Perception, 1989, volume 18, pages 55-68

Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects

Ken Nakayama, Shinsuke Shimojo, Gerald H Silverman
Smith-Kettlewell Eye Research Institute, San Francisco, CA 94115, USA
Received 8 June 1988, in revised form 3 October 1988

Abstract. Image regions corresponding to partially hidden objects are enclosed by two types of bounding contour: those inherent to the object itself (intrinsic) and those defined by occlusion (extrinsic). Intrinsic contours provide useful information regarding object shape, whereas extrinsic contours vary arbitrarily depending on accidental spatial relationships in scenes. Because extrinsic contours can only degrade the process of surface description and object recognition, it is argued that they must be removed prior to a stage of template matching. This implies that the two types of contour must be distinguished relatively early in visual processing and we hypothesize that the encoding of depth is critical for this task. The common border is attached to and regarded as intrinsic to the closer region, and detached from and regarded as extrinsic to the farther region. We also suggest that intrinsic borders aid in the segmentation of image regions and thus prevent grouping, whereas extrinsic borders provide a linkage to other extrinsic borders and facilitate grouping. Support for these views is found in a series of demonstrations, and also in an experiment where the expected superiority of recognition was found when partially sampled faces were seen in a back rather than a front stereoscopic depth plane.

1 Introduction

One of the major feats of human vision which has yet to receive an adequate scientific explanation is the phenomenon of learned pattern recognition. How is it that we can identify so many three-dimensional objects, so quickly and so effortlessly, all from an infinite variety of two-dimensional views? To accomplish this task, the concept of template matching is frequently invoked, roughly envisioned as a correlation process between portions of the image and stored templates in visual memory. The process remains a mystery, even though many have considered the problem from a variety of perspectives. These include contributions by Neisser (1967), Marr and Nishihara (1978), Biederman (1985), Ullman (1986), and Nakayama (1988), to name just a few.

Here we consider a related problem which is essential to this process of pattern recognition. How can the visual system recognize a two-dimensional view of an object when it is partially hidden behind other objects? It is our aim to show that the occlusion of an object by another object is more complicated than the simple removal of information about the hidden object. Two additional problems are introduced by occlusion. First, spurious edges are introduced at the occlusion boundaries and a distinction has to be made between the real edges of the more distant object and these spurious edges. Second, occlusion can often divide a single object into several image fragments and it becomes of importance to link parts of the same object preferentially whilst maintaining the segregation of image fragments of separate objects.

To illustrate some of the problems associated with occlusion, we present a demonstration in figure 1 originally introduced by Kanizsa (1979) and adapted from Bregman (1981). Figure 1a shows a set of uppercase letters (Bs), unoccluded. In figure 1b they are partially occluded by a snake-like figure. Finally, in figure 1c the occluder is rendered invisible, leaving just the image fragments of the letters. Despite the identical areal exposure of the uppercase letters in figures 1b and 1c, there is a clear difference in the visibility and clarity of the occluded letters. When the remaining letter fragments are present on their own without visibility of the occluder, they are not appropriately

segmented or grouped and their recognition as letters is difficult. Yet when the occluder is visible, their identity is more obvious.

We have found that very similar effects can be observed when both the occluder and the background are random-dot patterns and the occluder is defined in a purely 'cyclopean' fashion. When the top pair of stereo images in figure 2 is fused by crossing the eyes or the bottom pair is fused by uncrossing the eyes, the cyclopean contour of the snake-like occluder is clearly defined in a front depth plane relative to the letter fragments. Most important is the fact that the letters are much more recognizable when the occluder itself is visible and in front. Because the visibility of the occluder has such a decisive effect, it underscores our view that there is more to occlusion than the simple deletion of information.

What is it about the visibility of the occluder than enables the occluded objects to be more easily recognized? A number of possibilities come to mind.

First is the possibility that the explanation is to be found at a primitive level of object recognition, and that it is the encoding of the occluder as a coherent visual entity which provides information as to the grouping of the remaining image fragments. As such, it is necessary that the occluding portion of the image be 'recognized' as a coherent visual object and from this primitive identification the system can link and parse the remaining image fragments.

A second possibility, and the one that we favor and develop in this paper, is that the problem can be handled at an even earlier level, well before the process of pattern recognition. The basic idea rests on the notion that occluded objects contain and are bounded by two types of contour. First are the *intrinsic* contours of the object itself; these are the only contours which are inherently related to the object and they alone can provide valid information as to its identity and shape. Second are the set of *extrinsic* contours, formed accidentally by the interposition of another object in the line of sight. These extraneous edges have no intrinsic relation to the object itself and vary in position depending on the relative location of the object, the observer, and other occluding objects in the three-dimensional scene. As such they can only provide spurious input to pattern recognition systems. Thus, an occluding object not only hides information about the object which is behind, but it also adds extrinsic or extraneous

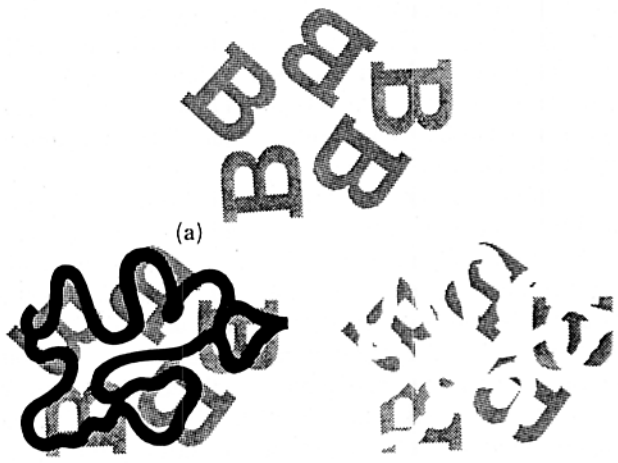


Figure 1. Problems associated with occlusion. (a) Uppercase letter Bs; (b) the same letters except partially hidden by a snake-like occluder; (c) the same, except the occluder has been rendered invisible. Note that although the exposed portions of the letters are identical in (b) and (c), the visibility of the letters is superior in (b). (Adapted from Bregman 1981.)

