

# The iconic bottleneck and the tenuous link between early visual processing and perception

K. Nakayama  
Psychology Department  
Harvard University  
Cambridge, MA 02138

In, *Vision: Coding and Efficiency*, Edited by Colin Blakemore, Cambridge  
University Press, pp.411-422, 1990.

## The iconic bottleneck and the tenuous link between early visual processing and perception

K. Nakayama

### Introduction

Late 19th century studies of the brain provided evidence that part of the cerebral cortex was made up of primary sensory receiving areas and primary motor areas. These comprised a relatively small portion of the total surface area of the cortex and with the exception of some specialized regions (such as Broca's area), the functional relevance of the other parts of the cortex remained a mystery. Vision was relegated to a small portion of the human cortex, occupying about 15 per cent of the total surface. Surrounding this primary area were secondary and tertiary zones, often referred to as 'association cortex'.

Very recent advances in neuroanatomy and neurophysiology, however, have changed this picture dramatically. Thanks to the pioneering work of Allman & Kaas (1974), Zeki (1978), and Van Essen (1985), we now know that monkey visual cortex contains at least 19 separate maps of the visual field and according to a recent review by Maunsell & Newsome (1987) visual processing occupies about 60 per cent of the cortical surface!

This overwhelming dominance of vision in relation to other functions should serve as a reminder that, as generally practiced, the current subdisciplines of visual perception and psychophysics may be too narrow to capture the wealth of processing involved. Threshold psychophysics, especially, has been pre-occupied with just the earliest aspects of vision. It has neglected the seemingly intractable questions such as the nature of visual experience, pattern recognition, visual memory, attention, etc.

Meanwhile the neurophysiologists have been making recordings from diverse regions of the visual cortex which could be closely related to these higher functions. For example, in V2, just one synapse

beyond the primary receiving area, it appears that the firing of some neurons is related to the formation of 'illusory' contours (von der Heydt *et al.*, 1984). In area V4 the receptive field organization of cells is very specifically and profoundly altered by the attentional state of the monkey (Moran & Desimone, 1985). Finally, in infero-temporal cortex, many laboratories find that some cells only fire when the complex image of a face appears in the visual field (Gross, 1973; Perrett *et al.*, 1982, 1987). So now it is the physiologists who seem to be leading the way, at least as far as higher visual functions are concerned; their observations show that many complex functions are being performed in these newly identified regions of visual cortex.

To begin to redress this imbalance between psychology and neurophysiology we offer a frankly speculative theory as to the overall functional organization of the visual system. It postulates an associate memory for image fragments (icons) adapted from cognitive psychology (Lindsay & Norman, 1977; Rumelhart & McClelland, 1986) and couples this with the emerging notion of a multi-resolution pyramid representation for early vision as suggested by workers in psychophysics, physiology and artificial intelligence. Because it is so very general and covers such a large range of phenomenon and possible mechanisms, the theory will probably resist verification or falsification. We present it nonetheless, mainly because of the paucity of plausible ideas in this area. Hopefully such a skeletal framework will open the door for more precisely formulated ideas, and ones that can be more easily tested.

In essence, our theory divides the visual system in two: early vision consisting of a feature pyramid followed by visual memory. We describe each in turn.

## The feature pyramid – an overview

Closest to the input is a massively parallel feature pyramid which comprises striate cortex and those portions of extrastriate cortex which are organized more or less retinotopically. This system is organized as a multi-resolution multi-feature pyramid. For example, such a system contains neurons sensitive to a variety of features, including disparity, motion, color, line orientation, line termination etc., each of which can be represented at a variety of scales.

The usefulness of a generic and multi-purpose pyramid has been suggested by Rosenfeld (this volume) and the specific notion of a multi-resolution pyramid for early cortical representation has been proposed on empirical and theoretical grounds (Burt & Adelson, 1983; Sakitt & Barlow, 1982). Moreover, the general idea is consistent with the physiological findings (DeValois *et al.*, 1982). The essence of the idea is a retinotopic representation of the image at varying degrees of scale or coarseness. So with each ascending level of the pyramid (as shown in Fig. 36.1), the image is represented with less and less spatial precision and resolution. It should be obvious that the different levels differ in information content with the coarsest representation of the image requiring fewer bits of information than the finest level.

