Neural Networks, Vol. 1, pp. 119–130, 1988
Printed in the USA. All rights reserved.

0893-6080/88 $3.00 + .00

ORIGINAL CONTRIBUTION

Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition

KUNIHIKO FUKUSHIMA

NHK Science and Technical Research Laboratories

(*Received and accepted* 15 *September* 1987)

Abstract—*A neural network model for visual pattern recognition, called the "neocognitron," was previously proposed by the author. In this paper, we discuss the mechanism of the model in detail. In order to demonstrate the ability of the neocognitron, we also discuss a pattern-recognition system which works with the mechanism of the neocognitron. The system has been implemented on a minicomputer and has been trained to recognize handwritten numerals.*

The neocognitron is a hierarchical network consisting of many layers of cells, and has variable connections between the cells in adjoining layers. It can acquire the ability to recognize patterns by learning, and can be trained to recognize any set of patterns. After finishing the process of learning, pattern recognition is performed on the basis of similarity in shape between patterns, and is not affected by deformation, nor by changes in size, nor by shifts in the position of the input patterns.

In the hierarchical network of the neocognitron, local features of the input pattern are extracted by the cells of a lower stage, and they are gradually integrated into more global features. Finally, each cell of the highest stage integrates all the information of the input pattern, and responds only to one specific pattern. Thus, the response of the cells of the highest stage shows the final result of the pattern-recognition of the network. During this process of extracting and integrating features, errors in the relative position of local features are gradually tolerated. The operation of tolerating positional error a little at a time at each stage, rather than all in one step, plays an important role in endowing the network with an ability to recognize even distorted patterns.

1. INTRODUCTION

Visual pattern recognition, such as reading characters or distinguishing shapes, can easily be done by human beings, but it is very difficult to design a machine which can do it as well as human beings do. We believe that the best strategy is to learn from the brain itself. We are studying the mechanism of visual information-processing in the brain, and trying to use it as a design principle for new information processors. More specifically, we are studying how to synthesize a neural network model which has the same ability as the human brain. As a result of this approach, a pattern-recognition system called the "neocognitron" has been developed (Fukushima, 1980; Fukushima & Miyake, 1982).

In the visual area of the cerebrum, neurons are found to respond selectively to local features of a visual pattern, such as lines and edges in particular orientations (Hubel & Wiesel, 1962). In the area higher than the visual cortex, it has been found that cells exist which respond selectively to certain figures like circles, triangles, squares, or even to a human face (Bruce, Desimone, & Gross, 1981; Sato, Kawamura, & Iwai, 1980). Accordingly, the visual system seems to have a hierarchical structure, in which simple features are first extracted from a stimulus pattern, and then integrated into more complicated ones. In this hierarchy, a cell in a higher stage generally receives signals from a wider area of the retina, and is more insensitive to the position of the stimulus.

Such neural networks in the brain are not always complete at birth. They gradually develop, adapting flexibly to circumstances after birth. Sophisticated brain functions, such as learning, memory, and pattern-recognition, are believed to be acquired through the growth of the neural network, in which neurons extend branches and make connections with many other neurons.

This kind of physiological evidence suggested a network structure for the neocognitron. The neocognitron is a hierarchical multilayered network consisting of neuron-like cells. The network has variable connections between cells, and can acquire the ability to recognize patterns by learning. It can be trained to recognize any

Requests for reprints should be sent to Kunihiko Fukushima, NHK Science and Technical Research Laboratories, 1-10-11, Kinuta, Setagaya, Tokyo 157, Japan.

set of patterns. After finishing the process of learning, the response of the cells of the highest stage of the network shows the final result of the pattern-recognition: only one cell, corresponding to the category of the input pattern, responds. Pattern recognition of the network is performed on the basis of similarity in shape between patterns, and is not affected by deformation, nor by changes in size, nor by shifts in the position of the input patterns.

In this paper, we discuss the mechanism of the model in detail. In order to demonstrate the ability of the neocognitron, we also discuss a pattern-recognition system which has been designed using the principle of the neocognitron. The system has been implemented on a minicomputer and has been trained to recognize handwritten numerals.

2. THE STRUCTURE AND BEHAVIOR OF THE NETWORK

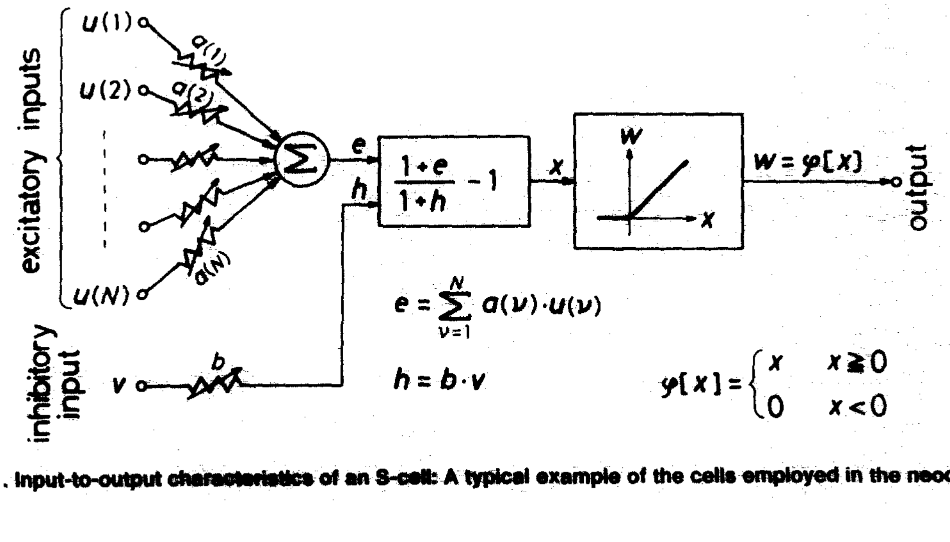
The neocognitron is a multilayered network consisting of a cascade of many layers of neuron-like cells. The cells are of the analog type; that is, their inputs and outputs take non-negative analog values, corresponding to the instantaneous firing-frequencies of biological neurons. Figure 1 shows a typical example of the cells employed in the network.

The hierarchical structure of the network is illustrated in Figure 2. There are forward connections between cells in adjoining layers. The initial stage of the network is the input layer, called U_0, and consists of a two-dimensional array of receptor cells u_0. Each of the succeeding stages has a layer of "S-cells" followed by a layer of "C-cells." Thus, in the whole network, layers of S-cells and C-cells are arranged alternately. Notation U_{Sl} and U_{Cl} are used to denote the layers of S-cells and C-cells of the lth stage, respectively. Incidentally, each U_S-layer contains subsidiary inhibitory cells, called V-cells, but they are not drawn in Figure 2.

S-cells are feature-extracting cells. Connections converging to feature-extracting S-cells are variable and are reinforced during a learning (or training) process. After finishing the learning, which will be discussed later, S-cells, with the aid of the subsidiary V-cells, can extract features from the input pattern. In other words, an S-cell is activated only when a particular feature is presented at a certain position in the input layer. The features which the S-cells extract are determined during the learning process. Generally speaking, in the lower stages, local features, such as a line at a particular orientation, are extracted. In higher stages, more global features, such as a part of a training pattern, are extracted.