edge information which can degrade the process of pattern recognition. In figure 3 we show an expanded view of the fragments of two letters which were obscured in figure 1, noting the intrinsic contours in black and the extrinsic borders in white. It is our view

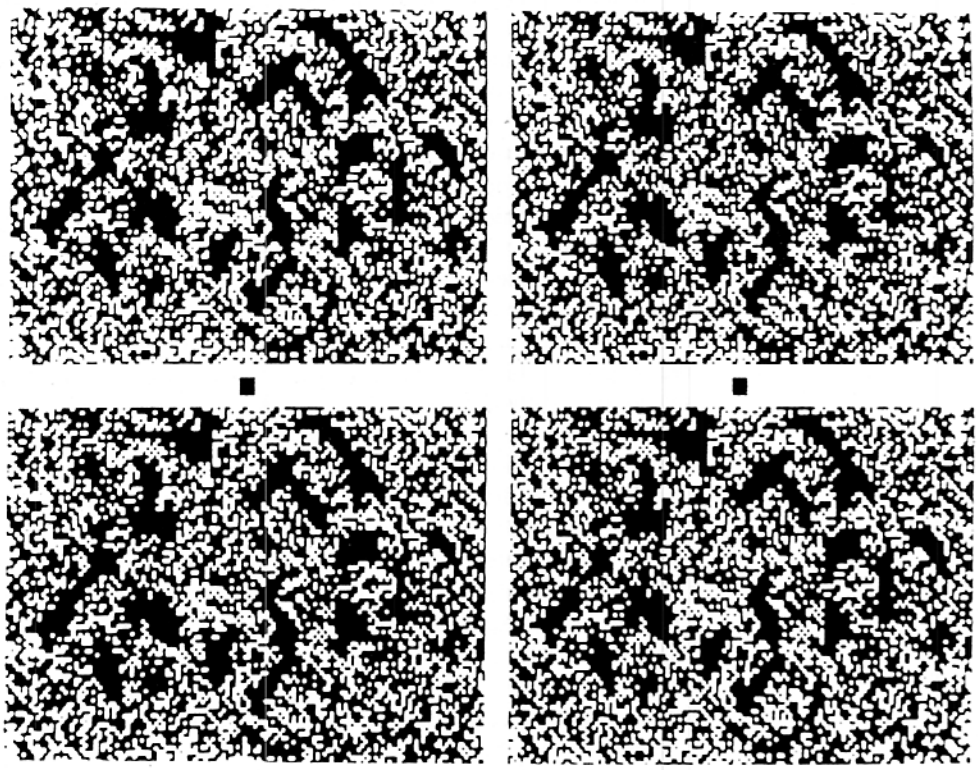


Figure 2. The same letters as seen in figure 1, occluded by an object having the same random-dot texture as the background. Thus the occluder is camouflaged when each half-image of the stereogram is viewed alone and not fused. The occluder has a different binocular disparity than the remaining portion of the display and the identity of the latter Bs is more evident when this snake-like figure emerges in front of the background.

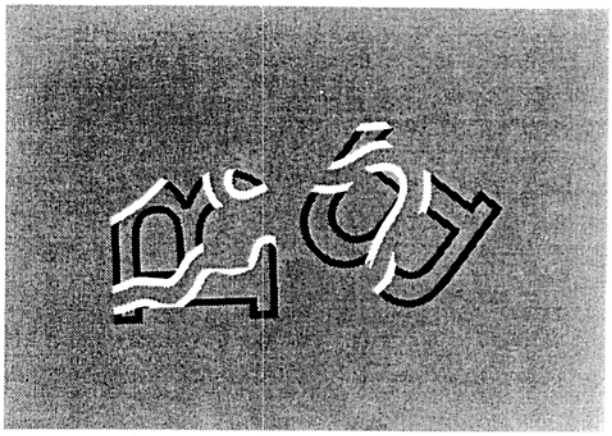


Figure 3. Expanded view of the two bottom letters seen in figure 1, depicting the extrinsic contours (white) and intrinsic contours (black).

that in order to prevent the extrinsic edges from becoming an input to pattern recognition templates, the visual system can and must make a distinction between these two types of contour. An important question remains. On what basis can the visual system make such contour classifications?

Given the real-world geometrical relations between occluded and occluding visual objects, we think that the answer could be relatively simple. Our hypothesis is that the classification could be made on the basis of depth alone. When two regions of an image share a common border, the natural constraints of the real world dictate that the border always 'belongs to' the region corresponding to the closer object. Similarly, it does not belong to the farther object and is thus 'extrinsic' to it.

To obtain supporting evidence for this depth-based classification hypothesis, we required a technique where relative depth can be easily manipulated without varying other aspects of the image. Stereopsis, particularly that occasioned by random-dot stereograms (Julesz 1971), provides a convenient approach. In this regard, the stereoscopic demonstration shown in figure 2 supports our 'depth' hypothesis because it is only when the snake-like occluder appears in the front plane that the letter Bs are most recognizable. When the stereogram is seen in its reversed configuration (with the upper stereogram viewed with uncrossed eyes or the lower stereogram viewed with crossed eyes), the Bs are very hard to discern, perhaps even more difficult to see than in the unfused case. This particular stereogram, however, does not provide an unbiased test of our depth hypothesis because one does not generally obtain a 'pure' reversal of depth by reversing binocular disparity in stereograms. This can be appreciated by observing the reversed configuration (top stereo pair with eyes uncrossed or bottom stereo pair with eyes crossed). Here the letter fragments can appear even more disjointed because they appear in both front and rear planes. Those visible to one eye alone are seen in the back plane whereas those sections visible binocularly are seen in the front plane. This follows from the fact that unpaired monocularly-viewed regions of three-dimensional stereograms are always seen in the back plane (see Julesz 1971; Nakayama and Shimojo 1988).

