MULTIPLE CHARACTER RECOGNITION - A SIMULATION MODEL

Sebastian Koebe & Gerhard Deffner[1]

University of Hamburg

In this article we describe a model which simulates the process of human recognition of handwritten characters. For pragmatic reasons, our first attempt is limited to the subset of the 10 digits. Work on this problem is not new, however, and the main requirements of successful recognition are well known: a model should account for the human ability to recognize characters:

(1) in all possible positions in a given display
(2) in a display containing multiple objects (numbers)
(3) of different sizes
(4) of varying shape and form
(5) if they are distorted (discontinuities of lines)
(6) in any orientation
(7) if they overlap

Our model can cope with problem 1 through 4, and 5 to some extent. It is a hybrid system in which the use of serial vs. parallel processes is contingent upon assumptions based on empirical data. In general, it can be described as an early selection system, recognizing numbers in a serial order, after parallel information about the whole visual field is utilized early in the process. This type of capacity limited recognition model first uses positional information about objects in the input as the basis for further object selection, then attention is focused on single items to perform the computationally expensive process of recognition.

The main idea is that requirements 1-3 are dealt with by a process of selecting and standardizing individual objects from the display. Recognition is achieved through a PDP-network. These procedures are called iteratively until all objects are recognized.

The input to the system is a 400 x 200 pixel array produced with a paint program tool. This matrix of binary units stands

---

[1] We want to thank the Fulbright Commission for providing a grant to the first author to spend a year at the Cognitive Science Institute at LaJolla and David Rumelhart who was supervisor during that year and who provided for ideas and encouragement to enter the PDP-paradigm.

for the output of the foveal part of the retina covering the centre of the visual field. Although the output from receptors in the retina is frequency-modulated with respect to the intensity of a stimulus (behavior over time), we assume a static pattern. The matrix of binary units can be thought of as a 'snapshot' of the firing pattern in the retina containing the relevant information for the process of pattern recognition. Another reason for working with a static array is the fact that only such information from the visual field is processed semantically which comes from the periods of relative rest of the eye during fixations and not from saccadic eye-movements.

The fact that the information processing capabilitity of the visual system is different for syntactic and semantic information (c.f. Rayner, 1975) has motivated our choice of two mechanisms in the model. The performance of the first is similar to the recognition of syntactic information by the human visual system. In the same vein that features such as word length or size of letters can be made out at some distance from the fixated point, we assume that objects can be picked out from the array we use. After this selection, another mechanism takes over - that of character recognition.

1. Component processes of the model

1.1. Deriving a positional map

The first step in analyzing a given binary array is to determine the number of objects and their position in the input. This is done by a fast parallel algorithm (assuming parallel hardware as in the brain), which computes the center of gravity for each (potential) object in the display. Consider two fully connected layers of units, with the first layer being the retina receptors and the second layer being a one-to-one map of the first layer with the same number of units. Every single receptor in the first layer, which can have only one of two possible states ('on' or 'off'), is connected to every unit in the second layer (Figure 1). When a receptor is active ('on' in the binary array), it sends activation to all of the units in the second layer according to the weights of the connections. The weight between a unit in the first layer and all other units in the second layer is set according to the Gaussian function of the distance between the units. Every unit in the second layer receives weighted input from all units in the first layer. All incoming activation is summed, thus yielding a total activation value for each unit in the second layer. This array of total activation can be illustrated as a map of activation values as in Figure 2. The local maxima in the landscape (the hilltops) represent the centers of gravity of objects in

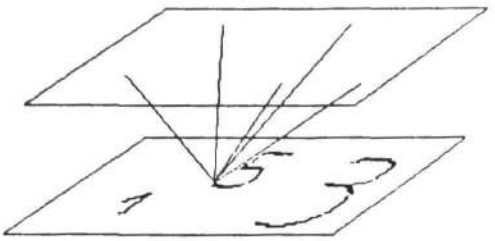

Figure 1: Weighted Connections

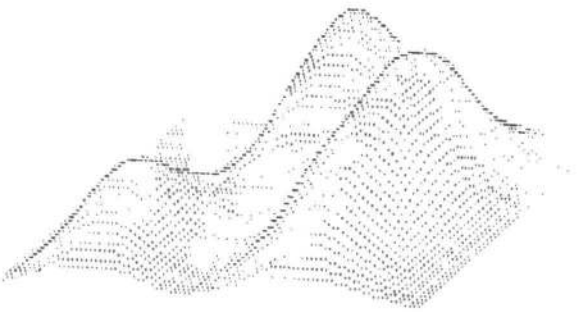

Figure 2: Activation landscape

the input, the number of local maxima is equal to the total number of objects. X- and Y-coordinates of these provide positional information for the next steps. By means of an attentional mechanism one of these is selected for further analysis, namely their recognition.