For illustrative purposes we will make this more concrete by estimating the relative information content at the various levels, recognizing that such numerical estimates are subject to error and acknowledging that they gloss over the actual details of the encoding process. The empirical basis of such estimates does not match the specificity suggested by the numbers themselves, and we do not mean to imply by assigning numbers that the entities denoted are physically discrete or quantized. Yet despite these limitations, it is possible that the use of such estimates can help us focus on otherwise difficult issues. For purposes of illustration we adopt the fiction that the system encodes the image in terms of pixels of varying size with a roughly constant number of bits per pixel. So the number of pixels becomes an intuitively reasonable index of information content. Again, we are aware that a very different type of coding than single pixel representation occurs. Thus in terms of contrast, it is likely that the system encodes the image in terms of oriented receptive fields of various sizes (DeValois *et al.*, 1982; Wilson & Bergen, 1979; Watson, this volume).

As an example, we provide an estimate as to the number of pixels required for the coding of contrast. First we need to remove the complication introduced

by the cortical magnification factor. We take a functional approach by noting how ordinary visual acuity varies with retinal eccentricity. Consistent with the complex logarithm description of the retino-cortical projection (Schwartz, 1977), the typical function relating letter acuity to eccentricity increases linearly with eccentricity (Weymouth, 1958). Using this data to calculate (by numerical integration) the number of such recognizable letters that could be squeezed into the visual field, we come up with a figure of about 1000. Assuming that each letter comprises a  $5 \times 5$  pixel array, roughly consistent with data indicating that 2 sinusoidal cycles/letter is adequate for letter recognition (Ginsburg, 1981), we arrive at a total of about 25 000 pixels in the whole visual field. This is roughly the equivalent of a  $160 \times 160$  pixel grid.

This describes the image as it is represented at the highest level of detail that can be encoded. But because it is a pyramid, the image is also represented at progressively coarser degrees of visual resolution. Thus, if at each level, we get coarser by a scale factor of two, we can see that a system of just five levels will have at its most coarse representation, a pixel grid of about  $10 \times 10$  pixels. To get a pictorial understanding of the hypothesized number of pixels, at least for the coding of achromatic contrast, five such representations are schematized in Fig. 36.1 where each definable square represents 100 pixels (a  $10 \times 10$  pixel grid).

So far, we have depicted the pyramid as if it operated at different scales analogous to banks of spatial frequency filters (Burt & Adelson, 1983; Sakitt & Barlow, 1982). This is misleading, however, since we would not want to exclude dots, edges and lines which only contain high spatial frequency information from being represented at the coarsest level of representation. Thus Craik–Cornsweet edges and difference of Gaussian dots (Carlson *et al.*, 1984) are represented similarly as ordinary edges and dots. So the early vision pyramid is far more abstract than simple spatial frequency scaling insofar as it represents edges, lines, etc. at different scales. Thus some form of appropriate communication between high spatial frequency mechanisms and the coarse level of representation in the pyramid is required. Interactions of this sort have been suggested by Rosenfeld (this volume) and Grossberg & Mingolla (1985) among others.

## Visual memory – an overview

At the other extreme, farthest removed from the eye, is visual memory. Such a system contains tiny pattern

