INTERNATIONAL JOURNAL OF CIRCUIT THEORY AND APPLICATIONS
Int. J. Circ. Theor. Appl., **27**, 25–41 (1999)

CNNUM STEREO ARCHITECTURE AND 3D TEMPLATE DESIGN TECHNIQUES

ANDRÁS G. RADVÁNYI[1]* LÁSZLÓ GÁSPÁR[2] AND GÉZA TÓTH[3]

[1] *Analogic and Neural Computing Systems Laboratory, Computer and Automation Institute of the Hungarian Academy of Sciences, H-1518 Budapest, POB 63, Hungary*
[2] *Ericsson KFT, H-1108 Budapest, Venyige u.3, Hungary*
[3] *University of Notre Dame, Department of Electrical Engineering, Notre Dame, IN 46556, U.S.A.*

SUMMARY

A CNN Universal Machine-based structure and analogic algorithm is proposed to extract 3D spatial information from stereo images. The CNN template values are obtained in two ways: from simplex optimization and from the energy function. The method is demonstrated on densely textured as well as grey scale stereograms. Copyright © 1999 John Wiley & Sons Ltd.

KEY WORDS: stereo depth detection; cellular neural networks

1. INTRODUCTION

The cellular neural networks (CNN) introduced in 1988[1,3] opened a new way in array computing; they have proved to be especially powerful in various image processing tasks. The application potential of CNNs in stereo image processing has been discussed in several papers.[12–16] In this one we propose a local neighbourhood multilayer CNN architecture of a specific complex cell and its possible realisation in CNN Universal Machine (CNNUM) framework.[2,4,6,17]

In Section 2, we briefly discuss the stereo depth extraction task and its representation in multilayer CNN. In Section 3.1, a massively coupled multilayer CNN structure is derived from rules borrowed from the computational theory of human stereo vision research. In Section 3.2, we convert the massive interlayer net of couplings obtained, to a local neighbourhood propagation mechanism and explain its 'time-sharing' realisation in a CNNUM structure of complex cells. In Sections 4 and 5, two template design approaches are detailed, the one applying simplex optimization of different goal functions, and the other based on an approximation of CNN energy function.

2. BASICS IN STEREO PROCESSING

Since Wheatstone's discovery of the stereoscope in 1838 it is known that images of a scene from two slightly different horizontal angles can be fused to induce spatial depth perception, consequently, such pairs convey 3D information. Taking snapshots — generally called left and right images of a stereo pair – from only slightly different angles the objects in the scene seem practically identical in the two snapshots. Due to the different position of objects in spatial depth, however, the relative position of their images among the images of other objects are different. The position difference of objects in the two snapshots is called *binocular*

* Correspondence to: A. G. Radványi, Analogic and Neural Computing Systems Laboratory, Computer and Automation Institute of the Hungarian Academy of Sciences, H-1518 Budapest, POB 63, Hungary. Email: radvanyi@sztaki.hu

Contract grant sponsor: OTKA; contract grant number: T012862

CCC 0098–9886/99/010025–17$17.50

Received 15 May 1997
Revised 19 November 1997

disparity. Comparing the snapshots, the relative spatial depth of objects can be extracted; for the larger the binocular disparity is, the closer to the viewer the corresponding object is located in depth.

Using artificial devices, however, we do not intend to recognise objects in a scene, we would rather like to label with spatial depth as much points in the snapshots as possible. Technically speaking, the main task in 3D stereo processing is to establish, without any identification or recognition, a one-to-one correspondence between the majority of points in the left and right images. In establishing a possible correspondence we can use the fact that the local texture or the form of any particular smaller object in the two images are practically identical, and always located on epipolar lines (that are horizontal with a good approximation in our case). Those objects that are located at different distances from the viewer will appear with different *disparity*, i.e. in different relative positions along the epipolar lines, a difference serving as a basis for depth detection placing image segments in the 3D space. For most of the little segments in either image, however, several matching counterparts can be found in the other one, i.e. the solution of the correspondence problem for a local approach is not unique. Depending on the type of 2D images, to decide whether a possible matching is a proper one corresponding to some real 3D feature, or merely a false one of a so-called false target, may pose difficulties. Moreover, nearby objects and edges might occlude 3D regions resulting in areas appearing in either 2D image only. Reliable decision should always be based on global evaluation, for a maximal set of one-to-one matchings can be considered as the reconstruction of the total 3D view except for the occluded regions.

Since the recognition and correspondence of parts in the left and right images are based on *global features*, on the one hand, and our aim is to solve the stereo correspondence problem within the CNN paradigm of basically *local processing capability*, on the other, we presume that the 3D scene and its images show enough texture for local evaluation[14, 15] or, alternatively, we do not intend to determine the spatial location of textureless parts, a task left for further high-level processing as interpolation, recognition or object identification.

The extensive research in human stereo vision in the 1970s, stimulated also by the invention of random-dot stereogram by Julesz[6] in 1960, led to the *computational theory* and some comprehensive structured functional model of stereopsis.[7–11] Due to their structural resemblance, the class of co-operative stereo models are practicable in multilayer CNN implementation.[3] In addition to the structure of facilitation and inhibition mechanism, we can adopt some underlying reasoning and rules as well, in developing a CNN stereo model and determining its parameters.

2.1. Representation in 3D

As always in computer-aided image processing, our images are represented as arrays of pixels. The dimension of pixels are in close connection to image resolution. In stereo representation of a 3D scene, the 2D pixel dimensions together with the exact orientation of left and right images determine the 3D pixelization of the scene. In theory, the 3D pixels are spherical cubes with a depth dimension depending on their distance from the cameras. In practice, when using narrow angle cameras and adapting the concept of parallel projection, the third dimension is pixelised by fronto-parallel planes. Thus, the task for depth detection is to determine the attributes of as much as possible 3D pixel-cubes in those locations where an object can be found. If, for simplicity, we exclude translucent objects, most of 3D pixels are either behind or in front of an object in the scene, rendering the attributes indeterminable in the first group and transparent in the second. In our approach, we represent the investigated portion of the 3D space with a multilayer cell structure, each cell standing for a 3D pixel-cube and each layer for a fronto-parallel plane. In practice, based on some prior knowledge (e.g. darkness histogram[15]) about the 3D arrangement, whole depth layers can be left out from investigation, thus saving resources by reducing the number of required layers. In the investigated layers, the hopefully extensive contiguous patches of cells in active state are to represent the visible surface segments of 3D objects in the scene.

From inherent properties of stereo image pairs discussed above follow, that with the left and right images sliding horizontally on each other, as shown in Figure 1, time-to-time matching segments in the two images will glide exactly to cover each other. By some appropriate *equivalence check* these exact matches can be detected in a series of successive shift positions that can directly be associated with fronto-parallel depth

layers. This way the equivalence operations produce a set of so-called equivalence patterns, binary maps of possible *target* positions for further processing. Figure 1 shows a simple depth extraction example. The 3D location of square □ and cross + are unique, but the location of the two circles ○ are ambiguous. They can be either at positions **A** or at positions **B**; but two of the four are false targets. Figure 1 also shows the two symbolic paths along which the pixels of left and right images 'go' in producing the series of equivalence patterns for the successive depth layers.

In Figure 1, the left and right images play a symmetrical role, going along such symmetrical paths that cannot be well implemented in a multilayer CNN of rectangular cell structure, where the cells of neighbouring layers are above each other. Shifting each layer in Figure 1 by half of the grid to the left with respect to the lower next one as shown in Figure 2, we obtain a 3D rectangular structure, with left pixel path perpendicular to the layers well suited to multilayer CNN.

