Perception, 1999, volume 28, pages 1347 – 1360

DOI:10.1068/p2987

Surface formation and depth in monocular scene perception

Marc K Albert
Vision Sciences Laboratory, Harvard University, Cambridge, MA 02138, USA;
e-mail: malbert@wjh.harvard.edu
Received 1 March 1999, in revised form 9 August 1999

Abstract. The visual perception of monocular stimuli perceived as 3-D objects has received considerable attention from researchers in human and machine vision. However, most previous research has focused on how individual 3-D objects are perceived. Here this is extended to a study of how the structure of 3-D *scenes* containing multiple, possibly disconnected objects and features is perceived.

Da Vinci stereopsis, stereo capture, and other surface formation and interpolation phenomena in stereopsis and structure-from-motion suggest that small features having ambiguous depth may be assigned depth by interpolation with features having unambiguous depth. I investigated whether vision may use similar mechanisms to assign *relative* depth to multiple objects and features in sparse monocular images, such as line drawings, especially when other depth cues are absent. I propose that vision tends to organize disconnected objects and features into *common surfaces* to construct 3-D-scene interpretations.

Interpolations that are too weak to generate a visible surface percept may still be strong enough to assign relative depth to objects within a scene. When there exists more than one possible surface interpolation in a scene, the visual system's preference for one interpolation over another seems to be influenced by a number of factors, including: (i) proximity, (ii) smoothness, (iii) a preference for roughly frontoparallel surfaces and 'ground' surfaces, (iv) attention and fixation, and (v) higher-level factors. I present a variety of demonstrations and an experiment to support this surface-formation hypothesis.

1 Introduction

In figure 1a most naive observers simply see two tilted ellipses lying in the frontoparallel plane. However, in figure 2a these ellipses appear to be slanted in depth and attached to the top faces of the blocks. A number of authors have suggested explanations for why 3-D blocks are seen in this kind of display (eg Waltz 1975), but why do the ellipses appear to be attached to the blocks? More surprisingly, the zigzag line in figure 1b tends to look flat and frontoparallel, whereas the same line in figure 2b appears to conform to the shape of the cross section of the cylinder above it.

(a) (b)

Figure 1. In (a) most observers simply see frontoplanar ellipses or ellipsoids. In (b) most observers see a flat, frontoplanar zigzag line.

In general, line drawings depicting multiple 3-D objects and features are ambiguous in the sense that each object or feature could lie anywhere along its visual ray to the observer. So there is a continuous infinity of logically possible 3-D scene interpretations. Although in many images other depth cues are available (eg T-junctions, height-in-the-field, size constancy, perspective, familiarity), observers nevertheless often assign rather definite scene interpretations in images without such cues.

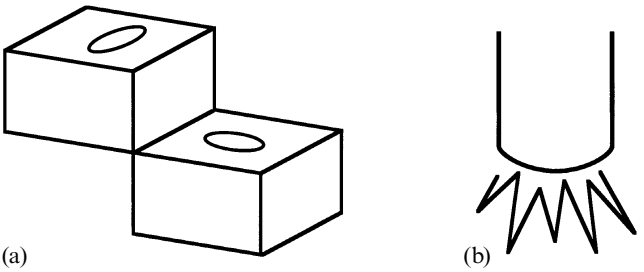


Figure 2. In (a) most observers see the ellipses as lying in the top faces of the blocks, and as having different shapes than the ellipses in figure 1a. In (b) most observers see the zigzag line as partly conforming to an imaginary extension of the cylindrical surface (cf figure 1b).

I suggest that *surface perception* is crucial to understanding the processes by which vision constructs relative depth in 3-D scene perception. For example, T-junctions are generally considered to be one of the most local and context-independent cues to depth order. However, figure 3 suggests that the efficacy of T-junctions can be quite dependent on global image structure: The contours leading into a T-junction must appear to be the boundaries of *surfaces* for the T-junction to function effectively as a cue to depth order. By reducing the 'closure' of these contours we can weaken the perception that they 'belong' to a surface. Figure 3 shows that this also weakens the perception of 'interposition' created by the T-junction.

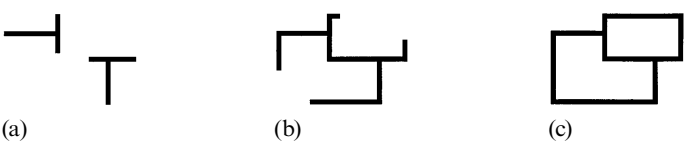


Figure 3. In (a) most observers do not perceive the T-junctions as points of occlusion. In (b) and especially in (c) they do. This suggests that T-junctions are not purely local cues to occlusion. T-junctions are cues to occlusion only when the contours are interpreted as boundaries of *surfaces*. In this example the greater closure of the contours in (b) and (c) supports this interpretation.

I propose that in displays such as figure 2a the block's 3-D representation inputs to a *surface-formation mechanism*. Although each ellipse has ambiguous depth relative to the top face of the block, it can still participate in this surface-formation process as long as its internal perspective information is consistent with the orientation of the interpolated surface. The ellipses appear to be attached to the top faces of the blocks because 'good' interpolated surfaces (ie planes) containing them and the 3-D edges that bound the top faces of the blocks can be constructed. This example suggests that surface formation may play an important role in depth and form perception of 3-D scenes.

2 Line drawings depicting multiple objects and features: Cue interactions and depth assignment

In figure 4 the bottom edge of the small block is colinear with the bottom edge of the large block, and they also appear to be colinear in 3-D (also see Jepson and Richards 1993). This is consistent with the *generic-viewpoint assumption*, since the observer would have to have an 'accidental' or 'special' viewpoint of the scene in order for noncolinear edges in 3-D to project to colinear edges in the image.

