Visual form perception

6041 words

Peter U. Tse and Howard C. *Hughes*

In the absence of any apparent effort, the human visual system recovers the 3 dimensional form of the visible environment from inherently ambiguous 2 dimensional retinal images. How this feat is accomplished is perhaps the most fundamental problem faced by vision science. Despite the impression that vision seems effortless, a vast amount of processing is involved in the construction of an internal representation of the visible scene. The central computational problem is one of correctly and rapidly interpreting inherently ambiguous patterns of retinal activation. Moreover, in order for it to guide navigation and all other interactions with the physical world, these massive computations must be accomplished very quickly. The retinal image does not directly specify the absolute or relative distances of visible objects, the orientations of their component surfaces, their surface color or whether they are stationary or in motion. Countless possible 3D worlds could have produced the light that enters our eyes and creates a specific retinal image. For example, an individual photoreceptor that responds to red light will respond equally to a red surface illuminated with white light as well as to a white surface illuminated with red light. By itself then, this cell cannot specify surface color, much less the shape of objects in the world.

The computations involved in interpreting this cacophony of ambiguous neural signals generated by 100's of millions of visually responsive neurons occurs within circuits within the retina, thalamus, midbrain tectum, and an extensive network of visual processing areas that together constitute about half of the total surface area of the neocortex in visually sophisticated primates, including humans. Although our understanding of this complex function has witnessed great advances in recent decades, it remains rudimentary in many key respects. Here we will give a brief overview of progress that has been made in the area of visual form perception in recent decades, with an eye on what remains to be accomplished. For heuristic purposes, the problem of how we see can be broken into two subproblems, one concerning how different types of information is detected in incoming light, and the other concerning how detected information is processed to construct the 3D surfaces, motions, and objects that we experience.

The detection of visual primitives used for processing form

Precortical processing.

The processing of form information begins in the retina. Ganglion cells in the retina have center-surround and color-opponent receptive fields which confer upon these cells a sensitivity to edges (abrupt changes in luminance or color) while at the same time rendering them relatively insensitive to regions of uniformity. In effect, the retina filters out uniformity because regions of uniformity convey little useful information. The information the retina transmits to the brain for further processing is a compressed version of the image that emphasizes border information at multiple spatial scales. Data compression is necessary to efficiently encode and transmit the information available in 6-7 million cones and 120-130 million rods along an optic nerve comprised of only about 1.1 million ganglion cell axons for each eye. Efficient data compression requires preservation of those aspects of an image that are most informative, while less informative information is discarded or left implicit. The fact that retinal processing emphasizes contour information suggests that contours play a crucial role in generating later representations of 3D form.

These retinal signals reach the cortex by several parallel pathways originating within the thalamus - including the dorsolateral division of the lateral geniculate nucleus (LGNd), the pulvinar nuclear complex, the intralaminar thalamic nuclei and the thalamic reticular nucleus – as well as the basal ganglia. Until the mid 1960s, it was believed that projections from the LGNd provided the only pathway through which visual information could gain access to the cortex and that the LGNd functioned primarily as a relay station for information passing from the retina en route to the cortex. Because the anatomical evidence indicated that in the primate, the cortical projections of the LGNd were confined to the striate cortex (area 17), this cortex, with its fine-grained retinotopic representation of the visual world, was viewed as the first and crucial stage in the cortical processing of visual information. Extrastriate areas beyond area 17 (e.g. areas 18 and 19, etc.) were considered association areas, meaning that they received their sole visual input from area 17, operating on this input to produce higher order visual function.

The extrastriate cortex is now known to receive a subcortical afferent supply independent of the geniculostriate system. This input, first described in cat and tree shrew, exists in all species examined thus far, includes the retinal projection to the superior colliculus, an ascending projection from the superficial laminae of the colliculus to the pulvinar complex of the thalamus, and from there to the extrastriate cortex. In some mammals including the cat, extrastriate cortex also receives a sizable afferent supply from the geniculate itself.

The parallel geniculocortical and colliculopulvinar cortical projections are not strictly independent. For example, the superior colliculus projects to the LGNd as well as to the pulvinar, and there exist extensive feedforward and feedback cortico-cortical pathways between striate and extrastriate cortical areas. In addition, all of these cortical areas contribute corticofugal projections to the LGNd, the pulvinar and the superior colliculus.

