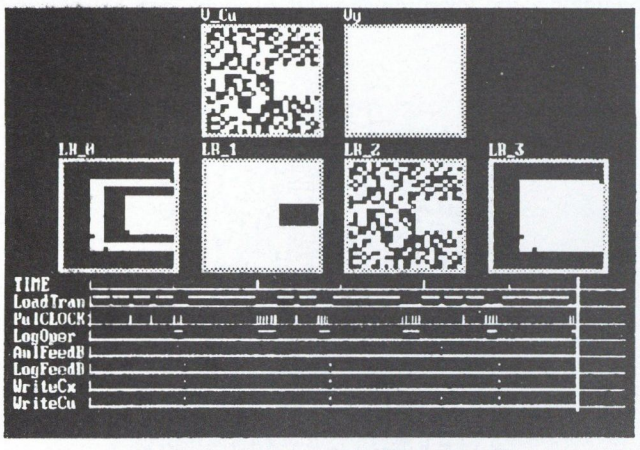


Fig. 8. Typical screen of CNNL Simulator

8.2. The CNN-HAC Simulator

Due to the insufficient computing power, to perform CNN type image processing experiments on realistic, large, high resolution pictures using merely PC software tools is not feasible. With CNN-HAC Simulator a hardware accelerator board (CNNHAC10, see Fig. 9) accomodating four Texas TMS320C25 fixed point digital signal processors and 8 Mbyte on-board RAM memory together with due interprocessor communication and control logic can be summoned to quickly and efficiently perform CNN tasks on images of up to one million pixels. Depending on the actual task and template structure, it may perform CNN calculations at the speed of 1 millon iterations/sec/cell.

To run the CNN-HAC Simulator a CNN-HAC board is required in the PC. The CNNHAC10 Simulator provides a hierarchical menu driven environment to select images, download them onto and upload from the board, to set up displaying parameters etc. The Simulator can accept CNI, CNB, IMG and PIC format images and produce CNB, IMG and PIC format image files. The different CNN algorithms running on CNNHAC10 board, written in Texas TMS320C2x Assembly Language, can be selected and downloaded too. In CNN Workstation a "hidden" database called HAC, accesible solely through CNN-HAC Simulator is maintained to store the different on-board software modules.

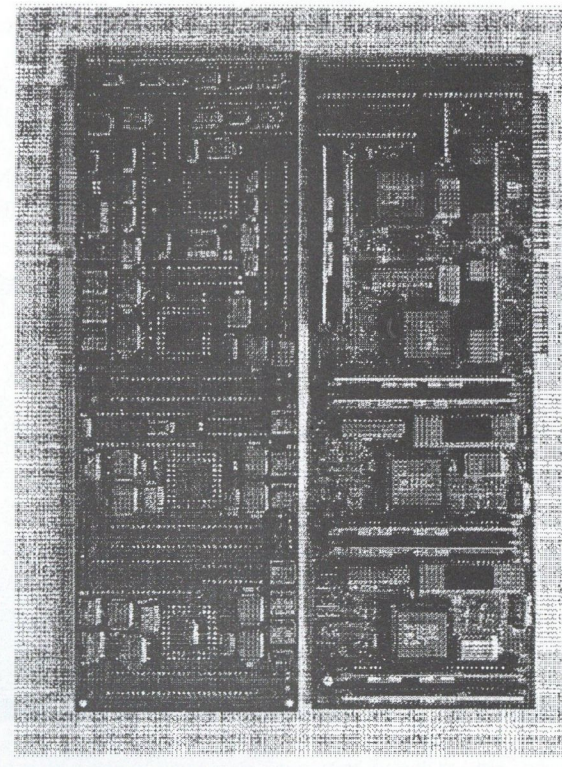


Fig. 9. The CNNHAC10 Hardware Accelerator Board

Further details about CNN-HAC Simulator and its usage can be found in [11]. A typical screen is shown in Fig. 10.

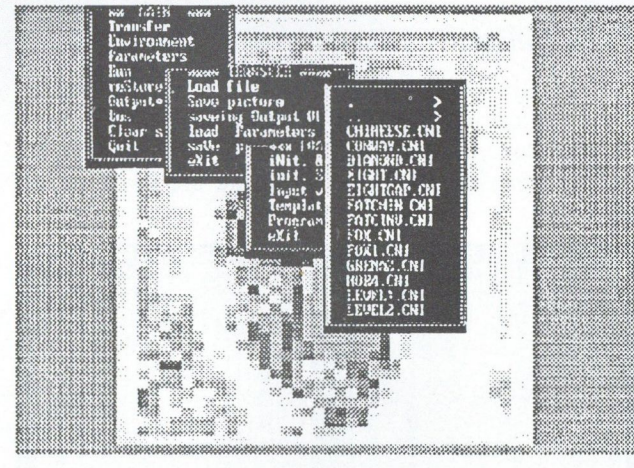


Fig. 10. Typical screen of CNN-HAC Simulator

8.3. The Multi-Layer CNNM Simulator

Small and medium size images up to some ten-thousand pixels, depending on the available memory and acceptable time consumption, can be processed with CNNM Simulator [12].

The CNNM Simulator provides a hierarchical menu driven user friendly experimentation environment capable of simulating a wide variety of possibilities related to the theory of CNN, brought up so far. The most important feature is the capability of simulating multi-layer networks, even with different cell densities on different layers. Since most applications are based on single layer networks, this feature was implemented in a way not making single-layer applications more complicated to use. There is no

restriction imposed on which layers are interconnected. The templates describing the interconnections can be linear, non-linear or delay-type. Besides the analog CNN, its digital version, the discrete-time CNN [4] can also be simulated with the program. Different types of configuration files are provided to make possible to restore the environment or even a whole simulation at any time.

For motion detection experiments the CNNM Simulator can accept image sequences in adjustable sampling rate. Results of long, time-consuming transient calculations can be recorded in disk files and played back later with adjustable speed.

Although both the pixel luminescence and the template values are given as real numbers, for the sake of computational speed they are converted into 2-byte fixed point format, which may very seldom result in limitations on the numerical resolution of input values and the range of CNN state values calculated by the Simulator. A typical screen of CNNM Simulator is shown in Fig. 11.

For the simulation of special optical CNN devices and performing some image feature extraction tasks, in addition to the two capacitor voltage values, which are the usual image inputs, an optional third input can also be supplied for each CNN cell. The cell current values, otherwise defined homogenously for the whole CNN as a template entry, can be given as an image in either database, for both the CNN-HAC and the CNNM Simulator.

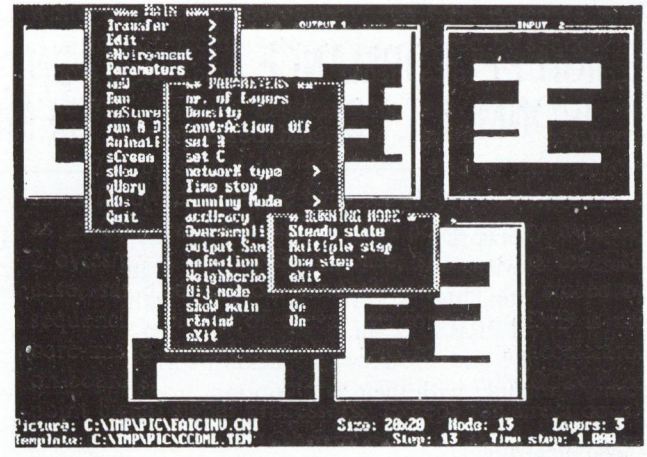


Fig. 11. Typical screen of CNNM Simulator

ACKNOWLEDGEMENTS

This work has been supported by the National Research Fund of Hungary (OTKA) under grant No.2578/1991 and by joint research grant No. INT-90-01336 of the National Science Foundation (USA) and the Hungarian Academy of Sciences. The various parts of the Workstation has been developed by dr. P. Szolgay, dr. T. Szirányi and our graduate students T. Boros, J. Csicsvári, T. Kozek, Zs. Ugray, P. Venetianer, Á. Zarándy and Tuomotapani Honkanen.

REFERENCES

[1a] L. O. Chua and L. Yang, "Cellular neural networks: Theory", *IEEE Transactions on Circuits and Systems*, Vol. 35, 1988.

[1b] L. O. Chua and L. Yang, "Cellular neural networks: Applications", ibid.

[2a] T. Roska and L. O. Chua, "Cellular neural networks with nonlinear and delay-type template elements", *Proc. IEEE CNNA-90*, 1990

[2b] T. Roska and L. O. Chua, "Cellular neural networks with nonlinear and delay-type template elements and non-uniform grids", *Int. J. Circuit Theory and Applications*, Vol. 20, No. 5, Sept – Oct 1992.

