1. Spatial Form as Inherently Three-Dimensional

Christopher W. Tyler

Perhaps the ultimate goal of visual processing is the understanding of the perception of objects, the accepted fundamental unit of our visual world. Among many recent treatments of this topic, the book on the visual perception of objects by Regan (2000) stands out as being the most analytically psychophysical. Its emphasis is on the coding of sensory information of various types into coherent object forms. This analysis is indeed a core issue in object perception. How does the visual system break down the sensory information into the discrete components of the object representation? In particular, this leads to the question of how the sparse information in each visual modality is integrated into the continuous percept of a coherent object. It is the process of recombination of the local sources of object information, which is often called the "binding problem" that is the topic of this overview. The binding problem is typically conceptualized in terms of the temporal binding of different stimulus properties or object features into a coordinate whole (e.g., Singer, 2001). Here, however, emphasis is placed on a spatial binding principle provides an entirely different insight into the binding problem.

Objects in the world are typically defined by contours and local features separated by featureless regions (such as the design printed on a beach ball, or the smooth skin between facial features). Leonardo's 1498 depiction of a dodecahedron (Figure 1.1) illustrates the point. The surface between the edges is perceptually vivid, and yet its location is not defined by any features in the image. The shading does not define this surface, because it is not homogeneous although the surface is perceived as flat. The inhomogeneity of the shading is interpreted as the painter's brush-strokes lying in the surface defined by the edges alone. The mean differences between the shadings on different surfaces are interpreted as consistent with the angles of the surfaces, helping to support the 3D interpretation, but the surfaces themselves are interpolated from the locations of the edges without regard to the details of the shading.

Surface representation is thus an important stage in the visual coding from images through to object identification. Surfaces are a key property of our interaction with objects in the world. It is very unusual to experience objects, either tactilely or visually, except through their surfaces. Even transparent objects are experienced through

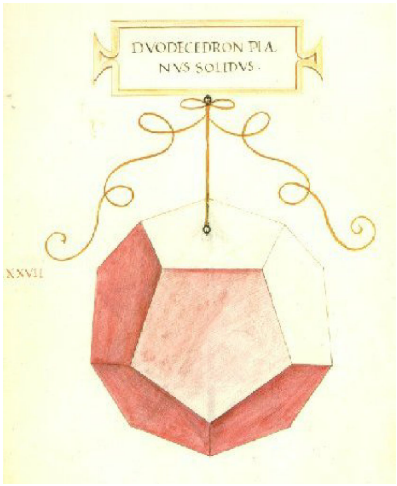


Figure 1.1: Illustration of a dodecahedron by Leonardo da Vinci from the book *Divina Proportione* by Luca Pacioli (1498).

their surfaces, with the material between the surfaces being invisible by virtue of the transparency. Really the only objects experienced in an interior manner are translucent objects, through which the light passes in a manner to illuminate the density of the material. Developing a means of representing the proliferation of surfaces before us is therefore a key stage in the processing of objects.

A very useful paradigm for the exploration of surface perception is the illusory overlay concept introduced by Schumann (1904). The basic paradigm is to overlay one set of objects by a background-colored mask of another object. The simplest version is the illusory bar (Figure 1.2A), consisting of two disks with sectors cut out of them generating the illusion of clear edges in the form of a vertical bar overlaid on the two disks (although the illusory edges fade if stared at directly). The triangular version developed by Kanizsa (1976) (Figure 1.2C is even more vivid.

The illusory contours can be interpreted as the result of a Bayesian 'bet' the most likely interpretation of the Kanizsa figure is as a triangular surface overlaying three disk-shaped surfaces, with the consequent enhancement of the edges dividing the triangular surface from the background of the same color. Rotating the pacmen elements by 90° to the right (Figure 1.2D) makes the bet implausible because of the lack of alignment of corresponding edges. The figure is now seen as three isolated pacmen with no illusory contours connecting them. On looking back at the original of Figure 1.2C, it may also be seen as isolated elements and some time may be required to regain the original percept of a triangular surface.

Surfaces may be completed not just in two dimensions, but also in three dimensions. A compelling example was developed by Tse (1999). The amorphous shape wrapping a white space gives the immediate impression of a three-dimensional cylinder filling the space (Figure 1.3). This example illustrates the flexibility of the surface-completion mechanism in adapting to the variety of unexpected demands for shape reconstruction.

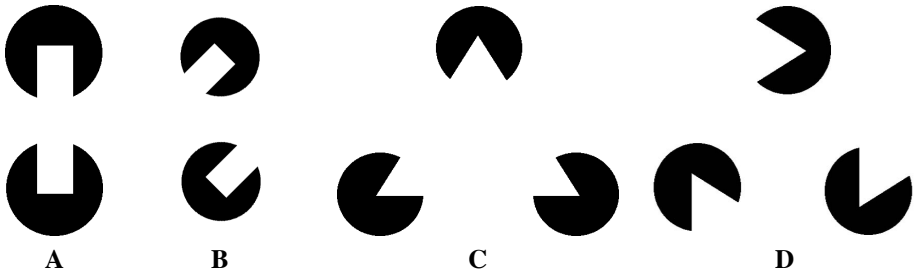


Figure 1.2: **A**. The original Schumann figure in which the alignment of the edges produces an illusory white bar. **B**. The same figure with the slots rotated to the right by 45°. Although the figure elements are identical, this manipulation destroys the coherence of the bar and degrades the percept to two isolated disks with no illusory contours. **C**. The Kanizsa version of the occlusion contours, based on a triangle. **D**. The Kanizsa triangle with 90° rotated elements, again destroying the subjective contours.

Figure 1.3: Volume completion of a cylinder (Tse, 1999).

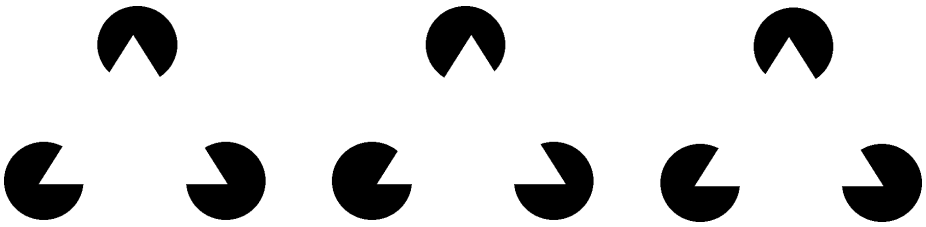


Figure 1.4: The stereoscopic Kanizsa figure. On crossing or uncrossing the eyes, two stereoscopic versions of the figure are seen, flanked by two monocular versions. In the version with the triangle in front, the illusory contours complete the straight sides of the triangle in the same way (though more strongly) as they did in the original. However, in the other stereoscopic version with the triangular region behind the disks, they are seen as open portholes. Now the illusory contours switch to complete the circular edges of the disks and disappear from the triangular edges, emphasizing the active nature of the object reconstruction process.