Thus, the front faces of the two blocks are perceived to be coplanar and at roughly the same depth from the observer. The blocks also appear to be resting on a common illusory 'ground surface.' However, for many observers the interpretation of this image changes when the figure is rotated clockwise by 90°. The small block now appears to be somewhat closer than the large block, and the rear face of the small block appears

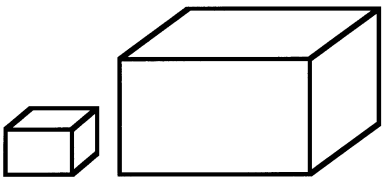


Figure 4. The perceived relative depths of the blocks depends on the orientation at which the stimulus is viewed. In the given orientation the front faces of the two blocks generally appear to be roughly coplanar in 3-D. If the figure is rotated clockwise, and particularly if it is rotated anticlockwise, this percept changes for many observers. Now the front face of the small block generally appears to be roughly coplanar with the rear face of the large block. See text for further details (modified from Jepson and Richards 1993).

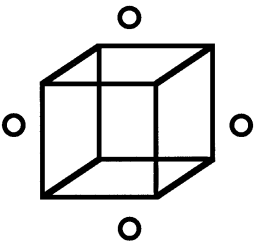


Figure 5. Each of the circles surrounding the Necker cube appears coplanar with either the front or the rear face of the cube. The circles that are adjacent to an edge belonging to the front face of the cube appear coplanar with the front face. Those that are adjacent to an edge belonging to the rear face of the cube appear coplanar with the rear face.

to be approximately coplanar with the front face of the large block. On the other hand, if the figure is rotated 90° anticlockwise relative to the original orientation, then the small block again appears to be closer than the large block, and the depth difference between the blocks is now more pronounced than when the figure was rotated 90° clockwise.

I suggest that a variant of the *proximity* rule of depth assignment is one factor affecting our perception of this display. According to this rule, features that are close to each other in an image should be interpreted as being close to each other in 3-D space. The proximity rule is another instance of the generic-viewpoint assumption (in the sense of 'small' probabilities): If two features are widely separated in space, then only a small range of viewpoints would place them near each other in an image. Thus it is more likely that these features are near each other in 3-D.[(1)] This rule is closely related to the 'equidistance tendency' of Gogel (1956), which states that objects that are nearby in an image tend to be perceived as nearby in 3-D. In figure 5 the apparent depths of the small circles (or bubbles) relative to the edges of the block are consistent with the proximity rule. In figure 4 the *features* on the blocks that are nearest to each other in the image are the right rear edge of the small block and the left front edge of the large block. A 'generalized' proximity rule would therefore predict that these features should be seen to be at similar depths, and so the small block should appear to be in front of the large block.

In addition, the well-known 'height-in-the-field' rule (eg Gibson 1950) appears to be influencing our perception of this display. According to this rule, objects that are higher in the visual field tend to be perceived to be at a greater distance from the observer, at least those objects perceived to lie below the horizon. In the original orientation of figure 4, the bottom edges of the blocks are at the same height in the visual field, so the height-in-the-field rule supports the inference that they are at the same depth in space. As discussed above, this interpretation is also supported by the colinearity of the bottom edges, but it conflicts with the (generalized) proximity rule. Thus, it appears that proximity has been overruled in this case by the combined effect of the height-in-the-field rule and the colinearity rule.

When figure 4 is rotated clockwise by 90°, the height-in-the-field rule might be expected to be inoperative since the observer is looking up at the blocks from below rather than down at them from above. In other words, the bottom faces rather than the top faces of the blocks are now visible. Although an 'aerial' viewpoint interpretation in which the observer is looking at the scene from 'overhead' is also possible here, the height-in-the-field rule might be expected to be less effective under this interpretation as well,

(1) This inference assumes that the features are, say, equally likely to be separated by all possible distances, up to some maximum distance D.

since from an aerial viewpoint the differences in the heights of objects in the visual field represent relatively small differences in distance from the observer. In this orientation, the influence of proximity appears to be at least as strong as, if not stronger than, colinearity.

When the original display is rotated anticlockwise by 90°, the height-in-the-field rule predicts that the small block should appear to be closer than the large block, in agreement with proximity. The combination of these rules now appears to strongly overrule colinearity. On the other hand, in general colinearity might be expected to be a less significant factor for displays in which the colinear edges are interpreted as belonging to distinct, independent objects.

3 Line drawings depicting multiple objects and features: Surface formation and depth assignment

3.1 *The theory of Jepson and Richards*

Many researchers have suggested that the perception of monocular line drawings is primarily the result of relatively high-level or cognitive processing. The analyses of figures 4 and 5 presented in the previous section might be seen as consistent with this view, although it is certainly possible that the interpretation 'rules' used in those analyses could be implemented by low-level mechanisms.

Jepson and Richards (1993) and Richards et al (1996) studied line drawings of a simplified blocks world, and developed a Bayesian framework for vision based on constraints and preferences that reduce the ambiguity of visual inference. Their general procedure is as follows: First, they construct a state space of the possible qualitative percepts of an image using certain 'hard' (ie inviolable) constraints. For example, they require that the faces of blocks be solid and impenetrable. Then, the qualitatively different percepts are organized as a partially ordered set with the use of various perceptual preference rules. The maximal elements of this partial order correspond to the set of possible perceptual interpretations of the image. If there is more than one maximal element, then perception may be multistable.

One of the most important perceptual preferences used in their theory is a preference for *physical stability*: Interpretations in which objects and features are stably resting on other objects or features (under the influence of gravity), or are in some other way 'attached' to them, are preferred over interpretations containing 'floating' objects or features. The impenetrability constraint for blocks mentioned above permits an unambiguous definition of what constitutes a 'physically stable' arrangement.