The C-cells are inserted in the network to allow for positional errors in the features of the stimulus. Connections from S-cells to C-cells are fixed and invariable. Each C-cell receives signals from a group of S-cells which extract the same feature, but from slightly different positions. The C-cell is activated if at least one of these S-cells is active. Even if the stimulus feature is shifted in position and another S-cell is activated instead of the first one, the same C-cell keeps responding. Hence, the C-cell's response is less sensitive to shifts in position of the input pattern.

This network structure is illustrated in Figure 2 in more detail. S-cells or C-cells in a layer are divided into subgroups according to the kinds of feature to which they respond. Since the cells in each subgroup are arranged in a two-dimensional array, we call the subgroup a "cell-plane." In Figure 2, each quadrangle drawn with heavy lines represents a cell-plane, and each vertically elongated quadrangle drawn with thin lines, in which cell-planes are enclosed, represents a layer of S-cells or C-cells. As schematically illustrated in Figure 3, all the cells in a cell-plane receive input connections of the same spatial distribution, and only the positions of the preceding cells are shifted in parallel from cell to cell. Although cells usually exist in numbers, only one cell



$$e=\sum_{\nu=1}^{N}a(\nu)\cdot u(\nu)$$

$$h=b\cdot v$$

$$\varphi[x]=\begin{cases}x & x\geqq 0\\ 0 & x<0\end{cases}$$

FIGURE 1. Input-to-output characteristics of an S-cell: A typical example of the cells employed in the neocognitron.

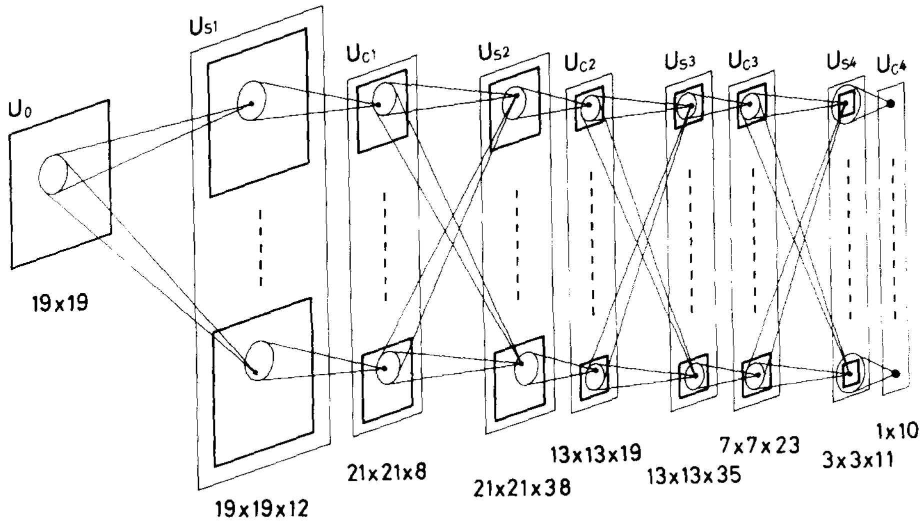


FIGURE 2. Hierarchical network structure of the neocognitron. The numerals at the bottom of the figure show the total numbers of S- and C-cells in individual layers of the network which are used for the handwritten numeral recognition system discussed in Section 4.

is drawn in each cell-plane in Figure 2. Incidentally, each ellipse in the figure represents the area from which a cell receives input connections.

The density of cells in each layer is designed to decrease with the order of the stage, because the cells in higher stages usually receive signals from larger areas of the input layer and the neighboring cells come to receive similar signals. Hence, in the highest stage, only one C-cell exists in each cell-plane.

Thus, in the whole network, in which layers of S-cells and C-cells are arranged alternately, the process of feature-extraction by S-cells and toleration of positional shift by C-cells are repeated. During this process, local features extracted in a lower stage are gradually integrated into more global features. Figure 4 illustrates this situation schematically. Finally, each C-cell of the highest stage integrates all the information of the input pattern, and responds only to one specific pattern. In other words, in the highest stage, only one C-cell, corresponding to the category of the input pattern, is activated. Other cells respond to patterns of other categories. Thus, the C-cells of the highest stage may be called "gnostic cells," and their response shows the final result of the pattern-recognition of the network.

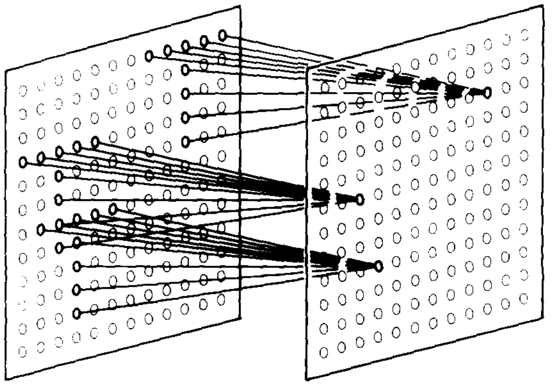


FIGURE 3. Illustration showing the spatial arrangement of the connections converging to single cells of a cell-plane.

The operation of tolerating positional error a little at a time at each stage, rather than all in one step, plays an important role in endowing the network with an ability to recognize even distorted patterns. Since errors in the relative position of local features are tolerated in the process of extracting and integrating features, the same C-cell responds in the highest layer, even if the input pattern is deformed or changed in size or shifted in position. In other words, the neocognitron recognizes the "shape" of the pattern independent of its size and position.

3. SELF-ORGANIZATION OF THE NETWORK

The connections converging to S-cells are variable, and are reinforced gradually in accordance with stimuli given to the network during the process of learning. Both processes, "learning-*without*-a-teacher" and "learning-*with*-a-teacher" can be used to train the neocognitron to recognize patterns.

3.1 Learning without a Teacher

We will first discuss the case of learning-without-a-teacher (Fukushima, 1980; Fukushima & Miyake, 1982). The repeated presentation of a set of training patterns is sufficient for the self-organization of the network, and it is not necessary to give any information about the categories in which these patterns should be