Cortical processing.

The study of perception has focused on the visual cortex and the multiple areas in which the retinal image is processed. The so-called "early" visual cortical areas which receive a direct projection from LGNd contain neurons selectively sensitive to changes in certain properties of the stimuli. For example, their levels of activity depend upon features of the retinal image such as contour orientation, contour scale (or spatial frequency), binocular disparity, and the direction and velocity of movement. Thus area 17 in tree-shrews, lemurs and primates and 17-18 in cats are considered "early" or "primary" visual areas which further filter aspects of the compressed messages transmitted by the retina, and extract a set of functional features or primitives. Early cortical processing thus appears to consist of a neural description of various image primitives and their locations within the scene. This description is a simplified version of the original retinal image, but it is still a long way from explicit identification of the 3D structure of the visible world. A great deal of additional computation is required. Trying to understand how such a complex system operates is a formidable task. As a result, simplifying concepts have emerged as guides. A keystone in thinking about the neural mechanisms of visual perception is the concept of hierarchical processing of the details of the visual image. A widely held view is that this processing occurs in a number of stages, the first of which performs an analysis or filtering of the retinal image by extracting different, elementary features (primitives), or classes of image "energy." It has been argued that different primitive features may be processed by relatively independent modules that specialize in extracting and interpreting particular classes of visual information.

It is commonly held that later or "higher" stages of visual processing combine aggregates of primitive features into progressively more complex representations. Two generally dichotomous characterizations of these elements should be mentioned. One is that contours are primitives that are used to define surfaces and object boundaries. Interpretations of receptive fields in terms of trigger "features," such as oriented bars, provide a potential neural implementation of this conceptualization. An alternative view is that receptive fields operate as localized, spatial frequency filters. The basic idea here is that cells with different receptive field dimensions are tuned to different spatial frequencies, such that cells tuned to regions spanning small visual angles are said to be tuned to high spatial frequency information, while cells tuned to regions spanning large visual angles are said to be tuned to low spatial frequency information. Ganglion cells with similar tuning characteristics are distributed throughout the retina, and can be thought of as bandpass channels that accomplish a decomposition of the image using something analogous to Fourier analysis or in more recent conceptualizations, a wavelet decomposition. Convolution of the image with a variety of bandpass channels operating in parallel thus produces multiple bandpass filtered representations of the image. As shown in Figure 1, each channel thus provides information at a different spatial scale or level of resolution.