[2c] T. Roska and L. O. Chua, "Dual CNN analog software", *Report DNS-1-92, Dual and Neural Computing Systems Res. Lab.*, Comp. Aut. Inst., Hung. Acad. Sci., 1992.

[3] J. A. Nossek, G. Seiler, T. Roska and L. O. Chua, "Cellular neural networks: Theory and circuit design", *Report No. TUM-LNS-TR-90-7, Inst. Network Theory and Circuit Design*, Technical University of Munich, Munich December, 1990 and *Int.J. Circuit Theory and Applications* Vol. 20, No. 5, Sept – Oct 1992.

[4] H. Harrer and J. A. Nossek, "Time discrete cellular neural networks: architecture, applications and realization", *Report No. TUM-LNS-TR-90-12*, Technical University of Munich, November, 1990

[5] K. Halonen, V. Porra, T. Roska and L. O. Chua, "VLSI Implementation of a reconfigurable CNN containing local logic", *Proc. CNNA-90*, or in *Int J. Circuit Theory and Applications* 1990.

[6] A. Radványi, K. Halonen and T. Roska, "The CNNL Simulator and some time-varying CNN templates", *Report DNS-9-91, Dual and Neural Computing Systems Res.Lab.*, Comp. Aut. Inst., Hung. Acad. Sci., 1991.

[7] T. Roska, "Analog events and a dual computing structure using analog and digital circuits and operators", in *Discrete Event Systems: Models and Applications (eds. P. Varaiya and A. B. Kurzhanski)*, Springer-Verlag, Berlin, 1988.

[8] Poceedings of the *IEEE International Workshop on Cellular Neural Networks and their Applications*, CNNA-90, Budapest, 1990, IEEE Cat.No. 90TH0312-9

[9] CNND Simulator, Cellular neural network embedded in a simple Dual computing structure, *Users' guide Version 3.0*, Report No. 37/1990 (ed.T. Roska and A. Radványi), Computer and Automation Institute, Hungarian Academy of Sciences (MTA SzTAKI), Budapest, 1990.

[10] T. Roska, A. Radványi, T. Kozek and T. Boros, "Dual CNN software library", Report DNS-7-1991, *Dual and Neural Computing Systems Res.Lab.*, Comp. Aut. Inst., Hung. Acad. Sci., 1991.

[11] A. Radványi, Á. Zarándy, "CNN-HAC: Cellular Neural Network Simulator Using Digital Hardware Accelerator Board, Version 5.0, User's Guide", Report DNS-11-1992, *Dual and Neural Computing Systems Res.Lab.*, Comp. Aut. Inst., Hung. Acad. Sci., 1992.

[12] A. Radványi, P. L. Venetianer and Á. Zarándy, "CNNM Multi-Layer Cellular Neural Network Simulator, Version 2.4, User's Guide", Report DNS-12-1992, *Dual and Neural Computing Systems Res.Lab.*, Comp. Aut. Inst., Hung. Acad. Sci., 1992.

[13] A. Radványi, T. Roska and Á. Zarándy, "DUALCOMP – Dual CNN Compiler to CNN-HAC1 Board, Version 2.0, User's Guide", Report DNS-13-1992, *Dual and Neural Computing Systems Res. Lab.*, Comp. Aut. Inst., Hung. Acad. Sci., 1992.

[14] P. Szolgay and T. Kozek, "Optical detection of layout errors of printed circuit boards using learned CNN templates", Report DNS-10-1991, *Dual and Neural Computing Systems Res.Lab.*, Comp. Aut. Inst., Hung. Acad. Sci., 1991.

A. G. RADVÁNYI and T. ROSKA
MTA SZTAKI

HIGH PERFORMANCE SIMULATION ENVIRONMENT FOR DIGITAL SYSTEMS

The SPRINT (Simulation Program of Response of Integrated Network Transients) and its X Window-based graphics environment SIGHTS (Standard Interface for Graphics Handling of Transient Signals) with their integrating shell PRESTO, all from High-Design Technology, Torino, Italy, and the Hewlett-Packard series 54120 digital oscilloscopes with their TDR options implement a new conception of digital system simulation for advanced high-speed electronics.

SPRINT and SIGHTS show fully complementary features with respect to all products in the market today, such as SPICE or SABER-like simulators, transmission line simulators. The unique capabilities of SPRINT and SIGHTS can be fully exploited in the area of digital electronics, as well as in broad range of applications ranging from analog to very high-frequency, and optical fiber applications. They can solve critical problems efficiently where other products fail.

The computing engine of the software environment is SPRINT. SPRINT uses revolutionary ways to simplify the modelling and simulation task, hence retain the required accuracy.

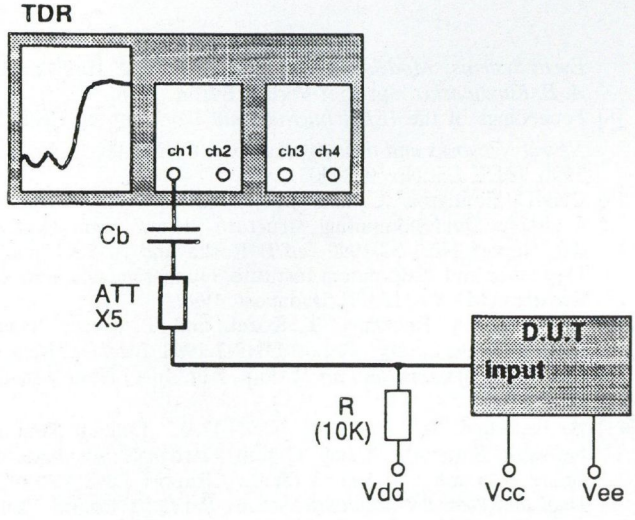


Fig. 1. Measurement setup with the Hewlett Packard Time-Domain Reflectometry unit. Reflection coefficient data on the measured port are collected.

Modelling of components is done by BEHAVIORAL TIME DOMAIN MODELLING (BTM). The component which is attached to a specific node in the circuit, is represented by its measured response at that node, no knowledge or additional data is required on the internal structure and/or operation of the device. This one-port approach greatly simplifies the models themselves and speeds up the simulation significantly. The Hewlett-Packard Time-Domain Reflectometry setup is used to collect the measured reflection coefficient data on the device. Fig. 1 shows the measuring setup. Then, to reduce the number of data points, using the MODEL CAPTURE SYSTEM (MCS) of SIGHTS, a piece-wise linear approximation of the measured response is generated. An example is shown in Fig. 2, where the collected data are shown together with their piece-wise linear approximation.

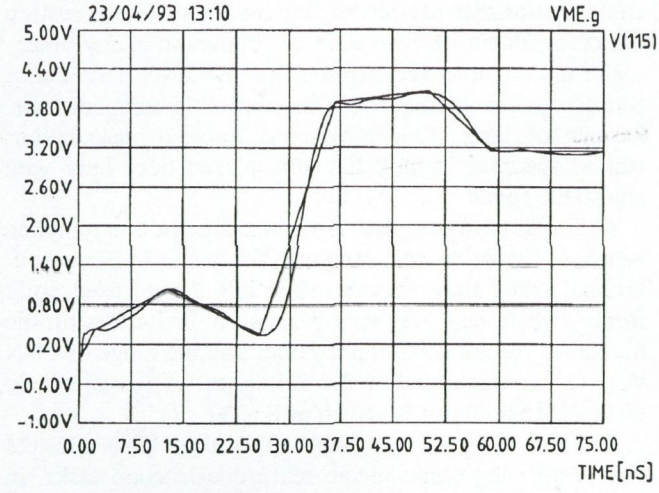


Fig. 2. Measured TDR response of a device with its piece-wise linear approximation.

Using this technique, generating models for complex active devices is an easy task. Fig. 3a is an example of how the input pin on a PGA package can be modelled with three simple components. The series transmission line represents the packaging, a piece-wise-linear (PWL) resistor, R1, takes into account the static nonlinearities of the input, while the PWL s11 reflection coefficient describes the dynamic behavior. If the transfer function of the active device is also of importance, the input model can be extended to incorporate the effect of packaging on the ground and supply pins, as well as the effects of static and dynamic transfer characteristics. A model of a CMOS circuit is shown in Fig. 3b.