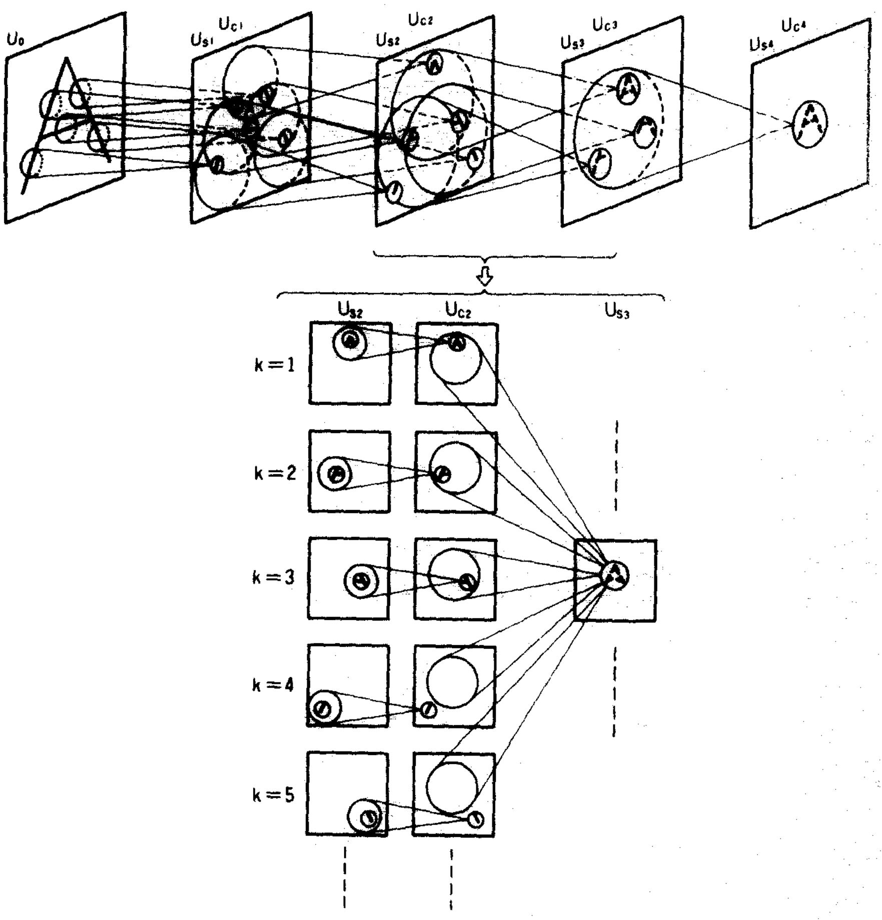


FIGURE 4. Illustration of the process of pattern recognition in the neocognitron.

classified. The neocognitron by itself acquires the ability to classify and recognize these patterns correctly on the basis of similarity in shape.

3.1.1 Reinforcement of Maximum-Output Cells. Self-organization of the neocognitron is performed with two principles. The first has been introduced for the self-organization of the "cognitron" (Fukushima, 1975, 1981) proposed earlier by the author. Specifically, the first principle is as follows:

The variable connection between two cells is reinforced if and only if the following two conditions are simultaneously satisfied:

1. The cell receiving the connection is responding the strongest among the cells in its vicinity.
2. The cell sending out the connection is also responding.

This principle can also be expressed as follows:

Among the cells situated in a certain small area, only the one which is responding the strongest has its input connections reinforced. The amount of reinforcement of each input connection to this maximum-output cell is proportional to the intensity of the response of the cell from which the relevant connection is leading.

In the neocognitron, this principle is applied to the variable input connections converging to feature-extracting S-cells. It should be noted that both excitatory and inhibitory connections are reinforced following this principle.

Figure 5 illustrates the connections converging to an S-cell. The S-cell receives variable excitatory connections leading from a group of C-cells of the preceding layer. It also receives a variable inhibitory connection leading from a subsidiary inhibitory cell, which is called a V-cell. The V-cell receives fixed excitatory connections from the same group of C-cells as this S-cell does, and is always responding with the average intensity of the output of the C-cells.

As the result of this network structure and the learning principle, the variable excitatory connections to the maximum-output S-cell grow so as to work as a "template" which exactly matches the spatial distribution of the response of the cells of the preceding layer. Thus, the maximum-output S-cell comes to acquire the ability to extract the feature of the stimulus which has been presented during the training period. In other words, through the excitatory connections, the S-cell receives signals indicating the existence of the relevant feature to be extracted. If an irrelevant feature is presented,

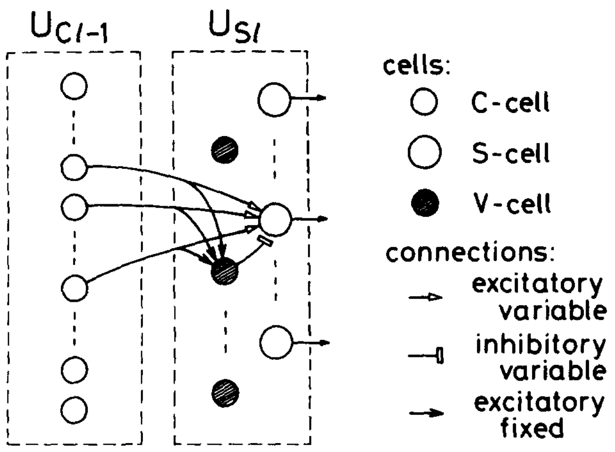


FIGURE 5. Connections converging to a feature-extracting S-cell.

however, the inhibitory signal from the V-cell becomes stronger than the direct excitatory signals from C-cells, and the response of the S-cell is suppressed. Thus, the S-cell is activated only when the relevant feature is presented. Incidentally, the V-cell can be said to be watching for the existence of irrelevant features. Thus, inhibitory V-cells play an important role in endowing the feature-extracting S-cells with the ability to differentiate irrelevant features, and in increasing the selectivity of feature extraction.

According to this principle, among the S-cells in a certain small area, only one cell which happens to yield the maximum output is selected to have its input connections reinforced. Because of the "winner-takes-all" nature of this principle, the duplicated formation of cells which extract the same feature does not occur, and the formation of a redundant network can be prevented. This situation resembles, so to speak, "elite education": Only the one cell which gives the best response to a training stimulus is selected, and only that cell is reinforced so as to respond more appropriately to the stimulus.

Once a cell is selected and reinforced to respond to a feature, the cell usually loses its responsiveness to other features. When a different feature is presented, usually a different cell yields the maximum output and has its input connections reinforced. Thus, "division of labor" among the cells goes on automatically.

With this principle, the network also develops a self-repairing function. If a cell which has been strongly responding to a stimulus is damaged and ceases to respond, another cell, which happens to respond more strongly than other cells, starts to grow and substitute for the damaged cell. Incidentally, the growth of a second cell has been prevented until then, because of the larger response of the first cell.

3.1.2 Development of Iterative Connections. The second principle introduced for the self-organization of the neocognitron is that the maximum-output cell not only grows, but also controls the growth of neighboring cells. In other words, the maximum-output cell works, so to speak, like a seed in crystal growth, and neighboring cells have their input connections reinforced in the same way as the "seed cell." The process of selecting seed cells will be discussed below in more detail.

Here, we define a term "hypercolumn": a hypercolumn is defined here as a group of S-cells in a layer whose receptive fields are situated at approximately the same position. In other words, each hypercolumn contains all kinds of feature extracting cells in it, and these cells extract features from approximately the same place in the input layer. Incidentally, if we rearrange the cell-planes of a layer and stack them in a manner shown in Figure 6, the cells of a hypercolumn constitute a columnar structure. Each hypercolumn contains cells from all the cell-planes.

Now, let a training pattern be presented to the network. From each hypercolumn, the S-cell which happens to respond the strongest is chosen as a candidate for seed cells. When two candidates or more appear in one and the same cell-plane, only the one whose response is the largest is selected as the seed cell of that cell-plane. When only one candidate appears in a cell-plane, the candidate automatically becomes the seed cell of that cell-plane. If no candidate appears in a cell-plane, no seed cell is selected from that cell-plane this time.

Thus, at most one seed cell is selected from each cell-plane of S-cells at a time. Usually, a different cell becomes a seed cell when a different training pattern is given.