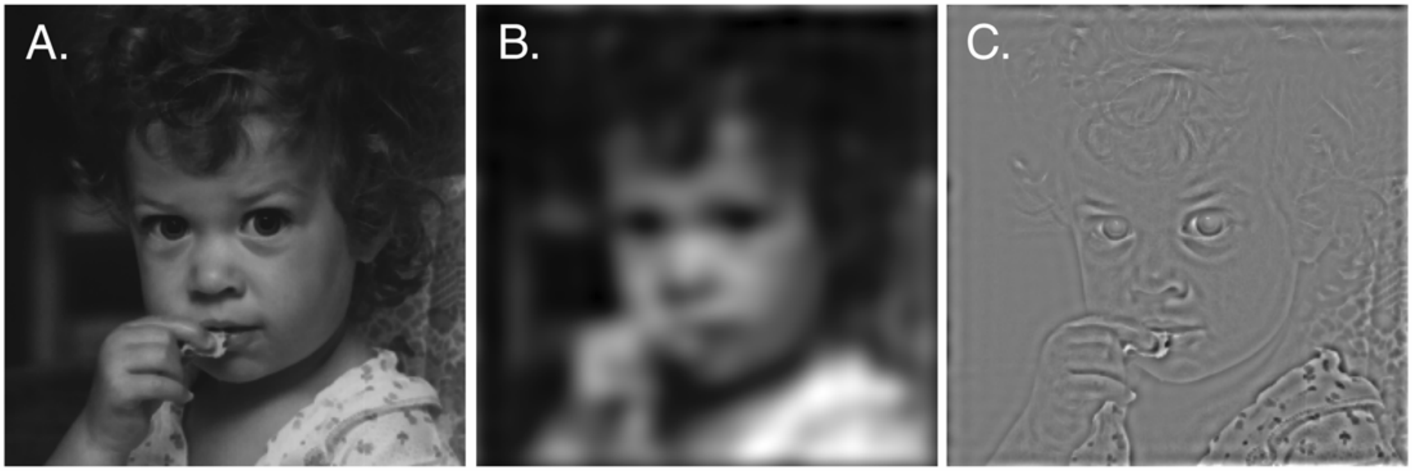

Figure 1. This is an illustration of the effects of filtering images into different bandwidths of spatial frequency. A. Original image of Carly Hughes. B. Low-pass filtered version of original image. C. High-pass filtered version of original image. Summing B and C yields A.

Low-pass filtering can convey information about overall image structure, and may contribute to certain Gestalt-like operations such as grouping, closure and good-continuation. High-pass filtered information emphasizes details within the image, particularly contours. Grouping procedures that compare, for example, contour orientations across the image, might take the high-passed output as input. Subsequent processes then operate on these multiple representations. It is not known which of these two general formulations (edge detection or spatial frequency analysis) is more accurate. Both types of detectors appear to exist and may constitute the extremes of a spectrum of processing types. It is important to recognize however that neither point of view suggests specific solutions to the most difficult conceptual problems raised by human pattern and form perception.

It is widely believed that early visual cortical areas are involved in grouping local information across the image into aggregate wholes. Gestalt Psychologists described the heuristics or criteria used by the visual system to group parts into wholes. They suggested that the combination of evolution and perceptual learning has produced mechanisms that are sensitive to the statistical regularities of real-world images. For instance, grouping procedures capture the fact that image regions that covary in certain ways tend to arise from common surfaces, objects, regions, and collections of objects in the world. Perceptual grouping is essential to the process of image segmentation – the process of determining which contours and textures belong to the same object. Because grouping involves decisions about what belongs with what in an image, it is tantamount to an inference about the state of the world. As such, grouping procedures do not merely extract information from the image, they create or construct new information. A particularly impressive example of the constructive nature of grouping and completion processes is the formation of illusory contours. When visible contours suggest the existence of a surface that is camouflaged against a similar background, the visual system creates contours even when there are none in the image, as shown in Figure 2A. This construction of information is known to happen at early stages of cortical processing, because cells have been found in areas 18 and 19 that respond to illusory contours even when no actual contours fall within their receptive fields.

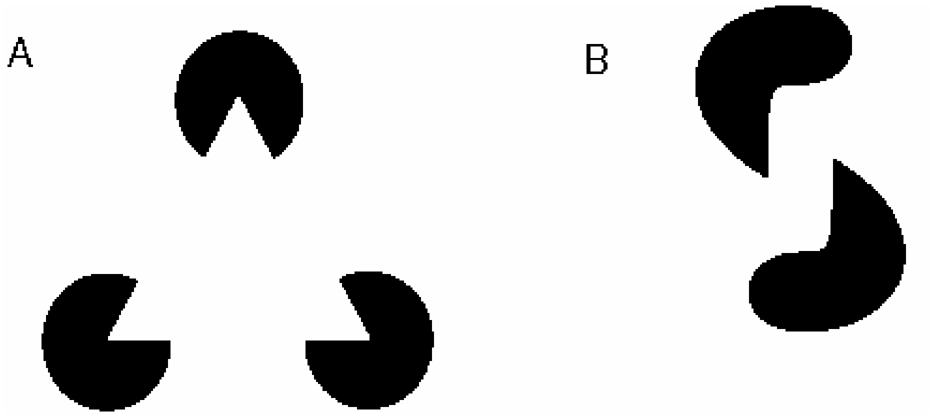

Figure 2. (A) An illusory surface with illusory contours (by G. Kanizsa); (B) An illusory volume wrapped around an illusory pole (by P. U. Tse).

