Visual Search as Constraint Propagation

Peter A. Sandon
Berrin A. Yanikoglu

Computer Science Program
Dartmouth College

Abstract

A handful of prominent theories have been proposed to explain a large quantity of experimental data on visual attention. We are developing a connectionist network model of visual attention which provides an alternative theory of attention based on computational principles. In this paper, we describe aspects of the model relevant to the dependence of visual search times on display size (number of objects in the stimulus image). Duncan's stimulus similarity theory provides the characterization of the experimental data which we use in simulating and evaluating our model. The characteristics of the network model that support the continuously varying dependence of search time on display size are the constraint propagation search implemented by a winner-take-all mechanism in the attention layer, and the lateral inhibition network within each primitive feature map, which provides the feature contrast needed to filter out background textures. We report the results of simulations of the model, which agree with experimental data on visual attention in human subjects.

Introduction

Visual attention refers to the phenomenon of perceiving relevant parts of a visual stimulus, while ignoring irrelevant parts. Treisman's feature integration theory [Treisman & Gelade 1980] provides one explanation of attentional phenomena in terms of primitive feature maps and how they are combined in recognizing complex patterns. Duncan's stimulus similarity theory [Duncan & Humphreys 1989] is an alternative explanation which is based on two similarity measures applied to relevant and irrelevant parts of the stimulus. We have proposed a network based theory of these same phenomena [Sandon 1989,1990], which combines aspects of both feature integration and stimulus similarity, while providing a framework for achieving computational efficiency.

In this paper, we discuss a particular aspect of our theory, which is most relevant to Duncan's similarity based theory. In particular, while Treisman's theory involves discrete distinctions between parallel and serial visual search, Duncan's theory involves a continuously varying dependence of search times on display size. In our network model of attention, this dependence is explained in terms of feature maps with mutually inhibiting activations and the time course of the winner-take-all (WTA) mechanism responsible for selecting the attentional focus.

In the next section we review the details of the feature integration and stimulus similarity theories, and mention related network models of attention. The following sections describe our own network model and the results of some relevant computer simulations.

Background

Feature integration theory. Treisman and her colleagues [Treisman & Gelade 1980; Treisman & Schmidt 1982] have collected data on human visual performance for a variety of tasks where attention is implicated. In particular, her data suggest that in a visual search task, when the target to be detected differs from non-target distractors along a single primitive feature dimension, the target can be detected in an amount of time that does not depend on the number of distractors. Thus, it appears that the search for the target proceeds in parallel, with all objects, target and non-target alike being processed simultaneously. This is referred to as feature search. On the other hand, when the target is distinguished from the distractors by a combination of values in two feature dimensions, the amount of time required to detect the target increases linearly with the number of distractors in the input. Thus, it appears that the search for the target proceeds serially, with all objects being processed in sequence. This is referred to as conjunction search.

Stimulus similarity theory. Duncan provides an alternative theory to explain the reaction time dependencies of visual search [Duncan 1985; Duncan & Humphreys 1989]. Setting aside the parallel versus serial distinction, Duncan claims that the search times depend only on two measures: the similarity between the target and the nontarget distractors (T-N similarity), and the similarity among the nontargets (N-N similarity). Higher T-N similarity increases the search time for the target, while lower N-N similarity has an even more pronounced effect in increasing the search time. This is referred to as the stimulus similarity theory. Duncan also emphasizes the importance of object size in obtaining the reported attentional effects. In particular, the search time dependence on number of objects is itself dependent on the ratio of object size to retinal eccentricity of the object within the image.

Previous connectionist models. Prior to this work, a handful of network models have been proposed to explain individual pieces of the attentional data. Hinton and Lang [1985] used a network similar to Hinton's [1981] mapping network to simulate illusory conjunctions. Using a winner-take-all (WTA) array of processing elements to represent the attentional focus, they found that illusory conjunctions could be made to occur if a random input pattern was presented prior to settling of the WTA process in the attention array. Mozer [1988] used a similar network structure to simulate a probabilistic attentional mechanism having variable focus size in his MORSEL system. Sandon & Uhr [1988] used a network similar to Hinton's mapping network, but implemented the representation of location hierarchically, as a means of more efficiently representing and learning translation invariant object recognition. This hierarchical representation of location leads naturally to the idea of representing the attentional focus in a hierarchy.