When a seed cell is selected from a cell-plane, all the other S-cells in the cell-plane grow so as to have input connections of the same spatial distribution as the seed cell. As the result, all the S-cells in a cell-plane grow to receive input connections of the identical spatial distribution where only the positions of the preceding

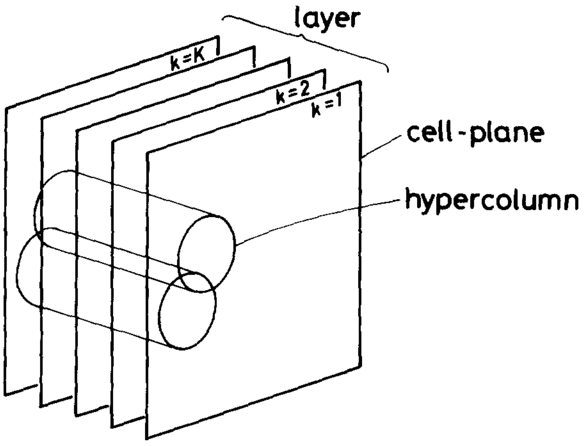


FIGURE 6. Relation between cell-planes and hypercolumns within a layer.

C-cells are shifted in parallel from cell to cell, as illustrated in Figure 3. In other words, connections develop iteratively in a cell-plane. Hence, all the S-cells in the cell-plane come to respond selectively to a particular feature, and differences between these cells arise only from difference in position of the feature to be extracted.

If the strength of all the variable connections is zero at the initial state before learning, self-organization of the network cannot start, because no cell can respond to the training pattern and maximum-output cells (or seed cells) cannot be selected. Hence, it is made that all the variable excitatory connections unconditionally get a very small value only when self-organization is going to start. In other words, each S-cell temporarily has very weak and diffused excitatory input connections only at the initial period of the self-organization. Once a reinforcement of the input connections begins, these weak and diffused initial connections are made to disappear. Incidentally, this situation coincides with the anatomical observation that, in the developing nervous system, synaptic connections between neurons are overproduced initially and the redundant axons are gradually eliminated afterwards.

If the period of generation of these temporary weak diffused connections is delayed a little for the cells of higher stages, self-organization of the network can be performed efficiently. Specifically, it is desirable to delay it until the growth of the cells of the preceding stage has been settled.

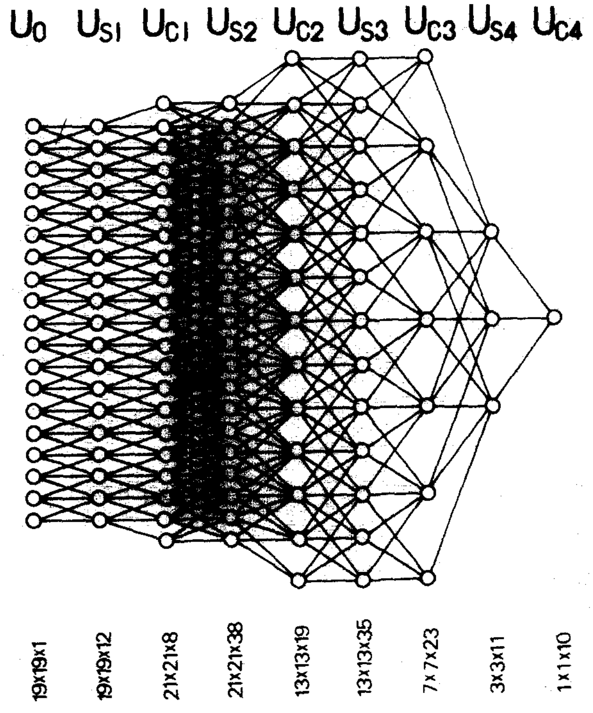


FIGURE 7. One-dimensional view of interconnections between cells of different cell-planes. Only one cell-plane is drawn in each layer.

3.2 Learning with a Teacher

As has been discussed above, in the case of learning-*without*-a-teacher, maximum-output cells are selected automatically as "seed cells." In the case of learning-*with*-a-teacher, however, "teacher" points out which cells should be the seed cells for each training pattern. The other process of learning is identical to that of the learning-without-a-teacher. It is, of course, not necessary to perform such a complicated procedure as calculating and adjusting the strength of all the connections one by one, but it is enough to point out which patterns or features should be extracted by which cells.

Learning-with-a-teacher is useful when we want to train a system to recognize, for instance, hand-written characters which should be classified not only on the basis of similarity in shape but also on the basis of certain conventions. For example, the geometrical similarity between "O" and "*Ø*" is about the same as that between "O" and "*Q*," but "O" and "*Ø*" must be recognized as the same character, while "O" and "*Q*" must be classified into different categories. It is impossible to train the system to recognize these characters by learning-without-a-teacher only, by which characters are classified only on the basis of geometrical similarity.

4. HANDWRITTEN NUMERAL RECOGNITION

In order to demonstrate the ability of the neocognitron, we have designed a system which recognizes hand-written numerals from "0" to "9." This system, a modification from an old system (Fukushima, Miyake, & Ito, 1983), has been implemented on a minicomputer (micro VAX-II) with an array processor (FPS-5105). The same system has also been implemented on a microcomputer (NEC PC-9801) which has a 16-bit main processor 8086 (with 384 kBytes memory) and a co-processor 8087 (Fukushima, Miyake, Ito, & Kouno, 1987).

The system has been trained by learning-with-a-teacher. Incidentally, experiments for learning-without-a-teacher have been reported elsewhere (Fukushima, 1980; Fukushima & Miyake, 1982).

4.1 Detailed Structure of the Network

The network has four stages of layers of S- and C-cells. The number of S- or C-cells in each layer is indicated in Figure 2. Layer U_{C4} at the highest stage has ten cell-planes, each of which has only one C-cell. These ten C-cells correspond to ten numeral patterns from "0" to "9."

Figure 7 shows how the cells of different cell-planes are spatially interconnected. This figure, in which only one cell-plane is drawn for each layer, illustrates a one-dimensional cross-section of the connections between

S- and C-cells. From this figure, we can read, for example, an S-cell of layer U_{S3} has 5 × 5 (excitatory) variable input connections from each cell-plane of layer U_{C2}. Since layer U_{C2} has 19 cell-planes, the maximum possible number of the variable input connections to each S-cell of layer U_{S3} is 5 × 5 × 19. It is important to note, however, that all of these 5 × 5 × 19 variable connections are not necessarily reinforced by learning. On the contrary, most of them usually remain at the initial state of strength of zero even after finishing learning. Since the variable connections of strength of zero need not be actually wired in the network, the effective number of connections are far less than the value directly read from this figure.

The output of each cell in the network is mathematically described below. In the following equations, notation $\mathrm{u}_{Sl}(\boldsymbol{n}, k)$, for example, is used to denote the output of an S-cell in the lth stage, where $\boldsymbol{n}$ is a two-dimensional set of coordinates indicating the position of the cell's receptive-field center in the input layer U_0, and k is a serial number of the cell-plane. For S-cells, k is in the range of $1 \leqq k \leqq K_{Sl}$, and for C-cells it is in the range of $1 \leqq k \leqq K_{Cl}$.