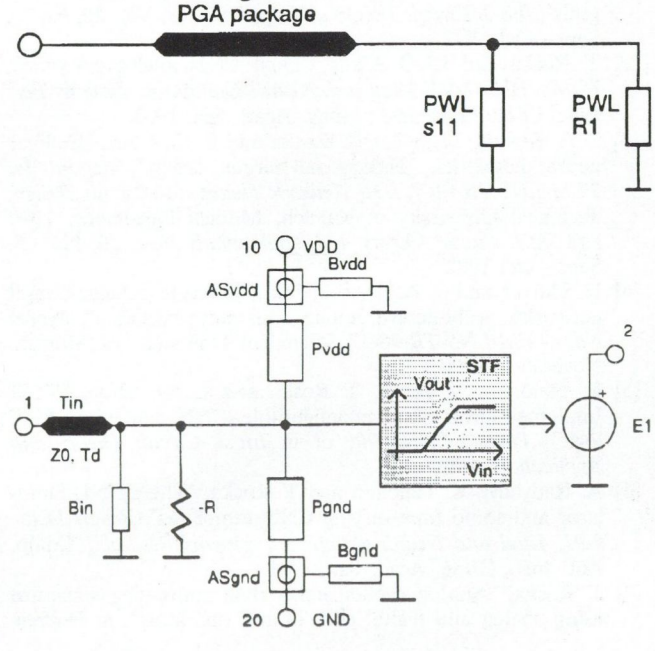


Fig. 3. a) Model of a PGA input; b) Model of a CMOS circuit from input to output.

From the single-device models, circuits and systems of any complexity can be built. A typical VME backplane model is shown in Fig. 4. The bus is modelled with bus cells, where the interconnection paths on the backplane, the connector pins, and the daughterboard traces are all modelled by individual transmission lines. The very fast operation of SPRINT will enable the design engineer to make many runs to detect marginal behaviour of the system. In Fig. 5 the result of multiple runs on the VME backplane is shown. SPRINT can also take models of lossless, lossy, and coupled transmission lines.

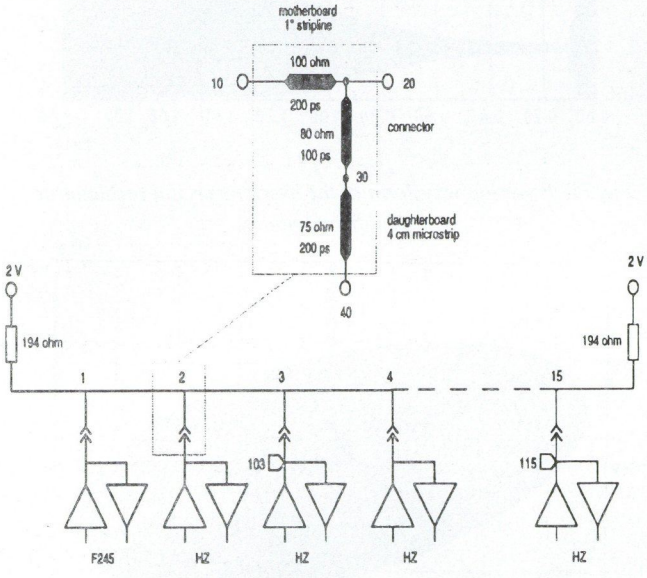


Fig. 4. Model of a VME backplane bus.

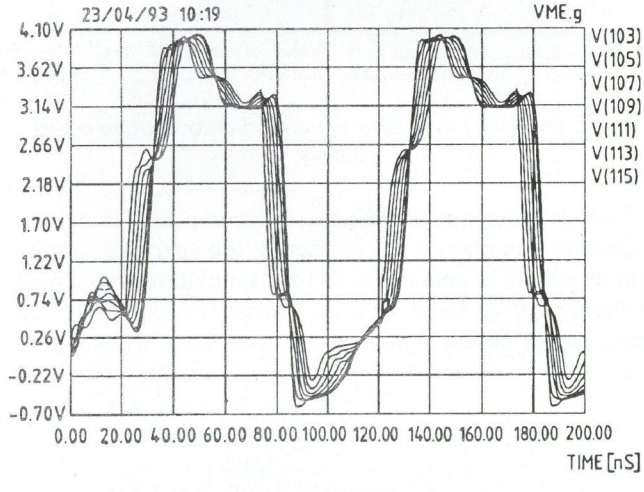


Fig. 5. Simulated time-domain waveform of the VME backplane bus with multiple runs.

Large communications systems can also be modelled by the Hewlett-Packard TDR unit and can be simulated by the High Design Technology SPRINT, SIGHTS and PRESTO. In Fig. 6, an IEEE 802 LAN is shown with the source, transmission line segments, and terminations. Fig. 7 is an example of a lossy transmission line response. Both the original response, collected by the Hewlett Packard TDR instrument, and the simulation result from its PWL approximation are shown. On the magnified scale, the tap load effect is clearly seen.

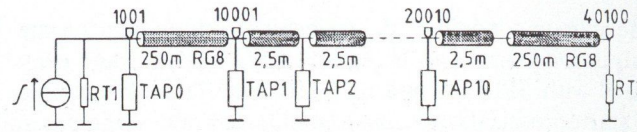


Fig. 6. Model of an IEEE 802 Local Area Network.

SPRINT and SIGHTS will output not only simple time-domain response waveforms, but presentation of simulation result in a form which is more convenient or typical for communications engineer is also possible. Fig. 8 illustrates the performance of the Local Area Network shown in Fig. 6 by its eye diagram for three different data speeds. SPRINT and SIGHTS are very useful in finding marginal operating conditions. Multiple simulation runs, or simulation runs with long pseudo random binary data sequences (PRBS) will reveal weak points of a design. SIGHTS will identify for the user the bit pattern which corresponds to the worst case eye closing. The eye diagram of Fig. 9 is an example for a PRBS simulation.

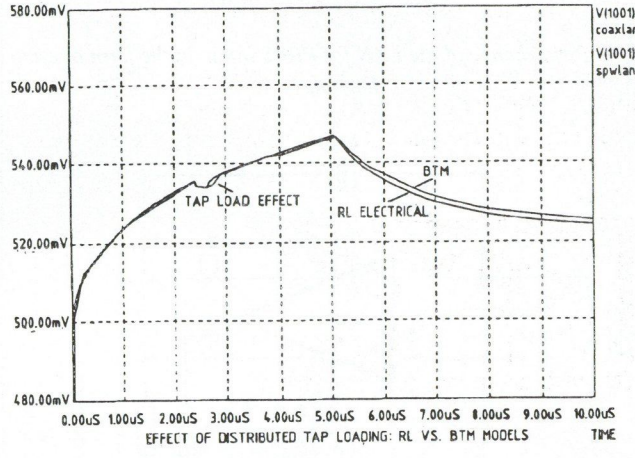


Fig. 7. Time-domain measured and simulated response of the lossy cable.

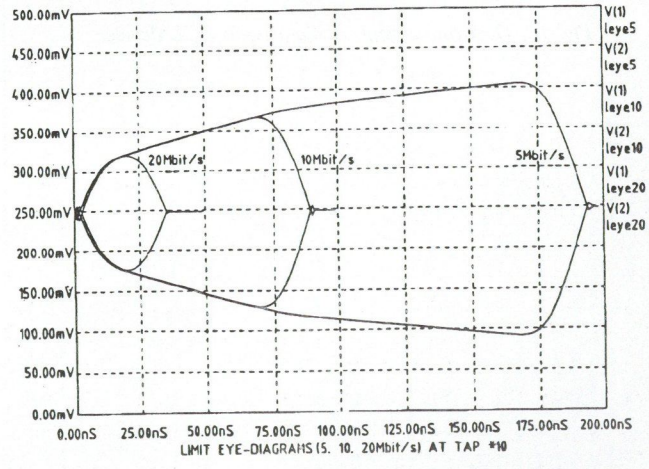


Fig. 8. Eye diagram of the LAN for 5, 10, and 20 Mbit/s data rate.

Besides of digital circuits, SPRINT and SIGHTS are also useful in analogue circuit simulation. Fig. 10 shows an overtone crystal oscillator with an ECL active device.

The startup conditions, the trajectories, phase jitter on the output waveform can be conveniently analysed and evaluated with SPRINT and SIGHTS. Fig. 11 is an example of startup oscillations. At three different nodes, the initial waveforms after power up are shown. The powerful features of SIGHTS can fully be exploited when trajectories are shown (see Fig. 12).

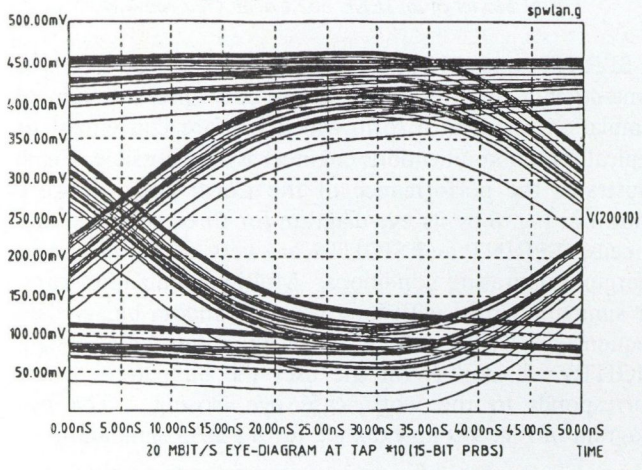


Fig. 9. Performance of the LAN for PRBS input, in the form of eye diagram.