No such problem exists for our demonstration shown in figure 4. Here we show two stereograms containing a partially visible letter C. A small section of the image containing a central part of the letter has been removed and replaced by random dots. So when viewed monocularly, and not fused stereoscopically, one does not see a C. Instead, two U-shaped figures, mirror-symmetric about a horizontal axis, are seen. If the stereograms are fused and seen in depth, this replaced area is clearly delineated and can be seen either in front or in back depending on which stereogram is viewed (and whether the eyes are crossed or uncrossed). It should be clear that the perception of a C, as opposed to seeing two U-shaped segments, is most apparent when this replaced area is seen in front rather than in back. Perceptually it is as if the letter flows behind or is continuous behind an occluder. When the occluding region is in back, however, the 'cut ends' of the letter fragments resist closure and the C cannot be seen.

The large difference in the perceived unity and clarity of the letter C in the two stereograms cannot be fully explained by the first interpretation mentioned above (ie the visual coherency notion) since the replaced square region is obviously and equally visible in both cases. Yet the C is very difficult to recognize when the square is 'behind' and is no clearer than for the case where the region is indistinguishable from background. Such is the case when the pattern is viewed monocularly. This provides considerable plausibility for our notions of edge classification based on depth relations alone.

To make this argument from a slightly different perspective we present a second demonstration, where the visibility of the occluder is even less apparent and explicit

(defined only by stereoscopically generated subjective contours),[1] yet where local depth relations remain and where they are decisive in determining what is perceived. In figure 5a we show a logo-like configuration consisting of three nearly identical segments or image fragments, each of which can be defined as the intersection of two imaginary circles, one having a much greater radius than the other. Our demonstration is designed to show that the perception of this configuration can be very different, depending on disparity relations and whether the system regards a given bounding arc as intrinsic or extrinsic. In terms of perception, the observer can see either a single large disk which has been partially occluded *or* portions of three smaller disks. The unoccluded representations of this single large disk and the three smaller disks are depicted as figures 5b and 5c, respectively. The demonstration can be seen by comparing the two stereograms in figure 6 where, if one cross fuses, one sees just a single occluded disk in the top stereo pair, whereas one sees portions of three occluded disks in the bottom stereo pair. The opposite will apply if one diverges to obtain fusion.

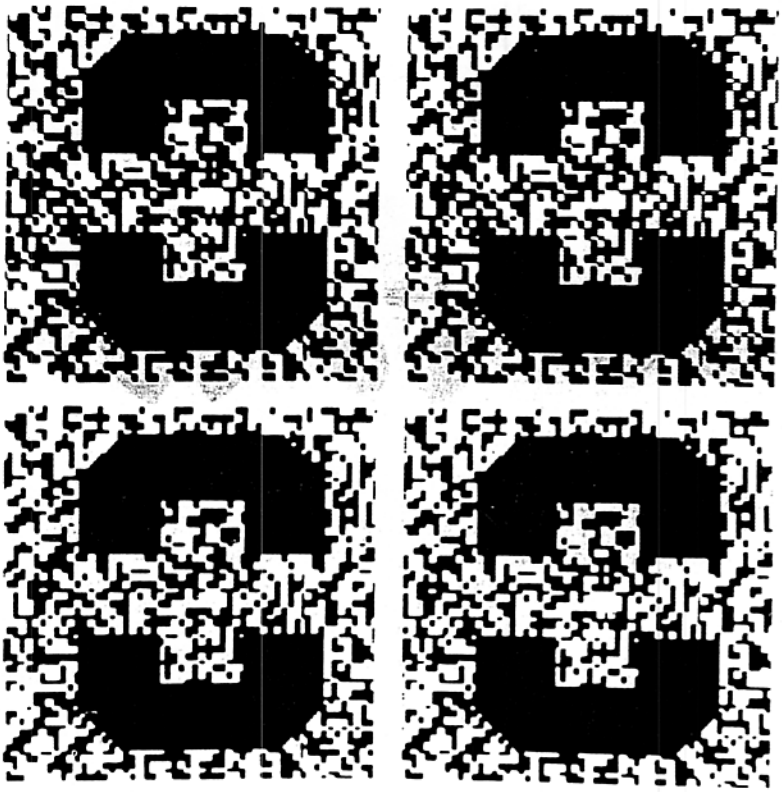


Figure 4. Letter C embedded in a pair of random-dot stereograms. A portion of the image which includes the middle portion of the C has been excised. It is replaced by a set of random dots which have either crossed or uncrossed disparity relative to the rest of the display. If the targets are binocularly cross-fused such that the right image is presented to the left eye, then a small region will be seen as in front in the top stereo pair and in back in the lower stereo pair. This relation will be reversed if the stereograms are fused with eyes diverged. In either case, it should be evident that the C is more discernible when this excised region is seen as 'in front'.

[1] Different yet important demonstrations of the role of stereoscopically defined subjective contours in perceptual organization can also be seen in papers by Gregory and Harris (1974) and Lawson et al (1974).

From this demonstration it should be clear that the sign of binocular disparity determines whether a given isolated blob remains segmented and separate from its neighbors or, alternatively, whether it becomes grouped with these same neighbors. This segmentation and grouping, in turn is decisive in determining 'what' is seen. Thus, this demonstration, together with previous ones, shows that depth information is crucial for the categorization of edges, and this in turn determines the grouping and segmentation of image fragments which then determine what object is recognized.

These conclusions, however, rest on 'subjective' demonstrations only and it is important to supplement them with objective performance measures of pattern recognition. Because capital letters and geometric figures are highly simplified and overlearned, it would be of interest to use more complex forms, similar to those encountered in everyday life. So to see whether these ideas could be applied to more complex images, we tested the ability of observers to recognize partially visible faces in photographs.