The output of an S-cell is given by

$$\mathrm{u}_{Sl}(\boldsymbol{n}, k) = r_l \cdot \varphi\left[\frac{\sigma_l + \sum_{\kappa=1}^{K_{Cl-1}} \sum_{\boldsymbol{\nu}\in A_l} a_l(\boldsymbol{\nu}, \kappa, k) \cdot \mathrm{u}_{Cl-1}(\boldsymbol{n} + \boldsymbol{\nu}, \kappa)}{\sigma_l + \frac{r_l}{1 + r_l} \cdot b_l(k) \cdot \mathrm{u}_{Vl}(\boldsymbol{n})} - 1\right] \tag{1}$$

where

$$\varphi[x] = \begin{cases} x & \text{if} \quad x \geqq 0 \\ 0 & \text{if} \quad x < 0. \end{cases} \tag{2}$$

In the case of $l = 1$ in (1), $\mathrm{u}_{Cl-1}(\boldsymbol{n}, \kappa)$ stands for $\mathrm{u}_0(\boldsymbol{n})$ or the output of a receptor cell of the input layer, and we have $K_{Cl-1} = 1$. Parameter σ_l is a positive constant determining the level at which saturation starts in the input-to-output characteristic of the S-cell.

$a_l(\boldsymbol{\nu}, \kappa, k)$ ($\geqq 0$) is the strength of the variable excitatory connection coming from C-cell $\mathrm{u}_{Cl-1}(\boldsymbol{n} + \boldsymbol{\nu}, \kappa)$ in the preceding layer, and A_l denotes the summation range of $\boldsymbol{\nu}$, that is, the size of the spatial spread of the input connections to one S-cell. $b_l(k)$ ($\geqq 0$) is the strength of the variable inhibitory connection coming from subsidiary V-cell $\mathrm{u}_{Vl}(\boldsymbol{n})$. As discussed before in connection with Figure 3, all the S-cells in a cell-plane have an identical set of input connections. Hence, $a_l(\boldsymbol{\nu}, \kappa, k)$ and $b_lk)$ do not contain argument $\boldsymbol{n}$ representing the position of the receptive field of the cell $\mathrm{u}_{Sl}(\boldsymbol{n}, k)$.

As can be seen from (1), the inhibitory input to this cell acts in a shunting manner. The positive constant r_l determines the efficiency of the inhibitory input to this cell.

The subsidiary V-cell which sends an inhibitory signal to this S-cell yields an output equal to the weighted root-mean-square of the signals from the preceding C-cells, that is,

$$\mathrm{u}_{Vl}(\boldsymbol{n}) = \sqrt{\sum_{\kappa=1}^{K_{Cl-1}} \sum_{\boldsymbol{\nu}\in A_l} c_l(\boldsymbol{\nu}) \cdot \{\mathrm{U}_{Cl-1}(\boldsymbol{n} + \boldsymbol{\nu}, \kappa)\}^2}\,, \tag{3}$$

where $c_l(\boldsymbol{\nu})$ represents the strength of the fixed excitatory connections, and is a monotonically decreasing function of $|\boldsymbol{\nu}|$, which satisfies

$$\sum_{\kappa=1}^{K_{Cl-1}} \sum_{\boldsymbol{\nu}\in A_l} c_l(\boldsymbol{\nu}) = 1. \tag{4}$$

The role of the root-mean-square cells in feature extraction is discussed elsewhere (Fukushima, 1981; Fukushima & Miyake, 1982).

The variable connections $a_l(\boldsymbol{\nu}, \kappa, k)$ and $b_l(k)$ are reinforced depending on the intensity of the input to the "seed cell," which is pointed out by the "teacher." Let $\mathrm{u}_{Sl}(\hat{\boldsymbol{n}}, \hat{k})$ be selected as a seed cell at a certain time. The variable connections $a_l(\boldsymbol{\nu}, \kappa, \hat{k})$ and $b_l(\hat{k})$ to this seed cell, and consequently to all the S-cells in the same cell-plane as the seed cell, are reinforced by the following amount:

$$\Delta a_l(\boldsymbol{\nu}, \kappa, \hat{k}) = q_l \cdot c_l(\boldsymbol{\nu}) \cdot \mathrm{u}_{Cl-1}(\hat{\boldsymbol{n}} + \boldsymbol{\nu}, \kappa), \tag{5}$$

$$\Delta b_l(\hat{k}) = q_l \cdot \mathrm{u}_{Vl}(\hat{\boldsymbol{n}}), \tag{6}$$

where q_l is a positive constant determining the speed of reinforcement. In the case of learning-with-a-teacher, a sufficiently large value is given to q_l so that the reinforcement of the input connections to each seed cell can be completed in a few steps of training-pattern presentation.

The output of a C-cell, which is inserted in the network to allow for positional errors, is given by

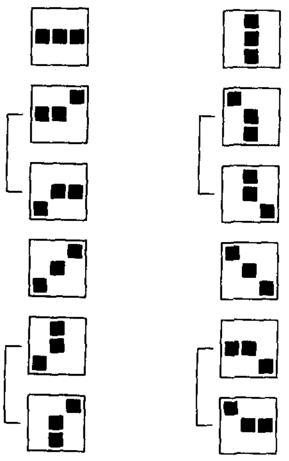


FIGURE 8. Training patterns used to train the 12 cell-planes of layer U_{S1}. Each hooked line shows that the outputs of the corresponding S-cell-planes are joined together and converge to a single C-cell-plane at U_{C1}. Only the central 3 × 3 area of each training pattern is shown, because the outside of this area is not effective for the training of layer U_{S1}.