For example, in figure 6 most observers report that the two blocks appear to be resting on a table, which is perhaps assumed to be stable owing to some structure outside the observer's field of view. A much less preferred interpretation is that the smaller block is attached to the back face of the larger block, although this interpretation becomes somewhat easier to see if the figure is turned upside down. But note that the blocks almost never appear to be just freely floating in the space above the table. They generally appear to be either directly or indirectly attached to the table. Similarly, the ellipses in figure 2a appear to be stably attached to the top faces of the blocks.

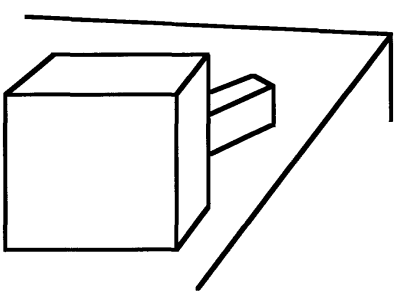


Figure 6. Most observers see both blocks as resting on a table. However, some see the small block as attached to the rear face of the larger block, particularly if the figure is viewed upside down (after Jepson and Richards 1993).

On the other hand, figure 4, when it is rotated 90° clockwise, and figure 5 seem to produce percepts of 'floating' objects that are not constrained by any requirement for physically stable interpretations (see below). So perhaps these effects are not really high-level after all, at least not completely.

4 An alternative explanation: The surface-formation hypothesis

I would like to suggest an alternative, more cognitively insulated explanation. I propose that 3-D scene interpretation in these displays is primarily determined by a relatively automatic surface-formation/interpolation mechanism. The input to this mechanism includes the 3-D structures of the perceived individual objects, if information for computing their 3-D structure is available. For example, in figure 2a the edges of the blocks strongly suggest a definite 3-D structure, whereas the ellipses do not. These input representations might be constructed by using an algorithm of the kind proposed by Waltz (1975).

I propose that surface formation or interpolation can occur between surfaces, lines, and points, as in figure 2a, as well as between surfaces, edges, and vertices of objects and features. This includes interpolations between partly or wholly occluded surfaces which are 'inferred' by the visual system (ie amodal surfaces), such as the bottom or back face of a block. Note that each 'hidden' face of a block still has two visible *edges* in the image. So, if the observer can correctly perceive these edges in 3-D, then the visual system could amodally 'interpolate' the surface spanned by these edges and thereby 'construct' the hidden amodal faces. Thus, a symmetry assumption for object shape is not necessary here.

In addition, the perceptions of figure 2b, figure 4 when it is rotated 90° clockwise, and figure 5 suggest that surface formation may sometimes operate without generating a strongly *visible* interpolated surface. It may generate only a barely perceptible transparent surface, or none at all. In other words, the activity threshold for the perception of a visible interpolated surface may be significantly higher than the threshold for influencing the perceived relative depths of the objects and features themselves.

Of course, I do not claim that surface formation *always* determines depth in 3-D scene perception of monocular line drawings. I believe that it is primarily a default, similar to the role that Gogel (1956) claimed for the equidistance tendency: Surface formation will be decisive only if it is not contradicted by other strong depth cues.

For example, the rotated square on the block in figure 7 is difficult to see as lying on the top face of the block. The symmetrical shape of the square apparently suggests that it lies in the frontoparallel plane. As a result, the surface interpolation between the edges of the top face of the block and the square is not very strong. (Compare this with figure 2a.) Instead, the square appears to be approximately flush with the inside rear face of the block. The upper edge of the square appears to be at approximately the same depth as the rear upper edge of the block, consistent with the equidistance tendency and the generalized proximity rule. This percept may be the result of a compromise between the surface-formation mechanism and the square's shape cue. More generally, some percepts may be the result of a compromise between competing surface processes, or between surface attraction/attachment and other depth cues.

Jepson and Richards (1993) proposed that human vision prefers 'physically stable' interpretations. However, such a preference may not require knowledge-based reasoning

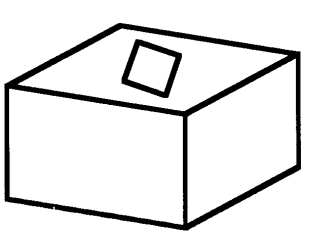


Figure 7. The diamond-like figure does not attach well to the top face of the cube because its shape suggests that it is frontoparallel to the viewer. Instead it appears to either sit on the block's top face, resting on its bottom corner, or to be inside of the top of the block (ie the block's top face is absent from the percept), perhaps attached to the inside surface of the block's rear face.

processes. It seems plausible that a mechanism based on surface formation and attraction might implement a useful 'heuristic' that would generally favor physically stable interpretations, since interpolation of a visible surface between features implies that the features are seen as 'attached' to the interpolated surface. In fact, the kind of mechanism which I have outlined might be implemented in a biologically plausible way with cooperative processing, similar to the cooperative algorithms proposed for stereopsis.

A critical test of the surface-formation hypothesis (ie a display in which surface attraction can explain the percept but physical stability cannot) is shown in figure 5. Each small circle (or bubble) appears to be approximately coplanar with either the front or the rear face of the Necker cube. Each circle appears to be coplanar with the face that it is closest to in the image, ie the surface interpolation determined by generalized proximity is favored. And note that when the Necker cube 'flips', the circles flip with it. This percept is difficult to explain in terms of physical stability, since most observers do not *see* a surface extending beyond the faces of the cube. Similar considerations apply to figure 2b.

Also, in figure 8 observers sometimes report that the front small block appears to be somewhat inside the large block, particularly when fixating the lower left corner of the large block. Notice that the front face of the large block is missing in this percept. Similarly, in figure 9 the small block can appear to be somewhat inside the left face of the large block when fixating the lower rear corner of the large block. Both of these percepts are inconsistent with a 'hard constraint' that the faces of blocks are solid and impenetrable. On the other hand, they are consistent with assigning depth on the basis of surface formation and attraction, and with a preference for surfaces that have a 'ground plane' orientation.