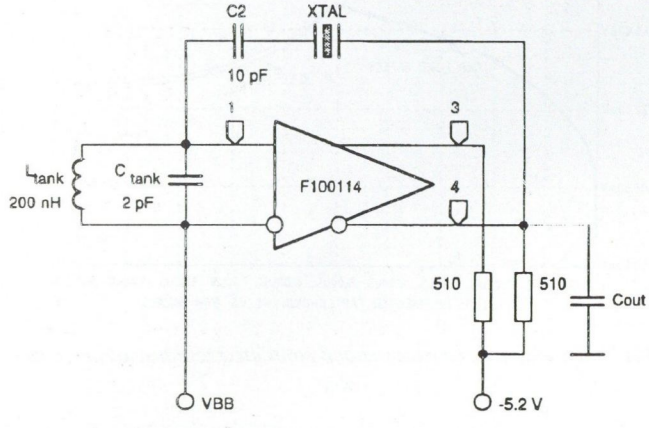


Fig. 10. Overtone crystal oscillator with ECL device.

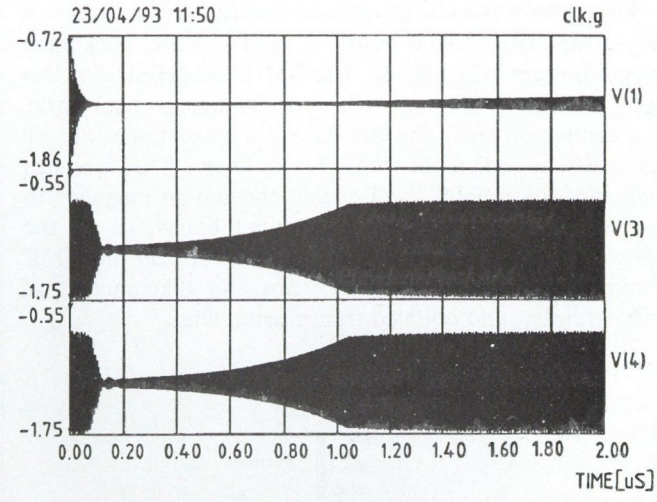


Fig. 11. Power-up waveform of the overtone crystal oscillator at three different nodes.

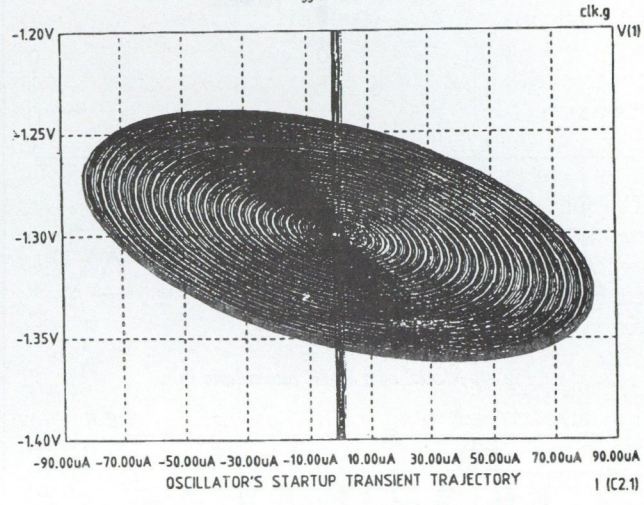


Fig. 12. Oscillation trajectories of the third-order overtone crystal oscillator.

The time-domain waveforms are computed using the convolution method, thus increasing the speed of simulation by orders of magnitudes. Fast simulation is essential in the design of large systems, where the only practical limitation to circuit complexity comes from the memory size.

High Design Technology
and
Hewlett-Packard

TECHNOLOGY EXCHANGE SERVICE

1. INTRODUCTION

There is a growing recognition of the importance of technological innovation in the competitiveness of companies and countries. There is also a need for extending and developing the managerial skills necessary to develop and exploit technological capability to meet the newly emerging challenges. To fulfill the above requirements technology transfer is a process of promoting technical innovation through the transfer of ideas, knowledge, devices and artefacts from leading companies, R& D organisations and academic research to more general application in industry and commerce. J. Budinszky in his paper entitled "Innovation and Small Enterprises" (Journal on Commcations, January 1993) gave a detailed survey of the importance of the subject.

All the industrialised countries provide financial incentiveness to encourage technology transfer. Different EC programmes (EUREKA, SPRINT, STRIDE, Third Framework Programme) deal with innovation and technology transfer, improving competiveness of European products by developing transfer, strengthening research facilities.

The activity of technology transfer intermedieres have frequently been dominated by either "technology push" or "demand pull" approaches. In Hungary the Technology Exchange Service founded by the National Committee for Technological Development (OMFB) has started its activity with pushing the new R& D results, where the research was supported by OMFB itself. The scope of the service includes promotion of other R& D results too. The balance between "technology push" and "demand pull" is realised by a good contact with organisations for innovation of SMEs (small and medium sized enterprises). The service uses all means of public relations: publicity, improvement of image of Hungarian technological results, consulting, organising exhibitions, building connections to national and international programmes and organisations. At present the costs of the service are covered by OMFB, in the future, however, royalty will be the source to finance this activity, similarly to the usual practice abroad.

For international networks the office uses the HUNTECH database, where short description of the specific topics as well as all data necessary for the interested partners are available. The topics include: agriculture and food industry, basic research, biotechnology, chemistry and new materials, construction industry, electronics and electrotechnics, environmental engineering, informatics, instrumentation and measurement technics, mechanical and production engineering, pharmacology and medical applications and power engineering.

Here we are pleased to introduce three interesting R& D results of electronics form the offer of the Service.

2. DUAL BEAM OPTICAL COMPARISON SPECTROMETER

Optical spectroscopy is one of the most sensitive methods of material testing. The comparison spectrometer developed at the Research Institute for Solid State Physics of HAS (project leader: Z. G. Horváth) is a special simple version of the "imaging spectrometers", using two independent light fiber inputs and a computer controlled CCD detector matrix. Evaluation is performed by calculating the ratio of two detected spectra. The result is independent of the actual optical transmission path.

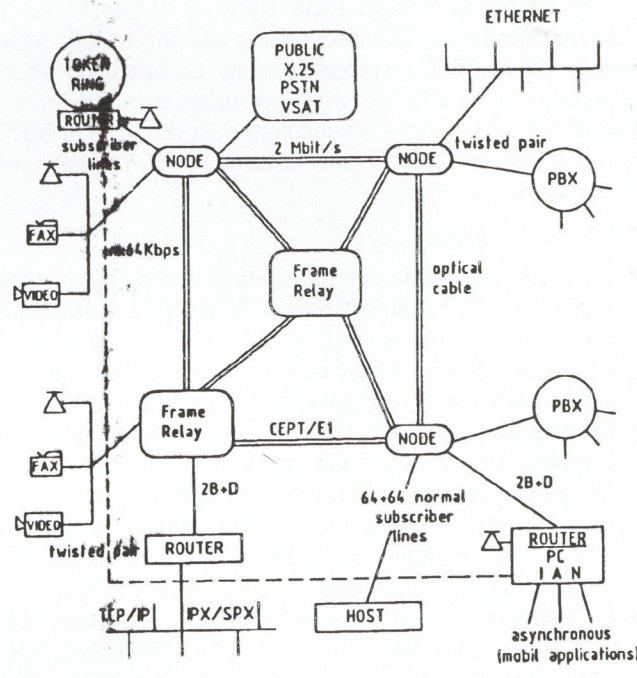


Fig. 1. MSDN system model

A PC card digitalises and stores in real time the whole measured emission, absorption or reflection optical spectra of a target and an etalon. The light is detected by a computer CCD camera at the output of the dual beam spectrometer. The software controls the exposition time, displays the spectra and the result of a sum or ratio of the two spectra with freely selectable parameters and finally determines the standard deviation of the result of calculation. The value of the standard deviation can be used as threshold of a given alarm level in pass/stop production line control applications.

Fields of application:

In *emission mode*: comparison of spectra of LEDs, laser diodes, light sources, gas discharges, TV and other screens, optical radiation of plasmas, fluorescence (foods, chemical and medical products, contaminations, art products, banknotes, links, stamples, special documents etc.) measurements (needs special illumination).

In *reflection mode*: colour analysis and control of illuminated painted industrial products, chemical , medical and agricultural quality control, optical mirror testing and the above fluorescence examples, in its original colours.