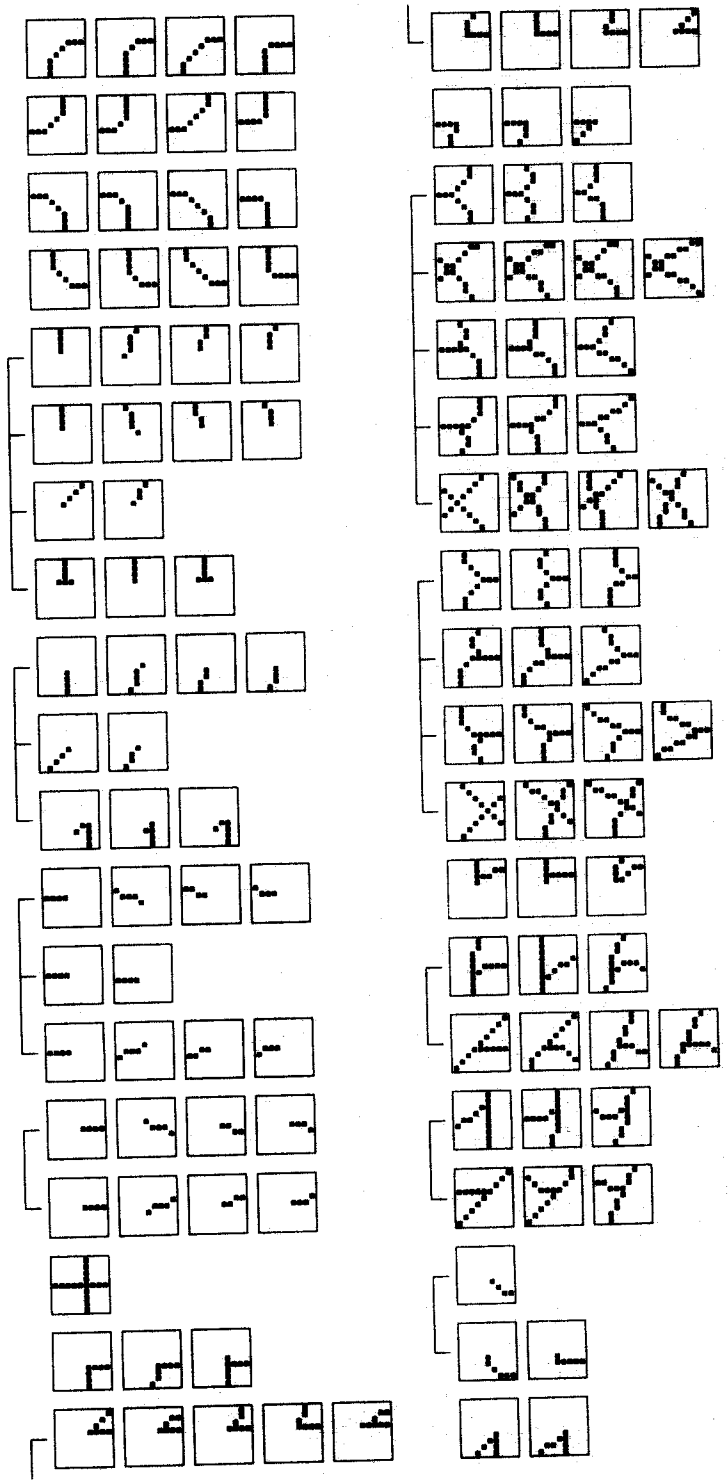


FIGURE 9. Training patterns used to train the 38 cell-planes of layer U_{S2}.

$$u_{Cl}(\boldsymbol{n}, k) = \psi\Big[\sum_{\kappa=1}^{K_{Sl}} j_l(\kappa, k) \sum_{\nu \in D_l} d_l(\nu) \cdot u_{Sl}(\boldsymbol{n} + \nu, \kappa)\Big], \tag{7}$$

where $\psi[\ \ ]$ is a function specifying the characteristic of saturation of the C-cell, and is defined by

$$\psi[x] = \frac{\varphi[x]}{1 + \varphi[x]}. \tag{8}$$

In this network, which is to be trained by learning-with-a-teacher, output of several numbers of S-cell-planes sometimes converges together to a single C-cell-plane. This condition of convergence, which will be discussed later, is represented by $j_l(\kappa, k)$.

Parameter $d_l(\nu)$ denotes the strength of fixed excitatory connections, and is a monotonically decreasing function of $|\nu|$. Hence, if $j_l(\kappa, k) > 0$ and if at least one

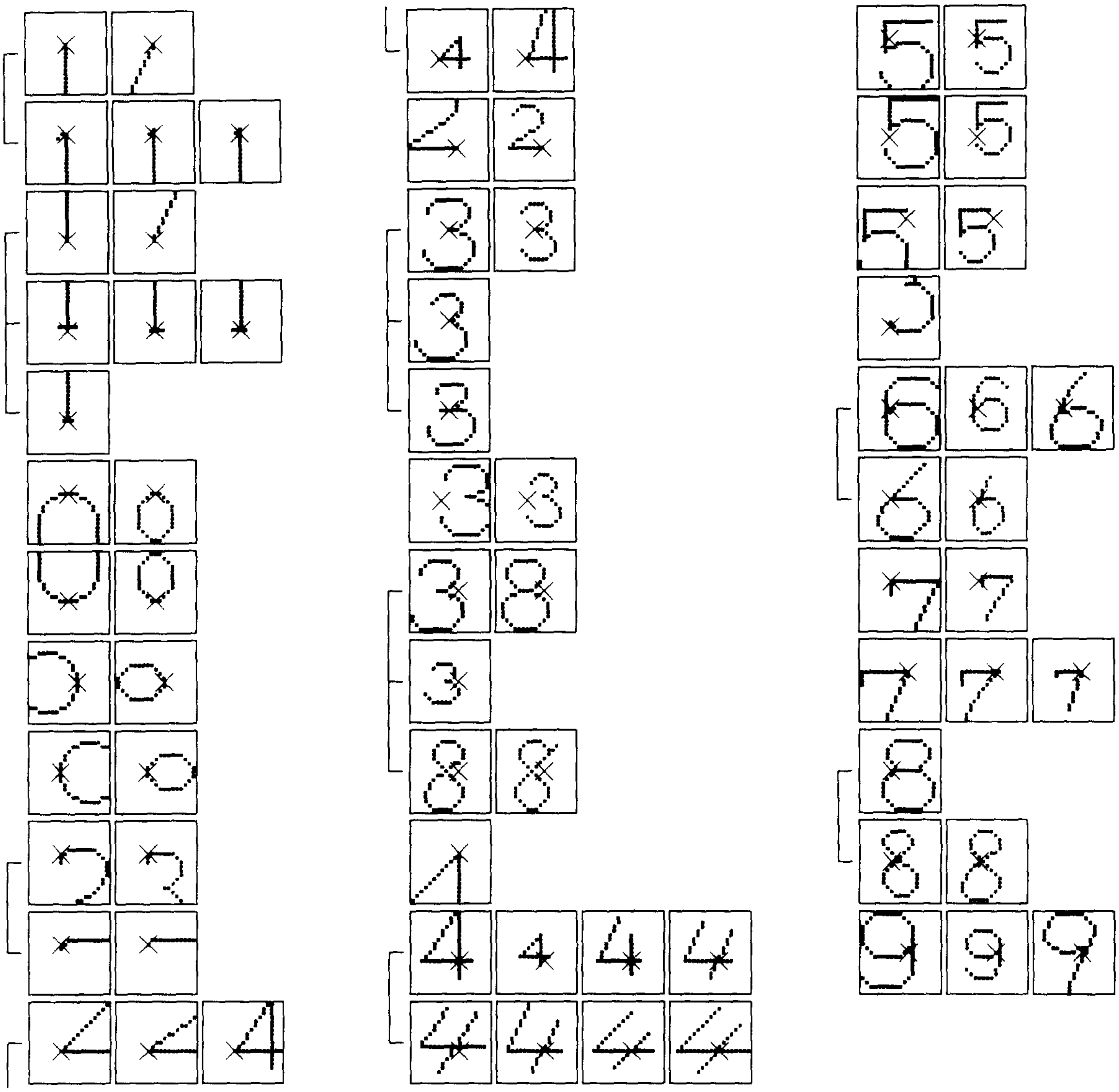


FIGURE 10. Training patterns used to train the 35 cell-planes of layer U_{S3}. Receptive field centers of the seed-cells are marked by crosses in the figure.

S-cell is activated in the area D_l of the κth cell-plane, to which these connections spread, this C-cell is also activated.

Now we will discuss $j_l(\kappa, k)$ in more detail. In the case of character recognition, even characters of different styles of writing have to be correctly recognized. In other words, input characters have to be classified not only on the basis of geometrical similarity but also on the basis of customs by which some particular kinds of large deformation are admitted. Sometimes when such deformation is too large, a single S-cell-plane is not enough to extract deformed versions of a feature. In such a case, another S-cell-plane is used to extract a deformed version of the feature, and the output from these S-cell-planes are made to converge to a single C-cell-plane. It is $j_l(\kappa, k)$ in (7) that represents this joining process[1]. Depending on whether or not the kth C-cell-plane receives signals from the κth S-cell-plane, $j_l(\kappa, k)$ takes a positive value or zero, respectively. Hence, for each κ, $j_l(\kappa, k)$ is usually zero except for one particular value of k.