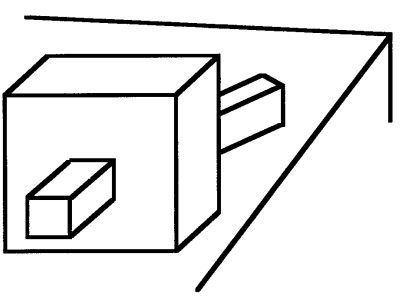


Figure 8. Sometimes the front small block is seen as attached by its rear face to the front face of the large block. However, at other times the small block can appear to be *inside* the large block, particularly on fixating the lower left corner of the large block. In the latter percept the small block may seem to hover somewhere between the former percept and a percept in which the small block is resting on the inside surface of the bottom face of the large block. In this case the front face of the large block is absent.

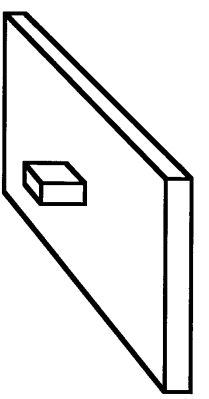


Figure 9. Sometimes the small block is seen as attached by its right face to the left face of the large block. However, at other times the small block can appear to be *inside* the large block, particularly on fixating the lower rear corner of the large block. In the latter percept the left face of the large block is absent.

In many of these displays it appears that proximity is influencing the selection of preferred surfaces. However, figure 10 shows that proximity is not always the determining factor (also see Jepson and Richards 1993). For many observers, when the Necker cube is perceived as being viewed from above, the features on the handle and the cube that are nearest to each other in the image (ie the leftmost corner of the handle and the leftmost vertical edge of the cube) are not perceived as being nearby in 3-D. So this percept cannot be explained by the equidistance tendency or by the influence of proximity on surface formation. As discussed above, figure 4, when viewed in the original orientation, also

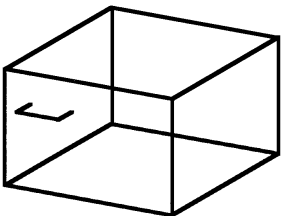


Figure 10. When the Necker cube is perceived as being viewed from above, the 'handle' is often seen as attached by its 'feet' to the outside surface of the front left face of the cube. In this percept proximity is not the determining factor in the visual system's preferred surface interpolation. If it were, then the handle should be seen as attached by its long edge to the inside surface of the front left face of the cube. Other percepts are also possible. (after Jepson and Richards 1993)

violates proximity. The surface-formation hypothesis may be viewed as a generalization of Gogel's equidistance tendency, allowing for non-frontoparallel surface formation and for factors other than proximity to affect surface formation.

Surface formation could also be combined with a depth-'anchoring' rule to infer the absolute depth of objects and features in 3-D scenes. Such a rule might fix the depth of the most salient object in the scene, or perhaps the geometric center of the scene, by using a 'specific distance tendency'. Then the depths of the other objects could be determined relative to that anchor point.

5 Factors affecting surface formation

According to my hypothesis, the inputs to the surface-formation process are representations of the 3-D structures and orientations of the objects and features in the scene (eg blocks, handles), if they are available, as in Jepson and Richards (1993). These representations could be constructed, in part, with the use of a Waltz-style algorithm (Waltz 1975). Then, the visual system would construct its 3-D scene percept by finding the 'best' interpolation between them, if reasonably 'good' surfaces can be constructed.

Which surface is best, appears to be determined by a number of factors: Smooth surfaces seem to be favored (planar surfaces being ideal). Surfaces that have orientations that approximate either a frontoparallel plane or a ground plane appear to be preferred over other orientations. Surfaces that include parallel or colinear contours are favored, as in figure 11. Proximity seems to be an important factor. The observers' focus of attention and point of fixation also influence surface formation, as in figures 8 and 9. In addition, higher-level factors like perspective, size constancy, and instructions to subjects can play a role.

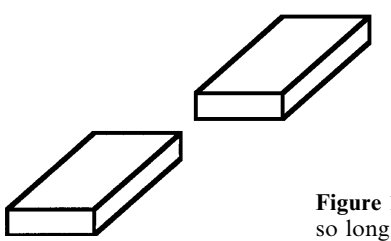


Figure 11. Colinearity of image contours suggests colinearity in 3-D, so long as no other depth cues contradict this.

So, in the surface-formation account of these 3-D scene percepts, the various rules of depth assignment, such as those considered here earlier as competing to determine our perceptions of figure 4, are now seen as top–down influences on a *surface-formation process*. This suggests a more parsimonious treatment of the perception of a large class of displays. For example, according to the surface-formation hypothesis the 'height-in-the-field' rule and Gogel's 'equidistance tendency' influence perception by means of the preference for 'ground' surfaces and frontoparallel surfaces, respectively. Since the height-in-the-field rule and the equidistance tendency can make very different predictions about the perception of a given stimulus, placing them within a common framework in which these and other factors have their effect by influencing the surface-formation process might permit a more unified theoretical account of 3-D scene perception.

In addition, the surface-formation hypothesis is supported by the observation that illusory ground planes are often perceived in line drawings depicting objects which *could* be resting on a ground plane (eg trees or cubes). According to the surface-formation hypothesis, illusory ground planes are seen in these displays because the surface-formation mechanism is very strongly activated by these configurations, strongly enough to generate a *visible* illusory surface. Similarly, surface-formation mechanisms are strongly activated by many random-dot stereograms (RDSs) and structure-from-motion (SFM) displays.