The network model

We have designed a network model of visual attention based on constraints drawn from three fields: computational principles developed in the machine vision literature, knowledge of the neurophysiology of vision, and behavioral data. In this section, we summarize the overall model. Sandon [1990] discusses the considerations motivating the network design, and presents the results of other simulation experiments.

The underlying structure of the network is based on the original network used by Hinton [1981] to model translation (as well as rotation) invariant object recognition. The idea is to represent both shape features and spatial location of

objects in separate arrays, and to combine these two sources of information multiplicatively (using so-called conjunctive connections) in recognizing objects. This design provides a computational structure that allows for the representation of multiple competing hypotheses about the location and identity of objects, and produces an interpretation of the image which is most consistent with the dual set of constraints (location features and shape features).

In previous work [Sandon & Uhr 1988], we augmented a pyramid structured shape network with a location subnetwork to efficiently perform translation-invariant object recognition. We used a two layered hierarchical representation for the location subnetwork, which reduced the required connectivity, allowing us to train the network to recognize objects in given positions and then generalize to novel positions. The model proposed here uses a hierarchical representation of spatial location as the basis for an attentional mechanism. This structure automatically provides translation-invariant recognition of objects, since its underlying structure is that of the translation-invariant network. To this basic structure, we add a capability for multiple scale analysis, by providing separate pathways for processing the image at different levels of resolution.

For the purpose of recognizing relatively simple geometric shapes, as are generally used as stimuli for psychological and neurophysiological studies, a shape feature hierarchy such as that used by Uhr [1978] or Sabbah [1985] is appropriate. Since the evidence for various shape feature detectors in the human visual system is still a matter of debate [Treisman & Gelade 1980; Sagi 1988; McLeod, et. al. 1988; Duncan & Humphreys 1989], this model makes no *a priori* commitment to a particular set of features. Instead, we develop our feature set incrementally as we simulate more of the behavioral data.

We must also specify the set of features used to activate the attention layers. There are two aspects to be considered in addressing this problem. First, what are the attentional features themselves, and second, how do these features interact to produce the attentional activation? Regarding the features themselves, we again choose those that produce simulation results in agreement with the behavioral data. At the lowest layer of attention, oriented edges and lines are used. At the higher layer, perceptual grouping features, such as parallel and collinear lines, symmetry, and adjacent line terminations are appropriate [Lowe 1987; Witkin & Tenenbaum 1983]. To combine features for attentional input, we implement an interaction among like features prior to their introduction to the attention array. In particular, a central-excitatory, peripheral-inhibitory interaction is applied to each of the feature maps used as input to the attention array. This contrast enhancement of features produces input to the attention array only when a given feature occurs in the image in relative isolation from other features of the same type.

The attentional focus is determined by a WTA competition within the attention array. The effect of the attentional activity is to gate the features from a particular region of the image up to higher layers of the network, where object recognition occurs. As previously noted, the features comprising the input to this recognition process are location invariant. Similarly, these features are made scale invariant, by transforming each possible scale to a normalized size. The individual data paths representing the different processing scales are combined prior to recognition processing using a policy of global precedence [Hughes et. al. 1990]. This policy assures that, in the absence of other information, the lowest resolution data path that exhibits significant attentional activity is gated to the higher processing levels.

Figure 1a summarizes the attentional model described in this section. The leftmost data path is for fine scale processing and includes two levels of attention. Data paths are bidirectional, providing a pathway for attentional priming, and other task-directed responses. The middle data path is similar, but starts with a lower resolution intensity image, and requires only one attention layer to select features for processing by the recognition processor. The rightmost data path involves the coarsest resolution intensity image, whose features are passed directly to the recognition processor. These three data paths provide processing at three scales. The choice of which scale to process is made by the scale arbitrator, which implements the global precedence policy.