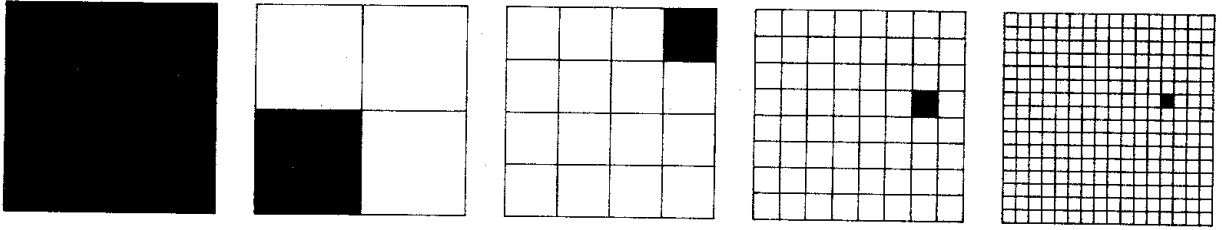
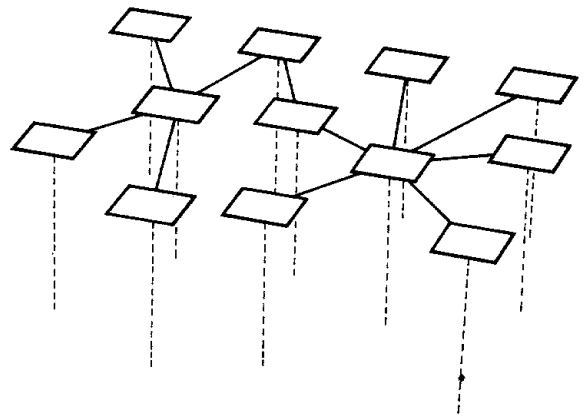


Fig. 36.1. Schematic representation of the multi-level pyramid. Coarsest representational level is shown at left. Finest representation shown on extreme right. Three intermediate representations are also shown. Note the existence of one shaded square at each level of the pyramid. This represents the size of the visual field that can be utilized for a hypothetical elementary pattern recognition operation at that level of representation in the pyramid.

recognition templates or icons<sup>1</sup> which are associatively linked. Thus they can be activated or potentiated by two different routes. First, by the process of pattern matching with incoming visual signals from the feature pyramid. Second, by the activation of other icons through associative learning. This memory system is situated in regions of visual cortex which shows the least evidence of retinotopic order and is most likely to be localized in the temporal lobes (Mishkin & Appenzeller, 1987). A small schematized subsection of this memory is shown in Fig. 36.2, illustrating at least two types of connections to these icons: afferent (from the pyramid) and associative (from within the memory itself).

The experience of seeing is dependent on the activation of these nodes or icons in visual memory. *Without such activation, visual perception cannot exist.* Essential to the theory as it is proposed is the assertion that these icons or templates contain surprisingly small amounts of information and that they capture the essential properties of an image fragment with very few bits. To keep our argument as numerically concrete as possible and to emphasize their small size, we assert that such icons contain only 100 pixels. Thus if such a template were to be roughly square it would comprise about  $10 \times 10$  pixels.



----- Afferent connections  
 ——— Associative connections

Fig. 36.2. Schematic description of a very small subsection of visual memory. Each icon or node has two types of possible connections, one set from within visual memory itself (solid lines) and one set to the output of the early vision pyramid (dashed lines).

### Evidence that icons are low resolution

Perhaps the most seemingly arbitrary single aspect of the theory is this assertion that the templates or icons have very low information content. Less controversial is the assertion that visual memory is made up of an associative network of such elements. To provide some plausibility to this idea of the very small icon size, we note data obtained from reading performance. If the visual system could pattern process only a small amount of pixels per unit time, then one should be able to drastically reduce the effective information available in certain visual tasks, and one should see no decrement in performance. This is a difficult experi-

<sup>1</sup> Note that our use of the term 'icon' is very different from that used in the past in cognitive psychology. Neisser (1967) coined the term to refer to short term visual storage (as originally described by Sperling, 1960) and it is synonymous with what could be called early cortical after discharge, specifically residual activation in the early vision pyramid after a brief flash. Our use of the term establishes the icon as a very small learned visual template, a constituent of visual memory.

ment to design in most free viewing tasks because one does not know where in the array the person is attending nor at what level of resolution. Reading, however, provides a stereotyped visual procedure which requires systematic attentional fixations along a line of print. Rayner (1978) has developed a computerized technique to limit the amount of intelligible text that is on a page by monitoring eye movements and replacing all but a small number of letters around the fixation point with meaningless symbols. They find that if one makes more than about 5 letters visible, then reading is not substantially improved. This tallies reasonably well with our 100 pixel limit since 5 letters comprise approximately 125 pixels. A separate study on reading by Legge *et al.* (1985) reaches a similar conclusion, finding that reading did not improve beyond the point where more than 3.5 letters were visible. So the visual system cannot process more than a small number of letters at a time and this number is not inconsistent with an icon size of 100 pixels.