In addition to the factors affecting surface formation mentioned above, there also may be a preference for minimizing the depth assigned to objects and features (see Rock 1983), and opaque surfaces may produce stronger surface percepts than transparent surfaces. For example, in figure 10 most observers initially see a Necker cube viewed from above with a handle attached by its 'feet' to the front face on the left. However, when the Necker cube reverses, the handle generally appears to be attached to the rear face on the left, and there is a tendency to see it as attached to the *interior* side of this face: It appears to be attached by its 'top' rather than its feet. These observations are consistent with possible preferences for minimizing the depth of the handle and for attaching it to a (locally) opaque rather than a transparent surface.

6 Experiment

I have already shown phenomena which cannot be explained by the height-in-the-field rule or by physical stability (eg figure 5). In the following experiment I show phenomena which cannot be explained by the equidistance tendency or by a surface-formation process that operates exclusively for 'ground' surfaces or frontoparallel surfaces.

6.1 *Methods*

Observers were seated approximately 80 cm in front of a 15-inch AppleVision color monitor driven by a Power Macintosh 8500/120 computer. They were presented with a display similar to figure 12. They were asked if they perceived the figure on the right as 3-D blocks arrayed in depth. If they responded "yes" (as all subjects did), then they were instructed to imagine changing their viewpoint on this scene by imagining themselves moving around and to the left in the scene until the faces of the blocks with gray dots on them were frontoparallel relative to their viewpoint. That is, they were asked to imagine looking at the faces with the gray dots 'head on.' They were then told that the tall black rectangle on the left represented the (now frontoparallel) face of the tall block with the gray dot on it, and that the smaller gray rectangle represented the face of the small block with the gray dot on it. They were instructed to adjust the position of the small gray rectangle on the left so that the relative position

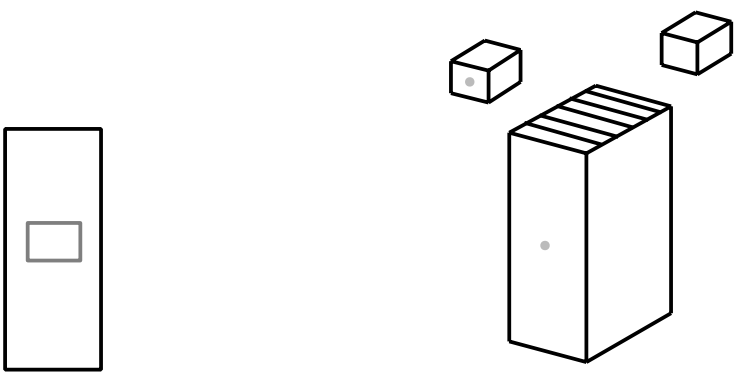


Figure 12. A stimulus display used in the experiment.

of this rectangle vis-à-vis the tall black rectangle was the same as what they imagined the two faces of the blocks with gray dots on them would look like from the imagined viewpoint. They adjusted the position of the gray rectangle using either a mouse or the four 'arrow' keys on a Macintosh keyboard.

Observers also made a similar judgment for figure 13, except that in this case they had to imagine changing their viewpoint by imagining that they moved around and to the right in the 3-D scene.

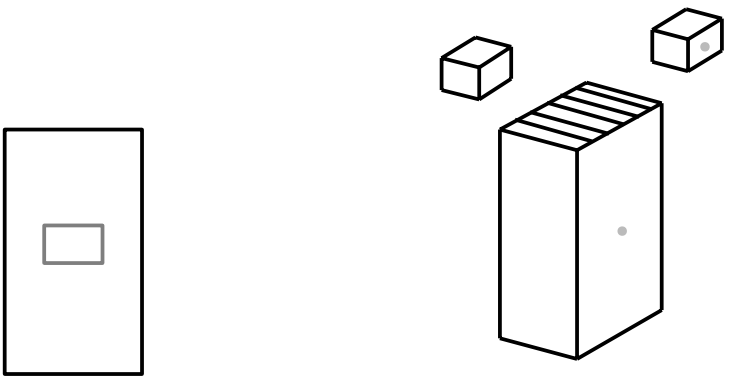


Figure 13. A stimulus display used in the experiment.

6.2 *Results and discussion*

The raw data of the placements of the small gray rectangle by the twenty-five observers are shown in the scatter graphs in figure 14. The tall rectangle shown on each graph represents the tall black rectangle in the experimental display. Each data point in figure 14a represents the position of the bottom right corner of the small gray rectangle as set by one observer when viewing the scene in figure 12. Similarly, each data point in figure 14b represents the position of the bottom left corner of the small gray rectangle,