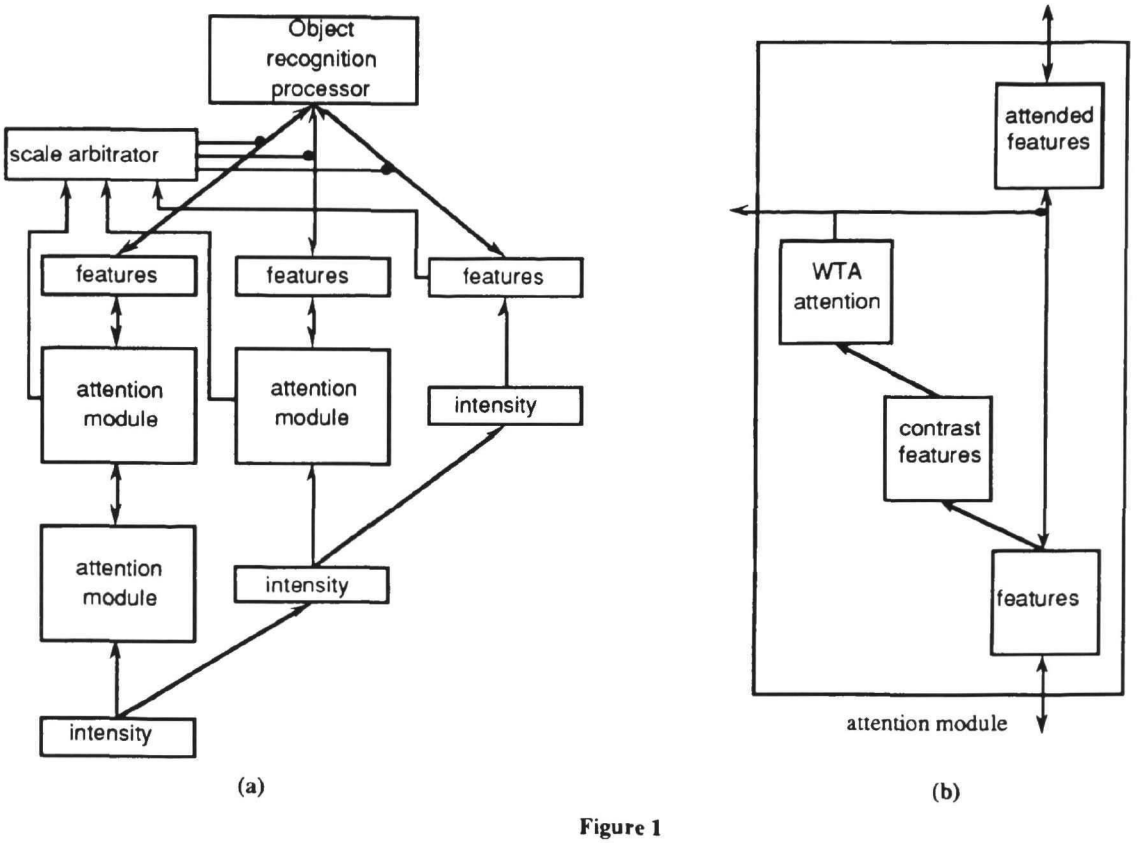

Figure 1

Simulation results

We have simulated a number of the experiments described by Duncan and Humphreys [1989] using a subnetwork of the model just described. The simulated network uses a 64 x 64 pixel intensity array to represent the input stimulus. We use only a single layer attention mechanism, but at two different scales. The primitive feature maps implemented in the simulation are four orientations of lines, and four orientations of 'L' corners. Each primitive feature map has an associated contrast feature map, which is computed by applying an on-center, off-surround lateral inhibition operator to the feature map. The activations of all contrast features are summed to provide the bottom-up initial activation of the attention arrays. To simulate the task-directed component of attentional activation, we weight the contributions of feature arrays associated with the known target object more than the remaining feature arrays when summing them. A WTA operation is applied to the

initial attentional activation, based on the following inhibition rule [Koch & Ullman 1985]:

$$dy_i / dt = y_i (x_i - \Sigma x_j y_j)$$
$$y_i (0) = (1/N) + \eta \quad ; \quad \eta \text{ is zero mean noise}$$

In this rule, the x_i are the attention layer activations, while the y_i are the activations of a set of auxiliary nodes. The rule constrains the y_i always to sum to 1. We have used a termination criterion that requires one of the y_i to exceed a value of 0.6, which is sufficient to guarantee that the corresponding region will win the competition. We report the number of WTA iterations required to reach criterion, as an indicator of the response time of the network. We do not include the recognition subnetwork in the simulation.