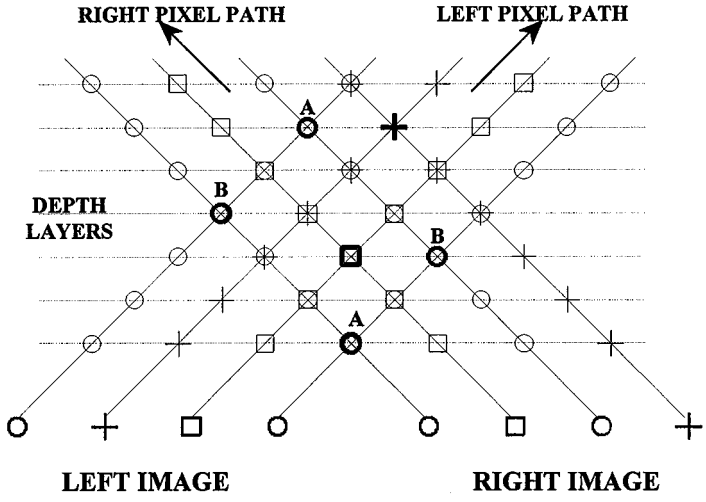

Figure 1. Sliding elements in a stereo-pair

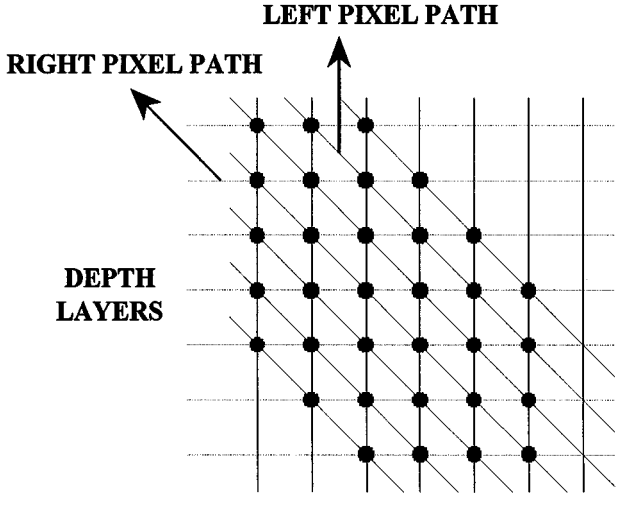

Figure 2. Rectangular transformation of 3D representation

3. DEPTH DETECTION IN MULTILAYER CNN ARCHITECTURE

Owing to the *global* nature of the stereo correspondence problem the amount of possible *false targets* is very high when using *local* equivalence check available in CNN processing. Except for some infrequent, 'luckily' unique cases, in principle, there is no local property in the images to decide on whether a matching is a real one or it corresponds to some false target. Based on common sense assumptions, however, the computational theory of stereo vision gives guidelines, algorithmic rules to meet simultaneously for most of local matches in order to form real objects when put together.[7,8]

3.1. Depth detection rules

One rule, called *continuity*, follows from the observation that everyday objects are usually smooth on most of their surfaces, therefore disparity also changes smoothly. Another rule, called *uniqueness*, is the consequence of the fact that there is a unique correspondence of each image point to some surface point in the 3D scene. There are two algorithmic interpretations of uniqueness; the stronger one is to ensure that neither point in either image should take part in more than one matching. The other and simpler one is to force a single-valued depth map, being valid only for opaque objects.[10] The rules of continuity and uniqueness provide for a sound methodical basis in deriving the neighbourhood relation for a massively coupled multilayer cellular neural structure and the *algorithmic core* implemented by it, to obtain a 3D solution as its equilibrium, for the stereo correspondence problem.[14,15]

The continuity rule is transformed into a single layer *facilitation* template for all layers. The uniqueness rule in its weaker form yields uniform local *inhibition* templates for all pairs of layers. To implement uniqueness in stronger form requires symmetrical pairs of horizontally shifted inhibitions for all pairs of layers along the two pixel paths (see Figure 1), where the amount of shift is the depth difference of the inhibiting and inhibited layers. When we process real images and not computer-generated ones, the pixel-to-pixel correspondence between the left and right images cannot be guaranteed, therefore we will implement the rules in a certain neighbourhood to perform a filtering of noise and inaccuracies of the imaging process. Displaying a cross-section perpendicular to CNN layers, Figure 3(a) shows the inhibiting influence received by a single cell from cells on other layers along the left pixel path. Figure 3(b) shows in 3D the same mechanism for two upper layer cells of a three-layer CNN, displaying it along the left pixel path for the one and along the right for the other.

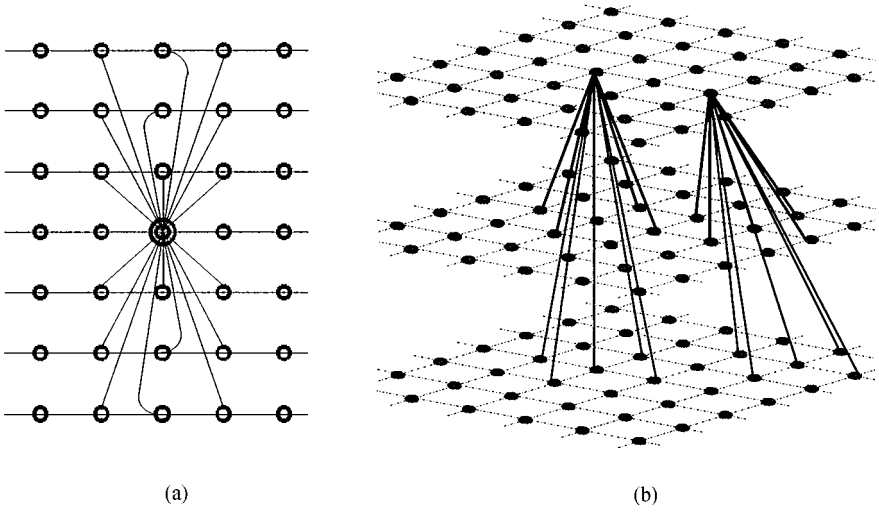

Figure 3. Stucture of interlayer inhibition with one-neighbourhood filtering

3.2. *Interlayer propagation mechanism*

The multilayer CNN implementing the above *core* of rules is supplied with the set of equivalence patterns, one for each layer, as initial state, and having the transient decayed, its steady state would yield the 3D solution. In this CNN structure, however, the feedback connections derived from the uniqueness rule create two massively dense interlayer nets of couplings, the first one belonging to the left pixel paths perpendicular to the CNN plane, and the second one belonging to the right pixel paths slanted to it. Consequently, in this CNN structure the local neighbourhood concept is not preserved concerning the depth of interlayer connections. Analysing the interlayer mechanism of either pixel path we find that each cell receives the sum of inhibitions from cells above it and that of from below as well. Both the upper and lower inhibitions are collected from the local neighbourhood of the pixel path. We can exert the same inhibition if we install two fast propagation lines along each pixel paths for collecting and transmitting inhibitions upwards and downwards. Forming a complex cell, each CNN cell is furnished with four adders, one for each propagation line, to add the inhibiting cell values to the value of the adder in the previous level. The sum of four in-cell adder values will exert cell inhibition locally. Since the neighbourhood of complex cells is limited to the adjacent layers, with the introduction of propagation mechanism of complex cells the local neighbourhood concept of CNNs is restored. The simplified structure of a complex cell is shown in Figure 4, with displaying only one inhibiting connection per propagation line. The thick upward and downward arrows from the central cell indicate inhibition towards the other layers, the bi-directional arrows stand for lateral connections.