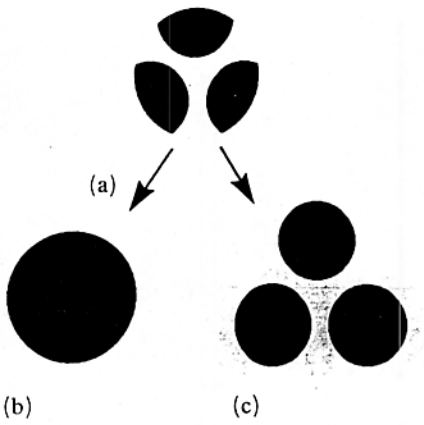


Figure 5. A three-segment logo-like figure (a) can appear as a partially visible single disk (b) or as three disks (c), depending on local depth relations and the consequent formation of subjective contours.

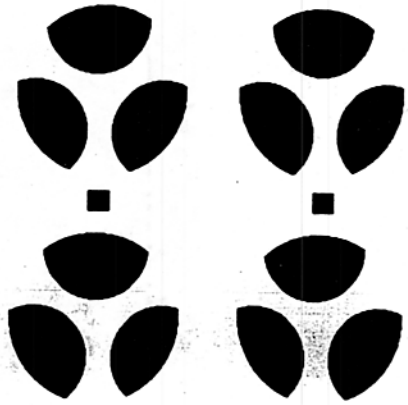


Figure 6. When the upper stereo pair is cross fused, the subjective contour will be adjacent to the more curved inner arcs of the blobs and the configuration will be seen as a single disk behind a Y-like occluder. In the uncrossed viewing, the subjective contours will be adjacent to the less curved outer arcs of each of the three blobs and the observer will see separate disks viewed through a circular aperture. The same applies to the bottom stereo pair, but reversed because the right and left images have been swapped.

2 Face recognition experiment

2.1 *Method*

We constructed partially visible samples of human faces by interposing horizontal strips of equal width which had no information about the face. These strips could appear in either a front or a rear stereoscopic plane relative to the faces. Figure 7 schematizes the two situations in a pictorial representation. Consider the case where the visible portions of a face are in front (figure 7a). According to our depth-based classification hypothesis, the horizontal edges will remain attached to the front portion of the binocular image which contains the face. As a consequence, these horizontal edges will act as a spurious input to hypothetical pattern recognition templates in visual memory. In other

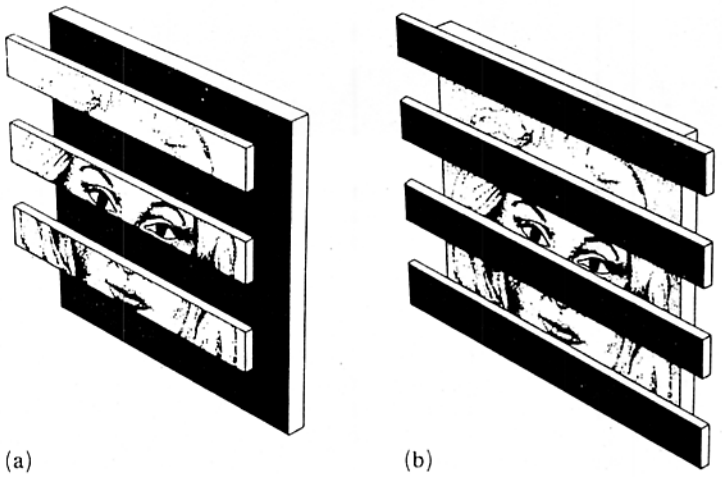
words, the intrinsic edges of the face will be indistinguishably mixed with these spurious edges and each horizontal panel will be processed separately as an independent object. Performance should be degraded in this case.

Consider the opposite case where the face is in the rear stereoscopic plane (figure 7b). In this situation, our hypothetical edge classifying mechanism would again attach the horizontal borders to the front plane, thereby removing it from the plane containing the face which is in back. But since the pattern recognition process is sampling from the rear stereoscopic plane, the image information in this plane is shielded from the extraneous input and better recognition performance is expected.

Figure 8 shows an actual sample of one of the stereograms used in the testing phase of our experiments. The upper and lower stereograms differ only insofar as one is the left-eye/right-eye reversal of the other. Despite the same 'informational' content in each stereogram, the perceived clarity of the face is very different. When the face is seen as in front as a consequence of disparity, all observers with normal stereoscopic vision report that the face is perceived with considerably less clarity and unity in comparison to the case where the face is perceived in back.

To test our hypothesis using objective performance measures of pattern recognition, we used the following procedure. Eight volunteers who had stereopsis and who were naive as to the goals of the experiment were recruited. They were given a set of sixteen unoccluded faces to examine in a pretest period, with the understanding that they would later be given twice as many faces in a subsequent testing period and would be asked to report whether a given face had been presented previously. The pretest faces were presented in blocks with a constant time exposure for all subjects. In the test period they were to make a yes/no decision as to whether the face seen had been shown in the previous pretest exposure period.