4.2 Training the Network

In order to train the network, "teacher" presents a set of training patterns and points out which cells should

[1] For a network which is to be trained by learning-without-a-teacher (Fukushima, 1980; Fukushima & Miyake, 1982) joining of the output of different cell-planes are not made. Hence, U_{Sl} and U_{Cl} has the same number of cell-planes (or $K_{Sl} = K_{Cl}$), and

$$j_l(\kappa, k) = \begin{cases} 1 & \text{for} \quad \kappa = k \\ 0 & \text{for} \quad \kappa \neq k. \end{cases}$$

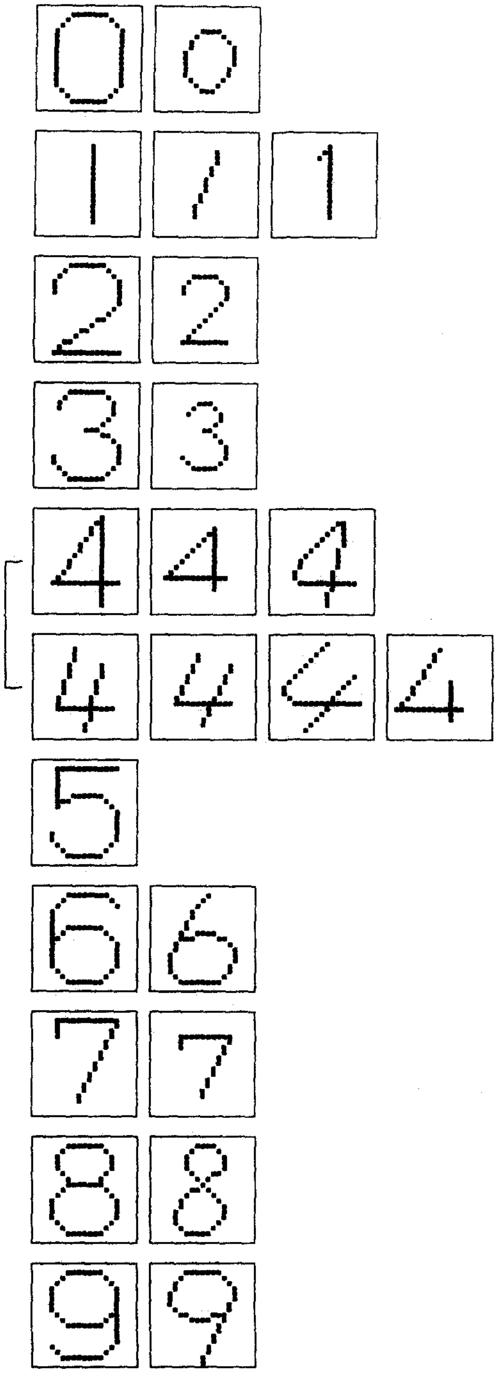


FIGURE 11. Training patterns used to train the 11 cell-planes of layer U_{S4}.

be the seed cells for each training pattern. The other process of learning goes on automatically.

Training has been performed step by step from lower stages to higher stages. In other words, training of a higher stage is performed after completely finishing the training of the preceding stages.

In the following example, the network is trained to recognize handwritten numerals from "0" to "9."

4.2.1 Training of Layer U_{S1}. Layer U_{S1} is trained to extract line components of different orientations. Figure 8 shows the 12 training patterns used to train the 12 cell-planes of layer U_{S1}. Each of the training patterns is presented to the network only once. The cell at the center of the cell-plane to be trained is always appointed as the seed cell. As can be seen from Figure 7, each cell of this layer has a receptive field of 3 × 3 in size. Hence, only central 3 × 3 area of each training pattern is effective for training, and only this central area is shown in Figure 8.

Since the size of receptive fields of S-cells is as small as 3 × 3, it is difficult to extract all parts of a line by only one cell-plane when the line has an inclination of 1:2. Hence, two cell-planes are used to extract such a slanted line component, and the output from these two S-cell-planes are joined together and made to converge to a single C-cell-plane. Each hooked line drawn to the left of the training patterns in Figure 8 shows how the outputs of the corresponding S-cell-planes are joined together.

4.2.2 Training of Layer U_{S2}. Figure 9 shows the training patterns used to train the 38 cell-planes of layer U_{S2}. Only the central 9 × 9 areas of the training patterns are shown here, because S-cells of this layer have receptive-fields of 9 × 9 in size. Again, the cell at the center of the cell-plane to be trained is appointed as the seed cell.

Sometimes, a single cell-plane is trained with more than one training pattern. This is effective to increase S-cell's ability to extract deformed features. A group of patterns arranged in a horizontal line in Figure 9 represents such a set of training patterns.

Similarly, as for layer U_{S1}, output of several S-cell-planes are joined together as indicated by hooked lines in Figure 9.

As can be seen from Figure 9, each training pattern consists of a part of a numeral pattern which is supposed

FIGURE 12. Experiment of handwritten numeral recognition on a minicomputer.

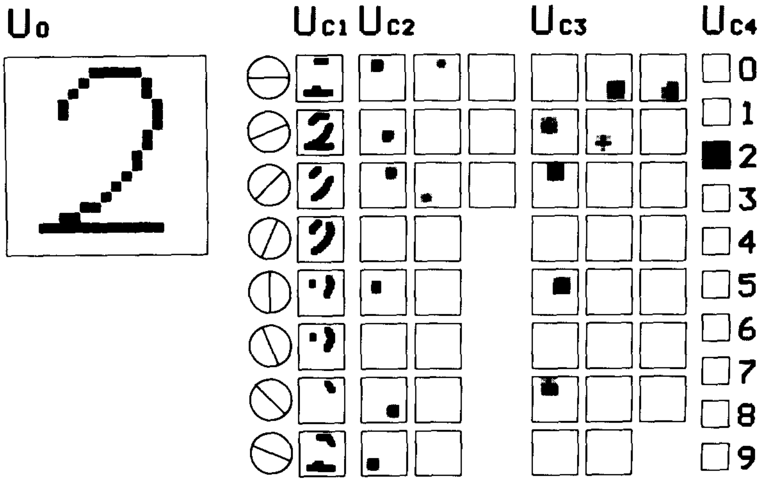


FIGURE 13. An example of the response of the C-cells in the network trained to recognize handwritten numerals.

to appear during the process of pattern-recognition. In other words, typical examples of deformed patterns are presented to the network as training patterns.

Generally speaking, good selection of training patterns is the most important for layer U_{S2} among all layers. If the training patterns for layer U_{S2} are properly selected, the network usually acquires a considerably high ability of pattern recognition, even though the selection of the training patterns for other layers is not so complete.

4.2.3 Training of Layer U_{S3}. Figure 10 shows the training patterns used to train the 35 cell-planes of layer U_{S3}. Since the receptive fields of S-cells of this layer are larger in size than the input layer U_0, the cell at the center of a cell-plane cannot always be appointed as the seed cell. Then, the position of the seed cell (that is, the receptive-field center of the seed-cell) is marked by a cross in each training pattern in Figure 10.