A process closely related to grouping and completion procedures is the segmentation of regions into component subregions on the basis of borders defined by abrupt changes in texture, luminance, motion, and other low-level image properties. Integral to the processes underlying segmentation and grouping is a phenomenon called "visual pop-out," where one area of the image automatically segments itself away from the background to define an object, figure, or region (see Figure 3). Visual pop-out is in turn inseparable from the automatic allocation of attention to salient visual events and objects. It appears that relatively automatic grouping and segmentation procedures, carried out principally in extrastriate and occipito-temporal visual cortical areas, work in concert with procedures dedicated to the allocation of attention, perhaps realized in complex neural circuitry located in the frontal eye fields, the superior colliculus, and parietal cortex. The aim of cooperation among cortical areas involved in form processing, motion processing, and attentional allocation appears to be the common goal of monitoring or tracking one or a few salient figures against the non-salient background.

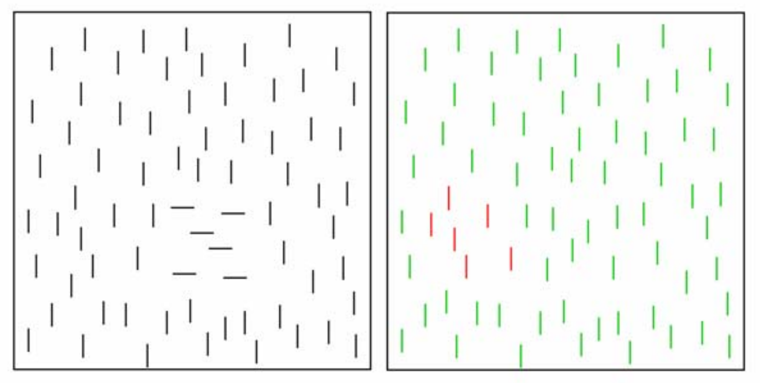

Figure 3. Here are two illustrations of the 'pop-out' effect, which refers to the immediate and effortless segmentation of a region from the background on the basis of differences in local features, in this case orientation and color.

Implicit in either the "primitive feature extraction" or "spatial frequency decomposition" views is the idea that individual cells can "code for" or are "tuned to" the presence of a specific feature or spatial frequency component. A literal interpretation of this idea is that activity in individual visual neurons signifies the presence of the feature to which that neuron is tuned. In order for neural activity to code for the presence of particular features, the cells would have to respond to that particular feature (such as orientation) independently of other properties (such as contrast). However, it has been recognized that, at least at the earliest stages of neural processing, individual neurons cannot uniquely encode specific features because their patterns of discharge vary across several dimensions simultaneously (for example, contrast, orientation, direction of motion, stereo disparity and/or spatial frequency). This has led to a greater appreciation of the fact that the neural representation of images and visible objects must somehow involve the profile of activity across populations of neurons. Computational studies have illustrated that this kind of "population coding" can generate very precise representations – much more precise than the tuning characteristics of any of the individual neurons contributing to the profile. Population coding appears to be a common means of specifying information with great precision in the nervous system, and additional examples can be found in oculomotor control, skeletal control and a number of other domains.

The notion that pattern and form perception involves a processing hierarchy has a very long history. There is a compelling intuition that suggests that perception is a progressive construction of elementary components or features. These features exist in all scenes, and it is only their relationships between one another that differs in different scenes and different objects. However, it is important not to embrace the notion of a hierarchy of visual processing too dogmatically because visual processing is not indisputably modular. Examination of area V4, for example, which had been thought to specialize in color vision, indicates that in addition, V4 contributes to several general aspects of visual information processing which include form perception, visual learning, spatial generalization, visual attention and stimulus selection. It cannot be overemphasized that these general functions are the result of extensive interactions among many cortical (and subcortical) areas. These secondary and tertiary cortices feedback onto the primary cortex, and their role in vision is well illustrated by the following experiments. In both cat and macaque, the classical, well studied visual areas have been isolated by extensive ablation of adjacent, frontal, parietal and superior temporal cortex. Although the spared, retinotopically organized areas were largely intact anatomically and functionally, the animals were blind for the duration of their lives. This dramatic finding indicates that areas of cortex not usually thought of as primarily visual in function play a significant role in visual perception.