In the testing phase the subjects viewed thirty-two occluded faces in consecutive trials. Each trial consisted of a 2 s presentation of a fixation frame presented binocularly and at zero disparity with respect to the CRT face. This was followed by a 600 ms presentation of the sampled face at the same disparity as the fixation plane. The horizontal slats not containing any facial information were presented either at 8 min visual angle of crossed disparity (in front) or 8 min of uncrossed disparity (in back). It should be clear that by briefly presenting the face at the same disparity as the pretrial fixation frame, the need for, and dependence on, vergence eye movements was removed. The faces subtended an angle of 4 deg × 5 deg. Sparse binocular noise (light



(a) (b)

Figure 7. Pictorial representation of a partially visible face as it would appear in either (a) a front or (b) a rear stereoscopic depth plane.

and dark dots) with the same local disparity of the other points in the same horizontal strip was added to the stimulus to make pattern recognition less than 100% and also to ensure that an appropriate stereoscopic segregation of the two depth planes was established.[2]

Photographs of thirty-two famous individuals taken from the fields of entertainment or politics were obtained from books and magazines, digitized using a TV camera, and stored in the microcomputer (Commodore Amiga 1000). The selection of sixteen faces out of the thirty-two was counterbalanced across the subjects to ensure that all faces were exposed with equal probability in the pretest and with equal probability of appearing in the front or rear plane in the test situation. The trial sequence was also constrained such that not more than three 'fronts' or three 'backs' were presented in a row. To obtain separate inputs to the two eyes as required for stereopsis, a phase haploscope was used. This alternated CRT images destined for the left and right eyes in synchrony with the opening and closing of very fast PLZT electrooptic shutters mounted on spectacles (Model SDC-105, Stereoptic Systems, San Diego, CA).

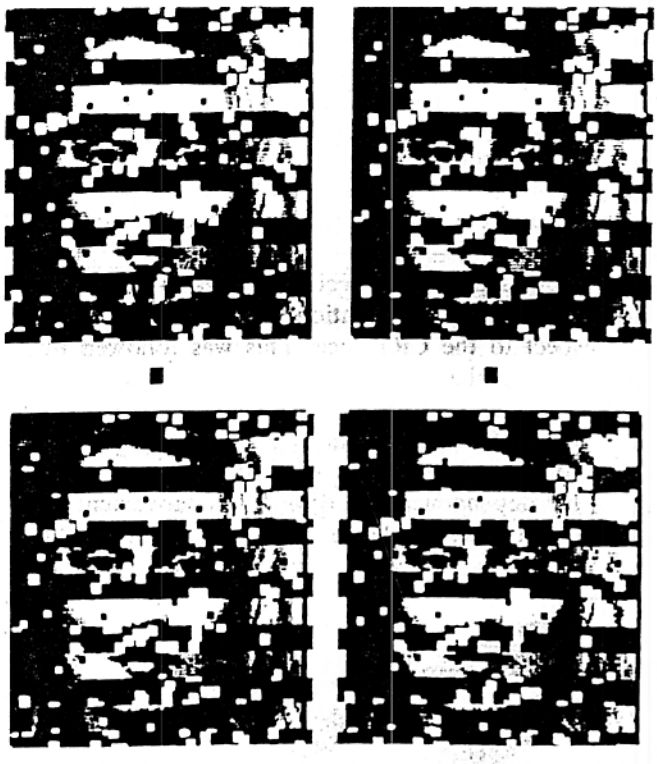


Figure 8. Example of stereograms presented in the face recognition experiment. Under crossed fusion, the face will appear in the rear plane for the upper stereo pair and in the front plane for the lower pair. Depth relations are reversed if fusion is obtained by divergence.

[2] The horizontal gradients of luminance intensity across any given face were not very abrupt. As a consequence, the sampled face which alternated with a series of blank occluders (as schematized in figure 7) did not yield a reliable segregation into two depth planes. Only with the addition of some randomly sprinkled 'pixels' in each depth plane which supplies more localized disparity cues (see Schor et al 1984) was the stereoscopic segregation adequately robust for the experiment.

2.2 *Results*

The results of the experiment can be seen in figure 9, which shows the number of errors for each of the eight subjects for both conditions. In all but one of the eight observers (who had just one error in each case), the number of errors was greater when the faces were in front. A chi-square test of this distribution of subjects was significant at the 0.05 level, and a separate chi-square test of the pooled scores indicated that the difference was significant at the 0.01 level. Thus, the results support our hypothesis by showing that a complex partially visible form is much more recognizable if it is presented in a back rather than in a front stereoscopic plane.

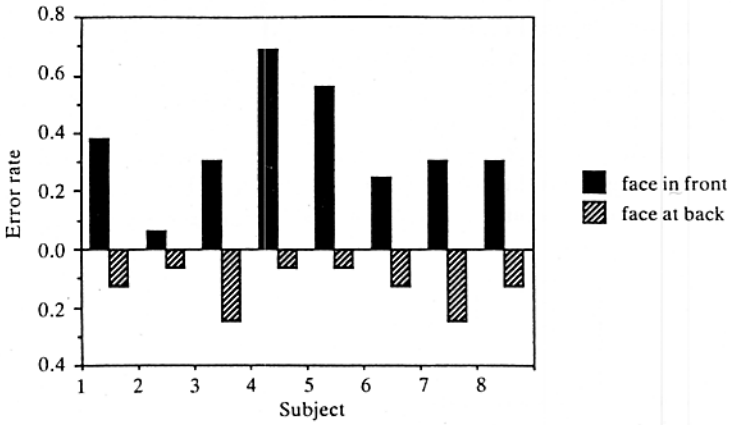


Figure 9. Individual error rates for all eight subjects in the face recognition experiment.

3 General discussion

3.1 *Our findings cannot be explained in terms of known stereoscopic phenomena*

Before we draw additional conclusions from the present results, we consider the findings in the context of contrasting psychophysical results where small patterns are presented in multiplanar stereoscopic displays. Fox and colleagues have reported the clear superiority of the front stereoscopic plane for the detection and identification of small (unoccluded) targets. For example, in metacontrast-like displays where the mask is in one depth plane and the target is in another, performance is always superior when the target appears in front (Lehmkuhle and Fox 1980; Fox and Patterson 1981). We have also witnessed a similar phenomenon when conducting experiments on visual search (Nakayama and Silverman 1986). We found that the reaction time to find a target amongst a set of distractors was shorter if the target was in a front rather than a rear stereoscopic plane. Thus for the detection or identification of small targets, it is clear that there is a bias towards better performance in the front stereoscopic plane.