Similar to Figure 3, Figure 5(a) and 5(b), respectively, shows the cross-section and the 3D view of CNN layers built of complex cells. In Figure 5(b), for the sake of clarity, only the contribution of one cell per layer is depicted.

To implement the expected effect, the propagation of inhibition should be fast, i.e. the propagation delay should be negligible with respect to the time constant of CNN cells. The problem of propagation speed can be circumvented by a 'software' realisation of the complex cell in a multilayer CNN universal machine framework operating in sampling mode.[17] The CNNUM program has to implement five phases cyclically in time-sharing; one for the CNN transient, the other four for performing the simple arithmetics of inhibition propagation. During the propagation phases each state value sampled at the end of transient time-slice is

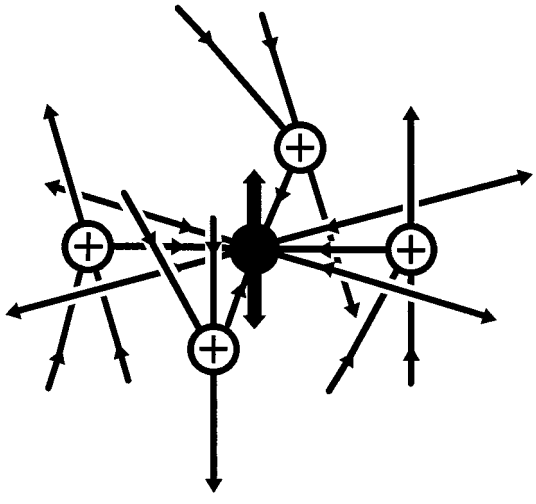

Figure 4. The complex cell of propagation mechanism with four address

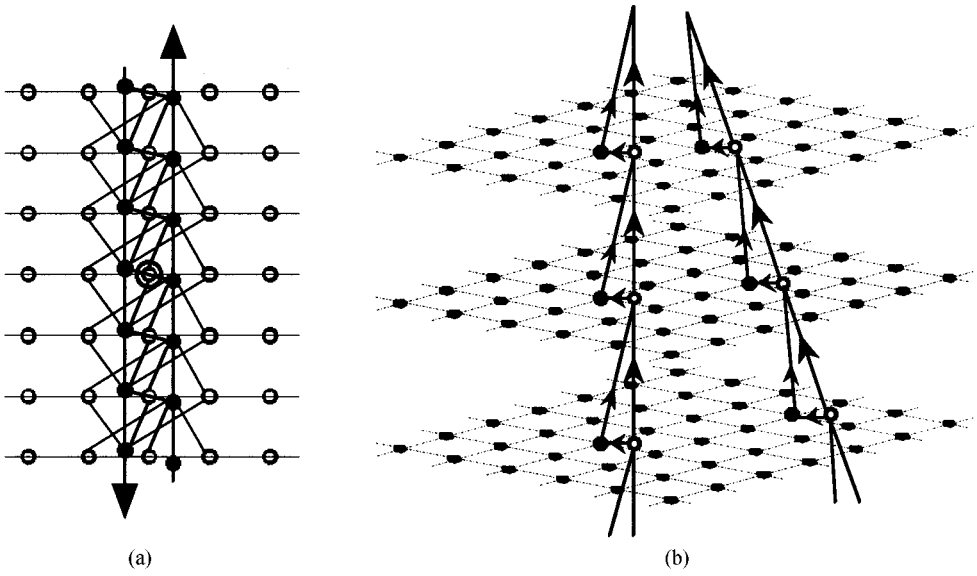

Figure 5. Interlayer inhibition through propagation mechanism

stored in an ARAM of the cell, and is reloaded when proceeding with the next transient time-slice. The length of the transient time-slice, i.e. the inhibition sampling rate is to be set as to avoid false trajectories.

4. TEMPLATE DESIGN—OPTIMIZATION APPROACH

In the previous sections we have developed a CNNUM architecture of complex cells to implement the analogical algorithm for stereo depth detection, in the following two main steps:

(i) in a succession of horizontal shifts calculating a pixel-by-pixel equivalence of the left and right images produce a series of binary (black on white) equivalence patterns, where the actual meaning of equivalence may change depending on the features of input images,

(ii) associating the equivalence patterns with depth layers, find maximum contiguous black areas that together would give maximum covering over the whole image area with minimum overlaps.

Now, we concentrate on how to obtain proper template values for the above task. Being the propagation mechanism transparent in this respect, in discussing template design techniques we will consider the original, massively coupled multilayer CNN structure of non-local depth neighbourhood.

We have tried several methods to obtain templates for depth detection in the multilayer CNN described above. The first sets of satisfactory templates quickly were found simply by error and trial based on our expectations and some reasoning. Secondly, we used simplex optimization[18] in two different ways to look for templates. Finally, we derived template values from the comparison of CNN energy function and a second-order goal function expressing our expectations. In most cases the templates were robust, giving good results in wide range of template values. Surprisingly enough, however, the templates obtained in different ways were basically different. Not even the optimization remained close to its initial value, when we started it from a properly working result of another method. We presume that our differently formulated requirements expressed in the different methods specify different non-linear dynamics that all converge to the same equilibrium from the same initial condition.

The application of optimization methods is always promising, when several contradictory requirements are to be met in a limited dimension parameter space. The advantage of simplex method is its simplicity

concerning the goal function of requirements and its ability in avoiding local minima. We tried it with several goal functions for different 3D depth detection applications embedded in a program written in Analogic CNN Language ACL[19] using Hardware Accelerator Board HAB[20] for bulk CNN calculations.

4.1. Set of optimization parameters

The execution time of a multidimensional optimization depends strongly on the number of variables, thus it must be reduced as possibly based on certain assumptions, exploiting e.g. symmetries. Our network can be described by three template matrices — lateral control from equivalence patterns, lateral facilitation and interlayer inhibition — and a current value. Owing to its central symmetry each template can be defined by a centre and an off-centre value, leading to the parameter set $(e_c, e_o, f_c, f_o, i_c, i_o, I)$. Depending on the application in mind this set can even be simplified further. Since off-centre values in control and inhibition serves mainly for low-pass filtering, in experimenting with computer-generated stereopairs they can be omitted. In the case of real camera images, when little spots bear more importance than individual pixels, those centre and off-centre entries are better to take equal values. In the following examples the optimization with five parameters (e_c, f_c, f_o, i, I) will be demonstrated. The corresponding templates are as follows:

$$\underset{\text{facilitation}}{\mathbf{A} = \begin{bmatrix} f_o & f_o & f_o \\ f_o & f_c & f_o \\ f_o & f_o & f_o \end{bmatrix}} \quad \underset{\text{inhibition}}{\mathbf{A}_{pq} = \begin{bmatrix} i & i & i \\ i & i & i \\ i & i & i \end{bmatrix}} \quad \underset{\text{control}}{\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & e_c & 0 \\ 0 & 0 & 0 \end{bmatrix}} \tag{1}$$

4.2. Optimization with heuristic considerations

If we confine our approach to the weak interpretation of uniqueness rule applying only one pixel path perpendicular to depth layers and also take into account that both the equivalence patterns and the output layers are binary, then the depth detection task with the above templates is reduced to handling $3 * 3$-pixel sized black-and-white pieces of stones. Fixing the number of layers it is fairly simple to scan through all the possibilities if we exploit the fact that due to the symmetry of templates, besides the central pixels, only the difference of black and white off-centre pixels in the investigated layer and that of together in all other layers are to be considered. Technically speaking, this approach led to five embedded cycles running in small integers in calculating the goal function for a template set. For each pixel configuration we defined the expected direction of change in the cell state. Also, we used three categories in the size of the expected change: no, small and large change. The goal function calculated the cumulated absolute difference between our expectations and the real change belonging to the template just evaluated. The only difficulty in defining the goal function was that while for some patterns the expected change is obvious, there exists a lot of ambiguous others. To circumvent this problem we introduced a measure to qualify the strength of our expectations. The advantage of this approach was its high speed, although the templates obtained were far from being satisfactory; rather we used them as an initial guess for our optimization based on training, that follows.