Another reason not to uncritically accept the notion of a hierarchy of visual processing is that cortical circuitry is not solely feed-forward in nature. We must bear in mind that most visual areas are heavily interconnected. For example, we have thus far emphasized the feed-forward pathways that distribute information from area 17 to multiple visual centers beyond 17. It is just as important to recognize that many feedback pathways exist; pathways that originate in "higher" extra-striate areas that project back to "lower" areas, including the striate cortex. These feedback pathways permit the so-called "higher" visual centers to exercise various kinds of control over "lower" cortical areas. Such feedback control could enable higher cognitive processes involving expectancies, memories, the current goals of the perceiver, and attention to guide or otherwise influence even the earliest stages of cortical visual processing. Cognitive and information processing approaches to visual perception refer to these higher-level influences over visual processing as "top-down" processing, and many classic perceptual phenomena point to its importance in a wide variety of perceptual functions including attention, reading, patterns of visual fixations, bistable percepts (see Figure 4), and image recognition itself.

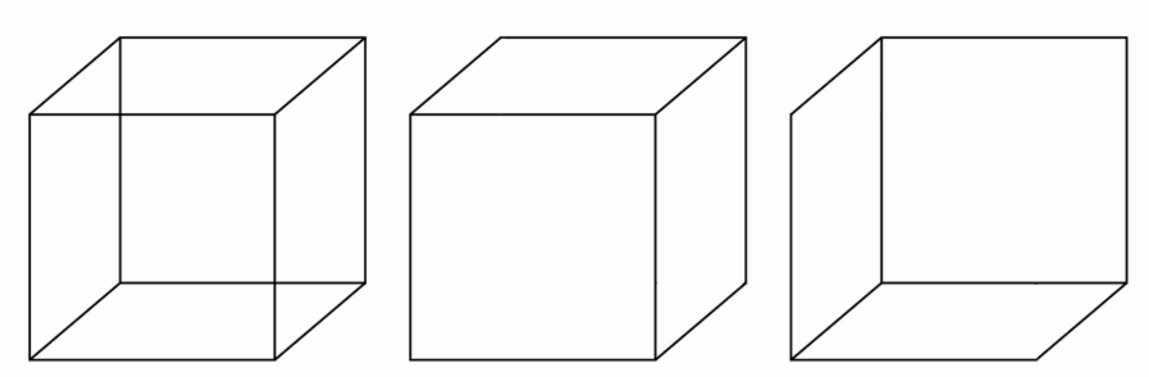

Figure 4. The figure on the left is the Necker cube, probably the most famous example of a bistable percept. The two alternative percepts this cube can produce are also illustrated. Bistable percepts such as this clearly demonstrate that visual percepts need not be entirely specified by the retinal image, but are also subject to endogenous influences.

The problems inherent in determining an object’s 3D form are inseparable from problems associated with recognizing that object. The perception of form does not necessarily involve making matches to representations of things seen in the past, but form *recognition* does. Once even a part of an unfamiliar object has been recognized, the object's representation in memory can be used to constrain possible shape solutions for its other, still unrecognized parts. For example, in the image shown in Figure 5 (photograph by R. C. James), it is hard to discern a Dalmatian standing among many black spots scattered on a white background because the part of the image corresponding to the dog lacks contours that define the edges of the dog, and the dog's spotted texture resembles that of the background. Many observers find that they first recognize one part of the dog, say the head, which then makes the whole dog's shape apparent. This is an example of top-down feedback from a model of a dog's shape in memory that constrains shape processing.