The effect of the Bayesian interpretation may be enhanced by adding a supporting cue to the spatial interpretation (Ramachandran, 1986). If the triangle is given a stereoscopic disparity to support the interpretation of the overlaid triangle, the need for edges dividing the triangle from the background becomes paramount. Figure 1.4 constitutes a three-element stereogram that provides the binocular disparity cues when fused by crossing the eyes (or by diverging them). The disparity is added only to the 'corner' regions of the pacmen, not to their circular boundaries. In direct viewing of the figure without binocular fusion, it is clear that these small shifts are almost unnoticeable, and have no effect on the quality of the illusion. However, the left and right pairings in Figure 1.4 provide near- and far-disparity versions of the identical figure, allowing one to contrast the perceptual effects of merely changing the sign of disparity at the triangular points. In the version with the triangle in front of the disks, the illusory edges are seen very strongly. The triangle standing out in depth appears substantially brighter than its background, and can be inspected much more extensively without loss of the illusion. The disparity cue provides extra confirmation that the corners are in front of the black disks, enhancing the percept that they are overlaid by a coherent object, which further requires that its edges must stand out from the background in the region between the disks.

However, Figure 1.4 also provides a version with the disparity consistent with a triangle lying behind the disks. This cue now interdicts the interpretation of an overlaid triangle and forces a completely different surface configuration because the triangular sectors are now behind the disks. It is striking that our visual systems immediately come up with a plausible alternative. The disks are now seen as open 'portholes' in a uniform surface, behind which the triangle is hidden except for its corners. In order to achieve this interpretation, two changes are required in the edge structure. The original illusory edges have to evaporate to provide for the uniform surface, and the

portholes require a curved rim completing the circle around each corner. These changes are achieved perceptually in dramatic fashion. Despite the fact that the monocular images are identical in the two cases (only the left- and right-eye images are switched), the perceptual interpretation is strikingly different. Both the depth structure and the edge brightness are reorganized to new spatial locations. This immediate perceptual reorganization attests to the power of the interpretation, in terms of a configuration of surfaces in space, to generate vivid perceptual experiences.

The version of the stereoscopic image in Figure 1.4 with the triangle behind also illustrates the principle of what Kanizsa (1976) termed 'amodal completion'. The surface interpretation is focused on the flat surface out of which the three portholes are cut. However, we are perceptually aware that the three points seen through the portholes belong to the same triangle. There is a connection between them that is felt spatially rather than just known logically. This connection does not give rise to the illusory contours of the 'modal completion' of the triangle seen visually in front of the surface (although some viewers see a blurred version of the underlying triangle semi-transparently through the surface). In terms of the perceived 3D structure, this connection between the points 'should' be invisible because it is hidden by the surface containing the portholes. But yet the points are perceived as part of a single triangle. This connection takes the form of an implicit perceptual knowledge that, if there was movement in the figure, the three points would move together because they belong to the same triangle. The completion is 'amodal', in the sense that it is mediated by implicit knowledge of the spatial structure, but is (usually) not seen in the visual modality. (Note that, as originally described by Kanizsa, 1976, these percepts may be seen as emergent interpretations with prolonged non-stereoscopic viewing of Figure 1.2 or Figure 1.4, for those who have difficulty in attaining the stereoscopic view.)

The examples of Figures 1.2–1.4 illustrate that surface reconstruction is a key factor in the process of making perceptual sense of visual images of black shapes. It is easy to talk about such processes verbally, but there is a large gap between a verbal description and a process that can be implemented in neural hardware. The test of neural implementation is to develop a numerical simulation of the process using neurally- plausible computational elements. The feasibility of a surface reconstruction process being capable of generating accurate subjective contours is illustrated in Figure 1.5-1.6 for the classic Kanizsa figure in the computational technique of Sarti, Malladi and Sethian (2000). The edge-attractant properties of the Kanizsa corners progressively convert the initial state of an isotropic spindle into a convincing triangular mesa with sharp edges. The resulting subjective surface is developed as a minimal surface with respect to a Riemannian metric of metrical distortions induced by the features in the image (analogous to the gravitational distortions of physical space in the Theory of General Relativity). The computational manipulation of this Riemannian surface reveals how the interactions within a neural network could operate to generate the subjective contours in the course of the 3D reconstruction of the surfaces of the world.

The algorithm first convolves the image with an edge detector to generate a potential function whose representation of the image corresponds to the raw primal sketch, as introduced by Marr (1982), which encodes image gradient, orientation of structures, T-junctions and texture. The minimum lines of this potential function denote the position of edges and its gradient is a force field that always points toward the local

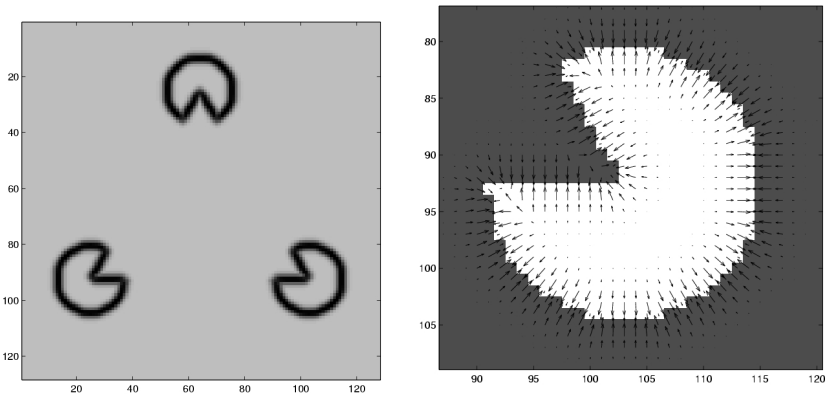


Figure 1.5: Local edge detection in the Kanizsa figure. A) edge map of original figure (Figure 1.2A); B) edge gradient map for one of the "pacmen". The gradient of this potential function is computed as a force field that always points towards the local edge.

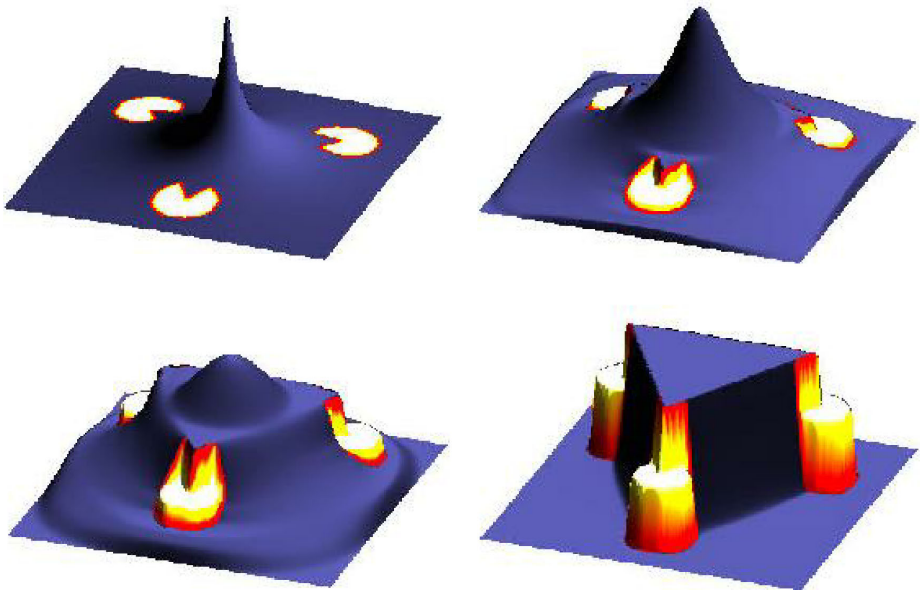


Figure 1.6: Development of the surface towards the subjective surface. The original features are mapped in white against a blue background, while yellow through red map the low values of the Riemannian metric indicating the presence of boundaries.