4.3. Optimization on training images

In the second method, we computed the error between the obtained depth results and the *a priori* known, required results for real images. In the simplest case the function was

$$\text{error} = \sum_{p=1}^{L} \sum_{(i,j)_p} |y_{ij}^p - \hat{y}_{ij}^p| \tag{2}$$

where L is the number of layers, y_{ij}^p and $\hat{y}_{ij}^p$ are the cell output and the desired output, respectively, at position (i, j) on layer p.

Depending on the application, it proved to be useful to include further terms in the error function to limit the range of cell values. To avoid going into deep saturation an additive term of extra 'punishment' was

applied for cases when *abs*(*cell state*) > *saturation limit*. Similarly, when expecting binary output we punished the grey cells meeting the condition *abs*(*cell state*) < *grey limit*. In case of applying either or both punishments the unwanted cell values were almost totally eliminated as shown in Figure 6 by the histograms of cell values.

The template values obtained were tested and found satisfactory both on computer generated stereograms and saturated black-and-white camera image-pairs produced by projecting a random pattern onto the scene, as shown in Figure 7.[14, 15] We have also tested modified versions of the optimization method with three further cases: real grey-scale images, images with large uniform areas and moving images. With these cases, our aim was to gain experience in tuning the optimization method and the templates to meet special requirements.

4.3.1. Detecting grey scale images. In everyday circumstances each object has some colour, shade and/or texture that are the basis even for human depth perception. The matching points or areas in the left and right images are searched for based on their common attributes. As shown in Figure 1, the matching can be unique or ambiguous, when more than one area exhibit similar attributes. A periodic pattern leads to especially

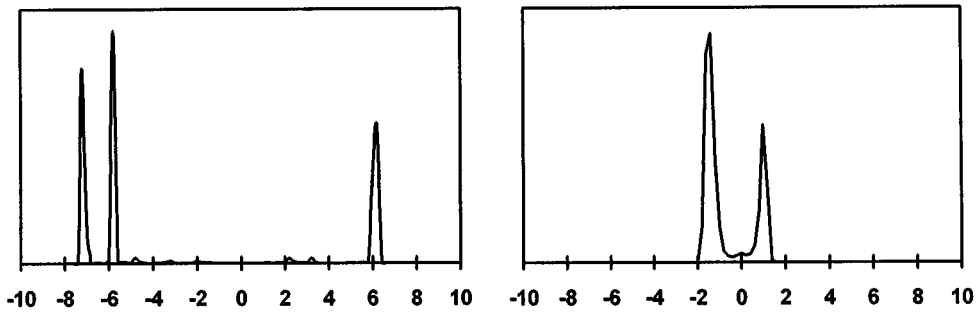

Figure 6. Histograms of cell state values without and with punishing deeply saturating cells

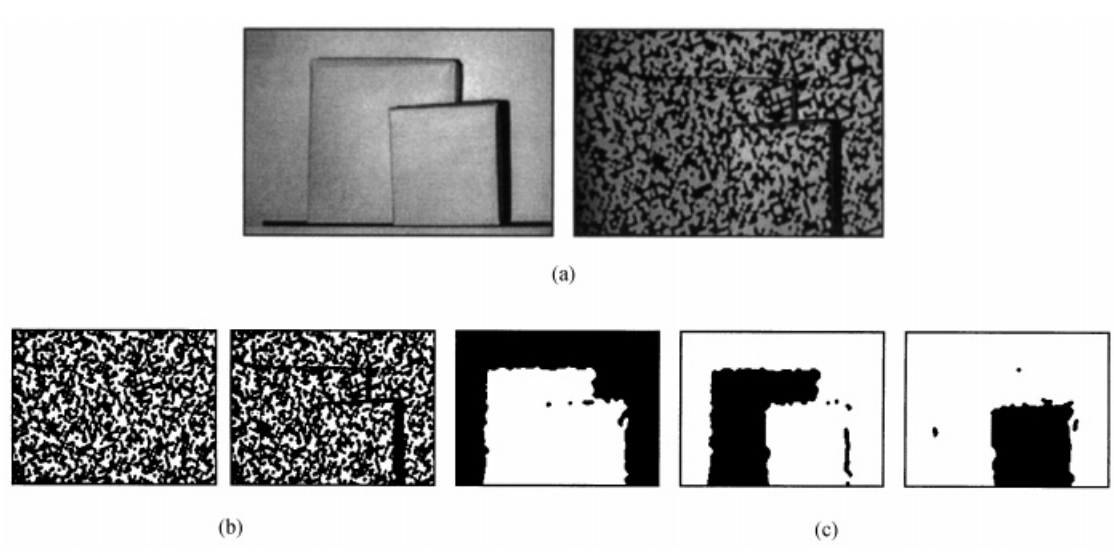

Figure 7. Depth detection in real scene: (a) camera image of two boxes with and without projected pattern; (b) its thresholded stereopair; (c) depth layers found

much false target, and its special case, the uniformly coloured areas continuously can fill depth ranges with ambiguous findings, that can only be positioned in a global processing. Obviously, matching more than one attributes, i.e. applying a complex equivalence criterion can clear some ambiguous cases.

In our experiment we used two boxes of continuously changing grey shade and an equivalence check of narrow grey window. The left and right images are shown in Figure 8, together with the two box faces found as depth layers.

4.3.2. Detecting depth with large uniform areas. As was already mentioned, in real life cases of 3D objects with large, uniform areas the depth cannot be determined in the lack of points to match. In the special case, when we know of some sort of backgrounds (or even a few of extensive uniform areas) we can exclude their attributes from depth detection and use only the different ones. Obviously, the 3D position of areas qualified as 'background' or as 'uniform' are not determined, but the abundance of false targets would not wreck the whole depth detection process. The separation of attributes is to take place already in the equivalence phase. The method is demonstrated in the following simple example with two layers of a step-like folded newspaper. Most part of the black-and-white scene is white, thus white is the background attribute. For depth detection only the thin black shapes of letters can be used. In this case a simple logical AND operation will do the separation of background areas and at the same time determine the equivalence patterns for the others. Figure 9(a) shows the left and right camera image of the folded newspaper. At the beginning of depth processing the images were thresholded to black-and-white. At the end of the depth detection the resulting depth pattern was flattened to produce a depth-mask, shown in Figure 9(b), for the input images.