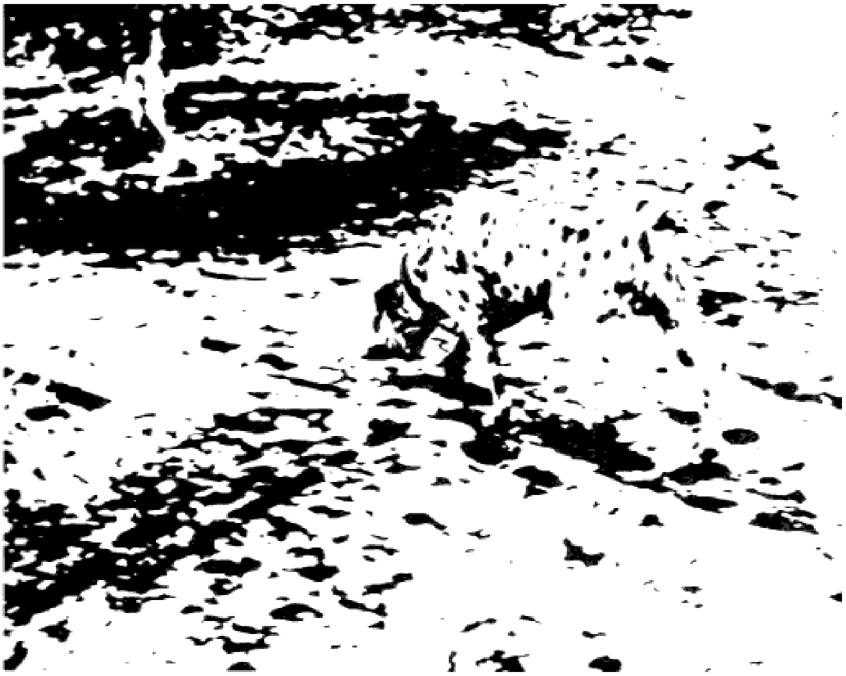

Figure 5. Can you see a Dalmatian dog here?

How newly formed representations of shape are matched to existing representations of shape in memory constitutes the contentious problem of how objects are recognized. While an in depth discussion of recognition takes us too far from our topic of form processing, suffice it to say that there does not appear to be a monolithic, multi-purpose general recognition system. Instead, multiple representations of an object appear to be formed, each specialized for the purpose and coordinate transformation required for some perceptual or behavioral task. For example, face processing appears to take place in a region on the underside of the cerebral hemispheres called the fusiform gyrus. Bilateral damage to this area results in prosopagnosia: the inability to recognize familiar faces. Prosopagnosics are typically able to recognize non-face classes of objects, and indeed the person suffering from prosopagnosia can recognize a face as a face – they can often tell the emotional expression, approximate age and gender of the face. But they cannot identify the people they know by viewing their face. This condition implies that the human brain contains neural circuits that may be specialized for the recognition of specific classes of objects (like faces). It is also important to realize that there are other perceptual disorders (agnosias) that impair object perception in general. Damage within the temporal lobes and/or occipital lobes can produce profound deficits in object recognition. In some cases, the patients cannot name an object presented visually, but can name the object if they feel it. In many cases they cannot copy drawings of objects, but sometimes can name the same object they cannot draw. This suggests that these areas of cortex are specialized for high-level shape processing, object categorization, and matches to memory. In particular, patients with associative agnosia, who have such lesions, probably fail to recognize objects because they fail to process shape in a normal way. It appears that elements of their local shape processing remain intact, while global analyses of shape and configural analyses of relationships among local shapes have been lost.

The constructive nature of visual form processing

Thus far we have focused on information that is available in the image that can be extracted or detected by filters or detectors tuned to useful image primitives. At some point these primitives must be used to construct representations of objects and 3D surface layout. In a sense, construction already begins at the stage of retinal ganglion cells, where uniformity is deemphasized in favor of information about borders. However, retinal processing is still confined to extraction of local information in the image. Further form processing requires the construction of information that is not given in the retinal image. For example, the retinal image is 2D not 3D. Any inference of a 3D surface from 2D image cues is tantamount to the addition, creation, and construction of information not present in the image. The stage of construction was first explored by Hermann von Helmholtz, the great 19th century physicist/physiologist/psychologist and later by the early 20th century Gestalt psychologists, who emphasized that visual perception must be subserved by rapid grouping procedures that link information across the image in a global fashion on the basis of heuristics such as contour continuity. While there is strong evidence that such grouping procedures are carried out in the course of visual form processing, little progress has been made in discerning how such global grouping procedures can be realized by local neuronal computations. Rather than give a necessarily speculative and sketchy description of how 3D form is computed at the level of neuronal 'hardware,' it will be more productive at this stage to consider how the construction of visual form may take place at the more abstract level of the information processing algorithms used to construct information about 3D form.