Duncan describes four extreme cases with respect to his similarity measures:

Case	T-N similarity	N-N similarity	dependence on display size
A	low	high	low
B	high	high	intermediate
C	low	low	low
D	high	low	high

We have simulated the experiments reported in [Duncan & Humphreys 1989] for each of the above four cases, using display sizes of 4 and 6. In addition, we have repeated these experiments for display sizes of 15 and 20. The displays in Figure 2 show the stimulus pattern(i) and initial attention activation(ii) for the following experiments: (a) Case A, Size 15; (b) Case B, Size 6; (c) Case C, Size 4; (d) Case D, Size 20. In the table below, we report the number of iterations of the WTA rule applied to the attention layer, required for one of the auxiliary nodes to exceed the value 0.6. The first value corresponds to the high resolution attention array, the second value to the low resolution array. We take the response time to be the lower of the two values.

	Display Size			
Case	4	6	15	20
A	22 - 15	27 - 15	12 - 15	12 - 13
B	22 - 17	27 - 18	24 - 99+	25 - 89
C	22 - 13	27 - 13	13 - 15	13 - 13
D	22 -*17	27 -*19	36 - 99+	*15- 99+

Discussion

The simulation results for this subnetwork are in agreement with the experimental data reported by Duncan and Humphreys. For Case A, our model achieves a search time independent of display size due to the strong response in the 'L' corner feature map corresponding to the target, which is directly reflected in the attentional activation, and the inhibition among the nontargets in the other feature maps, due to their high similarity. Case C gives similar results, though the inhibitory interactions among nontargets is weaker, due to their distribution among multiple feature maps. These two cases are not distinguished in Treisman's theory, both corresponding to the feature search condition.

For Case B, we get a moderate dependence of search time on display size. This is due to the somewhat weaker activation of the attention array corresponding to the

target that results from the inhibitory interaction of the target and nontargets in the feature maps. Finally, for Case D, the difference in attentional activity corresponding to target and nontargets is low, as a result of weaker inhibitory interactions among nontargets due to their low similarity, and of strong interactions between target and nontargets due to high similarity. This lack of contrast in some cases leads the WTA procedure to choose a nontarget as the selected focus of attention (as indicated by an asterisk in the table). This behavior would, in a more complete simulation, lead to a truly serial search for the target, mediated by multiple WTA settlings with inhibitory tagging [Klein 1988] between settlings. Again, Treisman's theory does not distinguish these cases, but presumes a mechanism like the multiple settlings as the source of the serial search times.

The connectionist network described above is intended to provide an alternative model of attentional behavior to that characterized by Treisman or by Duncan. We now consider the relations among the three theories. Treisman's theory appears to be a special case of Duncan's, since her feature-conjunction search dichotomy is subsumed by Duncan's T-N similarity measure, while the N-N similarity measure allows additional data to be modelled. On the other hand, despite Duncan's insistence that visual search times depend not on which features in target and nontarget objects are primitive, but only on the similarity among these objects, our model suggests that this 'similarity' measure is highly dependent on the identity of the primitive features. Since the interactions among features occur within the primitive feature maps, it is in the primitive feature dimensions that similarity is determined.

Treisman identifies two different search complexities, parallel and serial, which correspond to search times that are independent of display size and linearly dependent on display size, respectively. Duncan rejects these distinctions, preferring to describe a varying dependence of search time on display size. In our model, we observe three different search modes. When the target produces a strong activation in the attention layer, while distractors produce weak or no activation, the WTA procedure converges to an isolated activation corresponding to the target in a time that is virtually independent of the number of distractors. This corresponds to parallel search. When the target and nontargets produce approximately equal attentional activations, the WTA procedure converges much more slowly. More importantly, the region selected is as likely to correspond to any image object as any other, so multiple WTA settlings may be required to locate the target. This corresponds to serial search. When the target produces an attentional activation that is only moderately stronger than that produced by distractors, the dependence of search time on display size is determined by the WTA procedure. The resulting constraint propagation, or relaxation, search is parallel, in that it considers all object locations simultaneously, but is dependent on both the number and magnitude of the non-maximal activations, yielding a display size dependency that is different than either of the other two cases.