We mentioned these results because they are exactly opposite to those obtained in the present study and demonstrate that the specific phenomenon reported here cannot be attributed to these previously reported effects. As such, it strengthens our argument and supports our conclusion that the rear plane superiority is a distinct phenomenon related to the occlusion of objects and to the grouping of separated image fragments in the more distant depth plane.

3.2 *Functional considerations*

At this point we would like to summarize our interpretation of the results by making a number of propositions:

(i) Partially visible surfaces are bounded by two types of edges, those inherent to the surface itself (intrinsic) and those defined by occlusion (extrinsic).

(ii) Encoded depth provides a computationally plausible basis for making the distinction between extrinsic versus intrinsic edges because in the real world the border in common between the two regions always 'belongs' to the front region.
(iii) If an edge is classified as intrinsic, it belongs to that surface and provides a useful input for pattern recognition. Conversely, if it is classified as extrinsic, it does not belong to that surface. Thus, it can and should be shielded from the process of pattern recognition.[3]
(iv) Image regions containing extrinsic edges which face each other tend to link with other image regions similarly bounded such that these regions appear to link behind an occluder.

In figure 10 we illustrate these points in a more diagrammatic manner. Consider the regions labeled upper (U), middle (M), and lower (L). If a region, say M, is labeled by the visual system as being in front, then according to our hypothesis the shared borders between region M and the other regions (contours C and C′) will be attached to area M and detached from areas U and L. Conversely, if M is labelled as 'in back', then the common borders will be 'detached' from M and 'attached' to U and L, respectively. As suggested earlier, such a classification by depth will make a crucial difference in terms of the inputs to a hypothetical pattern recognition process.

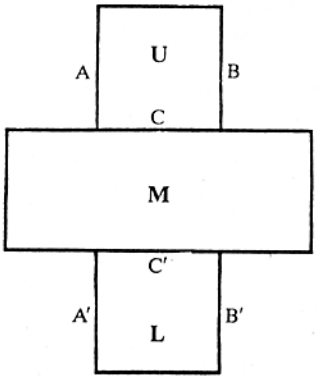


Figure 10. Schematic diagram of contour relations. Portion of an image containing surface patches bounded by contours is shown. We suggest that if area M (middle) is coded as in front, then contours C and C′ are intrinsic to area M but extrinsic to areas U (upper) and L (lower) respectively. Furthermore, we hypothesize a neural representation of the linkage of collinear contours A and A′ as well as B and B′ if M is coded as in front (see text for further details).

[3] At first glance it may seem that the situation may be more complicated when the object is self-occluding (eg when one portion of an object covers another region, leading to an 'intrinsic versus extrinsic' classification problem within a single target). In such a case, both the occluded and occluding contours might be considered 'intrinsic' to the object. However, note that the processes of disparity-dependent edge classification and of amodal completion behind an occluding surface can still be applied, with or without prior knowledge as to whether the occluded and the occluding edges are part of the same single object. To decide whether an object occludes another object or whether they are two parts of a single object requires further elaboration of perceptual processes, making use of other cues such as three-dimensional curvature of contours and surfaces. Further, it is often the case that the self-occluded and the self-occluding surfaces can be readily regarded as two separable parts of one object (eg a human arm partially occluding the trunk), which may be treated as two 'subobjects' at a more detailed level of the hierarchical object representation (Hoffman and Richards 1985; Marr and Nishihara 1978). Thus, the possibility of self-occlusion does not invalidate the kind of algorithm that we propose here, and the terminology of edge classification (intrinsic/extrinsic) can be preserved in the self-occlusion case as well.

In the present paper we have restricted our study to stereoscopic depth and have argued for its role in the categorization of the edges of adjacent surfaces. We would like to think, however, that this labeling is based on more general depth mechanisms, in particular those mediated by monocular cues as well (see also Nakayama et al 1989; Shimojo et al 1989). For example, it is reasonable to suppose that the increase visibility of the letters in figure 1b as opposed to figure 1c can also be the result of monocular depth processing. In particular, the presence of the T-junction in figure 1b provides support for occlusion (recognized by Helmholtz 1909/1962), while only 'L'-junctions are present in figure 1c. We argue that L-junctions provide support for the interpretation that the edges are intrinsic rather than extrinsic (see Guzman 1968).

Consistent with the phenomenology associated with figures 1, 2, and 4, we also speculate that gaps in collinear pairs of contours, eg A and A′ as well as B and B′ (in figure 10), will be differently represented in the nervous system depending on whether the middle region (M) is labeled as in front or in back. If M is in front, we suggest that the collinear lines are linked and could provide, at some level of representation, information regarding the existence of the hidden contour, one that could conceivably form an input for later template matching in the pattern recognition process. We think this linking process as embodied in the representation of the hidden contour is of possible importance in correctly grouping the remaining image fragments. This notion has been foreshadowed by Kanisza (1979), who made the distinction between two types of contours which are not explicit in the raw gray-level image: 'amodal' contours (the invisible contours hypothesized here) and 'modal' contours [as exemplified by the well-known subjective contours; see also Michotte (1954) and Grossberg and Mingolla (1985)].

Earlier, we considered a number of possibilities to explain the phenomenon seen in figure 1. The first was a higher level explanation requiring that the region be encoded as a distinct entity independent from the figure to be recognized. Second, and the hypothesis that we favor, is the simpler view that edge classification is based on relative depth. We argue that this edge classification precedes complex pattern recognition because *its* very purpose is to aid in the implementation of the recognition process under conditions of partial visibility. At this point it is of some interest to consider the possible level of the visual system at which the hypothesized processes could be implemented.

3.3 *The question of neural loss and mechanism*

We would like to suggest that such a primitive process could occur very early in the chain of visual processing, much earlier than the stage of template matching. Two lines of evidence support the plausibility of such a view.