As we stated at the outset, recovering 3D form from the inherently ambiguous 2-D retinal image is perhaps the most fundamental problem faced by the visual system. To solve this problem, multiple systems have evolved to recover 3D shape from the various cues to form found in the image. Examples are the perception of shape from shading, the perception of shape from motion (often called 'structure-from-motion') and the perception of shape from retinal disparity cues (such as random dot stereograms and 'magic eye' books). Solving for shape using multiple strategies and cues has numerous advantages. Multiple circuits can compute shape solutions in parallel, reducing overall computational time. If one subsystem should reach a solution before others, that shape solution can constrain and be constrained by the computations being carried out by other subsystems. Parallel and concurrent processing also reduces the likelihood of attaining non-veridical solutions because all shape subsystems must come to mutually consistent solutions, permitting a form of error-checking and redundancy. Thus, under normal viewing conditions, the problem of recovering 3D shape from the 2-D image can be solved using multiple mutually constraining cues. In some cases, however, just a single cue can be used to generate a distinct percept of 3D form. For static images, contours probably offer the strongest constraints on 3D form because other cues, such as shading or texture, appear to be captured or dominated by contour cues.

There are several shape formation models in the literature. The primary ones are (1) codon theory, (2) a parts-based or structural approach, (3) a medial axis approach, (4) an approach premised on the complete recovery of visible surface orientations and depths, and (5) a contour propagation approach. The theories are not mutually exclusive, and each is inadequate in different ways. Each could supplement versions of any of the other major theories. It is important to keep in mind that the representation of 3D shape used by the visual system may not be monolithic. Shape codes could involve aspects of more than one of these theories, depending on the particular problem to be solved. It is therefore instructive to compare the major shape theories.

The (1) codon theory (Richards & Hoffman, 1985; Richards *et al*., 1987) imposes useful constraints on which 3D shapes can be inferred from image contours. The key insight underlying this theory is a theorem by Jan Koenderink that proves that there is a law-like correspondence between the sign of contour curvature (positive, negative, or zero) and the sign of surface curvature (bulging, saddle-shaped, or flat). The codon approach offers useful constraints on possible shapes but does not predict which specific shape of all possible shapes will be perceived. The codon theory makes mistakes because it is built upon the flawed assumption that all contours arise from volumetric objects. Thus an elliptical silhouette is claimed to look like an ellipsoid when to most observers it in fact looks like a flat hole or disc lying on a ground plane.

(2) Geon theory (Biederman; 1987; see also Marr, 1982) is built on the idea that objects can be represented in view-invariant terms as an assemblage of primitive parts called “geons” or geometric cones. Geons are constructed by sweeping a basic shape through space to define a volume. A strong version of this theory claims that all objects can be represented as a combination of geons, which function as a sort of shape alphabet from which more complex objects can be constructed. Segmenting a complex object at regions of deep surface concavity is thought to give rise to geons when no further segmentation is possible. A confusion may arise because geons have been regarded as a solution to two independent problems: (1) shape formation, on the one hand, and (2) object recognition, on the other. While 3D parts may be a useful way to index a shape in memory in order to recognize something, 3D shapes need to be constructed before they can be segmented into 3D parts. It is therefore unlikely that shapes are themselves constructed from a small alphabet of primitive simple shapes. We can after all, enter a cave and see all kinds of strange shapes, none of which is reducible to a geon or collection of geons. However, once surfaces and volumes have been constructed, it is reasonable that these are segmented at regions of local minima of surface curvature (compare Hoffman & Richards, 1985; Hoffman & Singh, 1997) and that these parts may serve as an index for the matches to memory that underlie recognition. A solution to the primary problem of shape formation should not be limited to combinations of primitive volumes, because many shapes lack a distinct volume entirely, such as the surface of the ocean or swirls of smoke.