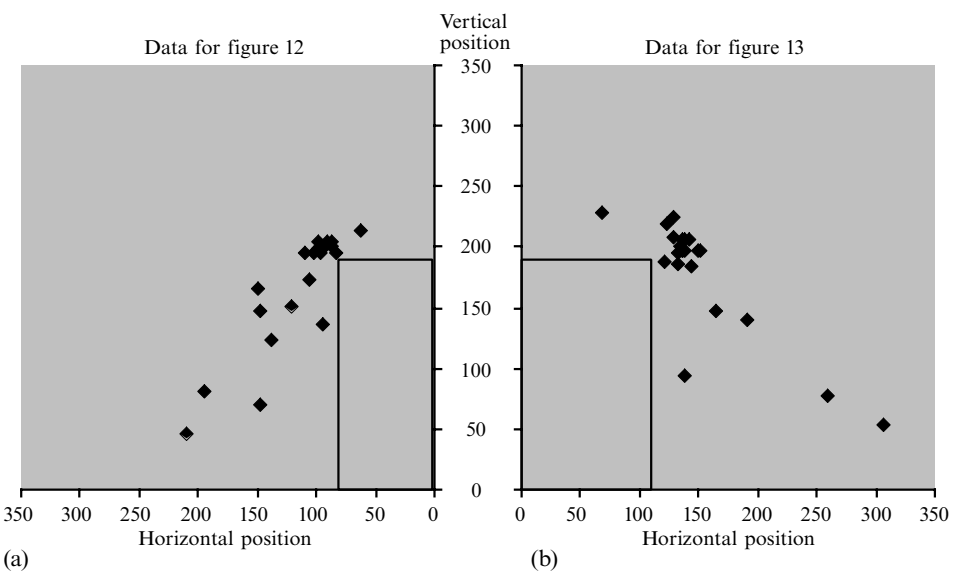


Figure 14. Scatter graphs representing the raw data of observers in the experiment. (a) Observer probe settings for figure 12. The data points represent the position of the lower right corner of the gray rectangle in figure 12 as set by one observer. The rectangle in the graph represents the large rectangle in figure 12. (b) Observer probe settings for figure 13. The data points represent the position of the lower left corner of the gray rectangle in figure 13 as set by one observer. The rectangle in the graph represents the large rectangle in figure 13.

as set by one observer when viewing the scene in figure 13. As can be seen, the data points systematically tend to have greater X values than the corner of the large rectangle, showing that the equidistance tendency is not decisive here. The data also tend to have Y values greater than zero, showing that the tendency to see the small block as attached to an inferred ground surface is not strong either. Instead, there seems to be a tendency to see the bottom faces of the small blocks as roughly coplanar with the top face of the large block.

Application of the binomial test confirmed these observations. The null hypothesis for the X values was that they were equal to that of the top corner of the large rectangle (the equidistance tendency). This hypothesis was rejected with confidence $p < 0.001$ in both displays. The null hypothesis for the Y values was that they were equal to that of the bottom corner of the large rectangle (the height-in-the-field rule, so that both blocks are seen as resting on a common ground plane, approximately). This hypothesis was also rejected with confidence $p < 0.001$ in both displays. Therefore, these phenomena cannot be explained by either the equidistance tendency or by a surface-formation process that generates ground surfaces and frontoparallel surfaces exclusively. Note that the physical-stability constraint also cannot explain these results. Finally, the hypothesis that the Y values were normally distributed about the height of the top corner of the large rectangle (ie that the bottom face of the small block appears to be roughly coplanar with the top face of the large block) could not be rejected with confidence $p < 0.05$ in either display.

The large blocks in figures 12 and 13 were made relatively tall in order to reduce the tendency for observers to see the small blocks as resting on a common ground plane with the large block. Thus, I tried to create stimuli which would maximize the influence of the top face of the large block on the perceived depth of the small blocks. Clearly this was only partly successful. Some observers still saw the small blocks as very far away, almost in contact with an inferred common ground surface that might also be supporting the large block. Other observers saw the depths of the small blocks as intermediate between the prediction of the equidistance tendency and the prediction of coplanarity of the bottom faces of the small blocks with the top face of the large block. This percept may have received additional support from certain 'flatness cues' in these displays.

7 Related research: Surface formation in stereopsis and structure-from-motion

Surface formation is a familiar concept in vision research. Recent psychophysical and computational studies of stereopsis and SFM suggest that depth can be assigned to objects and features by surface formation and attraction. Such mechanisms *assign* depth to features having ambiguous depth by finding their best interpolation with features having unambiguous depth.

For example, Mitchison and McKee (1985, 1987a, 1987b) suggested that surface formation, in particular a planarity constraint, can guide the correspondence solving process for binocular features, and can even assign depth values that are inconsistent with any possible binocular correspondences, at least for short viewing times.

In da Vinci stereopsis (Nakayama and Shimojo 1990) binocular features generate a surface which is extrapolated to the monocular features, thereby assigning them unambiguous depth in the background (also see Takeichi et al 1992).

Collett (1985) used RDSs containing binocular dots in some regions of the image and monocular dots in other regions. He found that the monocular dots often appeared to have depths that were consistent with smooth interpolations or extrapolations of the surfaces defined by the binocular dots. This study involved not only frontoparallel surfaces, but also surfaces slanted in depth. These findings suggest that Gogel's (1956) equidistance tendency may be partly due to the operation of a more general surface-formation mechanism.

Hildreth et al (1995) and Treue et al (1995) suggest that surface interpolation is fundamental to the perception of depth in SFM displays. They propose that surface interpolation explains the ability of observers to build up 3-D representations over long periods of viewing time ($\sim$1 s) in SFM displays in which the individual features have very short lifetimes ($\sim$ 80 ms). They propose that when a new dot appears, it is assigned an initial depth value with the use of the 3-D surface structure built up from previous frames of the motion. This allows the visual system to further refine its current estimate of 3-D structure using the newly created dot.

As discussed above, in the case of monocular line-drawing interpretation, *all* of the objects and features may have ambiguous depth. However, a similar surface-formation mechanism could be used to assign *relative* depth, especially when other depth cues are absent. In other words, vision may tend to organize disconnected objects and features into common surfaces. Indeed, when the 'height-in-the-field' rule is operative, observers often perceive a 'ground plane' on which the objects appear to rest. In addition, among the interpretations that observers report in some of the displays presented in this article, many tend to involve 3-D coplanarity (Mitchison 1988) of various image features. However, since visible interpolated surfaces often do not accompany these percepts, I suggest that relatively weak interpolations, in terms of their ability to produce a visual impression of a continuous surface, might still exert a strong influence on perceived relative depth in 3-D scenes.