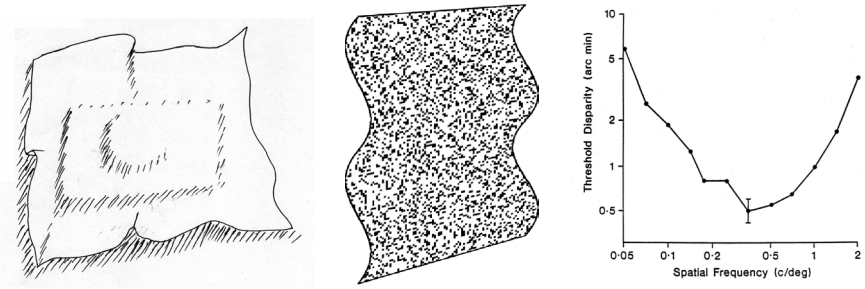
edge (Figure 1.5). This potential function defines a metric of the embedding space in which the perceived surface is developed. During its evolution (Figure 1.6), the surface is attracted by the existing boundaries and progressively steepens. The surface develops towards the piecewise constant solution by continuation and closing of the boundary fragments and filling in of the homogeneous regions. A solid object is progressively delineated as a constant surface bounded by the existing and reconstructed shape boundaries (Sarti et al., 2000).

It is particularly interesting that the surface developed through the SMS Riemannian-metric algorithm has the apparently contradictory properties of sharp edges combined with a smoothness constraint. The smoothness constraint is a property of minimal surfaces, such as the surface of an aggregation of soap bubbles. The tensions within the surface of a soap bubble tend to minimize the local curvature, so it settles to the form of maximum smoothness. In the SMS algorithm, however, the implementation also allows sharp edges as a component of the solution, when they increase the smoothness of the rest of the surface. In these respects, the algorithm closely mimics the human visual system, which tends to identify edges of objects and to assume smooth surfaces extending between these edges. The SMS algorithm provides a neurally-plausible implementation of the reconciliation between these two apparently contradictory demands of the surface properties of object reconstruction.

1.1 Surface representation through the attentional shroud

One corollary of this surface reconstruction approach is a postulate that the object array is represented strictly in terms of its surfaces, as proposed by Nakayama and Shimojo (1990). Numerous studies point to a key role of surfaces in organizing the perceptual inputs into a coherent representation. Norman and Todd (1998), for example, show that that depth discrimination is greatly improved if the two locations to be discriminated lie in a surface rather than being presented in empty space. This result is suggestive of a surface level of interpretation, although it may simply be relying on the fact that the presence of the surface provides more information about the depth regions to be assessed. Nakayama, Shimojo and Silverman (1989) provide many demonstrations of the importance of surfaces in perceptual organization. Recognition of objects (such as faces) is much enhanced where the scene interpretation allows them to form parts of a continuous surface rather than isolated pieces, even when the retinal information about the objects is identical in the two cases. This study also focuses attention on the issue of border ownership by surfaces perceived as in front of rather than behind other surfaces. While their treatment highlights interesting issues of perceptual organization, it offers no insight into the neural mechanisms by which such structures might be achieved.

A neural representation of the reconstruction process may be envisaged as an attentional shroud (Tyler and Kontsevich, 1995), wrapping the dense locus of activated disparity detectors as a cloth wraps a structured object (Figure 1.7A). This depiction shows how the shroud may envelop an object to capture the broad features of its shape, although some degree of detail may be lost. This self-organizing surface is envisaged as operating in the manner of what Julesz (1971) called “the search for dense surfaces”, as instantiated in the stereopsis model of Marr and Poggio (1979). Both of these concep-



A **B** **C**

Figure 1.7: **A**. Cartoon of the attentional shroud wrapping an object representation. **B**. Depiction of a random-dot surface with stereoscopic ripples. **C**. Thresholds for detecting stereoscopic depth ripples, as a function of the spatial frequency of the ripples (from Tyler, 1990). Peak sensitivity (lowest thresholds) occurs at the low value of 0.4 cy/deg (2.5 deg/cycle). Thus, stereoscopic processing involves considerable smoothing relative to contrast processing.

tualizations were restricted to planar surfaces in the frontoparallel plane of the eyes. The attentional shroud, on the other hand, is proposed as a self-organizing connectictivity that spreads through the array of activated disparity detectors, known as the "Keplerian array", attracted by the closest sets of disparity detectors in the 3D metric encompassed by the Keplerian array. This process is what Tyler (1983, 1991) called "cyclopean cleaning", the simplification from the complexity of the activated Keplerian array of spurious correspondences to the single cyclopean surface of the final depth solution. At that time, the cleaning processes were envisaged as largely consisting of disparity (or epipolar) inhibition, together with lateral facilitation through neighboring fields of activation at similar disparities. The concept of the attentional shroud emphasizes that there is always a depth solution at every location in the field, and that it is based at the level of the generic depth representation rather than residing purely in the process of stereoscopic reconstruction.

The attentional shroud has inherent limitations with regard to the complexity of the surface that it can reconstruct. It cannot follow the 3D shape to the level of detail provided by the luminance information, but is restricted to depth gradients that have less steepness than may occur in the physical structure. Such a loss of detail is characteristic of the stereoscopic process, as may be established by studies of the resolution of ripples in sinusoidal stereoscopic surfaces of the sort depicted in Figure 1.7B. The graph in Figure 1.7C, reproduced from Tyler (1990), shows how the amplitude threshold varies with the spatial frequency of the stereoscopic ripples. This graph illustrates that the stereoscopic depth reconstruction of surfaces is limited to a maximum spatial bandwidth of only about 2 cy/deg (or 0.5 deg per ripple cycle). This limitation is as much as ten times less than the bandwidth for resolution of luminance information (grating acuity). The peak sensitivity is at an even lower frequency, requiring 2.5 deg

for each ripple cycle. Thus, the stereoscopic reconstruction of surface shape is capable of rendering depth variations only to a coarse scale of representation. This neural process operates as though the depth reconstruction were by a flexible material whose connectivity was too stiff to match sharp discontinuities in the depth information.