4.3.3. Detecting moving objects. In experimenting with moving objects we used the arrangement with projected pattern, shown in Figure 7. Unlike to the previous cases, now the CNN transient does not start with cells in filled-in zero state, it gets periodically a new input, instead. To produce output continuously, the transient must continue from the actual non-zero state of cells. If the state of a cell happens to be deeply in the saturation domain, much above $+1$ or below -1, then for a need of changing input, it takes a longer time to remove large amount of charges from the cell capacitor and find a new settled state in time, or in worst case

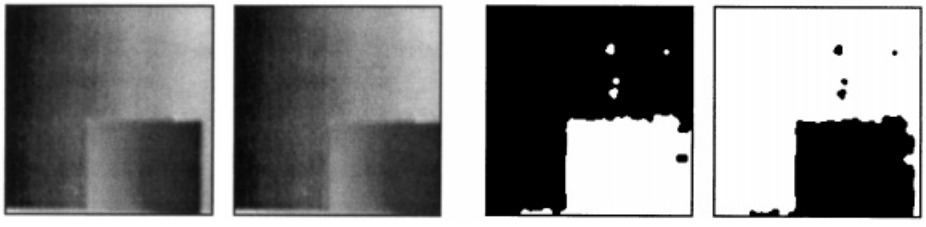

Figure 8. Two boxes with gradually changing grey shade and the extracted depth structure

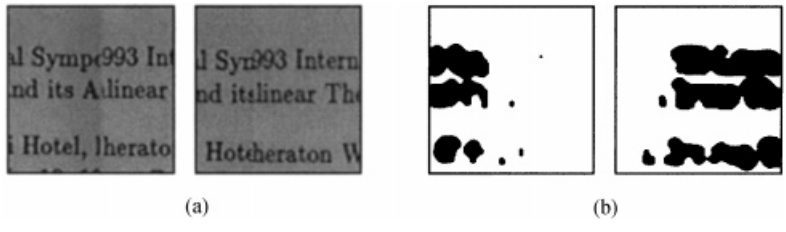

Figure 9. Photo pair of the folded newspaper sheet and the depth mask obtained

the network may not be able to find the proper output. Therefore, for following moving objects it is crucial to find a template that would not allow the cells to go into deep saturation.

For real scene camera images we always use templates that, in addition to extracting depth, would filter out little noisy spots in the image. According to our experience, however, the templates with good filtering properties usually drive the cells well into saturation, therefore a trade-off was to be found between these two contradictory requirements. The example below clearly shows the effect of poor filtering that, if necessary, can be improved by a cascaded CNN for noise removal. Downward in Figure 10, the camera image-pairs are shown as they followed each other at the input of the depth detecting CNN. Next to each pair of inputs, the extracted three layers of depth can be seen. It is worth mentioning, that in this case of moving objects the CNN, in parallel with extracting depth, at the same time has to perform the thresholding of incoming grey-scale camera images and also the equivalence operations, a task that can be considered as a pre-processing phase when dealing with still images.

4.4. Robustness of the network—Sensitivity on template values

In Table I, typical template values obtained at different goals in optimization are shown. Concerning the first two rows in Table I, in analysing several sets of templates, experimenting with numerous values of parameters changing around the optimum found and evaluating the error function for them, we met high

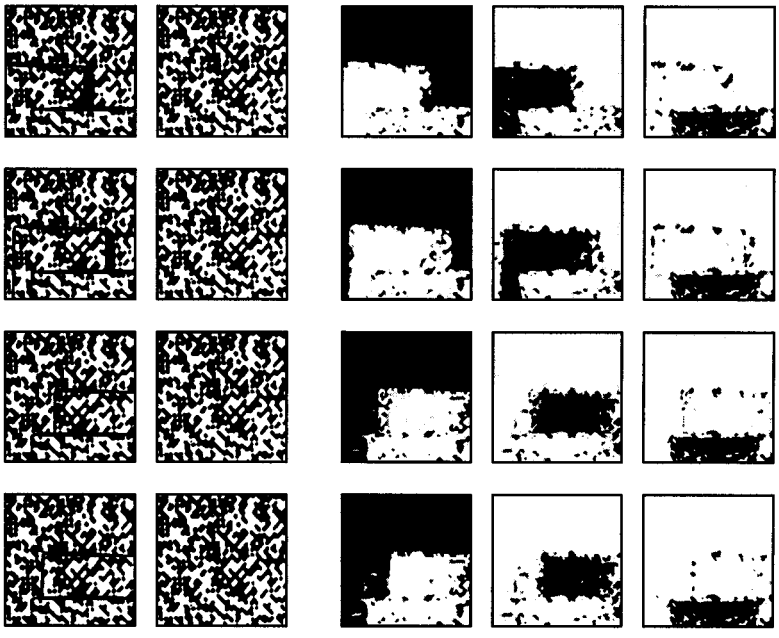

Figure 10. The input pattern sequence for moving object detection and the depth layers found

Table I. Typical template values for different optimization goals

Goal function	e_c	f_c	f_o	i	I
General optimization	0·658	0·542	0·348	− 0·321	− 3·231
Punishing saturated and grey cells	0·591	0·377	0·061	− 0·065	− 2·152
Optimization for moving objects	1·331	0·156	0·071	− 0·122	− 1·999

robustness in accordance with our expectations. In the moving object case with continuously changing input (in general, when solving partial differential equations) the sensitivity of solution as measured in the level of output noise was higher. In all the cases we tested, the inputs and the outputs are digital, except for the input of the experiment with the grey scale image. However, if we consider the equivalence operation as a preparation step then even in this case the input to the main network part is also digital.

The templates obtained well reflect our intentions expressed in the goal functions. If we punish the saturated cells then parameter $\mathbf{e}_c$ and $\mathbf{f}$'s controlling the influence from the equivalence patterns and the feedback, respectively, decrease. In the case of moving objects, however, the effect of the equivalence patterns becomes stronger because for a new input the cells must be able to follow it by changing even from one saturation to the other. For the same reason the feedback cannot be very strong, a cell should not 'adhere' to its momentary state, it ought to adapt flexibly to external changes. As an unpleasant side effect of this flexibility, we detected a degradation in the lateral filtering of results, to be properly set for a trade-off. In general, however, all the parameters must be considered together and not examined one by one, because many different template sets can give the same result due to the robustness of the network.

We tried the method with several goal functions for different 3D depth detection applications. The simplex optimization frame was implemented in an Analogic CNN Language ACL[19] program running on a Pentium 120 MHz PC. It invoked a multilayer CNN simulator using Hardware Accelerator Board HAB[20] for bulk CNN calculations, whenever a goal function value belonging to a new set of optimization parameters was required. Based on previous experience, 30 iterations was allowed for the simulator to find a solution for the three-layer test example of dimensions $174 * 143$ pixels in each layer. The calculation of a goal function value together with the data transfer time between HAB and its host was around 30 sec. The simplex process took typically 15–30 hour, i.e. several thousands of optimization steps to find satisfactory sets of parameters sufficiently close to each other.

5. TEMPLATE DESIGN—ENERGY FUNCTION APPROACH

Since neural networks, the same way as all physical systems, seek for steady state a local minimum, an equilibrium of their total energy, the classical method[2] for designing synaptic weights or CNN template values is, first, to compose a second-order goal function reflecting our expectations concerning the solution of the task to be solved and, second, to match symbolically that goal function to the energy function of the neural network expressed in physical parameters. The energy function of a CNN of unity value cell resistors and cell capacitors, working in the linear range of its (sigmoid) output function is as follows:

$$E = -\frac{1}{2}\sum_{p}^{L}\sum_{q}^{L}\sum_{i}^{n}\sum_{j}^{m}\sum_{kl\in N_r(i,\,j)} A_{pq}\,(ij,\,kl)\,V_{p,\,ij}\,V_{q,\,kl} - \sum_{p}^{L}\sum_{i}^{n}\sum_{j}^{m} I_{p,ij}\,V_{p,ij} + \frac{1}{2}\sum_{p}^{L}\sum_{i}^{n}\sum_{j}^{m} V_{p,ij}^2 \tag{3}$$

where L is the number of CNN layers, n and m are the number of rows and columns in a layer, respectively, and $V_{p,ij}$ is the cell state and also the output within saturation.