In *transmission mode* (dual line head with halogen light source is needed as option): control of optical filters and other coating (sunglases, TV or PC screen filters) lightguides, transparent objects (solids, like: different home car — or industrial — windows, crystals; gases or liquids: having absorption in the given optical range like sulphur, simple smoke solvents, and other chemical, biological or medical samples).

3. DATA COMMUNICATION SYSTEM BETWEEN PC AND LAN NETWORK

The Managed Service Data Network (MSDN) developed at Computing, Communication and Innovation Technology Corp. (SzKI), (project leader: G. Leporisz) offers a solution for establishing an "enterprise, corporate-wide" integrated data communication system in the form of PBX, and/or a Private Virtual Network to be built on digital back bone network is Hungary.

MSDN offers opportunities:

- for establishing data communication connecting of type POINT-POINT, and POINT-MULTIPOINT with bandwidth on demand, as well as with a data speed of 64 Kbps–384 Kbps, which can be produced on existing subscriber circuits (normal telephone cable),
- for establishing closed "private" data network inside a company or between companies with channel and/or fast packet switching operation of the network.
- for connecting complex computing centers, LANs, as well as independent PCs to the above system, and for establishing interconnections among these elements,
- for building up "extra" services inside the private network e.g., FAXING, MAILING, DOCUMENT, ARCHIVING etc.,
- for office automation at medium and large companies.

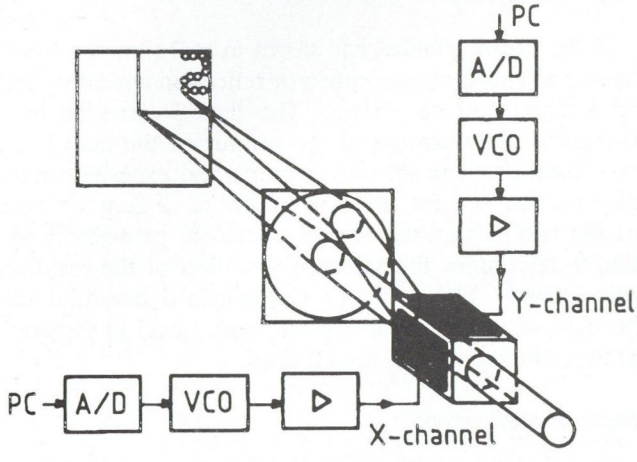


Fig. 2. XY-deflector system

Typical application examples:

- Integration of HOSTs into a nation-wide system.
- Connection of PC workplaces into a network.
- LAN-LAN extension among work sites.
- Multimedia and Video Conference.

At present a model system is operating in SzKI. By the end of 1993 a system for pilot service will be installed at Hungarian Telecommunications Company Ltd (MATÁV).

4. TWO-DIMENSIONAL (XY-) DEFLECTOR

The deflector system operates using Bragg diffraction of the incident light beam. This is induced by the refraction index grating coursed by a travelling acoustic wavefront.

The intensity of the diffracted light is dependent upon the power of the acoustic beam which is proportional to the input electric power, the angle of diffraction is determined by the grating constant that is proportional to the acoustic frequency. Therefore the light beam can be scanned by changing the frequency. This is the basis of the operation of acousto-optic deflectors.

A two-dimensioned acousto-optical deflector system has been developed by at the Department of Atomic Physics, TU Budapest (project leader: L. Jakab)

The xy deflector has the following specifications:

Optical wavelenght	633 nm
Center frequency	47 MHz±15 MHz
Active aperture	10×10 mm
Random access time	16 μs
Resolution	100×100 dots
Diffraction efficiency	50 %
Passive optical loss	5 % (with antireflection coating)
Maximum input electric power	1 W/channel
Interaction medium	Tellurium dioxide
Transducer	Lithium noibate
Frequency range	40–50 MHz
VCO tuning range	0–5 V (analog tuning)
Output power	1 W

Applications:

- Beam scanning
- Image processing
- Optical crossbar switching
- Optical matrix manipulation and operations (multiplication, transponation etc.)

A successful application has been accomplished in control of metal working machine tools at the laboratories of Thomson CSF, France.

KLÁRA R. SÁRKÖZY
Technology Exchange Service
OMFB

COST ACTIVITIES IN TELECOMMUNICATIONS

COST stands for 'Cooperation européenne dans le domaine de la recherche scientifiques et technique' (European Cooperation in the Field of Scientific and Technical Research).

COST is a framework and forum for technical and scientific cooperation among twenty five European countries joining common actions. It was established in 1971 by a Ministerial Conference. Their aims are to strengthen European industrial and scientific competitivness through cross-border collaborative research projects.

The main characteristics of COST operation is its voluntary, not-for-profit pre-competitive basis. The participants of the research join on their own decision and no special support is done for this type of activity. The COST is not an EC organization, however it works with the agreement of and the administrative support is offered by EC.

COST activities include the following areas of research: informatics, telecommunications, transport, oceanography, materials, environment, meteorology, agriculture, food technology, social sciences, medical research, civil engineering, forestry and chemistry. In the 1971–1991 period about 160 COST project have been successfully launched.

In the first twenty years the telecommunication sector proved to be one of the most active COST fields. About 50 projects, nearly one third of the whole COST activity were launched in this area. They cover the whole spectrum of telecommunications, emphasizing the following main topics:

- radiocommunications, radiosystems, antennas and propagation;
- optical communications, optical systems;
- signal processing, encoding of video signals;
- telecommunications networks, multiservice networks;
- human factors in telecommunications, services for disabled;
- telecommunication security.

The basic principles of COST cooperation are:

- participation on a voluntary basis,
- contribution based on gentlemen's agreement (no legal obligations),
- funding provided by participating countries (except if a team works on contract basis for a participant),
- moderate financial involvement, large participation of scientists,
- three to five years duration projects,
- participants mainly from government agencies PTTS, universities and industry,
- all the participant countries and EC may initiate research projects.

The organization of COST is the following:

The *Ministerial Conference* — of the members countries' ministers responsible for R& D is the highest body, funding the policy of COST. The hungarian member of the Ministerial Council is Prof. E. Pungor, Minister the President of the National Board for Technical Development (NBTD).

After 20 years of the funding Ministerial Conference a second one was held in November 1991 in Vienna. This conference enlarged the COST family accepting the application of Czechoslovakia, Hungary, Iceland and Poland. The member countries of COST now (April 1993):

- The 12 EC countries: Belgium, Denmark, France, Germany, Greece, Italy, Ireland, Luxemburg, The Netherlands, Portugal, Spain, United Kingdom.
- The 6 EFTA countries: Austria, Finland, Iceland, Norway, Sweden, Switzerland.
- Other European countries: Croatia, Czech Republic, Hungary, Poland, Slovak Republic, Slovenia, Turkey.

Each of these countries enjoys the same rights and privileges within COST. Participation in COST projects is also possible from European Countries which are not members, but only from individual entitles, organizations and institutes and not on a country basis as is the case of member countries.

The central body of COST is the *Committee of Senior Officials* (CSO). One of the representatives of CSO from each member country is the *National COST Coordinator* (NCC). The CSO formulates the general strategy of COST, appoints the Technical Committees, decides on their tenus of Preference, approves the proposed research projects and prepares the Memorandum of Understanding (MOU) for the projects.

The responsibility of NCC is mostly the connection between the researchers, institutions of the country and the CSD and COST secretariat in Brussels. The Hungarian NCC is P. Konc with the NBTD.

The COST committee dealing with legal, administrative and financial matters is *JAF* (Justice–Administration–Finance).

The *Technical Committee* (TC) is responsible for the activity on a larger research area. In the field of telecommunications the relevant body is the Technical Committee of Telecommunications (TCT). The TC-s are formed by members of the representatives of the member countries on the given area and appointed by CSO for a period of one to three years. The membership may be prolonged. They examine the proposals for the new projects to be forwarded to the CSO, control the implementation and the coordination of the running projects and evaluate the results achieved. The representative of Hungary in TCT is the author of this paper.

Each project is run by a *Management Committee* (MC) formed from national delegates of countries signing the relevant MOU. At least five signatory countries are requested to launch a project. The MC selects the research work, plans the details, exchanges the information among the participants, coordinates with other projects within and outside COST, discusses the project extension and precues the annual final reports.

The *Secretariat* is supplied by EC.

The operational projects of TCT with at least one more year duration are the following the deadline and the active Hungarian participation or interest is noted.