4.2.4 Training of Layer U_{S4}. Figure 11 shows the training patterns used to train the 11 cell-planes of layer U_{S4}. Similarly as U_{S1} and U_{S2}, the cell at the center of each cell-plane is appointed as the seed cell.

Since it is difficult to recognize "4" and "4" with a single cell-plane only, two cell-planes are used. For other numerals, however, one cell-plane is enough to recognize even deformed versions of the pattern written in different styles. For instance, both "9" and "9" are correctly recognized by one and the same cell-plane. This is because most of the distortions in shape of the input pattern have already been absorbed during the process in the previous stages.

4.3 Response of the Network

Now the response of the network which has finished learning is tested. In this experiment, the input pattern is drawn on a magnetic tablet, as shown in Figure 12. Although a tablet is used to input hand-written numerals, the system does not use any temporal information about the order of the strokes of the character. The character which has already been drawn is used as the input pattern for the network. With the progress of calculation in the computer, the response of the layers of C-cells is displayed successively on a graphic terminal. Figure 13 shows an example of this display. To the input layer U_0, a numeral "2" is presented. In the highest layer U_{C4}, shown at the extreme right, only cell "2" is activated. This means that the neocognitron recognizes the input pattern correctly.

Figure 14 shows some example of deformed input patterns which the neocognitron has recognized correctly. It is a matter of course that the neocognitron recognizes these patterns correctly even though they are shifted in position. When an input pattern is presented in a different position, the response of cells in intermediate layers, especially those near the input layer, varies with the shift. However, the higher the layer is, the smaller is the variation in response. The cells of the highest layer are not affected at all by a shift in position of the input pattern.

It has also been shown that even where the input pattern has been increased or diminished in size, or is skewed in shape, the response of the cells of the highest layer is not affected. Sometimes, when the input pattern has been distorted too much, the response of the cells in the highest layer is weak, but still a response is elicited from the correct cell. Even though the input pattern has some parts missing or is contaminated by noise, the neocognitron recognizes it correctly.

5. DISCUSSION

As has been shown here, the neocognitron has many remarkable properties which most modern computers

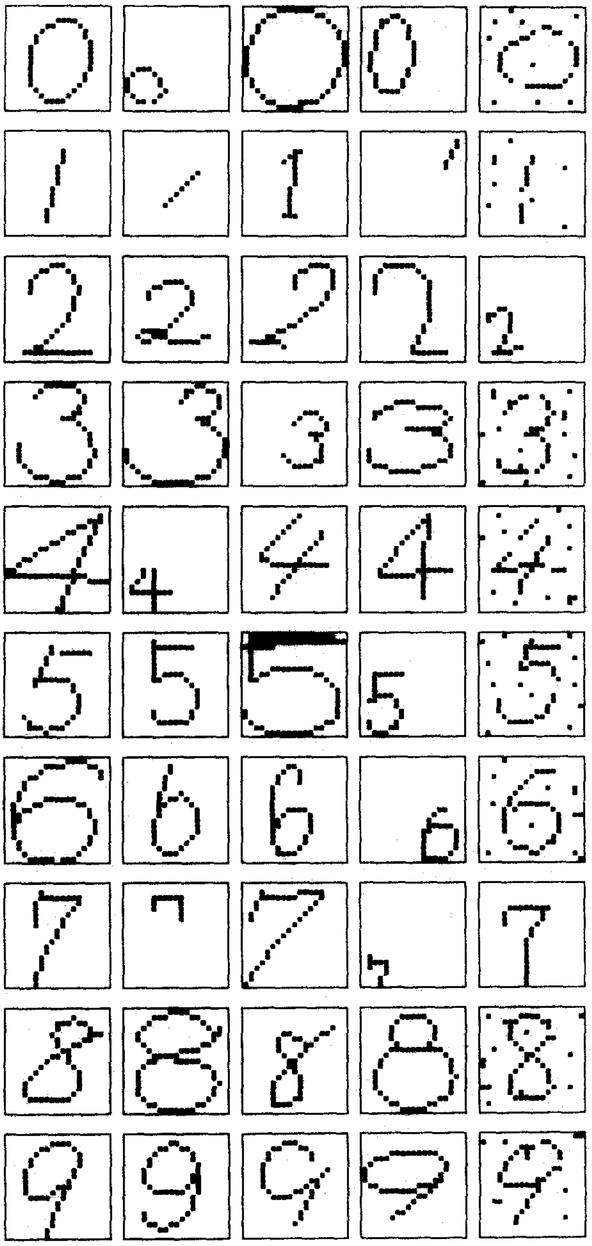


FIGURE 14. Some example of deformed input patterns which the neocognitron has recognized correctly.

and pattern-recognizers do not possess. Since the neocognitron can learn, it can be trained to recognize not only Arabic numerals, but also other sets of patterns, like letters of the alphabet, geometrical shapes, or others. Hence, it is possible to design a neocognitron as a universal pattern-recognizer, which can be used, after training, for an individual purpose.

If the number of categories of the patterns to be recognized is increased, the number of cell-planes in each layer of the network also has to be increased. The number of cell-planes, however, need not be increased in proportion to the number of categories of the patterns. It is enough to increase it in less than linear proportion, because local features to be extracted at lower stages are usually contained in common in patterns of different categories.

If we want to construct a system which can recognize more complex patterns like Chinese characters, it is recommended to increase the number of stages (or layers) in the network depending on the complexity of the patterns to be recognized.

The principles of the neocognitron are not restricted to the processing of visual information only, but can also be applied to other sensory information. For example, it would be possible to construct a speech-recognition system with a little modification.

Although the neocognitron has forward (i.e., afferent or bottom-up) connections only, the information-processing ability of the network can be greatly increased if backward (i.e., efferent or top-down) connections are added. The model of selective attention recently proposed by the author (Fukushima, 1986) is an example of such an advanced system. We are continuing the research, and we hope to develop an artificial brain closer to the human brain.

REFERENCES

Bruce, C., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology,* **46**(2), 369–384.

Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics,* **20**(3/4), 121–136.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics,* **36**(4), 193–202.

Fukushima, K. (1981). *Cognitron: A self-organizing multilayered neural network model* (NHK Technical Monograph No. 30), Tokyo: NHK Technical Research Laboratories.

Fukushima, K. (1986). A neural network model for selective attention in visual pattern recognition. *Biological Cybernetics,* **55**(1), 5–15.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition,* **15**(6), 455–469.

Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechnism of visual pattern recognition. *IEEE Transactions on Syst. Man Cybernetics,* **SMC-13**(5), 826–834.

Fukushima, K., Miyake, S., Ito, T., & Kouno, T. (1987). Handwritten numeral recognition by the algorithm of the neocognitron—An experimental system using a microcomputer (in Japanese). *Transactions of the Information Processing Society of Japan,* **28**(6), 627–635.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal de Physiologie,* **160**(1), 106–154.

Sato, T., Kawamura, T., & Iwai, E. (1980). Responsiveness of inferotemporal single units to visual pattern stimuli in monkeys performing discrimination. *Experimental Brain Research,* **38**(3), 313–319.