The first component of the energy function is a sum of the energy of current sources controlled in feedback by templates A. In the second component, $I_{p,ij}$ is standing for the sum of the constant current source I and the current of sources controlled forward from the input through templates B. The third component is the energy stored in cell capacitors.

5.1. *Criteria for finding a 3D surface*

In order to find proper template values for stereo depth extraction, we develop a goal function to enforce several criteria representing the continuity and uniqueness rules, in the form of second-order expressions of cell capacitor voltages, descending towards preferable solutions. Each criterion is to convey a particular point of view, sometimes contradictory in local neighbourhoods; their weighted competition is to terminate

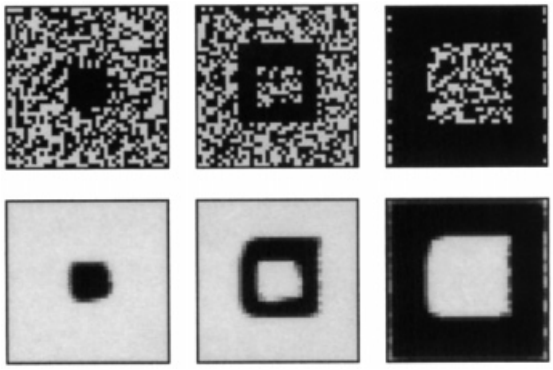

Figure 11. The equivalence pattern set and the depth layers found for a computer-generated random-dot stereogram of a three-level step-pyramid

at an equilibrium that we consider as the solution for our problem. The starting point is the set of equivalence patterns Q obtained from pixel-wise equivalence operations between the left and right images at a series of relatively shifted positions.

Criterion 1. Since for a human spectator the equivalence pattern set (Figure 11) vividly shows the surface segments looked for, indicating parts of the equivalence set belonging to the result well separated from the other, noisy areas, in E_1 we require a pixel-wise identity between the equivalence image set and the result.

$$E_1 = \sum_{p}^{L} \sum_{i}^{m} \sum_{j}^{n} (V_{p,ij} - Q_{p,ij})^2 = \sum E_{1p,ij} \tag{4}$$

where $Q_{p,ij}$ is value of pixel (i, j) in the pth equivalence pattern.

Criterion 2. From continuity rule it follows that we have to prefer the set of surface segments of maximum contiguous areas, resulting in the local requirement E_2 expressing that in the horizontal and vertical one-pixel neighbourhood of each pixel the number of pixels of the same type (black or white) should be maximum.

$$\begin{aligned} E_2 &= \sum E_{2p,ij} \\ &= \sum_{p}^{L} \sum_{i}^{m} \sum_{j}^{n} [(V_{p,ij} - V_{p,(i-1)j})^2 + (V_{p,ij} - V_{p,(i+1)j})^2 + (V_{p,ij} - V_{p,i(j-1)})^2 + (V_{p,ij} - V_{p,i(j+1)})^2] \end{aligned} \tag{5}$$

Depending on application the diagonal neighbours can also be included in E_2, however, experience shows that extending this criterion to the diagonal neighbours results in cutting of sharp edges.

Criterion 3. To obtain a binary (black-and-white) result we need a criterion E_3 with minima at state space points of co-ordinates composed of -1's and $+1$'s, as follows:

$$E_3 = -\sum_{p}^{L} \sum_{i}^{m} \sum_{j}^{n} V_{p,ij}^2 \tag{6}$$

Criterion 4. The first three criteria are to work in each depth layer without considering the state of other layers. The uniqueness rule postulates that each visible point in the 3D world has exactly one, unique left and right image (except for occluded regions), leading to the consequence that each pixel in the left image has only

one corresponding pixel in the right one. To enforce this type of uniqueness, in Section 3 we introduced inhibition within layers to prevent a pixel from contributing to surface segments in more than one depth layers.

As it was explained, when a pixel from the left image and another from the right one form a surface segment, the relative difference in their horizontal positions determines the 3D depth, i.e. the depth layer. On the other hand, for each 3D pixel uniquely can be found those two pixels, one from the left image, the other from the right (see Figures 1 and 2), that may form a surface segment there. For the ith depth layer, the difference in their original horizontal positions is just i, and they approach each other from layer to layer up to the ith one. Passing depth layer i, their paths will diverge. By the uniqueness rule each surface segment should be the only one along the paths of the contributing pixels, calling for the interlayer inhibition mechanism along the pixel paths, as was discussed. Moreover, not all the pixel in one image must find a corresponding pixel in the other; there might be occluded areas in the 3D sight, visible only on either image, therefore, it is not expected from the interlayer mechanism to guarantee the existence of any surface segment along each pixel path.

Since in a CNN cell the [black,white] range is mapped to $[1, -1]$, and at most one black pixel is to be allowed along each pixel path, we consider following products of the cell capacitor voltages along each pixel path:

$$(V_{p,ij} + 1) \sum_{q \neq p}^{L} (V_{q,(i+(p-q))j} + 1) \tag{7}$$

The value of this product increases with the number of $+1$ of cell capacitor values along the pixel path crossing layer p at pixel (i,j), if the cell capacitor value itself is $+1$ at that pixel too. Summing such products for all cells and pixel paths we obtain the inhibition criterion E_4.

$$E_4 = \sum_{i,j} E_{4,ij} = \sum_{i,j} \left[\sum_{p}^{L} (V_{p,ij} + 1) \sum_{q \neq p}^{L} (V_{q,(i+(p-q))j} + 1) \right] \tag{8}$$

5.2. Determining CNN templates

We associate the energy function to minimise by CNN with a weighted sum of expressions E_1, E_2, E_3 and E_4

$$E = a * E_1 + b * E_2 + c^* E_3 + d * E_4 \tag{9}$$

The role of weights is to give relative emphasis to the criteria and to help to avoid local minima. The weights that follow have been adjusted and tested in several application examples. It was found that the CNN operation is robust with respect to the values of weights.

The comparison and matching of the two expressions for the energy function, the one comprising template elements and the other composed of criteria, is to result in formulae for template values. Since both expressions are a sum of components, it is worth while to make the comparison by components, i.e. to produce the template formulae as sums too. To shorten lengthy expressions in the following comparison we use a relative notation for template entries instead of the generally used absolute ones:

$$A_{pq}(x, y) \Leftrightarrow A_{pq}(i, j; i + x, j + y) \text{ and } A(x, y) \Leftrightarrow A_{pp}(x, y)$$

1. Let us expand first the weighted component E_1 (4) and (9) coming from criterion 1 and substitute the corresponding terms for $V_{p,ij}$ and $Q_{p,ij}$ from the energy function (3)

$$\begin{aligned} aE_{1\,p,ij} = a(V_{p,ij} - Q_{p,ij})^2 &= a(V_{p,ij}^2 - 2V_{p,ij}\,Q_{p,ij} + Q_{p,ij}^2) \\ &= -\frac{1}{2} A(0,0)\, V_{p,ij}^2 - B(0,0)\, V_{p,ij}\, Q_{p,ij} + \frac{1}{2} V_{p,ij}^2 + a\, Q_{p,ij}^2 \end{aligned} \tag{10}$$

If, for simplicity, we choose $a = 1/2$, then the coefficient of cell capacitor voltage will be zero, and the contribution of this criterion to self-feedback will be zero too: $A(0,0) = 1 - 2a = 0$. (Although the self-feedback bears importance in CNN stability, being only an additive part to it, this zero value in itself cannot endanger the final stability.)