219 Future Telecommunications and Teleinformatics Facilities for Disabled People (Sept. 1993. H: Technical University of Budapest, BME)

220 Integrated Space /Terrestrial Networks (Feb. 1994. H: Central Physical Research Institute, KFKI)

227 Integrated Space/Terrestrial Mobile Networks (Apr. 1995. H: Post Office Research Institute, PKI)

228 Simulation for Satellite/Terrestrial Networks (Jan. 1996.)

229 Applications of Digital Signal Processing (Apr. 1994. H: BME)

230 Stereoscopic TV-Technology and Signal Processing (Apr. 1996.)

232 Speech Recognition over the Telephone Line

233 Prosodics of Synthetic Speech (Feb. 1996.)

335 Radio Propagation Effect on Next Generation Fixed-Service Terrestrial Telecommunication Systems (Oct. 1995. H: BME)

211 TER Redundancy Reduction Techniques for Coding of Video Signals in Multimedia Services (Oct. 1995)

237 Multimedia Telecommunication Services (Feb. 1997.)

238 PRIME: New Predictions and Retrospective Ionospheric Modelling over Europe (March 1995.)

230 Ultra High Capacity Optical Transmission Networks (Jan. 1997.)

240 Techniques for Modelling and Measuring Advanced Photonic Telecommunication Components (Apr. 1996. H: BME)

241 Characterization of Advanced Optical Fibres for the New Photonics Network (Jan. 1997. H: PKI)

242 Methods for performance evaluation and design of multiservice broad band networks (Apr. 1996. H: BME)

243 Electromagnetic Compatibility in Electrical and Electronic Apparatus and Systems (Sept. 1995. H: BME)

244 Biomedical Effects of E. M. Radiation (Sept. 1996. H: F. Joliot Curie Radiobiology Institute)

245 Active Phased Arrays and Array Fed Antennas

246 Material Reliability of Passive Optical Components and Fibre-Amplifiers in Telecommunication

Proposed projects to be approved by CSO:

247 Verification and Validation Methods for Formal Description (H: KFKI)

248 The Future European Telecommunication User

Further new proposals will be discussed in the next TCT meeting held at Budapest in June 1993.

All the projects listed above are open for new participants. The institutions interested should contact to the Hungarian representative of TCT.

The author whishes to express his appreciation for the help of Messrs. P. Koncz and K. Fazekas.

LÁSZLÓ ZOMBORY
Technical University of Budapest

REFERENCES

[1] J. M. Dwyer: COST Telecommunications Activities. EMC Symposium Wroclaw, Poland, September 1992.

[2] M. Monteiro, P. Costigan: The COST Telecommuncations Programme. 4th IFIP Conference, Liege, Belgium, December 1992.

[3] F. Fedi: Proceedings of the Symposium on New Frontiers for the European COST in Telecommunications, Rome, Italy, October 1992.

[4] F. Fedi: Foreward *ibid.*

[5] J. M. Dwyer: Major Achievements of 20 years of European Co-Operation in COST Telecommunications. *ibid.*

[6] G. Lajtha, G. Sallai: Perspective of the Opportunities of Development in Hungary: Scientific and Technological Aspects in Telecommunications *ibid.*

AN INTERESTING CENTENARY: THE "TELEPHONE JOURNAL"*

In 1893, on the 15th of February, the first "broadcasting"-service of the world was started in Budapest, Hungary. With the help of the telephone network and telephone headphones, a 14 hours long, edited program was transmitted to the subscribers. The program consisted of news, exchange news, entertaining music, as well as transmissions from theathers and concert halls. The paper includes details about the technical solutions, about the program, and about the inventor, Mr. Theodor Puskás.

For most people it seems to be an evidency that the first broadcasting system was the sound-radio introduced in the twenties. In fact however this is not true. The first broadcasting system precede the radio by a quarter of a century.

The sound-radio is namely the amalgamation of two brilliant ideas. One of them is the invention that the electromagnetic waves can be used as a carrier of messages. The transmission of messages by means of electrical current was already well known in the time of starting the radio as Edison's telegraph and Bell's telephone worked on this basis. The other essential basis of broadcasting is the ingenious idea that the electromagnetic waves — or other carriers — could transmit also edited programmes consisting of news, informations, entertaining and artistic performances etc. instead of simple messages. It means that every broadcast-system is composed of two parts: the hardware which is a complex of equipments of the delivery-system and the software which is called in this case radio-(or television)-programme. In principle any system which is capable to transfer programmes from one point to many other points can be used as hardware. If these reception-points are connected to a physical network we speak about a cable-system, if the programme can be received anywhere without physical connection to the source we speak about radio broadcasting.

The Hungarian Theodor Puskás was the first who recognized that the already existing telephone network can be used for programme-distribution. Hence he found the hardware, adapted it for this purpose, created an appropriate programme, the "software", and synthetizing them established the world's first broadcasting system. The new service got the name "Telephone-Journal" (or Telephone Courier, in Hungarian "Telefonhírmondó") and it started its operation on the 15th of February 1893 in Budapest. At the beginning the Telephone-Journal had 60 subscribers. This number increased in 1894 to 700, and in 1899 there were already more than 7000 subscribers.

In the following years this figure fluctuated between five and seven thousand. The maximum number was nearly 10.000. The system was in use until the end of the Second World War. From the end of 1925 it was operated together with the newly established Hungarian Radio as "Hungarian Telephone-Journal and Radio Company", and its programme was in bigger part identical with the radio-programme. In the final period the Telephone-Journal was used especially in hospitals, asylums etc. It is worth to mention that even after the start of the sound-radio service many subscribers sticked to the Telephone-Journal because its reception was more stable and of better quality than that of the radio with the very early crystal-receivers.

At the beginning Puskás used the normal telephone-lines and phone-sets for the transmission of the Telephone-Journal programme. In this system the subscriber had to ask the operator of the telephone exchange to connect him to the studio-center of the Telephone-Journal. The disconnection happened in the same way.

But soon the building up of a separate network for the Telephone-Journal was begun. The wires were fastened on isolators mounted on steel frames on the roofs of the houses. The wires were conducted to the flats from the facades through the walls and they ended in headphone sets.

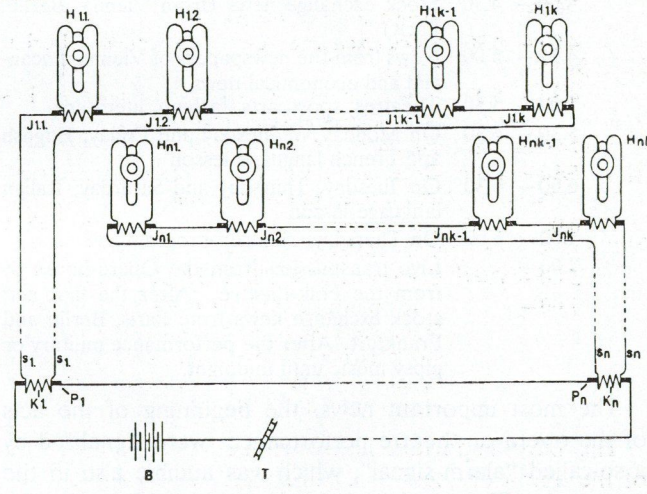


Fig. 1. Original circuit diagram of the Telephone-Journal

Fig. 1. shows the original circuit diagram of the system. The carbon type microphone (T) was connected in series with the battery (B) and with the primary windings of a number of transformers $(P_1 \ldots P_n)$. The secondary winding of each transformer $(K_1 \ldots K_n)$ was connected to a loop consisting of about 200 primary windings of other transformers placed near to the subscribers $(I_{11} \ldots I_{1k}, I_{n1} \ldots I_{nk})$. The secondary windings of these latter transformers supplied the headsets of the subscribers $(H_{11} \ldots H_{1k}, H_{n1} \ldots H_{nk})$. The advantage of the system was that only one transformer in the center and one wire-pair were necessary for about 200 subscribers living in the same direction. We must not forget that at that

*This paper was presented originally on the 94th AES Convention in Berlin on 18 March 1993.

time no amplifiers existed and the audio frequency signal induced by the microphone current alone had to supply all the five-ten thousand headphones. Therefore the speakers had to speak very loud. This was an exhausting task and for this reason the speakers were changed every 15 minutes.

Let us see the program of the Telephone-Journal in one day of the year 1897. It was called that time "Order of the day".