1.2 Interpolation of object shape within the generic depth map

Once the object surfaces have been identified, we are brought to the issue of the localization of the object features relative to each other, and relative to those in other objects. Localization is particularly complicated under conditions where the objects could be considered as "sampled" by overlapping noise or partial occlusion - the tiger behind the trees, the face behind the window-curtain. However, the visual system allows remarkably precise localization even when the stimuli have poorly defined features and edges (Toet and Koenderink, 1988). Furthermore, sample spacing is a critical parameter for an adequate theory of localization. Specifically, no low-level filter integration can account for interpolation behavior beyond the tiny range of 2-3 arc min (Morgan and Watt, 1982), although the edge feature of typical objects, such as the form of a face or a computer monitor, may be separated by many degrees. Thus, the interpolation required for specifying the shape of most objects is well beyond the range of the available filters.

Conversely, accuracy of localization by humans is almost independent of the sample spacing. For sample spacings ranging from 30 minutes to 3 minutes separation, localization is not improved by increasing sample density (Kontsevich and Tyler, 1998). This limitation poses an additional challenge in relation to the localization task, raising the 'long-range interpolation problem' that has generated much recent interest in relation to the position coding for extended stimuli, such as Gaussian blobs and Gabor patches (Morgan and Watt, 1982; Hess and Holliday, 1992; Levi et al., 1992; Kontsevich and Tyler, 1998).

Localization information is available from multiple visual cues, as indicated in Figure 1.1. Position information is available from luminance form, disparity profile, color, texture and other visual cues. Localization in the sampled stimulus might employ interpolation over many such cues. In a task in which the object shape is defined both by luminance and disparity, for example, the basic sources of noise determining the localization error are (i) early noise in each visual modality contributing to the position determination, (ii) late noise in the peak localization process.

To probe the nature of object processing by different cues, we may utilize a position task for which the threshold for localization of an object is determined for objects defined by various visual modalities (such as luminance and disparity). If localization is performed in separate visual modalities, the position thresholds might be expected to combine according to the their absolute signal/noise ratios, assuming that the signals form separate visual modalities have independent noise sources. The observers would be able to interpolate one estimate of the position of the profile from the luminance information alone and a second estimate from the disparity information alone. In this

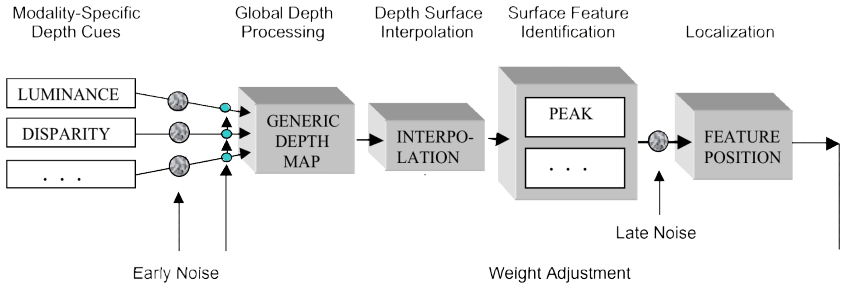


Figure 1.8: The generic depth modal of localization processing based on unitary interpolation input. Localization information is available from multiple visual cues from luminance form, disparity profile, color, texture and other visual cues. Object feature binding may be accomplished by the sensory information being fed into a *generic depth map*. The local cues in this map of depths would then be subject to a depth surface interpolation process operating over multiple visual cues to bind the various features into a coherent representation of the object, from which the generic localization information may be derived.

case, signals from the various modalities (L, D, ..., X) would combine to improve the localization performance. Adding information about the object profile from a second modality would always improve detectability and could never degrade it.

Likova and Tyler (2003) addressed the unitary depth map hypothesis of object localization by using a sparsely sampled image of a Gaussian bulge (Figure 1.9). The luminance of the sample lines carried the luminance profile information while the disparity in their positions in the two eyes carried the disparity profile information. In this way, the two separate depth cues could be combined or segregated as needed. Both luminance and disparity profiles were identical Gaussians, and the two types of profile were always congruent in both peak position and width. The observer's task was to make a left/right judgment on each trial of the position of the joint Gaussian bulge relative to a reference line, using whatever cues were available. Threshold performance was measured by means of the maximum-entropy Ψ staircase procedure (Kontsevich and Tyler, 1999).

Observers were presented the sampled Gaussian profiles defined either by luminance modulation alone (Figure 1.9A), by disparity alone (Figure 1.9B), or by combination of luminance and disparity defining a single Gaussian profile (Figure 1.9C). It should be noticeable that the luminance profile evokes a strong sense of depth as the luminance fades into the black background. If this is not evident in the printed panels, it was certainly seen clearly on the monitor screens. Free fusion of Figure 1.9B allows perception of the stereoscopic depth profile (forward for crossed fusion). The third panel shows a combination of both cues at the level that produced cancellation to flat plane under the experimental conditions. The position of local contours is unambiguous, but interpolating the location of the shape of the nose to locate its tip, for example,

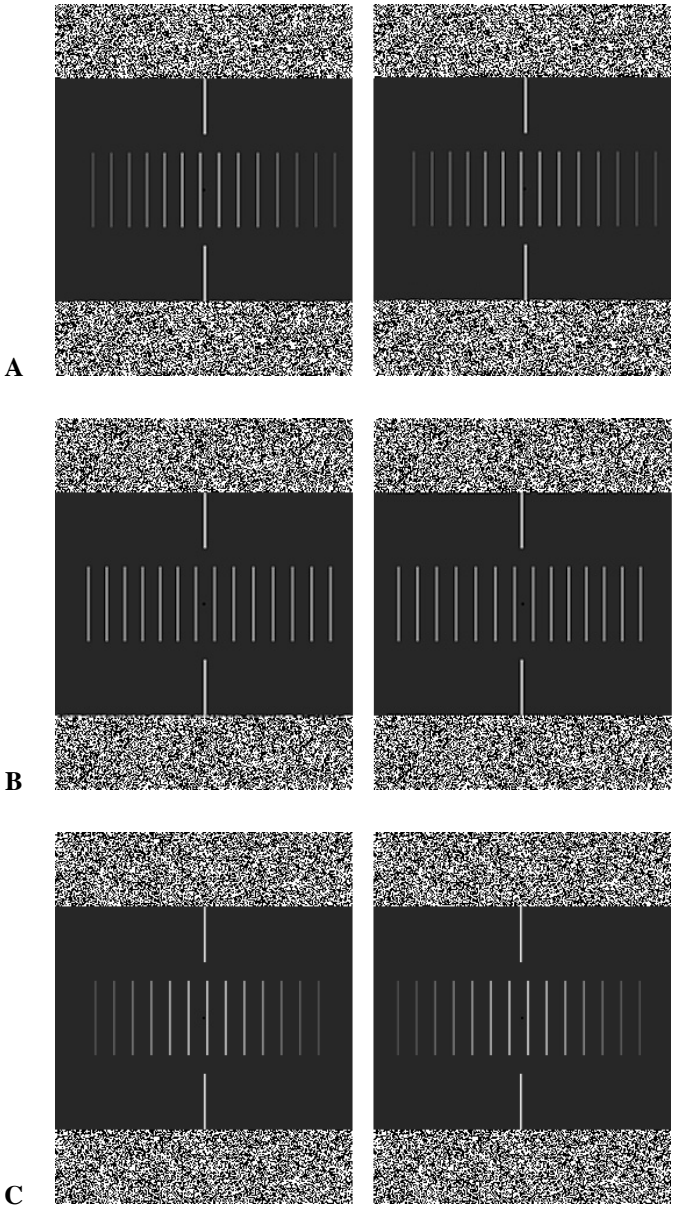


Figure 1.9: Stereograms showing examples of the sampled Gaussian profiles used in the Likova and Tyler (2003) experiment, defined by **A** luminance alone, **B** disparity alone, and **C** a combination of luminance and disparity. The paris of panels should be free fused to obtain the stereoscopic effect.

is unsupportable.