Finally, we mention some additional experiments reported by Duncan & Humphreys. In addition to the similarity measures, another critical determinant of search times is the size of the objects in the display. For a given eccentricity, larger objects reduce the display size dependency. In our model, this result is predicted by the reduced interaction of objects within the feature maps, due to the limited extent of the inhibitory connections. In another experiment which used 'L' corners of different orientations for targets and nontargets, a large dependence of search time on display size was found when the nontargets were CW and CCW 90° rotations of the target. The explanation given by Duncan is that the target is similar to the

nontargets, being the same object except for a rotation. Our current simulations could not produce this result, since there is no representation of rotation, nor of similarity across rotation. Our model would have to be elaborated to include rotation, as in Hinton's [1981] original model, in order to simulate this behavior.

References

Duncan, J., "Visual search and visual attention," in *Attention and Performance XII*, ed. M. I. Posner & O. S. M. Marin, pp. 85-105, Erlbaum, Hillsdale, NJ, 1985.

Duncan, J. and G. W. Humphreys, "Visual search and stimulus similarity," *Psychological Review*, vol. 96, pp. 433-458, 1989.

Hinton, G. E., "Shape representation in parallel systems," *Proc. 7th IJCAI*, pp.1088-1096, Vancouver, 1981.

Hinton, G. E. and K. J. Lang, "Shape recognition and illusory conjunctions," *Proc. 9th IJCAI*, vol. 1, pp. 252-259, Los Angeles, August 1985.

Hughes, H. C., R. Fendrich and P. A. Reuter-Lorenz, "Global versus local processing in the absence of low spatial frequencies," to appear, *J. Cognitive Neuroscience*, 1990.

Klein, R., "Inhibitory tagging system facilitates visual search," *Nature*, vol. 334, pp. 430-431, August 1988.

Koch, C. and S. Ullman, "Shifts in selective visual attention: toward the underlying neural circuitry," *Human Neurobiol.*, vol. 4, pp. 219-227, 1985.

Lowe, D. G., "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, vol. 31, pp. 355-395, 1987.

McLeod, P., J. Driver and J. Crisp, "Visual search for a conjunction of movement and form is parallel," *Nature*, vol. 332, pp. 154-155, March 1988.

Mozer, M. C., "A connectionist model of selective attention in visual perception," *Proc. 10th Conf. Cognitive Science Society*, pp. 195-201, Montreal, August 1988.

Sabbah, D., "Computing with connections in visual recognition of origami objects," *Cognitive Science*, vol. 9, pp. 25-50, 1985.

Sagi, D., "The combination of spatial frequency and orientation is effortlessly perceived," *Perception & Psychophysics*, vol. 43, pp. 601-603, 1988.

Sandon, P. A., "Simulating visual attention," to appear, *J. Cog. Neuroscience*, 1990.

Sandon, P. A., "An attentional hierarchy," commentary on "A solution to the tag-assignment problem of neural networks" by G.W. Strong and B. A. Whitehead, *Behavioral and Brain Sciences*, p. 414, September 1989.

Sandon, P. A. and L. M. Uhr, "An adaptive model for viewpoint-invariant object recognition," *Proc. 10th Conf. Cognitive Science Society*, pp. 209-215, Montreal, August 1988.

Treisman, A. and H. Schmidt, "Illusory conjunctions in the perception of objects," *Cognitive Psychology*, vol. 14, p. 107-141, 1982.

Treisman, A. and G. Gelade, "A feature-integration theory of attention," *Cognitive Science*, vol. 12, pp. 99-136, 1980.

Uhr, L. M., "Recognition cones and some test results," in *Computer Vision Systems*, ed. A. R. Hanson and E. M. Riseman, pp. 363-377, Academic Press, New York, 1978.