The second term in the above expression comes from the comparison of factors containing $Q_{p,ij}$. It gives a contribution to the value of forward control template: $B(0,0) = 2 * a = 1$ and shows, that the set of equivalence patterns will serve as the input to the network through unity weight. The last term with Q_{ij} can be ignored, since – with its constant value – it would not affect the location of energy minimum.

2. The expansion of E_2 (5) and the determination of template contributions follow similar considerations. To obtain a proper result with this criterion, the summation has to go through a whole layer and the effect of borders will be ignored.

$$\begin{aligned} bE_{2p,ij} &= b\left[(V_{p,ij} - V_{p,ij-1})^2 + (V_{p,ij} - V_{p,ij+1})^2 + (V_{p,ij} - V_{p,i-1j})^2 + (V_{p,ij} - V_{p,i+1j})^2\right] \\ &= b\left(8V_{p,ij}^2 - 2V_{p,ij}V_{p,ij-1} - 2V_{p,ij}V_{p,ij+1} - 2V_{p,ij}V_{p,i-1j} - 2V_{p,ij}V_{p,i+1j}\right) \\ &= -\frac{1}{2}\left(A_{pp}(0,0)\,V_{p,ij}^2 + A_{pp}(0,-1)\,V_{p,ij}V_{p,ij-1} + A_{pp}(0,1)\,V_{p,ij}\,V_{p,ij+1}\right. \\ &\quad \left. + A_{pp}(-1,0)\,V_{p,ij}V_{p,i-1j} + A_{pp}(1,0)\,V_{p,ij}\,V_{p,i+1j}\right) \end{aligned} \tag{11}$$

If we choose $b = 1/2$ and take into account that in the summation over the whole layer each pair of cells arises twice and therefore the coefficients are to be divided by 2, we obtain the following one-neighbourhood $(3 * 3)$ contribution to the feedback template:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{12}$$

3. To finish with single layer criteria, we continue with E_3 (6) of criterion 3 aimed to provide for binary output. It gives a simple contribution to self-feedback, as follows:

$$cE_{3p,ij} = -cV_{p,ij}^2 = -\frac{1}{2}A_{pp}(0,0)\,V_{p,ij}^2 \tag{13}$$

Similarly to the previous cases, choosing $c = 1/2$, we have the third contribution to self-feedback: $A_{pp}(0,0) = 2c = 1$.

We can compose now the final feedback facilitation template too, with and without substituting a value for weights:

$$A = \begin{bmatrix} 0 & 2b & 0 \\ 2b & 1 - 2a - 8b + 2c & 2b \\ 0 & 2b & 0 \end{bmatrix} \Rightarrow A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -3 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{14}$$

4. Finally, we discuss criterion 4 that was composed to control the 'distribution' of surface segments among depth layers as dictated by the uniqueness rule. From expression (7) contributing in E_4 (8), we pick a single product, keeping only a layer index, and compare with its counterpart in the energy function:

$$\begin{aligned} d(V_p + 1) * (V_q + 1) &= dV_pV_q + dV_p + dV_q + d \\ &= -\frac{1}{2}A_{pq}(x,y)\,V_pV_q - \frac{1}{2}A_{qp}(-x,-y)\,V_qV_p - I_pV_p - I_qV_q + d \end{aligned} \tag{15}$$

We obtain the feedback template connecting depth layer q to depth layer p and a contribution to constant current source as follows:

$$A_{pq}(x, y) = -d \text{ or choosing } d = 1,\ A_{pq}(x, y) = -1 \qquad (16)$$
$$I_p = -1$$

By running indices through all the layers we can collect the constant current contributions flowing into each single cell, owing to the inhibition along the left and right pixel paths, with the result $I = 2 - 2L$.

Based on the multilayer structure of CNN in Figure 2, the general indices in the template above can be replaced with actual ones considering the direction of pixel paths among layers. Owing to the rectangular representation, the left pixel path always goes through pixels of identical positions perpendicularly to the successive depth layers, resulting in feedback connections among all cells in the same position of different layers $A_{pq}(0,0) = -1$. The right pixel paths go slanted, keeping always one pixel to the left when stepping from one layer to the next one, i.e. $A_{pq}(q - p, 0) = -1$. These two nets of feedback connections coming from the uniqueness rule create those two massively dense nets of couplings, the first one belonging to the left pixel paths being perpendicular to the CNN plane, and the second one belonging to the right pixel paths slanted to it, that were already discussed in Section 3.

5.3. Generalisation of equivalence check

In developing the multilayer stereo CNN architecture and deriving its templates, we built upon the essentially accepted necessity of an equivalence operation. A more general interpretation of equivalence can be achieved by formulating the related expectation as a criterion to match with some part in the energy function. Supposing we know the solution of a depth extraction task, we can define a solution set S_p on each layer p, to contain pixels belonging to the solution. Proper solution sets S_p will minimise the following expression:

$$\sum_{p}^{L} \left(\sum_{i,\, j \notin S_p} Q(L_{ij}^p - R_{ij}^p) - \sum_{i,\, j \in S_p} Q(L_{ij}^p - R_{ij}^p) \right) \qquad (17)$$

where $Q(\cdot)$ is a symmetrical function with minimum at $Q(0)$, just as abs$(\cdot)$, L_{ij}^p and R_{ij}^p are the left and right images, respectively, shifted to level p.

If we calculate $Q_{ij}^p = Q(L_{ij}^p - R_{ij}^p)$ set of images and through control templates B connect to the input of the stereo CNN structure, the contribution of the involved current sources, a part of $I_{p,ij}$ in equation (1), to the CNN energy function (1) can be expressed as

$$\sum_{p}^{L} \sum_{i,j} B(0, 0)\ Q_{ij}^p\ V_{ij}^p \qquad (18)$$

Taking into account that in a binary non-saturating steady-state solution V_{ij}^p is either -1 or $+1$ depending on whether the pixel at (i, j) belongs to the solution set S_p or not, expression (18) will directly lead to equation (17) with a coefficient $B(0,0)$. Consequently, the stereo CNN architecture will minimise equation (17) that therefore can be considered as an alternative to criterion 1. Accepting this interpretation of equivalence check gives way to incorporating it into the depth detection dynamics, instead of considering it as a pre-processing phase, leading to structures necessary, for instance, in depth detection of moving objects.

5.4. Examples with templates

We made numerous experiments with weights a, b, c, d. Several sets of weights were tested on computer-generated random-dot black-and-white and grey scale stereograms, the same or similar to the previously discussed ones. Table II summarizes the outcome of experiments. The results with parameters shown in Table II were satisfactory; for computer-generated random-dot stereograms we calculated the rate of bad pixels, it was found to be below 1%. The strange negative value for self-feedback f_c can be accepted if we

consider its role as a counterbalance for the strong inhibition. Similar role is assumed by constant current I in keeping the dynamic range of operations close to the centre of the linear domain of CNN.

Figure 11 shows three layers of the equivalence pattern set and the resulting depth layers for a computer-generated random-dot stereogram of a step pyramid. In order to test the disturbing effect of empty layers, the depth detection was performed with five layers; in addition to the three layers of the step pyramid two layers were placed in such a depth where no surface segment was expected. The CNN found the three layers properly and left the additional two layers practically empty except for a small number of pixels at the pyramid steps, demonstrating that the effect of empty layers in the depth detection process is negligible. The ragged contours of pyramid steps came from the inherent local uncertainty of the random-dot technique, that can be verified by simple visual inspection of the original stereogram.