### **Focal attention: limited readout from the early vision pyramid**

Here we consider the implications of tiny icon size. In a previous section we suggested that at the highest level of resolution, the image representation comprised a large, say  $160 \times 160$  pixel, grid and this is far too much information to be effectively compared with pattern recognition templates having a small  $10 \times 10$  pixel extent. The amount of information that can be sampled from the pyramid in the process of pattern recognition cannot exceed the size of the templates or icons themselves.

From this it follows that pattern recognition from the whole pyramid is not feasible in one single step. Many such steps which we call attentional fixations or elemental pattern matching operations will be required. In quantitative terms and if our estimates and ideas are reasonable, the sampling of the whole field will require about 250 of these elemental pattern matching operations because the  $10 \times 10$  pixel arrangement can only cover  $1/256$ th of the high-resolution map (see Fig. 36.1). If one were to sample the representation of the visual field at the coarsest  $10 \times 10$  level of resolution, however, only a single elemental pattern matching operation will be required.

Because of these quantitative considerations, it would seem that for the purposes of elementary pattern recognition, the visual system is faced with a trade-off. It can sample from the pyramid at lower levels of spatial resolution to obtain an overview of the visual scene whilst sacrificing detail. Alternatively, it

can sample at a very high level of resolution to get detail but sacrificing the overview. The shading of squares in Fig. 36.1 illustrates the very different amounts of coverage of the visual field that can be obtained as one conducts an elemental pattern matching operation at different levels of the pyramid. The existence of selective attention to particular portions of the visual field has been well documented (Posner, 1980, 1987). See also Nakayama & Mackeben, 1989. Selective attention to one spatial scale vs. another is less well documented but preliminary evidence to support such mechanism has been obtained by Sperling & Melcher (1978).

Figure 36.3 schematizes the visual system as a whole. Closest to the input end is the massively parallel pyramid, comprising the machinery of early vision and receiving parallel input from the retina. Farthest from the input is visual memory, also a massively parallel system, associatively linked and composed of tiny icons having very low informational capacity. Although the connections and consequent quantity of information shared within each of these massively parallel systems is great, the connection *between* these systems is not. This link is extremely band-limited and constitutes a critical bottleneck in the visual system.

So how does vision occur in ordinary circumstances? We argue that for normal scenes, vision involves a serial sequencing of elementary pattern matches (attentional fixations) from different loci in the pyramid. The net result of such matches is residual activity in those icons which have been recently activated by feature pyramid output and also those which have been associatively linked or potentiated by such icons. Thus we argue that the conscious act of perception is directly related to aggregate activation of these icons in visual memory.

As an illustrative example, consider the visual system confronted with a mountain landscape scene which is very briefly presented in a tachistoscope but with sufficient time to allow for three attentional fixations. First the system does a pattern match to the whole scene at lowest resolution and gets a memory activation capturing the gross outline of the mountain. Then it makes a second more detailed attentional fixation at a lower level of the pyramid centered at the mountain peak. Finally, one other fixation is directed downward towards the house near the base. We argue, however, that other icons may also be partially activated, not through visual input but by associative linkage of those icons which have received visual input. So from the perspective of capturing specific input from the retina, only three very low resolution