AM.	9.30 – 10.00	Programme-information, informations from Vienna and abroad, telegrammes received during the night announcements of the official gazettes
	10.00 – 10.30	Stock exchange-news
	10.30 – 11.00	Press-review, interesting news, telegrammes
	11.00 – 11.15	Stock exchange news
	11.15 – 11.30	Local and theatre-news, sport
	11.45 – 12.00	Stock exchange news
	11.45 – 12.00	News form the House of Parliament, news from the provinces and from abroad
	12.00 – 12.30	News form the House of Parliament, from the royal court, political and military news
PM.	12.30 – 1.30	Stock exchange news (from Vienna and Berlin)
	1.30 – 2.30	News from the House of Parliament, announcements of the capital-community, telegrammes
	2.30 – 3.30	News from the House of Parliament, local news, telegrammes
	3.30 – 3.15	Stock exchange news
	3.15 – 3.30	Reading of interesting articles from newspapers
	3.30 – 4.00	News from the House fo Parliament, news form the Court of law
	4.00 – 4.30	Stock exchange news (from Vienna, Berlin, Paris)
	4.30 – 5.00	News from the newspapers of Vienna, fincancial and economical news
	5.00 – 5.30	Theatres, sport, arts, fashion, literature
	5.30 – 6.30	On Monday Wednesday and Friday; English and French language lesson
	6.00 – 6.30	On Tuesday, Thursday and Saturday; Italian language lesson
	5.00 – 6.00	On Thursday: children-programme
	7.00 –	Live transmission from the Opera-house or from the Folk-theatre. After the first act: stock exchange news from Paris, Berlin and Frankfurt. After the performance military or gipsy music until midnight.

The most important news, the beginning of the acts of the opera or theatre performances were signalized by a so-called "alarm-signal", which was audible also in the next room. The "Order of the day" was published in the "Bulletin of the Telephone-Journal".

And now some words about the inventor Mr. Theodor Puskás. He was a typical representative of his romantical century. He was born in 1844 in Budapest, Hungary and he had an eventful and adventurous life. He studied in the famous Theresianum in Vienna and in the Technical University of Vienna. But he couldn't finish his studies because of financial difficulties due to the death of his father. Since he could speak fluent English, French and German he went to London where he worked at a company. But soon the English Waring Brothers Company looked for Hungarian speaking engineers because they obtained a concession for railway building in the eastern part of Hungary. Puskás applied for the job and he has undertaken the task. This way he returned to Hungary. After having finished this work, he went back to Vienna where he opened a Travel Agency which prospered well during the World Exposition in 1873 in Vienna. Having earned here a bigger amount of money he crossed to the USA. That was the time of the "gold fever" in Colorado and Puskás tried also his fortune. But in this case he wasn't successful. Therefore he sold his mines and came back to Europe. But soon he was again in America because he wanted to study the recently invented telephone system. He thought that this system has fine prospects. He got acquainted with Edison and he became his collaborator. Later he went to Paris and was the European representative of Edison. According to the tone of the letters from Edison there was a very good and friendly relation between them. (After the death of Puskás his widow got a photograph of T. A. Edison with the following remark: "Theodor Puskás was the first man in the world to suggest the Central station for the telephone.")

In 1880 Puskás encouraged his younger brother Ferenc Puskás to establish a telephone network in Budapest. With his help the first telephone central was inaugurated on the first of May 1881, which was the first one in the whole Austro-Hungarian Monarchy. But the younger Puskás died in 1884 and Theodor came back immediately to Budapest to continue the work of his brother. He worked hard as he had to solve a lot of administrative, financial and technical problems. (For instance to get rid of the disturbances in the telephone-calls caused by the in the meantine started tramway traffic service.) In 1892 he patented the Telephone-Journal in the Monarchy and in some other countries. In 1893 he got the concession to start the service on the wires of the existing telephone network. The troubles and the hard work have weakened his health and he died suddenly on the 17th of March 1893, one month after the opening of the Telephone-Journal. He was 49 years old. David L. Woods wrote in his article [1] in 1969: "The real pioneering genius who created these concepts of modern radio (and television) programming was Théodor Puskás, who established the first operational broadcasting organization in 1893."

In the Broadcasting House of the Hungarian Radio there is a marble tablet in the memory of Theodor Puskás since the 50. years anniversary. The centenary was celebrated on the 15th of February this year by the successors of the Telephone-Journal: the Hungarian Radio, the Company Antenna Hungaria (the company operating the transmitters) and the Hungarian Telecom Company.

Last but not least let me mention the name of my collegue Mr. Miklós Szabó, a most enthusiastic collector of the historical documents and artifacts of the Telephone-Journal and those of Hungarian Radio, who died suddenly last year. Should this paper remind us also of his person.

[1] David L. Woods: The "first" Broadcasting Station. *Journal Broadcasting* 1967.

GÁBOR HECKENAST
Magyar Rádió (Hungarian Radio)
Budapest, Hungary

INTERNATIONAL CONFERENCE ON THE DEVELOPMENT AND LIBERALIZATION OF TELECOMMUNICATIONS IN EASTERN EUROPE AND THE FORMER SOVIET UNION —

The conference was held on 27th and 28th April, 1993 in Budapest under the auspices of the London based Adam Smith Institute, one of the world's leading policy think-thanks. The Institute is involved in various conferences on the development of the economies of the Eastern European countries and the republics of the Former Soviet Union.

This conference has been organized with the purpose of promoting the development and liberalization of telecommunications in the countries of Eastern Europe and the Commonwealth of Independent States.

The Adam Smith Institute has invited an outstanding panel of 40 top speakers from West and East including Telecommunications Ministers from Republic of Hungary, Czech Republic, Russian Federation, Republic of Poland, the Ukraine, and senior representatives of the European Commission, Deutsche Bundespost Telekom, Eutelsat, Eurotel project, EBRD, Bankers Trust, Nokia Telecommunications, Cable and Wireless, GPI, Sprint International, Ericsson Technika, Alcatel, Ameritech International.

The first session of the conference discussed the state of play in Europe: privatisation, deregulation and globalisation of telecommunications markets — an update on recent changes in Western Europe and their implications for Eastern Europe and the FSU. Ungerer, Head of Regulatory Affairs Commission of the European Communities outlined in his talk the deregulation process of telecommunications which was set out by the EC Green Paper in 1987. The basic principle of reform has been liberalization. In 1988 an EC Directive was issued opening up the community market for terminal equipment. In July, 1990 the EC Services Directive introduced competition for value added services. In the field of public voice telephony service the choice between monopoly and competition was maintained. Second major issue has been the separation of regulatory and operational functions. Thirdly, the establishment of Pan-European functions, as equipment approval, network access conditions have also a significant effect on telecommunications development. According to Ungerer, the market of telecommunications services in the EC could quadruple by the year 2010 from US$ 100 billion to about US$ 425 billion. To support this market an investment of US$ 550 is necessary. In Central and Eastern Europe and the countries of the CIS, to catch up with the West an estimated US$ 200 billion is needed. The investment requirement is high. The European Community, the EBRD and the World Bank has addressed the problem by providing funding for a series of targeted programmes. Revenue growth and a strong capital base are on the top of the development agenda. None of them can be done without further liberalization.

Bölcskei, Director General of Telecommunications Undersecretariat of State, Hungary outlined the results in separating the proprietory and regulatory functions in telecommunications. Expressed the view of the government that the provision of public telephone services is a basic commodity of which forms an obligation for the government. The necessary development cannot be supported from revenues or the state budget, therefore external sources of funding are inevitably required.

Chrudina, Minister of Telecommunications presented the first steps taken towards separation of the regulatory and operational functions in telecommunications of the Czech Republic.

The topics in the next session were the developments in specific infrastructure areas, including satellite, mobile and fixed networks. An overview of developments in telecommunications markets within Eastern Europe and the FSU during the past four years was given.

Significant date were given by Jonscher from Central Europe Trust Company. He has shown that basic networks are very underdeveloped by Western standards. However, the demand for service does not lag as far behind the West as does supply. Revenues from new telephone lines in the region can be excepted to reach over US$ 30 billion in the coming years (Table 1).

Table 1.

Country	Population (million)	Current Number of Exchange Lines (000's)	New Exchange Lines Required (000's)	Cost of new Exchange Lines△ (USD Billion)	Annual Revenue* (USD Billion)
CSFR	15.6	2,300	2,380	3.57	2.34
Hungary	10.6	918	2,260	3.39	1.59
Poland	38.0	3,280	8,120	12.18	5.70
Russia	148.0	19,460	17,540	26.31	18.5
Ukraine	51.8	7,710	5,240	7,86	6.48
Bylerus	10.3	1,730	845	1.27	1.29
Bulgaria	9.0	1,994	256	0.38	1.13
Romania	22.9	2,160	3,565	5.35	2.86

△ Cost to install an exchange line USD 1500

* Assume USD 500 / exchange lines

In Session 3 emerging Eastern telecommunications markets — risks and rewards — timetables for return on investment were discussed. Case studies on Russian and Romanian development programs have been presented by Pozshitkov, Deputy Minister of Posts and Telecommunications, Russian Federation and Stefanescu, Director General of Planning and Strategy, Ministry of Communications, Romania. Western experts from EBRD, Coopers and Lybrand, International Finance Corporation and Bankers Trust Company discussed the options for financing the modernisation of telecommunications infrastructure in Eastern Europe.