Thus, surface formation in stereopsis, SFM, and monocular line drawings not only affects perceived depth *between* image features (ie depth interpolation), it also can determine or modify the perceived depths of the features *themselves* relative to what would be predicted on the basis of local binocular disparity or relative-motion information alone. When the image contains a large number of features, the computational cost of constructing 3-D scene percepts by means of surface formation is likely to compare very favorably with the cost of determining the individual depths of large numbers of features from their binocular disparities or by SFM equations. I believe that surface formation may have important functional similarities in all three domains by providing a computationally efficient heuristic for constructing 3-D-scene percepts. This would be particularly useful in cases where an animal needs to compute and update its scene representations quickly, for example when it is running across uneven terrain trying to escape from a predator. Perhaps the information that the visual system can effectively process in such an emergency situation is in some ways similar to the sparse information provided by the line drawings investigated here. The idea that surface formation may be a useful heuristic for scene perception when computational resources are limited is also suggested by the findings of Mitchison and McKee (1985, 1987a, 1987b) that, for short viewing times, surface formation in stereopsis can 'overrule' depth information from binocular disparity.

I believe that the phenomena considered in this article are also related to the surface-formation phenomena associated with 'surface contours' (Stevens 1981), and to the line-drawing phenomena investigated by Barrow and Tenenbaum (1981).

8 General discussion

Possible advantages of the surface-formation approach are as follows: (i) It might be a largely bottom–up mechanism that could plausibly be implemented in neural hardware, although some top–down influences are likely. (ii) It is consistent with the fact that line drawings often contain visible interpolated surfaces, such as faces of blocks and 'ground' surfaces that are perceived to support depicted the objects. (iii) It can explain 3-D-scene perception in displays in which the relevance of higher-level cues is unclear (eg in figure 4, when it is rotated 90° clockwise, and in figure 5). The interpolation process may be proceeding without much top–down influence here, guided partly by

2-D proximity: interpolations between surfaces and contours that are nearby in the image are favored.

Albert and Tse (1999) showed that surface attachment can also affect the perceived shape of an individual object depicted by a silhouette. For example, in figure 15a the large black blob appears to be a Hershey's "Kiss" sitting on a ground plane. In figure 15b the large black blob appears to have quite a different intrinsic shape: perhaps a seal or the back of a whale protruding out of an inferred body of water. However, the large blob in figure 15a is identical to the one in figure 15b, except for a 90° 2-D rotation in the image plane. I suggest that the observed shape change is due to surface formation and attraction of the bottom contour of the large blob to the ground surface. This results in propagation of very different 3-D shape information over the complete surface of the blob. A similar effect can be seen in figure 16 (Albert and Tse 1999). In figure 16a we see a 3-D crescent-like figure with its outer ends arcing forward. Note that attachment of the crescent's bottom contour to the ground plane supports this shape percept. On the other hand, in figure 16b we see that, when the same silhouette is placed in optical contact with a cylinder, it tends to appear relatively flat and to bend backwards and wrap around the cylinder. In this case both the top and bottom contours of the silhouette seem to attach to the 3-D surface of the cylinder. The silhouette now looks something like a thin piece of tape.

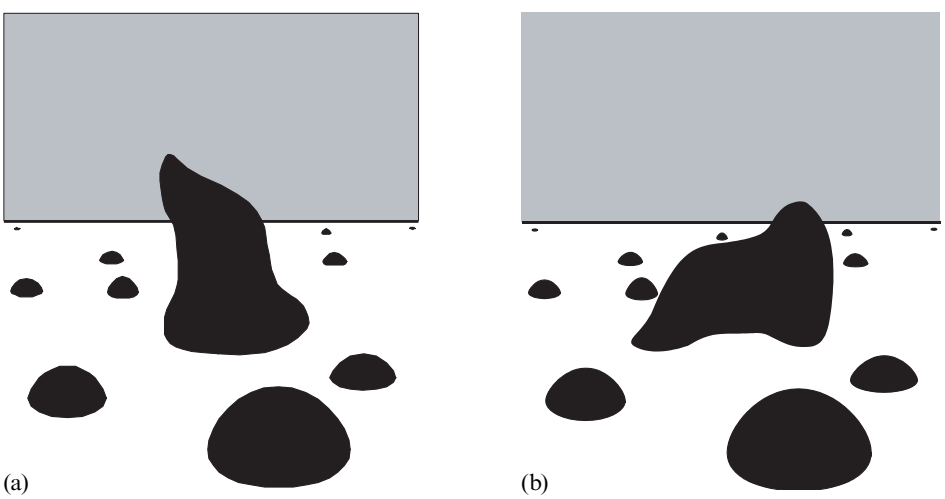
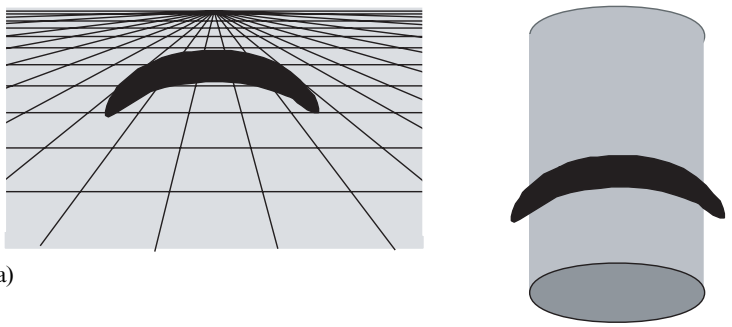


Figure 15. The intrinsic 2-D shape of the large black blob in (a) is the same as the one in (b). They differ only by a rotation of 90° in the image plane. However, the *perceived* intrinsic 3-D shape elicited by these figures is quite different, for most observers. In (a) most observers see something like a Hershey's "Kiss", having a roughly circular horizontal cross section. In (b) most observers see something like the back of a camel, or the back of a whale or a porpoise rising above the surface of a body of water as it prepares to dive. Thus, the corresponding cross sections (now vertically oriented) no longer appear circular. This may be explained by a tendency for the inferred ground surface to attract the bottom contour of the silhouette (which is in full optical contact with the ground surface), thereby changing the 3-D shape that these contours impart to the blob. (from Albert and Tse 1999)