Localization from disparity alone was much more accurate than from luminance alone, immediately suggesting that depth processing plays an important role in the localization of sampled stimuli (see Figure 1.10, green dots). Localization accuracy from disparity alone was as fine as 1-2 arc min, requiring accurate interpolation to localize the peak of the function between the samples spaced 16 arc min apart. This performance contrasted with that for pure luminance profiles, which was about 15 arc min (Figure 1.10). Combining identical luminance and disparity Gaussian profiles (Figure 1.10, red circles) provides a localization performance that is qualitatively similar to that given by disparity alone (Figure 1.10, green line). Rather than showing the hump-shaped function predicted by the multiple-cue interpolation hypothesis, it again exhibits a null condition where localization is impossible within the range measurable in the apparatus. Contrary to the multiple-cue hypothesis, the stimulus with full luminance information becomes impossible to localize as soon as it is perceived as a flat surface. This null point can only mean that luminance information *per se* is insufficient to specify the position of the luminance profile in this sampled stimulus. The degradation of localization accuracy can be explained only under the hypothesis that interpolation occurs within a unitary depth-cue pathway.

Perhaps the most startling aspect of the results in Figure 1.10 is that *position* discrimination in sampled profiles can be completely nulled by the addition of a slight *disparity* profile. It should be emphasized that the position information from disparity was identical to the position information from luminance on each trial, so addition of the second cue would be expected to reinforce the ability to discriminate position if the two cues were processed independently. Instead, the nulling of the luminance-based position information by the depth signal implies that the luminance target is processed exclusively through the depth interpretation. Once the depth interpretation is nulled by the disparity signal, the luminance information does not support position discrimination at all (null point in the red curve in Figure 1.10)..

This evidence suggests that depth surface reconstruction is the key process in the accuracy of the localization process. It appears that visual patterns defined by different depth cues are interpreted as objects in the process of determining their location. Only an interpolation mechanism operating at the level of *generic depth representation* can account for the data. Specifically, a depth interpolation mechanism accounts for the impossibility of position discrimination at the cancellation point and the asymmetric shift of the cancellation point by the luminance cue (Figure 1.10). The fine resolution of the performance when disparity information is present clearly implies that an interpolation process is involved in the performance, because it is about 8 times better than could be supported by the location of the samples alone (even assuming that the sample nearest the peak could be identified from the luminance information; see Likova and Tyler, 2003).

Evidently, the full specification of objects in general requires extensive interpolation to take place, even though some textured objects may be well defined by local information alone. The interpolated position task may therefore be regarded as more representative of real-world localization of objects than the typical Vernier acuity or other line-based localization tasks of the classic literature. It consequently seems remarkable that luminance information per se is unable to support localization for objects

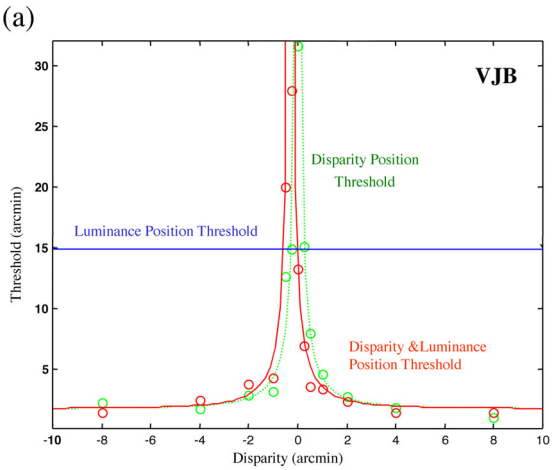


Figure 1.10: Typical results of the position localization task. The green circles are the thresholds for the profile defined only by disparity, the red circles are the thresholds defined by disparity and luminance. The dashed green line shows the model fit for disparity alone, the red line that for combined disparity and luminance. The blue line shows threshold for the pure luminance. Note leftward shift of the null point in the combined luminance/disparity function.

requiring interpolation. The data indicate that it is only through the interpolated depth representation that the position of the features can be recognized. One might have expected that positional localization would be a spatial form task depending on the primary form processes (Marr, 1982). The dominance of a depth representation in the performance of such tasks indicates that the depth information is not just an overlay to the 2D sketch of the positional information. Instead, it seems that a full 3D depth reconstruction of the surfaces in the scene must be completed before the position of the object is known.

1.3 Transparency

A major complication in the issue of surface reconstruction is the fact that we do not perceive the world solely as a set of opaque surfaces. There are many types of object that are partially transparent, allowing us to perceive more than one surface at different distances along any particular line of sight. The depiction of transparent objects was a particular obsession of the Dutch artists of the 17th century, but it is interesting to note that it extends as far back as Roman times. The fruit bowl and water jug in the wall-painting from the *House of Julia Felix* near Pompeii (Figure 1.11) illustrates that fine glassware and mirrored surfaces were appreciated at this epoch of civilization also.

At first sight, the perception of transparency seems at variance from the concept of the unitary surface reconstruction of the attentional shroud. A key feature of random-

dot stereograms is their ability to support the percept of transparent depth surfaces (Julesz, 1970; Norcia and Tyler, 1984). Here the depth tokens are assuming a primary role, for they first need to be specified at each point in the image before the construct of a surface running through each appropriate set of points can be developed. It is as though the surface is strung across the depth tokens to segregate the relevant sets of monocular dots, rather than the reverse. The visual system may be capable of supporting the simultaneous percept of up to three overlaid surfaces (Weinshall, 1990) from fields of randomly intermixed dots. Such multilayered percepts seem to make it difficult to maintain the perspective that construction of object surfaces is the primary process in spatial perception, because they emphasize the local depth tokens of each feature as the primary structure of visual 3D space, with the surface superstructure erected upon their scaffolding.

Before abandoning the view that there is a single surface representation at any point in the field, it is important to be sure that there is no interpretation under which the single surface can remain the primary vehicle of reconstruction, even for perception of multiple transparent surfaces. One such view is that, although only a single surface may be reconstructed at any one moment in time, transparent perception may be obtained by sequential reconstruction of each of the multiple surfaces in turn. Marr and Poggio (1979) followed the approach of the Automap model of Julesz and Johnson (1968) in proposing such sequential reconstruction of depth surfaces. The idea is that surface reconstruction was achieved within a fixed array of cortical disparity detectors by vergence eye movements that shifted the surface reconstruction to different distances in physical space. In each new physical location, the otherwise rigid stereo reconstructive apparatus could then find the densest disparity plane to form the singular local surface. Transparency would be perceived by sequential operation of the local surface reconstruction.