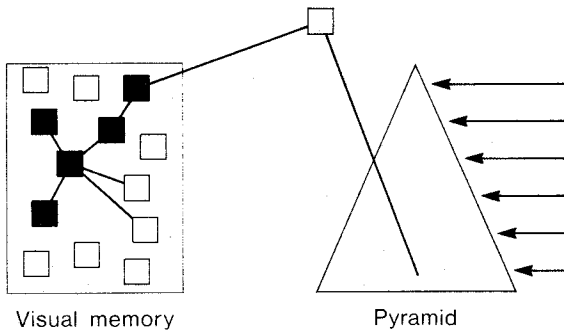


Fig. 36.3. Overview of the visual system. On the left is visual memory, containing tiny icons linked associatively. On the right is the multi-resolution pyramid which receives afferent visual input. The linkage between these two massively parallel systems is a narrow bandwidth channel having a capacity of the order of a single icon.

Shaded squares in visual memory provide a schematic representation of a number of icons which have residual activation from previous pattern matches from the early visual pyramid and which constitute the 'contents' of conscious visual experience.

'snapshots' at three different scales have been taken, yet it is sufficient for the observer to capture the essentials of the scene. The total set of activated icons is enough to convey the 'meaning' of the scene, and the observer is unaware of the essentially serial nature of the construction process nor of its tenuous relation to visual input. As far as the observer is concerned he was presented with a scene and he has grasped it all at once.

What we are saying is that our own introspective understanding of vision is somewhat of an illusion. We regard our visual world as 'just there', not as something which is only acquired after sequential sampling and reconstruction. It appears that vision occurs in parallel yet our actual contact with the world is essentially serial,<sup>2</sup> constructed by a sequence of low bandwidth pattern recognition matches. Thus the actual amount of visual information that is explicitly

<sup>2</sup> At the same time that we have the phenomenologically naive belief that visual perception is conducted in parallel, it can also be argued that the phenomenology associated with eye fixations supports something much closer to the present theory. As we make a set of eye fixations we know that very fine detail can only be made in central vision. Yet the scene remains remarkably 'seamless' and 'there'. Thus we are forced to conclude that the pick-up is serial yet something endures (the activation of visual memory), to preserve the scene.

used as part of the pattern recognition process is but a tiny fraction of the information available at any instant in the feature pyramid.

### Need for a controller?

The most distinguishing feature of the theory is the notion of a limited attentional bandwidth (limited pattern recognition capacity) coupled with the complementary notion of a multi-scale pyramidal representation of early vision. The organization of the pyramid as a data structure is well suited for the tasks we suggest because it enables the system to scan the image for its essential properties in an efficient manner, appropriately switching levels of resolution to get both the overview and the necessary details.

As described, however, the process might seem to require an 'agent' or 'genie', to direct these attentional fixations so as to optimize the pickup of information. This is likely to be the case for a certain fraction of the time, but at others, the control of attention could be determined at a very low level. This has been suggested by Julesz (1984) who concluded that texture or feature gradients draw attention. Koch & Ullman (1985) say much the same in their description of the saliency map which directs the spotlight of attention. In particular, Koch & Ullman suggest a winner-take-all network based on some plausibly hypothetical properties of early feature maps which is adequate to direct some aspects of attention. Beyond this selection, Koch & Ullman suggest that the system may shift to the next most salient feature, based on its proximity or similarity to the previous feature.

Such low level schemes will not be sufficient for many aspects of attentional control and other mechanisms will be required. Again this may not require as much centralized control as one might think. It is conceivable that attentional fixation instructions could be distributed and linked to the visual memory itself. One possibility is to attach the fixational routines to particular icon sets in visual memory. A low resolution icon representing the gross features of an object might contain 'pointers' to other appropriate fixations. Thus the outline of a face might activate attentional 'fixations' at finer levels of detail to recognize eyes, nose and mouth, thus providing information to recognize a specific face (see Noton & Stark, 1971). Such an approach might be analogous to object oriented algorithms more familiar to specialists in computer science.

In addition to the controller function we speculate that there also needs to be an 'addresser'. Such a mechanism will register the address or locus of sam-

pling from the pyramid and create a corresponding address and size for an activated icon in a more generalized body-centered coordinate system. Such organization is necessary to preserve the spatial relations of the sampled image fragments in the scene and also to provide a coordinate reference for motor behavior.