In Session 4 the topic was establishing the operating framework for telecommunications services and equipment in Eastern Europe and the FSU. The paper of Depczinsky, the Polish Telecommunications Minister confronted the problem of creating a successful telecommunications structure. He discussed the legal aspects of deregulation and presented data on the development projects. The financial sources were indicated and their relative importance were evaluated.

As a guest speaker at a luncheon Schamschula, Minister of Telecommunications, Hungary reviewed three fields of telecommunications reform: technical developments, institutional modernization, new legislature. Technical development can be characterized by more than 15 % increase in the number of main lines in 1992. The policy of the government gives preference to the concessional competition to provide suitable telecommunications in the non-profitable regions also. Privatization will be carried out in the next year, 49 % of the shares will be sold, about 30 % to foreign companies providing both financial and professional support.

In Session 5 experiences of joint venture partners were discussed. Alahuhta, President of Nokia Telecommunications stressed the fundamental importance that the parties share the same, realistic expectations for the joint venture. It is equally important that the parties agree on division of responsibility.

Wilson, from the Ameritech reported on the details of a joint venture for building and operating a cellular system in Poland with France Telecom and Polish PTT. Two Hungarian managers Sugar from WESTEL and Fodor from Ericsson Technika gave informations on the experiences of two newly organized joint venture companies. Radzikowski, Managing Director of Sprint Networks, Russia related about experiences in a successful joint venture in data communication which became profitable after 12 months. □

Information for authors

JOURNAL ON COMMUNICATIONS is published monthly, alternately in English and Hungarian. In each issue a significant topic is covered by selected comprehensive papers.

Other contributions may be included in the following sections:

- INDIVIDUAL PAPERS for contributions outside the focus of the issue,
- PRODUCTS-SERVICES for papers on manufactured devices, equipments and software products,
- BUSINESS-RESEARCH-EDUCATION for contributions dealing with economic relations, research and development trends and engineering education,
- NEWS-EVENTS for reports on events related to electronics and communications,
- VIEWS-OPINIONS for comments expressed by readers of the journal.

Manuscripts should be submitted in two copies to the Editor in chief (see inside front cover). Papers should have a length of up to 30 double-spaced typewritten pages (counting each figure as one page). Each paper must include a 100—200 word abstract at the head of the manuscript. Papers should be accompanied by brief biographies and clear, glossy photographs of the authors.

Contributions for the PRODUCTS-SERVICES and BUSINESS-RESEARCH-EDUCATION sections should be limited to 16 double-spaced typewritten pages.

Original illustrations should be submitted along the manuscript. All line drawings should be prepared on a white background in black ink. Lettering on drawings should be large enough to be readily legible when the drawing is reduced to one- or two-column width. On figures capital lettering should be used. Photographs should be used sparingly. All photographs must be glossy prints. Figure captions should be typed on a separate sheet.

INTERNATIONAL SEMINAR ON TELECOMMUNICATIONS SYSTEMS MEASUREMENTS

BUDAPEST, HUNGARY, NOVEMBER 9-10, 1993

Organized by the

SCIENTIFIC SOCIETY FOR TELECOMMUNICATIONS

and the

JOURNAL ON COMMUNICATIONS

The Seminar will give a comprehensive review of measurement problems encountered during installation and operation of modern telecommunications systems and of the measurement methods developed for controlling systems parameters to meet transmission requirements. The speakers will represent Hungarian companies and institutions and outstanding Western system measurement centers having significant relations in Hungary.

Seminar topics include measurement of

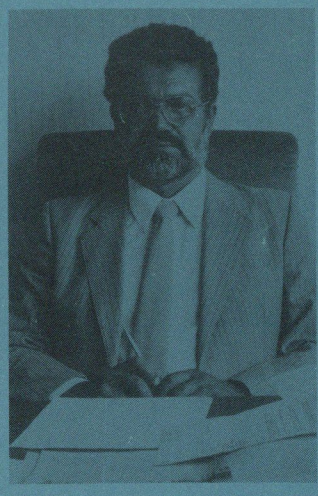
- Digital transmission systems
- Synchron digital hierarchies
- Telecommunication protocols
- Digital exchanges
- Digital networks
- Microwave systems
- Optical transducers

A round table discussion will be organized on new measurement methods in developing telecommunications sytems.
Measuring instruments will be presented at an exhibition organized concurrently with the seminar.
Language of the seminar will be English with simoultaneous translation.
Seminar lectures will be published in the October 1993 issue of the Journal on Communications, seminar participants will receive copies of the journal at the registration desk.

Further information

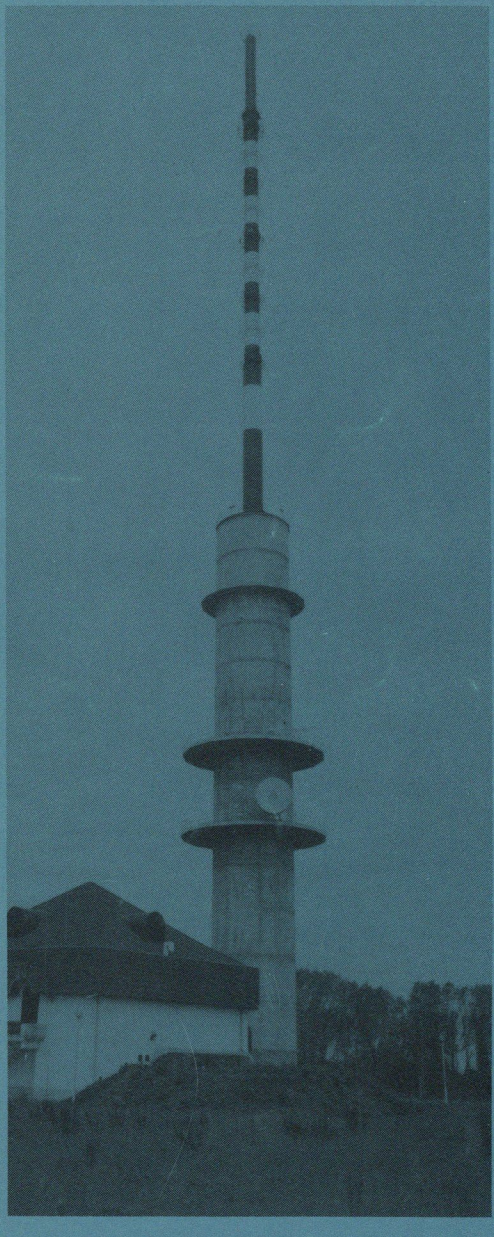
Scientific Society for Telecommunications
Kossuth Lajos tér 6-8.
1055 Budapest, Hungary

Ms Katalin Mitók, organizing secretary
Phone: 36 1 153 1027
Fax: 36 1 153 0451



antenna hungária

Hungarian Radiocommunications Corporation



The Antenna Hungária PLC has been in the service of wireless telecommmunication in Hungary for almost seventy years. Our former employees, who had worked for Antenna Hungária's legal predecessor, made a major contribution to the launching of Hungarian radio broadcasting, and also played an important role in founding the broadcasting of Hungarian Television in the early Fifties.

We consider our principal task to get to know and to apply the domestic and international scientific achievements of radio communication, to introduce every new of broadcasting service that contributes to the progress of the nation's culture and economy.
Our latest venerable challenge was to conduct the first successful test in satellite television broadcasting.

Antenna Hungária is engaged in the following principal activities and services:

- the broadcasting of radio and television programmes
- relaying/transmission of programme signals both domestically and internationally
- conducting of on-the-spot television broadcasts
- operating of main telephone network connections
- data transmission and radiogram services
- telecommunication planning services & telecommunication research, experimental development
- other services
- various industrial and auxiliary activities

We have forged close professional ties with leading companies both in Hungary and abroad, and have embarked on a number of joint ventures. One result of joint enterprise is a national person finder network, and we are also planning to operate a land surface data transmitting network, as well as VSAT and cellular mobile radio telephone services.

In recognition of our achievements, the Antenna Hungária PLC was one of the companies designated by the Hungarian Government to organise and service the Europa Telecom 92 exhibition in 1992.

The year 1993 might bring significant changes in the life of our company due to the possible privatisation process. The management hopes that through the privatisation of the company new domestic and foreign financial funds will arise giving the background of further dynamic development of the firm's services. By realising our plans we can strengthen our position in the market of radio and television broadcasting and other wireless services.

The second stage in 1993 is a multi-phase share issue, which will attract sizeable foreign interest according to our expectations. To gain access to and to utilise high tech and international expertise we have participated in the establishment of several Joint Ventures and we tend to continue these activities.

Public trading of Antenna Hungária shares is expected to commence by 1995.

Should you need further details please do not hesitate to contact us at:

Antenna Hungária PLC
Budapest, VIII., Trefort utca 2.
H-1088 Hungary
Tel.: 36/1/118-1233
Fax: 36/1/138-4008