1.2. Standardisation of selected objects

Next, the object associated with the selected center of gravity has to be isolated in the input array. Given the position of the center of gravity for an object, all the adjacent 'on' units surrounding the center will constitute the isolated object. The criterion for adjacency in our model is very strict. Adjacency is at first defined as a distance of one unit. This means that only objects which are built out of directly adjacent active units are isolated as one. Other adjacency criteria can be chosen to allow for discontinuity of lines. Information about the height of local maxima can also be used to adjust the criterion. At present, our model uses a static criterion, however.

Once an object is isolated, its size can be determined as the maximum number of pixels horizontally and vertically. Next, the object is mapped onto a square matrix (see Figure 3a). The resolution of this matrix is then reduced to yield a standardised 10 by 10 bit matrix (Figure 3b). Arriving at such a final matrix, we have satisfied the above requirements concerning position and size.

1.3. Character recognition

On an abstract level, this task can be described as that of detecting inter-character variation in the face of intra-character variation (different shapes of the same digits). This can be thought of as a mapping problem. All input matrices

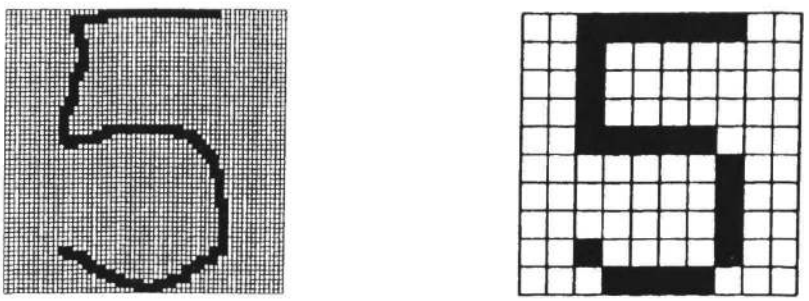

Figure 3: Selection at original (a) and reduced resolution (3)

containing the same digit (for instance "5") should be mapped onto the same output. Thus, there are 10 possible output states. The question is, what the representational form of the output should be. On the one extreme would be a relatively dense representation using 10 units, one of which is "on" for each digit ("grandmother cells"); on the other extreme would be 10 patterns of the same size as the input (10 x 10 units), each of which represents a prototypical digit. Dense discrete information is what we require as output from a recognition mechanism. At an intermediate level, however, the 10 x 10 representation is more plausible: for one, the number of possible output patterns is kept large, thus imposing no limits on the recognition capacity. Also, we want to separate a final decision phase (Sternberg, 1969) from the earlier process dealing with variations of shape that results in differential information about the presence of various features in the input array. In this way, a level is provided where featural (bottom-up) information can be combined with contextual information prior to the final decision about the identity of the stimulus (allowing for contextual effects like the Stroop phenomenon or context enhancment in reading). This distinct level is a prerequisite of more comprehensive models of human recognition as it is found for example in Rumelhart and McClelland's Interactive Activation Model of Context Effects in Letter Perception (1981).

Traditionally, shape variance has been dealt with through mechanisms of feature analysis (Selfridge & Neisser, 1960). A major problem has been that a comprehensive set of features must be defined by the designer of the system for each character set to be recognized. As a more flexible alternative, connectionist networks can be used which 'learn' features from material presented to them and then generalize to new input. After sufficient learning, the knowledge about relevant

features is embedded in the hidden layer of such networks and can be utilized in mapping the input matrix onto the 10 x 10 matrix of the intermediate level.

The connectionist network we use has three layers and employs backpropagation as the learning procedure (Rumelhart, Hinton & Williams, 1986). It consists of 100 input units, 50 hidden units and 100 output units with no direct connections from input to output. During learning, input patterns in the form of 100-element vectors (representing the 10 x 10 matrix) are mapped onto corresponding prototypical target vectors (standing for the ten prototypical digits). Since it is not practical to map all possible permutations ($2^{100} = 1.2676506 * 10^{30}$) of the input vector space on to the 10 target vectors, we took a sample of the input vector space, by asking 20 subjects to write digits on a computer screen. The same standardisation procedure as described earlier was used to transform their handwritten digits into the 10 by 10 standard form. In order to enlarge sample variability, white noise was added to the input vectors. A total of 1000 input vectors was used for learning.

After learning, the network can now be used to produce 100-element output vectors for any new input vector. The elements of the output vector can assume continous values between 0.0 and 1.0, thus reflecting the degree to which certain features are contained in the input material.