Witkin, A. P. and J. M. Tenenbaum, "What is perceptual organization for?," *Proc. IJCAI-83*, pp. 1023-1026, Karlsruhe, West Germany, August 1983.

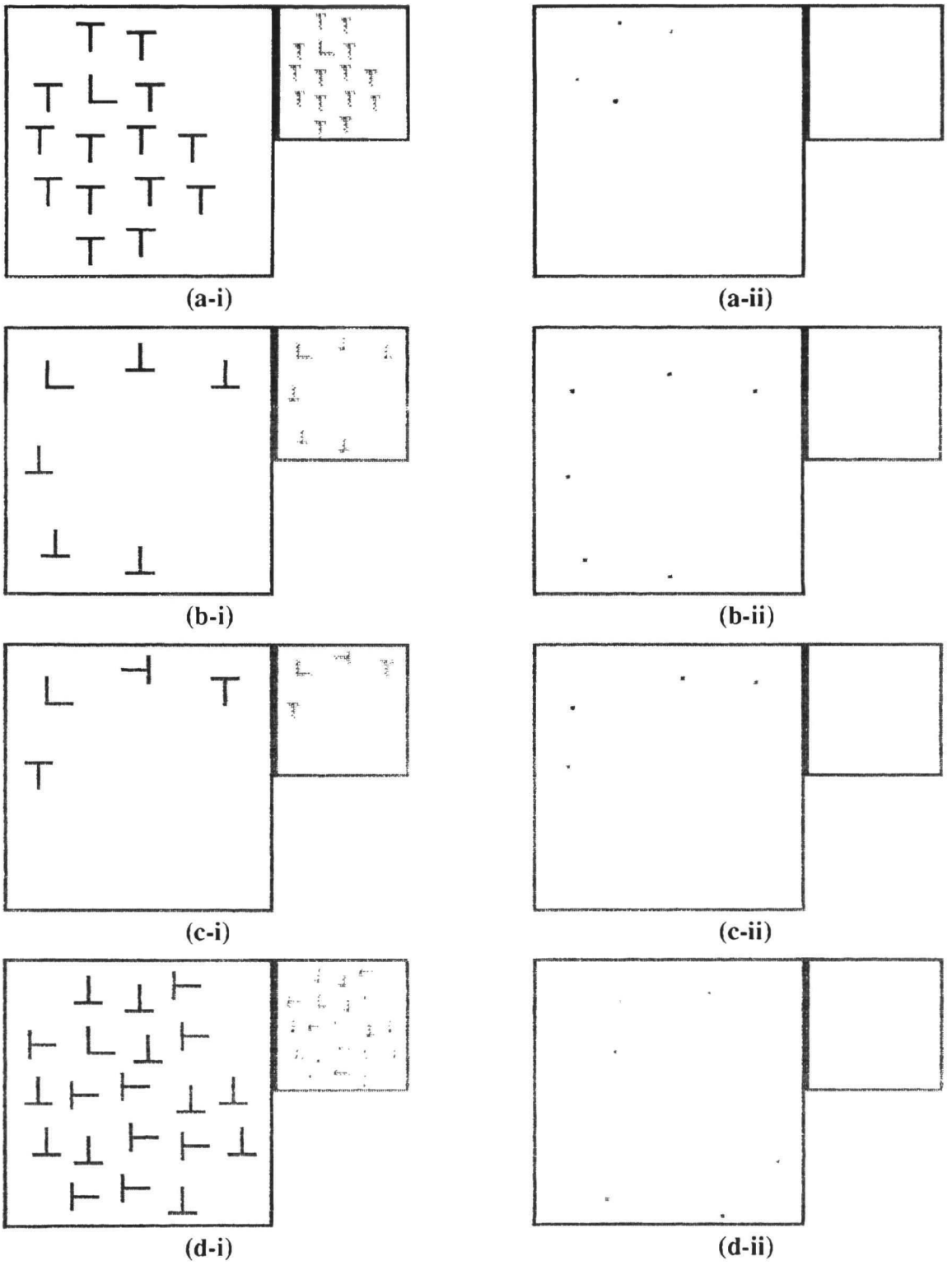

Figure 2