The hypothesis of sequential reconstruction by vergence eye movements makes two testable predictions. One is that the disparity range of depth reconstruction mechanism is, by postulate, limited to disparities near zero. Disparity images of flat planes near zero disparity therefore should be easier to detect than disparity images that cut through the zero disparity plane at a steep angle. Steep stereoscopic surfaces should require a sequence of several vergence positions before they can be fully reconstructed. Such a prediction was tested by Uttall, FitzGerald and Eskin (1975), who generated planes up to 80° from frontoparallel in dynamic-noise stereograms and presented them in brief exposures too short for vergence eye movements to occur. Two-alternative forced-choice experiments (with a monocularly indistinguishable null target of random depth information) indicated that the detectability of such depth planes was almost independent of angle of slant. This result makes it difficult to conceive how any model with based on purely frontoparallel surface reconstruction can be operating in human vision.

A second feature of the eye-movement reconstruction concept is that it does not include a mechanism of attentional enhancement of surfaces projecting within the array of disparity detectors; the only local focusing mechanism is presumed to be that of vergence tracking of the eyes through the 3D optical image. Since stereoscopic attention to a particular plane can be demonstrated (Tyler and Kontsevich, 1995 and Figure 1.8), it could explain the perception of transparent surfaces without vergence eye movements.

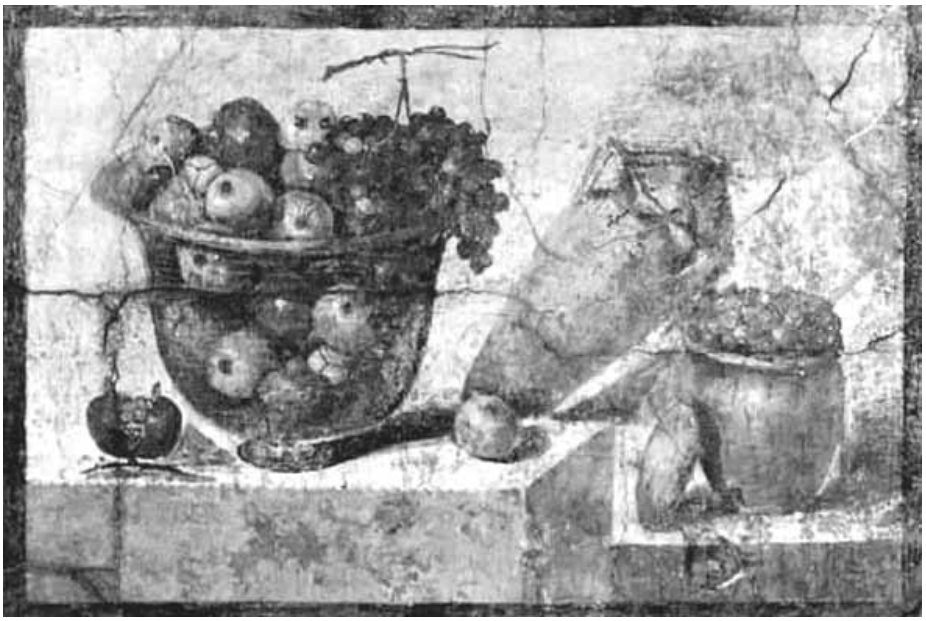


Figure 1.11: Wall painting from the House of Julia Felix, illustrating the transparent glassware and reflective vessels available to the Pompeiian aristocracy at the beginning of the Roman Empire.

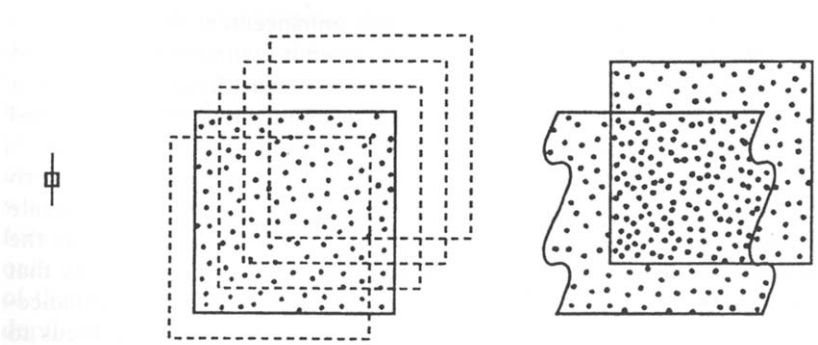


Figure 1.12: Frontoparallel stereoattention stimulus. Observers fixated on a stable fixation target (left). The test stimulus consisted of a transparent pair of depth planes (right). One of these depth planes, selected at random on each trial, had a sinusoidal depth ripple whose phase had to be identified. The transparent test target was preceded by a flat cueing plane (center) at one of five disparities selected at random. Thus, the cueing plane could be unpredictably at the same or different depths from the plane of the depth ripple on each trial.

To determine the nature of transparency perception, Tyler and Kontsevich (1995) presented a pair of transparent stereoscopic planes in front of and behind a fixation marker (Figure 1.12). To determine the visibility of the two surfaces, they added the modulation signal of a sinusoidal disparity corrugation that could appear in either of the two ambiguous planes on each trial, the other remaining flat. As in Figure 1.8, attention was drawn to one of the planes by presenting an attentional cueing plane immediately prior to the transparent stimulus. The corrugation itself could be in one of two phases (sine or sine relative to the fixation point) to form the forced-choice discrimination task for the observer. When the priming plane fell close to the disparity of either the front or back transparent plane, the phase of the corrugations became readily discriminable. But no information was available about the phase of the non-cued plane. Because the priming plane contained no corrugations, it added no information to the discrimination task. Its effect, therefore, must have been due to a non-feature-specific enhancement of the information processing capability in a limited disparity range, which may be described as the operation of disparity-specific attention.

The result from the transparent-plane experiment is that the shape discrimination that is easy in the attended plane is impossible (at this duration) in the other plane of the transparent pair. This result reveals that the transparent percept does not allow discrimination of detail in two planes simultaneously. Only the attended plane can be resolved. It appears, therefore, that the attention mechanism plays the same role as the vergence shifts in the vergence eye-movement hypothesis of depth reconstruction. Only one plane can be attended at a time, with the details of the other plane inaccessible to consciousness until attention is switched to that depth location. On this interpretation, the perception of transparency is an illusion akin to the illusion that we see the world at high resolution throughout the visual field. In fact, we see at high resolution only in the restricted foveal region, but we point the fovea to whatever we wish to inspect, so its high resolution is available at all locations in the field. So effective is this sampling mechanism that most people are unaware of the existence of the limited spatial resolution outside their fovea. In a similar fashion, we may be unaware of the surface reconstruction mechanism filling across the plane of current interest at a time.