It might be objected that if this surface-formation process can be influenced by higher-level factors, then it must also be a relatively high-level process, and therefore it must be very different from the surface-interpolation mechanisms operating in RDSs and in SFM displays. However, consider the theoretical status of illusory-contour perception (eg the Kanizsa triangle). It is well known that higher-level factors such as familiarity, perceptual set, and instructions to subjects can exert a strong influence on illusory-contour perception. So, although there exists considerable evidence



(a) (b)

Figure 16. In (a) most observers see a volumetric crescent-like figure lying on the ground plane. In (b) the same 2-D silhouette appears to be much flatter and to bend back around the cylinder. Thus the perceived 3-D shape of this silhouette appears to be strongly influenced by a tendency to be attracted to and to conform to nearby surfaces to which it might be attached. In (a) only the bottom contour of the crescent attaches to the ground plane, whereas in (b) both the top and bottom contours attach to the surface of the cylinder. This causes the silhouette to appear flatter in (b).

that processes of illusory-contour perception begin at a relatively low level in the visual system (von der Heydt 1986), most neural theories allow for higher-level influences as well (eg Grossberg and Mingolla 1985; Grossberg 1994).

One possible reason why surface formation has not previously been suggested as a basis for 3-D scene perception in line drawings may be that surface formation is usually discussed in connection with displays containing a fairly large number of features (eg RDSs, SFM displays, or 'surface contour' displays). In contrast, I suggest that surface formation may influence 3-D scene perception in line drawings containing relatively few features. Moreover, the illusory surface percepts in RDSs and SFM displays are often accompanied by illusory *contours* bounding the constructed surface. For example, RDSs often contain abrupt changes in disparity. I suggest that this produces more conspicuous, visible surface formation than is generally observed in the kinds of stimuli studied here.

9 Summary and future research

To summarize, I have proposed that surface formation can play an important role in 3-D scene perception when the visual information is relatively sparse and other cues are weak or ambiguous. These processes may strongly affect the perceived relative depths of objects and features even if conscious perception of these surfaces is weak or absent.

There is a great deal more that we need to know about the surface-formation process in order to develop this proposal into a detailed, quantitative theory of 3-D-scene perception. As suggested above, surface formation can be completely overruled by the presence of other strong depth cues. In addition, there appear to be a variety of factors that determine the visual system's preferences for one surface formation over another, and the relative importance of these factors remains to be determined.

Acknowledgements. Thanks to Ken Nakayama, Patrick Cavanagh, and Peter Tse for helpful discussions. MKA was supported by NIH 5 F32 MH11103-02.

References

Albert M K, Tse P U, 1999 "The role of surface attachment in perceiving volumetric shape" *Perception* in press

Barrow H G, Tenenbaum J M, 1981 "Interpreting line drawings as three-dimensional surfaces" *Artificial Intelligence* **17** 75 – 116 [reprinted in Brady J M (Ed.) *Computer Vision* (Amsterdam: North-Holland)]

Collett T S, 1985 "Extrapolating and interpolating surfaces in depth" *Proceedings of the Royal Society of London, Series B* **224** 43 – 56
Gibson J J, 1950 *The Perception of the Visual World* (Boston, MA: Houghton Mifflin)
Gogel W, 1965 "Equidistance tendency and its consequences" *Psychological Bulletin* **64** 153 – 163
Grossberg S, 1994 "3-D vision and figure – ground separation by visual cortex" *Perception & Psychophysics* **55** 48 – 120
Grossberg S, Mingolla E, 1985 "Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading" *Psychological Review* **92** 171 – 211
Heydt R von der, Peterhans E, Baumgartner G, 1984 "Illusory contours and cortical neuron responses" *Science* **224** 1260 – 1262
Hildreth E C, Ando H, Anderson R A, Treue S, 1995 "Recovering three-dimensional structure from motion with surface reconstruction" *Vision Research* **35** 117 – 137
Jepson A, Richards W, 1993 "What is a percept?", RBCV-TR-93-43, Department of Computer Science, University of Toronto, Toronto
Mitchison G J, 1988 "Planarity and segmentation in stereoscopic matching" *Perception* **17** 753 – 782
Mitchison G J, McKee S P, 1985 "Interpolation in stereoscopic matching" *Nature (London)* **315** 402 – 404
Mitchison G J, McKee S P, 1987a "The resolution of ambiguous stereoscopic matches by interpolation" *Vision Research* **27** 285 – 294
Mitchison G J, McKee S P, 1987b "Interpolation and the detection of fine structure in stereoscopic matching" *Vision Research* **27** 295 – 302
Nakayama K, Shimojo S, 1990 "Da Vinci stereopsis: depth and subjective occluding contours from unpaired image points" *Vision Research* **30** 1811 – 1825
Richards W, Jepson A, Feldman J, 1996 "Priors, preferences and categorical percepts", in *Bayesian Approaches to Perception* Eds D Knill, W Richards (Cambridge: Cambridge University Press)
Rock I, 1983 *The Logic of Perception* (Cambridge, MA: MIT Press)
Stevens K A, 1981 "The visual interpretation of surface contours" *Artificial Intelligence* **17** 47 – 74 [reprinted in Brady J M (Ed.) *Computer Vision* (Amsterdam: North-Holland)]
Takeichi H, Watanabe T, Shimojo S, 1992 "Illusory occluding contours and surface formation by depth propagation" *Perception* **21** 177 – 184
Treue S, Anderson R A, Ando H, Hildreth E C, 1995 "Structure-from-motion: Perceptual evidence for surface interpolation" *Vision Research* **35** 139 – 148
Waltz D L, 1975 "Understanding line drawings of scenes with shadows", in *The Psychology of Computer Vision* Ed. P H Winston (New York: McGraw-Hill) pp 19 – 91