Finally a comment about neural implementation. The model as proposed implies that the connections between visual memory and the outputs from early vision are constantly changing. At one moment, visual memory is connected to, and thus samples from, say, the lowest level of the pyramid. Then later it may sample from a restricted region of the visual field from a high-resolution section. As yet there is no obvious circuitry to mediate these processes which would seem to require the formation of temporary yet orderly sets of connections. But it is perhaps interesting to note that the existence of temporary synapses has been proposed (Crick, 1984) and that more recently 'shifter-circuits' have been suggested (Anderson & Van Essen, 1987) which could temporally connect one two-dimensional representation to another, preserving local retinotopic order.

### Extensive preprocessing in the pyramid

Our discussion so far has purposely oversimplified the nature of the multi-feature pyramid so as to stress the main features of the theory. Now, however, we must mention several properties of the pyramid which are of critical importance to guide the pattern recognition process. Thanks to the work of many, most notably Barlow (1960, 1961) and Marr (1982) it has become increasingly clear that the processing in early vision is highly sophisticated and captures visual information in a seemingly 'intelligent' manner without recourse to cognitive top-down processing. Two properties of the pyramid seem particularly important in this regard: feature differencing and feature grouping.

With respect to featural differencing, we envision that for each feature map, there exist inhibitory networks to enhance differences in that particular feature. Thus for the representation of motion, neural networks are organized so that velocity differences rather than absolute velocities are registered (Allman *et al.*, 1985a; Frost & Nakayama, 1983). Likewise, orientation (Blakemore & Tobin, 1972; Nelson & Frost, 1978), as well as other features, is organized so that spatial differences in that feature are emphasized rather than the features themselves. These mechanisms, consisting of connections outside of the classi-

cally defined receptive fields (Allman *et al.*, 1985b) are particularly evident in cortical area V4 (Desimone & Schein, 1987). They accentuate featural differences and are relatively insensitive to a whole field containing textures having the same features.

These neurophysiological properties support the general points raised by Julesz's texton theory which has outlined the importance of primitive features in early vision and has given particular emphasis to the notion of featural or texton density gradients (see also Beck *et al.*, 1983). Featural difference maps are useful to provide both the outlines of a two-dimensional image to be compared with templates in visual memory (such outlines may be analogous to Marr's place tokens) as well as providing potential loci for the direction of visual attention (as suggested by Julesz, 1984).

In addition to feature differencing, the pyramid must also support grouping algorithms. These have at least two major functions: (1) to appropriately link and enhance different portions of the image for the purposes of pattern recognition; (2) to suppress all other parts of the image so that pattern matching is only applied to the appropriate portion of the image. Grouping is a process which pre-organizes information in the feature pyramid to make it amenable for pattern recognition. This is the familiar figure-ground process and one that is essential if pattern recognition is to occur.

Many grouping laws are well known as they are embodied in various Gestalt laws of perception. Furthermore, they have also received some attention in recent times. The work of Julesz, Grossberg and others, for example, are partially devoted to characterizing the cooperative and competitive networks underlying this grouping process. One of the most important process is similarity of grouping, i.e. those elements which have the same color, orientation, disparity, motion, etc. are linked (see Barlow, 1981). It is suggested that grouping requires an excitatory linkage between the representation of like features and inhibiting coupling between unlike features and that the network parameters of excitation and inhibition can increase or decrease as a function of experience (plasticity) or the demands of the moment (modulation).