1.4 Object-oriented constraints on surface reconstruction

One corollary of this surface reconstruction approach is a postulate that the object array is represented strictly in terms of its surfaces, as proposed by Nakayama and Shimojo (1990a). The dominanace of a depth representation in the performance of the position interpolation and transparency tasks indicates that the depth information is a core process that must be completed before the configuration of the object is known. It is proposed (Tyler and Kontsevich, 1996; Likova and Tyler, 2002) that the depth representation is not simply an abstract pattern of neural firing, but an adaptive neural surface representation that links the available depth information into a coherent two-dimensional manifold in a process analogous to the mathematical one of Figure 1.7. Is such a mechanism neurally plausible?

The phenomenon of perceived (phantom) limbs after amputation provides a perceptual lesion that provides profound insight into the strata of perceptual representation in the somatosensory system (Ramachandran, 1998). Applying such insights to the visual system provides a radical view of its selforganizing capabilities. It is well known that amputees experience a clear and detailed sense of the presence of the limb in the space that it would have occupied before amputation. This implies that there is a cortical representation of the limb that is distinct from its sensory representation. The logic of this implication is that the sensory representation is no longer being supplied with consistent information, in the absence of the peripheral input. Any residual input will be disorganized noise, and therefore would not support a coherent representation of the pre-existing limb structure.

Less well known, but well established, is that the amputee is capable of manoeuvring the perceived phantom at will (but only if it was manoeuvrable before amputation; a paralyzed limb remains perceptually paralyzed after amputation; Ramachandran, 1998). This manipulable representation corresponds to the body schema of Sir Henry Head, a complete representation of the positions of the limbs and the body that is accessible to consciousness and manipulable at will (Head et al., 1920). Head proposes the body schema as a neurological construct that has some specific neural instantiation, but it has been largely dismissed as metaphorical in the succeeding century. The idea of a conscious manipulable body schema provides a challenging view of the self- organizing capabilities of the neural substrate, but one that is hard to dismiss when details of the phantom limb manifestation are taken into account (Ramachandran, 1998). It suggests that there are three levels of representation of the sensory world in the visual system:

1. The visual representation in striate cortex, which includes the neural Keplerian array of disparity detectors. The coordinate frame for this representation would be the retinal coordinates of the location on the retina (or the joint retinal coordinates of the two eyes for the stereoscopic aspect).

2. The spatial representation in parietal cortex (in object-centered coordinates). The site of Shepard's (1971) manipulable image, Julesz' dense planes and Tyler's attentional shroud. It also corresponds to Gregory's hypotheses of the spatial configuration tested during perceptual alternations. The representation is inherently self-organizing, with

   (a) local surface tension to bind it into a data-reducing form,

   (b) a tendency to self-destruct (autoinhibition) unless continually reinforced by sensory input.

   (c) conformity to amodal instruction from distant spatial regions.

3. the intended configuration of the manipulandum in frontal cortex (in egocentric coordinates for convenient manipulation). This attentional manipulation is endogenous, in the sense that it can be manipulated at will according to higher cognitive instruction.

Figure 1.13: Left: Inverted picture of Mirror Lake, Yosemite, with scattered leaves in 'sky'. Right: Reverted picture reveals that the leaves are floating on the water's surface, filling transparently across the space to the shoreline. The surface of the lake bottom is visible below and the mountains beyond, making a complex image with three levels of surface reality.

This conceptualization of space perception is a high-level, dynamic representation that may be termed "prehensile vision". The property that distinguishes the tail of the primates from that of all other species is that it is prehensile; it can be guided by neural signals to reach out and grasp objects like tree branches by wrapping around them, operating like a fifth hand. Miller (1998) has drawn attention to the ability of our vision to perform analogous feats. He describes a depiction of a lake mirroring a sky, with a few leaves scattered on the surface, as may be illustrated in a photograph of Yosemite's Mirror Lake (Figure 1.13). When viewed upside down, the reflected sky is upward, and appears distant, with the scattering of leaves seen as blowing through space. Right-side up, the cues are sufficient for the reconstruction of the reflective surface of the lake extending toward us, with the leaves floating in the perceptually-completed surface. Thus the same region of the picture is seen as distant in one orientation but transparently close in the other.

What it the mechanism for this reorganization? The triple scheme for a prehensile process of spatial reconstruction proposes that the neural surface representation is not merely a passive connection between local sources of activation, but a dynamic self-organizing search mechanism with guidance from top-down frontal-lobe influences as to where might be interesting to look and what sense a particular arrangement would make. For example, if viewed for sufficient time, the inverted picture of the lake can also elicit surface completion, once it has been conceptualized as an inverted picture in which the lake surface might extend upward over our heads rather than below us. This would be an example of a modified Bayesian constraint. Lake surfaces, by gravitational constraints, are always below us (in the non-scuba environment!). There should therefore be a strong Bayesian constraint against expecting a surface above us. But this constraint is eliminated for the case of *pictures* of the environment, if it is possible that the picture may be inverted.

Driven by such influences, the prehensile representation can reach out its surface

reconstruction network, or attentional shroud, to search for constellations of surface cues making up meaningful interpretations of the structure of the environment and the objects within it. The process is analogous to the way the hand of the blind person reaches out to feel the shape of objects within range, except that the visual 'hand' is infinitely extensible to wrap whatever form is encountered all the way to the far reaches of space. The concept of prehensile vision gives neural sinew to the exploratory perceptual experience that we have in a new spatial environment. It is a component of the attraction to the scenic view at a 'Vista Point' on the highway. We step out of the enclosed space of the vehicle and experience our prehensile reconstruction mechanisms probing the arrays of visual information reaching the retina to expand the scope and reach of the spatial representation across the forms of the distant landscape. This process is often conceptualized as a cognitive endeavor; "Oh, there's that lake we just passed and there's the famous mountain peak we are aiming for". The concept of prehensile reconstruction proposes that beneath this cognitive appeal is a level of dynamic perceptual reconstruction that probes and molds the visual information in a surface representation of the surrounding hillsides to experience them in a quasi-tactile manner that is neurally equivalent to feeling the curves of a bed-comforter.

1.5 Conclusion

The evidence assessed in this review triangulates onto the concept that the predominant mode of spatial processing is through a flexible surface representation in a 3D spatial metric. It is not until the surface representation is developed that the perceptual system seems to be able to localize the components of the scene. This view is radically opposed to the more conventional concept that the primary quality of visual stimuli is their location, with other properties attached to this location coordinate (Marr, 1982). The concept of the attentional shroud, on the other hand, is a flexible, self-organizing network that operates as an internal representation of the external object structure. In this concept, the attentional shroud is, itself the perceptual coordinate frame. It organizes itself to optimize the spatial interpretation implied by the complex of binocular and monocular depth cues derived from the retinal images. It is not until this process is complete that the coordinate locations can be assigned to the external scene. In this sense, localization is secondary to the full depth representation of the visual input. Spatial form, usually seen as a predominantly 2D property that can be rotated into the third dimension, becomes a primary 3D concept of which the 2D projection is a derivative feature. In this connection, it is worth noting that position signals have a delayed integration time relative to luminance integration (Tyler and Gorea, 1986). This is just an additional line of evidence that position is a derivative variable from the primary object representation, rather than the primary metric property implied by the graphical representation of optical space.