First, we (Shimojo et al 1988) have found that this same intrinsic versus extrinsic distinction also applies to the problem of motion encoding, a process generally considered to be much more primitive and earlier than learned pattern recognition. By using the Wallach barber-pole illusion (Wallach, 1935), we were able to show that the usual solution to the 'aperture problem' afforded by the long axis of a viewing window did not apply under specific stereoscopic manipulations. In particular, the vertical dominance of motion in a vertically elongated aperture disappeared when uncrossed disparity was added to the drifting stripes relative to the edges of aperture. Thus we made the analogous argument for motion as we do here for pattern recognition. Moving line terminators in the image which are thought to resolve the aperture problem must also be subjected to the same intrinsic versus extrinsic test. Those terminators which are formed by occlusion and occasioned when the stripes are seen as far behind the aperture plane do not assist in the solution of ambiguity. Because the encoding of motion is regarded as a relatively early cortical process and because we

have shown that it also requires this intrinsic versus extrinsic classification in accordance with occlusion constraints, it increases the likelihood that such a classification occurs relatively early.

Second, the physiological substrate for such a depth-based classification is well established at very early cortical levels. For example, stereoscopic depth relationships in the scene can be represented as early as V1 and are well elaborated by V2 (Fischer and Poggio 1979). Thus, in principle, the classification scheme that we suggest could begin very early, perhaps as early as V1 or V2.

Having made the case for a relatively early classification of edges and its possible role in the linking and segmentation of images for the purposes of pattern recognition, it seems natural to ask whether known properties of cells in striate or extrastriate cortex could act as a plausible substrate. In this regard the intrinsic horizontal connections of striate cortex come to mind. A recent study suggests relatively specific domain interactions over several millimeters across striate cortex, possibly mediated by horizontally directed axons at the supragranular level (Ts'o et al 1986; Ts'o and Gilbert 1988). Cells sharing a similar orientation preference or color preference are interconnected, even though their receptive fields are nonoverlapping. A hypothesis arising from our results is that such horizontal linkages could be gated by local depth signals, thus ensuring that properties are linked, but selectively according to whether an intervening image region is closer.

As a specific example, consider a line which has an end or terminator. We have indicated that the existence of a line terminator in an image is ambiguous, either indicating the real end of a line or the continuation of a line behind an occluder. If processes outlined in this paper can be accounted for by stages of cortical processing, that are relatively low level, then it becomes of interest to ask whether at least a fraction of cells having end-stopped properties, or alternatively those having extremely long receptive fields (Gilbert 1977), might have rather different susceptibilities to the encoded depth of regions which interrupt or break lines of an image. We would predict that if an end-stopped cell were to encode the real end of a line it would fire so preferentially when the line breaking region was in back rather than in front. Conversely, cells which might be thought of as coding hypothetical contours behind an occluder might respond preferentially if the line breaking regions were coded in front rather than in back.

3.4 *Implications for machine vision*

From the viewpoint of classical matching vision, the most important step towards object recognition is *image segmentation*. Once the image is correctly segmented into subparts which correspond to physical objects, then required processes for pattern matching are much easier. Despite considerable efforts over a long period, however, the theory and practice of segmentation have remained primitive for several reasons.

The most fundamental problem with this classical idea of segmentation, according to Marr (1982), is that 'objects' and 'desirable regions' (those to undergo the recognition process) are almost never visually primitive constructions, and hence cannot be recovered from any early representation unless specialized knowledge about the class of possible objects is available. In fact, the classical image segmentation approaches have tried to apply different sets of constraints on a segmentation algorithm, depending upon different types of scene and different types of objects to expect. This is presumably the reason why unrealistically large amounts of visual knowledge are required to obtain only a small increment of generality. [For example, rules such as 'runways are oriented parallel to terminal building' or 'runways do not have curved segments' can be useful only to interpret airport scenes: see McKeown et al (1985). See also Tenenbaum and Barrow (1976).]

Marr (1982) maintained that the basis for segmentation must be embedded in the early visual processes as general constraints, together with the geometrical consequences of the fact that surfaces coexist in three-dimensional space. The large gap between the two-dimensional viewer-centered early representation of visual input (the 'primal sketch') and the three-dimensional object-centered models of things to recognize, motivated him to propose the construction of a viewer-centered surface representation (the '$2\frac{1}{2}$-D sketch') as an intermediate stage necessary to bridge the gap.

We agree with Marr in that image segmentation should be based on early implementation of general constraints and should not depend on the knowledge of objects. However, side by side with a *surface* representation such as proposed by Marr, the kind of processing suggested by the present study offers an important and perhaps more efficient shortcut, one which only requires the encoding of relative depth in a crude *contour* representation.

As an example, consider a side view of a horse partially hidden by a tree. When the viewing angle is a preferable one in that the canonical axis of the object (the head-tail axis of the horse) is close enough to the frontoparallel plane, the two-dimensional projection carries sufficient information about the prototypical characteristics of the object (long face and neck, four skinny legs, a broom-like tail, etc) for it to be already sufficient for recognition of the object (see Marr and Nishihara 1978; Biederman 1985). Thus, as soon as the contour of horse is disambiguated from the contours of the tree by crude processing of binocular disparity or other depth cues, the object may be quickly recognized as a horse by simple processes such as feature detection or two-dimensional template matching (see Nakayama 1988) without the construction of a detailed representation of depth or orientation of local surfaces.

3.5 *An alternative role for stereoscopic vision*

It is generally accepted that stereopsis plays a major role in the metrical encoding of distances in the third dimension. Our results suggest, however, that this presumed role for stereopsis may be overemphasized. We would like to suggest that there are other more biologically fundamental and phylogenetically ancient roles for stereopsis satisfying the need to detect and recognize patterns. On the basis of comparative physiological findings in birds, for example, Pettigrew (1986) has suggested that one of the major functions of stereopsis is to break camouflage, especially as it seems to be present only in predatory birds which attack prey against the ground. Stereoscopic vision, for example, does not seem to be present in predators which attack prey which are flying above. Pettigrew argues that such targets would be clearly seen against the sky and would not require anticamouflage procedures for detection.