The existence of feature differencing and similarity grouping is particularly helpful in interpreting the results of visual search experiments, where it is the task of the observer to identify a target from amongst a set of distractors. Treisman (1985) found that the search for a target differing by a single feature was easy and conducted in parallel (search time

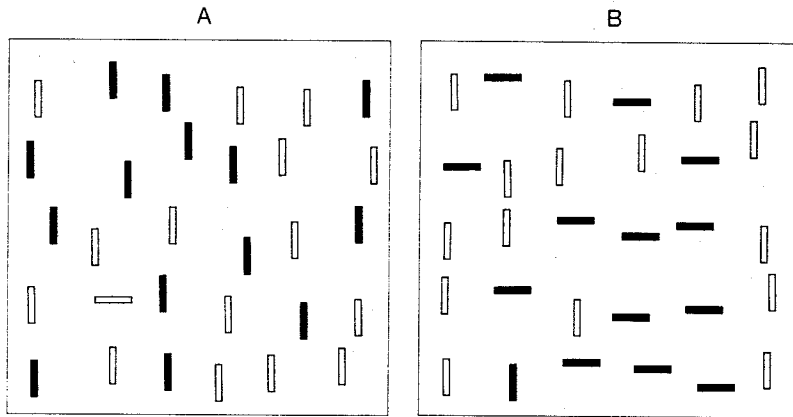


Fig. 36.4. Two types of visual search task. In (A) the observer has to find the target defined by a single feature difference, namely he needs to find the target having the horizontal orientation. We argue that feature differencing mechanisms (in the domain of orientation for this particular case) automatically mark the spot at which an elementary pattern matching operation will occur. Consequently, the search for such a single target will be rapid and will not be slowed by increasing the number of distractors (vertical bars). In (B) the observer must find the target defined by the conjunction of orientation and color. The observer must find either a white horizontal bar or a dark vertical bar amongst the distractors. We argue that, for this situation, feature differencing operations (either for color or orientation) will fail to select a unique site for pattern matching. As a consequence, pattern matching will be made at a number of wrong sites before the correct target is found. As such, search time will increase for greater number of distractors.

independent of number of distractors) whereas the search differing in a conjunction of two features was serial (search time increased for larger number of distractors). See Fig. 36.4 for an illustration of these two types of visual search displays.

We have confirmed this for a number of conjunctions (in particular, the conjunction of orientation and color), but for many other conjunctions including any dimension linked to disparity (Nakayama & Silverman, 1986a,b), the search can be conducted in parallel. Furthermore, with extended practice, it has been claimed that the conjunction of virtually any pair of dimensions can be made to occur in parallel (see Steinman, 1987; also Wolfe *et al.*, 1989 and personal communication).

The ideas proposed here provide an interpretative framework to understand these rather puzzling results. In such multi-element search arrays the system is faced with two problems. First is the capacity limitations of learned pattern recognition templates. We have postulated that such icons have only very limited information content. Thus it is not feasible to sample the whole target display with a single template match at the lowest level of resolution because the targets are too small and are thus indistinguishable at the lowest level of resolution. The pattern matching operation needs to be directed to a higher resolution level in the pyramid and to a particular locus. This

leads to the second problem. How is this site to be selected? In the case of a simple search for a single deviant feature, the problem is relatively easy. Feature differencing mechanisms can designate the single site for pattern recognition. For the case of feature conjunctions, however, the problem is more complex since feature differences on any given dimension are present in many sectors of the array and no single obvious site emerges for the more specialized pattern matching process. The system is forced to pattern match at a variety of wrong sites before finding the target. This could account for the increased search time for some conjunctions.

As mentioned earlier, however, the search for many feature conjunctions can be conducted in parallel (Nakayama & Silverman, 1986a,b). To explain this ease by which many conjunctions are searched, we invoke similarity grouping. This process takes like features, say those sharing a common disparity and links them, suppressing all others,<sup>3</sup> see Fig. 36.5. Then feature differencing operations on the remaining targets (those not suppressed) enable a single site to be marked for pattern recognition and the search task

<sup>3</sup> Although designed to solve the lower level problem of stereo-matching, the postulation of such a cooperative process has been suggested earlier (Nelson, 1975; Marr & Poggio, 1976).