The decision which completes recognition is accomplished by relating this vector of continous values to the 10 prototypical binary vectors. Similarities between continous and prototypical vectors are computed and the name of that protoype is used as a label for the item to be recognized, for which similarity is greatest and above a preset threshold. If all similarities are below threshold, a decision of "no known character" is made. These labels provide the desired discrete output states.

2. Putting it all together

The interaction of the various mechanisms is illustrated in Figure 4. The 400 x 200 input bit matrix (1) is made available as a positional map (2) also. The main control structure is that of a loop in which Attention selects one position at a time and feeds coordinates (3) to Object Standardisation. Object Standardisation isolates the corresponding object from the input matrix and transforms it to the standard 10 x 10 matrix (4). The matrix is then fed to Character Recognition which outputs discrete characters. This is repeated (the digit "5" is used as an example of the first cycle) until the positional map is exhausted.

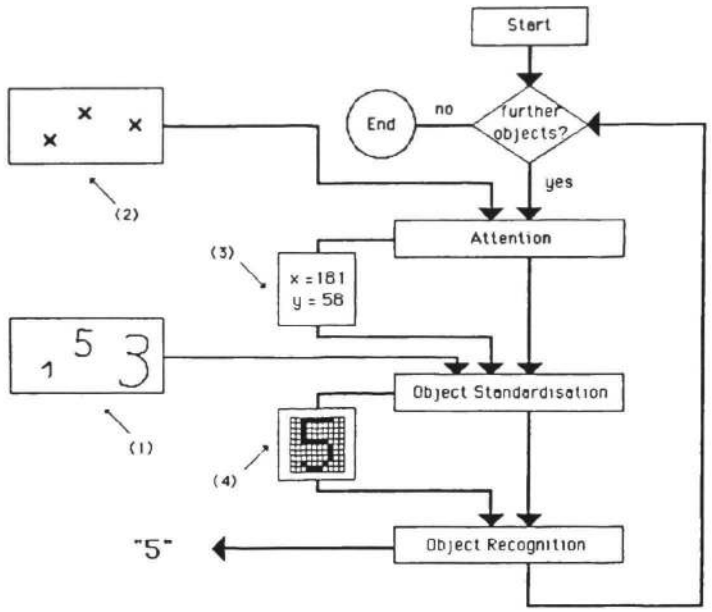

Figure 4: Interaction of the components and flow of control

3. Discussion

The main characteristics of the process of character recognition used in our model are: First, the limitation of recognition capacity available at a time (there are only 100 input units to the PDP-network). This limitation, imposing the requirement of an attentional mechanism for focussing on particular areas in the input, appears justified when we consider the enormous size of the initial input vector space of the 400 x 200 pixel matrix. Mappings from a space of 2^{80000} input vectors seem out of the question. Due to the selection of single items and subsequent reduction in resolution, the number of input units necessary for recognition can be reduced dramatically. The decision in favour of this practically inspired approach takes into account the trade-off between parallel processing capacity and time (fast parallel vs. slow serial recognition).

Second, we asssume automatic parallel processes providing information for the process of selecting characters from the visual field. This is the basis for early selection. Our model only computes a positional map of objects, but empirical data suggest that the human visual system also provides information like for instance colour and texture maps of the primary input. In an elaborate system, an attentional mechanism (which we did not detail) would have to integrate all such sources of information in order to control the selection of items in the input.

Third, in this model variance in the input is reduced stepwise. Before a featural analysis of characters is performed, position and size variations are eliminated. With regard to the recognition problems 6 and 7 (overlap and rotation of characters) the presented approach does not suggest obvious solutions.

In principle, the model can be extended to cover larger character sets. We would have to train the PDP-network on examples of new characters in order to generalize to the relevant structures[2]. It is interesting to note that the system learns through positive examples only. Information from the vast space of vectors not representing characters would not allow for a detection of regularities. The variation resulting from random sampling of this space is huge and unsystematic in relation to the variance of the extremely small percentage of vectors which do represent character.

References

Rayner. K.E. (1975) The perceptual span and peripheral cues in reading. Cognitive Psychology, 7, 65-81.

Rumelhart, D.E. & McClelland, J.L. (1981) An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect an some tests and extensions of the model. Psychological Review, 89, 60-94.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning internal representations by error propagation. In: D.E.Rumelhart & J.L. McClelland (eds.) Parallel distributed processing, Volume 1. Cambridge: MIT Press.

Selfridge, O.G. & Neisser, U. (1960) Pattern recognition by machine. Scientific American, 203, 60-68.

Sternberg, S. (1969) The discovery of processing stages: Extensions of Donders' method. In: W.G. Koster (ed.) Attention and Performance II. Amsterdam: North Holland.

---

[2] The resolution of a 10 x 10 standard matrix would probably be too small to represent letters with all important features. This could be overcome by increasing the size of the matrix.