The net result of this analysis is to offer a novel insight into the nature of the binding problem. The separate stimulus properties and local features are bound into a coherent object by the glue of the 3D surface representation. This view is a radical counterpoint to the concept of breaking the scene down into its component elements by means of specialized receptive fields and recognition circuitry. However, an important aspect

of the "understanding" of objects is the representation of the 3D spatial relationships among their components. This understanding cannot be achieved in full by a 2D map of the component relationships. The evidence reviewed in this overview points toward the key role of the surface representation in providing the "glue" or "shrink-wrap" to link the object components in their appropriate relationships. It also emphasizes the inherent three-dimensionality of this surface "shrink-wrap" in forming a prehensile matrix with which to cohere the object components whose images are projected onto the sensorium. While further details remain to be worked out, the simulations of the Sethian group (Figure 1.6) provide assurance that such processes are readily implementable not only computationally but with plausible neural components that could reside in a locus of spatial reconstruction such as the parietal lobe of the human cortex.

References

Breitmeyer, B., Julesz, B. and Kropfl, W. (1975). Dynamic random-dot stereograms reveal up- down anisotropy and left-right isotropy between cortical hemifields. *Science*, 187: 269-2-70.

Buckley, D., Frisby, J. P. and Mayhew, J. E. (1989) Integration of stereo and texture cues in the formation of discontinuities during three-dimensional surface interpolation. *Percept.*, 18: 563-5-88.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proc. Royal Soc. Lond. B*, 171: 179–196.

Gregory, R. L. (1980). Perceptions as hypotheses. *Phil. Trans. Roy. Soc. Lond. B*, 290: 181–197.

Head, H., Rivers W. H., Holmes, G. M., Sherren, J., Thompson, H. T. and Riddoch, G. (1920). *Studies in Neurology*, London: H. Frowde, Hodder and Stoughton.

Hess, R. F. and Holliday, I. E. (1992). The coding of spatial position by the human visual system: Effects of spatial scale and contrast. *Vis. Res.*, 32: 1085–1097.

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

Kanizsa, G. (1976). Subjective contours. *Sci. Am.*, 234: 48–52.

Kontsevich, L. L. and Tyler, C. W. (1998). How much of the visual object is used in estimating its position? *Vis. Res.*, 38: 3025–3029.

Kontsevich, L. L. and Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vis. Res.*, 39: 2729–2737.

Levi, D. M., Klein, S. A. and Wang, H. (1994). Discrimination of position and contrast in amblyopic and peripheral vision. *Vis. Res.*, 34: 3293-3313.

Likova, L. T. and Tyler, C. W. (2003). Peak localization of sparsely sampled luminance patterns is based on interpolated 3D object representations. *Vis. Res.*, (in press).

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

Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proc. Roy. Soc. Lond. B*, 204: 301–328.

Miller, J. (1998). *On Reflection*. Yale University Press: New Haven, Conn.

Mitchison, G. J. and McKee, S. P. (1985). Interpolation in stereoscopic matching. *Nature*, 315: 402–404.

Morgan, M. J. and Watt, R. J. (1982). Mechanisms of interpolation in human spatial vision. *Vis. Res.*, 25: 1661–1674.

Nakayama, K. and Shimojo, S. (1990) Towards a neural understanding of visual surface representation. In T. Sejnowski, E. R. Kandel, C. F. Stevens and J. D. Watson (Eds), *The Brain Cold Spring Harbor Symposium on Quantitative Biology*, Cold Spring Harbor Laboratory: NY, 55: 911–924.

Nakayama, K., Shimojo, S. and Silverman, G. H. (1989). Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Percept.*, 18: 55–68.

Norcia, A. M. and Tyler, C. W. (1984). Temporal frequency limits for stereoscopic apparent motion processes. *Vis. Res.*, 24: 395–401.

Norman, J. F. and Todd, J. T. (1998). Stereoscopic discrimination of interval and ordinal depth relations on smooth surfaces and in empty space. *Percept.*, 27: 257–272.

Pacioli, L. (1498/1956). *Compendium de Divina Proportione*. Fontes Ambrosiani: Milan.

Ramachandran, V. S. (1986). Capture of stereopsis and apparent motion by illusory contours. *Percept. Psychophys.*, 39: 361–373.

Ramachandran, V. S. (1998). Consciousness and body image: lessons from phantom limbs, Capgras syndrome and pain asymbolia. *Phil. Trans. Roy. Soc. Lond. B*, 353: 1851–1859.

Regan, D. M. (2000). *Human Perception of Objects*. Sinauaer and Associates: Sunderland, MA.

Sarti, A., Malladi, R. and Sethian, J. A. (2000). Subjective Surfaces: A method for completing missing boundaries. *Proc. Nat. Acad. Sci. USA*, 12: 6258–6263.

Schumann, F. (1904). Beitrage zur Analyse der Geishctswahrnemungen: !. Einege Beobachtungen über die Zusammenfassung von Gesichtseindrucken zu Einheiten. *Psychologische Studien*, 1: 1–32.

Shepard, R. N. and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171, 701–703.

Singer, W. (2001). Consciousness and the binding problem. *Ann. New York Acad. Sci.*, 929: 123–146.

Toet, A. and Koenderink, J. J. (1988). Differential spatial displacement discrimination thresholds for Gabor patches. *Vis. Res.*, 28: 133-143.

Tse, P. U. (1999). Volume completion. *Cog. Psy.*, 39: 37–68.

Tyler C. W. and Cavanagh, P. (1991). Purely chromatic perception of motion in depth: two eyes as sensitive as one. *Percept. Psychophys.*, 49: 53–61.

Tyler, C. W. and Gorea, A. (1986). Different encoding mechanisms for phase and contrast. *Vis. Res.*, 26: 1073–1082.

Tyler, C. W. and Liu, L. (1996). Saturation revealed by clamping the gain of the retinal light response. *Vis. Res.*, 36: 2553–2562.

Tyler, C. W. (1983). Sensory aspects of binocular vision. In *Vergence Eye Movements: Basic and Clinical Aspects*, pp. 199-295. Butterworths: Boston.

Tyler, C. W., Kontsevich, L. L. (1995). Mechanisms of stereoscopic processing: stereoattention and surface perception in depth reconstruction. *Percept.*, 24: 127–153.

Weinshall, D. (1991). Seeing "ghost" planes in stereo vision. *Vis. Res.*, 31: 1731–1748.

Wurger, S. M. and Landy, M. S. (1989). Depth interpolation with sparse disparity cues. *Percept.*, 18: 39–54.

Yang, Y. and Blake, R. (1995). On the interpolation of surface reconstruction from disparity interpolation. *Vis. Res.*, 35: 949–960.