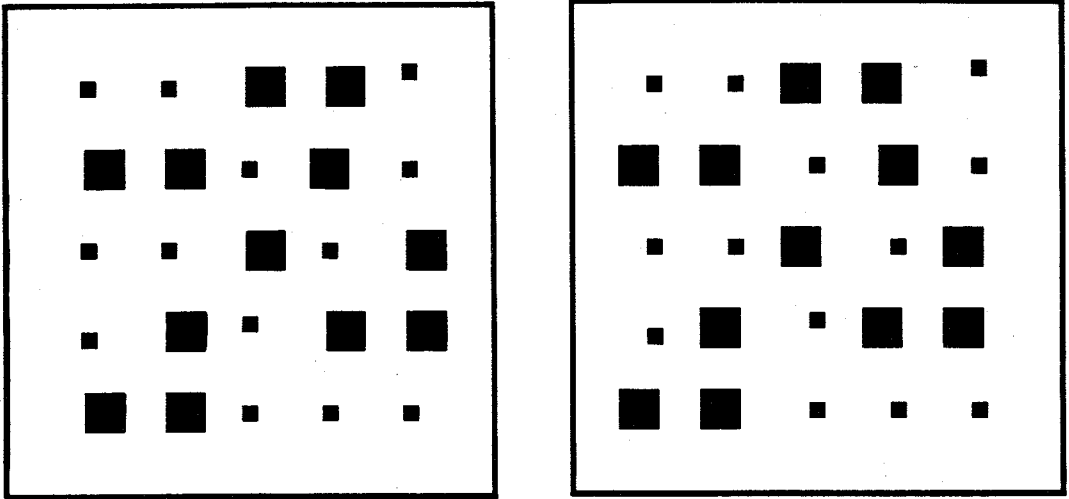


Fig. 36.5. Hypothetical feature grouping processes in the search for the conjunction of stereoscopic depth and size. If this pair of images are fused stereoscopically, all the distractors in one stereo-plane will be of one size and the distractors in the other plane will be of the other size. Task of the observer is to find the odd size in a given depth plane. For the usual case of crossing one's eyes (so that the right image is viewed by the left eye) the odd target is a large square in the front plane. We argue that feature grouping algorithms link targets of the same disparity and suppress targets of other disparities. Then feature differencing algorithms can pick out the odd size target from within a given stereoscopic plane. As such, search times are not influenced by the number of distracting targets (see text, also Nakayama & Silverman, 1986a).

appears as essentially effortless. To deal with the often marked improvement in performance with extended practice (Steinman, 1987), we suggest that the coupling parameters of the underlying neural networks can become modified to enhance grouping along particular feature dimensions.

### Object representation

The theory concentrates on the pick-up of information from the pyramid, emphasizing the very small size of the visual icon and the consequent bottleneck in visual information transmission. Thus vision proceeds by a set of sequential pattern matches from different levels of the pyramid, grabbing information from the pyramid at varying scale and position and activating icons corresponding to various sizes and position in the visual field. In this section we suggest how this process of sampling from a multi-resolution pyramid could dictate the basis of object representation in visual memory.

Most important to consider is the very small amount of information contained in an icon in comparison to the detailed visual knowledge that we have of most real objects. Thus the icon itself cannot be the fundamental unit of object representation but is only a

component. So we suggest that visual objects are assemblies of icons, associatively linked through visual experience. These correspond to the set of samplings or attentional fixations taken from the multi-level pyramid. For each object, therefore, there are various iconic snapshot representations taken at varying degrees of size (relative to the object). Thus, object representation consists of the aggregate of icons activated by a given object. For example the representation of an elephant might consist of some whole body icons showing typical side, rear, front and three-quarter views. Associatively linked to each of these views might be more detailed representation of head, trunk, tusks, mouth, eyes, feet, tail, etc. We suggest that these views at different scales (corresponding to attentional fixations) represent canonical representations of object parts and they are linked associatively. Hoffman & Richards (1985) suggest, for example, that distinct 'parts' of an object are almost invariably delineated by regions of negative intrinsic curvature in the object and are correlated with concavities in the image. It is possible that such 'parts' plus whole views of an object comprise the canonical views or canonical representations of an object. The plausibility of such canonical views and of their dominant role can be appreciated by introspection.