Five depth layers extracted from a real scene stereo-pair of two boxes (similar to the scene in Figure 7) are shown in Figure 12. On the contrary to the previous, computer-generated case, small segments in the second and fourth 'empty' layers were found at the edges of the two boxes. These false findings can be attributed to

Table II. Template values from energy function approach (for facilitation corner f_o zero values are assumed (see[14]))

Configuration	e_c	f_c	f_o	i	I
Template expressions	$2a$	$1-2a-8b+2c$	$2b$	$-d$	$2-2L$
Values for $a=b=c=1/2 \quad d=1 \quad L=5$	1	-3	1	-1	-8

Figure 12. The depth layers found for a real scene stereo pair of two boxes with projected pattern

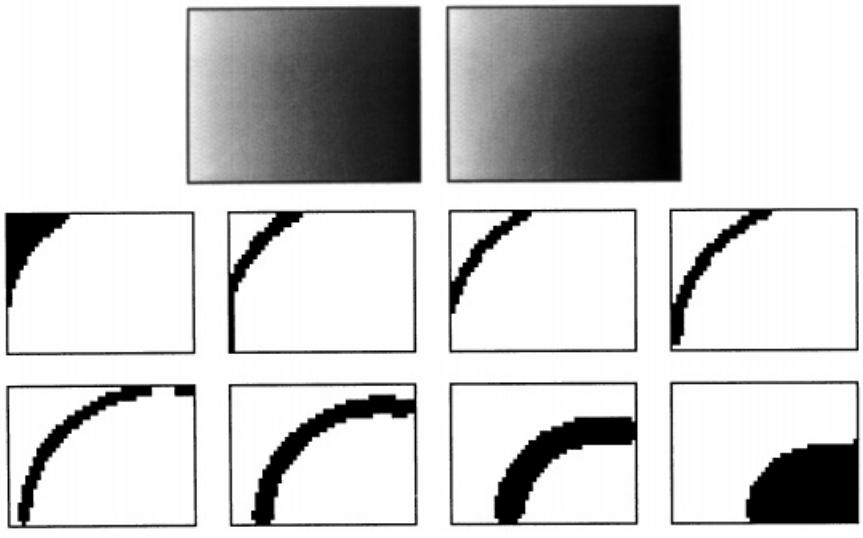

Figure 13. Continuous grey scale stereo pair of a spherical segment and the contour lines found at eight equidistant depth

local ambiguities in the projected pattern, since they are visually observable in stereoscope too, and therefore they cannot be considered as an improper functioning of the depth detection algorithm.

Figure 13 shows a continuous grey scale stereogram of a spherical surface, modelling a real scene stereo image situation. In producing the equivalence pattern set we simply used the difference of grey levels with low thresholding. Due to the continuously changing grey level and the slowly varying depth there was no danger of false targets. The evolution of CNN transient for several examples showed that most of the surface segments were dealt among layers quickly, already in the beginning of the transient. Most of the segments were light grey, almost white first. The final dark, mostly black tone was achieved only after a longer filtering phase. The typical time elapsed before an approximate steady state was reached was 10–50 RC, the CNN time constant.

6. CONCLUSIONS

Based on models and rules established in the computational theory of human stereo vision we developed a local neighbourhood multilayer CNNUM structure to extract 3D depth from stereo images. The CNN templates were derived in several ways using optimization and as well as a symbolic matching of CNN energy function to a second-order goal function. The templates obtained proved to be robust in a wide range of parameters.

ACKNOWLEDGEMENTS

This work was supported in part by the National Research Fund in Hungary (OTKA, grant T012862).

REFERENCES

1. L. O. Chua and L. Yang, 'Cellular neural networks: theory' and 'cellular neural networks: applications', IEEE Trans. Circuits Systems, **35**, 1257–1290 (1988).
2. J. J. Hopfield and D. J. Tank, 'Neural computation of decisions in optimization problems', *Biol. Cybernet*., **52**, 141–152 (1985).
3. T. Roska and L. O. Chua, 'The CNN paradigm', *IEEE Trans. Circuits Systems*, Special Issue on CNN, I: Fundamental Theory and Applications, **40**, 147–156 (1993).
4. T. Roska and L. O. Chua, 'The CNN universal machine: an analogic array computer', *IEEE Trans. Circuits Systems*, Special Issue on CNN, II: *Analog and Digital Signal Processing*, **40**, 163–173 (1993).
5. L. O. Chua, T. Roska, T. Kozek and Á. Zarándy, 'CNN universal chips crank up the computing power', *IEEE Circuits Dev*. 18–28 (1996).
6. B. Julesz, *Foundations of Cyclopean Perception*, Chicago University Press, Chicago, 1971.
7. D. Marr, '*Vision*', Freeman, San Francisco, 1982.
8. D. Marr, G. Palm and T. Poggio, 'Analysis of cooperative stereo algorithm', *Biol. Cybernet*. **28**, 223–239 (1978).
9. R. Blake and H. R. Wilson, 'Neural models of stereoscopic vision', *Trends in Neuroscience*,*TINS*, **14** (10), 445–452 (1991).
10. P. Dev, 'Perception of depth surfaces in random-dot stereograms: a neural model', *Int. J. Man–Machine Studies*, **7**, 511–528 (1975).
11. W. E. L. Grimson, 'A computer implementation of a theory of human stereo vision', *Phil. Trans. Roy. Soc. London*, **B292**, 217–253 (1981).
12. S. Park, S.-J. Min and S-I. Chae, 'Stereo correspondence with discrete-time cellular neural networks', *Proc. ISCAS'94*, London, 1994, 6.225–6.228.
13. M. Tanaka and M. Awata, 'Extraction of depth information by cellular neural networks', *Proc. ISCAS'94*, London, 1994, 6.281–6.284.
14. A. Radványi, 'Solution of stereo correspondence in real scene: an analogic CNN algorithm', *Proc. 3rd IEEE Int. Workshop on Cellular Neural Networks and their Applications* (*CNNA*-94), Rome, 1994, pp. 231–236.
15. A. Radványi, 'Spatial depth extraction using random stereograms in analogic CNN framework', *Int. J. Circuit Theory Appl.* **24** (1) 69–92, 1996
16. S. Taraglio, A. Zanela, 'CNN for the stereo matching problem', *Proc. 4th IEEE Int. Workshop on Cellular Neural Networks and their Applications* (*CNNA*-96), Seville, Spain, 1996, 93–98.
17. J. M. Cruz, L. O. Chua and T. Roska, 'A fast, complex and efficient test implementation of the CNN universal machine', *Proc. 3rd IEEE Int. Workshop on Cellular Neural Networks and their Applications* (*CNNA-94*), Rome, 1994, 61–66.
18. J. A. Nelder and R. Mead, 'A simplex method for function minimisation', *Comput J*., **7**, p. 308 (1965).
19. T. Roska, P. Szolgay, Á. Zarándy, P. L. Venetianer, A. Radványi and T. Szirányi, 'On a CNN chip-prototyping system', *Proc. CNNA'94—1994 3rd IEEE Int. Workshop on Cellular Neural Networks and their Applications*, Rome, Italy, 1994.
20. T. Roska, G. Bártfay, P. Szolgay, T. Szirányi, A. Radványi, T. Kozek, Zs Ugray and Á. Zarándy, 'A digital multiprocessor hardware accelerator board for cellular neural networks: CNN-HAC', *Int. J. Circuit Theory Appl*., **20** (5), 589–600 (1992).