The present study suggests yet another role for stereopsis, that of delineating and linking parts of an object which are partially hidden. It is of interest that such a mechanism would require only the most primitive form of stereopsis, one that codes the sign of relative disparity and not its magnitude.

Acknowledgements. This research was supported by Grant 83-0320 from AFOSR and Grants EY-05408 and EY-01186 from the National Institutes of Health to the first author (KN). The second author (SS) was supported by the Japanese Society for the Support of Junior Scientists and the Rachel Atkinson Fellowship Fund. The authors would like to thank Nance Wilson for assistance in preparing the manuscript.

References

Biederman I, 1985 "Human image understanding: recent research and a theory" *Computer Vision, Graphics, and Image Processing* **32** 29–73

Bregman A L, 1981 "Asking the "what for" question in auditory perception" in *Perceptual Organization* Eds M Kubovy, J R Pomerantz (Hillsdale, NJ: Lawrence Erlbaum Associates) pp 99–118

Fischer B, Poggio G F, 1979 "Depth sensitivity of binocular cortical neurons of behaving monkeys" *Proceedings of the Royal Society of London, Series B* **204** 409 - 414

Fox R, Patterson R, 1981 "Depth separation and lateral interference" *Perception & Psychophysics* **30** 513 - 520

Gilbert C D, 1977 "Laminar differences in receptive field properties of cells in cat primary visual cortex" *Journal of Physiology (London)* **268** 391 - 421

Gregory R L, Harris J P, 1974 "Illusory contours and stereo depth" *Perception & Psychophysics* **15** 411 - 416

Grossberg S, Mingolla E, 1985 "Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations" *Perception & Psychophysics* **38** 141 - 171

Guzman A, 1968 "Decomposition of a visual scene into three dimensional borders" *Fall Joint Conference* 33; reprinted (1984) in *Information Technology Series, Volume VI, Artificial Intelligence* Ed. O Fischein (Reston, VA: AFIPS Press) pp 310 - 335

Helmholtz H von, 1909/1962 *Physiological Optics* volume 3 (New York: Dover, 1962); English translation by J P C Southall for the Optical Society of America (1924) from the 3rd German edition of *Handbuch der physiologischen Optik* (Hamburg: Voss, 1909)

Hoffman D D, Richards W, 1985 "Parts of recognition" *Cognition* **18** 65 - 96

Julesz B, 1971 *Foundations of Cyclopean Perception* (Chicago, IL: University of Chicago Press)

Kanisza G, 1979 *Organization in Vision. Essays in Gestalt Perception* (New York: Praeger)

Lawson R B, Cowan E, Gibbs T D, Whitmore C D, 1974 "Stereoscopic enhancement and erasure of subjective contours" *Journal of Experimental Psychology* **103** 1142 - 1146

Lehmkuhle S, Fox R, 1980 "Effect of depth separation on metacontrast masking" *Journal of Experimental Psychology: Human Perception and Performance* **6** 605 - 621

Marr D, 1982 *Vision* (San Francisco, CA: W H Freeman)

Marr D, Nishihara H K, 1978 "Representation and recognition of three dimensional shapes" *Proceedings of the Royal Society of London, Series B* **200** 269 - 294

McKeown D M Jr, Harvey W A Jr, McDermott J, 1985 "Rule-based interpretation of aerial imagery" *IEEE Proceedings on Pattern Analysis and Machine Intelligence* **7** 570 - 585

Michotte A, 1954 *La Perception de la Causalité* (Louvain: Publications Universitaires)

Nakayama K, 1989 "The iconic bottleneck and the tenuous link between early visual processing and perception" in *Visual Coding and Efficiency, Festschrift for Horace B Barlow* Ed. C Blakemore (Cambridge: Cambridge University Press)

Nakayama K, Shimojo S, 1988 "Depth, rivalry and subjective contours from unpaired monocular points" *Investigative Ophthalmology and Visual Science (Supplement)* **29** 21

Nakayama K, Shimojo S, Ramachandran V S, 1989 "Transparency: its relation to depth, subjective contours, and neon-color spreading" submitted to *Perception*

Nakayama K, Silverman G H, 1986 "Serial and parallel processing of visual feature conjunctions" *Nature (London)* **320** 264 - 265

Neisser U, 1967 *Cognitive Psychology* (New York: Appleton-Century-Crofts)

Pettigrew J D, 1986 "The evolution of binocular vision" in *Visual Neuroscience* Eds J D Pettigrew, K J Sanderson, W R Levick (Cambridge: Cambridge University Press) pp 208 - 222

Schor C, Wood I, Ogawa J, 1984 "Spatial tuning of static and dynamic local stereopsis" *Vision Research* **24** 573 - 578

Shimojo S, Silverman G H, Nakayama K, 1989 "Occlusion and the solution to the aperture problem for motion" *Vision Research* (in press)

Tenenbaum J M, Barrow H G, 1976 "Experiments in interpretation-guided segmentation" Technical Note 123, Stanford Research Institute, Stanford, CA, USA

Ts'o D Y, Gilbert C D, 1988 "The organization of chromatic and spatial interactions in the primate striate cortex" *Journal of Neuroscience* **8** 1712 - 1727

Ts'o D Y, Gilbert C D, Wiesel T N, 1986 "Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis" *Journal of Neuroscience* **6** 1160 - 1170

Ullman S, 1986 "An approach to object recognition: aligning pictorial descriptions" Memo 931, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

Wallach H, 1935 "Über visuell wahrgenommene Bewegungrichtung" *Psychologische Forschung* **20** 325 - 380