Relevant here is a large body of empirical evidence that shows that object recognition is view-dependent, in contrast to the predictions of geon theory. Researchers who emphasize this evidence argue that objects are represented and stored as a series of views. However, what comprises a view is not clear. At one extreme, a view might just be a 2-D image. This extreme would have difficulty accounting for the various constancies (i.e. indifferences to image transformation) expressed by the visual system. For example, an object defined by contours alone, motion alone, or texture alone will tend to look like it has the same shape across these cues. Moreover an object viewed from various distances and under various lighting conditions will generally appear to have the same shape, although particular images will be very different from each other. A more moderate stance is that a view is a collection of features. Such features, even if they do not explicitly represent 3D shape or depth information, may implicitly capture 3D information, because viewpoint invariant recognition could emerge if all views of an object are matched to the same node in a distributed neural network. If network models can be built that match correctly, it may become difficult to experimentally distinguish whether the visual system constructs explicit representations of 3D shape or whether it only acts as if it did. At the other extreme, a view might include an explicit representation of 3D shape. Tarr & Kriegman (2000), for example, suggest that a view is a span of viewpoints over which the qualitative shape description, in terms of occluding contour relationships, does not change. This converges to a certain extent with the revised version of geon theory (Biederman & Gerhardstein, 1995; Hummel & Biederman , 1992), according to which recognition will be view-invariant only over a set of views for which a given collection of geons is visible.

3D structures (such as holes, protrusions, parts, corners, valleys, indentations, etc.) and the particular spatial relationships that hold among them (e.g. hole below pinnacle above bulge) which can be discerned from a given viewpoint are *intrinsic* to the object and can underlie a viewpoint-invariant representation of shape because these same structures will be visible from many other viewpoints. Even if a geon description *per se* is not utilized by the visual system for recognition, it is likely that some other structural description is utilized. In general, the visual system attempts to recover the intrinsic properties of objects (e.g. surface reflectance, material substance, 3D shape) because these are more or less constant, whereas extrinsic properties (e.g. lighting, shading, shadows, distance, orientation) are constantly changing. Both intrinsic and extrinsic information can be derived from the image, and probably both types are stored and utilized for various tasks, including recognition.

(3) Medial axis theory is built upon the insight that objects, such as human bodies, can be reduced to stick figures that are nonetheless recognizable. Perhaps such stick figures function as a rudimentary code for recognizing shapes and objects. This type of theory is inadequate because current algorithms calculate axes in the image, not in the world. A long stick can cast an image that looks like a disk, if viewed from the appropriate angle. Reducing such an image to an axis will be difficult if not impossible. And yet, it is not clear how to generate medial axes in a 3D sense. Interpreting some 2D images as 3D objects on the basis of medial axes alone may require that we have the 3D shape description of the object already, and this is just what we are trying to recover from the image. If we limit ourselves to determining axes in the image, then the medial axis approach can give wrong solutions.

(4) Metric surface recovery theories (e.g. Marr, 1982) maintain that perceived shape depends on recovering precise values of depth and surface orientation in viewer-centered coordinates for every point of a visible surface. This approach to shape recovery is incorrect because the shape code underlying visual perception does not obey a Euclidean metric. More recently, an extensive literature has emerged showing that there is substantial variance and inaccuracy when observers try to specify depth and orientation values for positions on a surface, even when given varied or multiple sources of visual information. Most cues for shape, including motion parallax, perspective, texture gradients, surface contours, occluding contours, highlights, shading, or shadows, can only provide information about the *sign* of surface curvature. These shape-from-x cues can perhaps provide information about relative surface orientation or curvature but cannot provide information about absolute surface orientation or curvature. Two shape cues, disparity and motion, can in principle provide information about absolute surface curvature at any visible point of an object provided that certain reasonable assumptions, such as object rigidity, are adopted. However, the visual system does not seem to fully exploit this image information because perceived shape is not coded metrically at least insofar as depth and surface orientation are not precisely represented. It appears that the visual system may be satisfied with a fairly inaccurate representation of 3D form rather than the precise one hoped for in Marr's (1982) surface recovery program.

After researchers rejected Marr’s (1982) program for the metric recovery of surface orientation and distance, it was not clear what type of shape description could or should replace it. Just because a metric description is not attainable does not mean that precise shape information cannot be recovered from the image. However, none of the major contour-based approaches to form perception besides Marr’s offer a program for the recovery of precise shape information from the image. Certainly geons, codons, or medial axes are not capable of uniquely or metrically specifying the curved internal structure of a surface. However, these other approaches cannot even give a precise ordinal or relational description of the curved internal structure of a surface. There seems to be a gap between the precise but metric description of shape that Marr sought and the imprecise, non-metric shape descriptions that have been offered in its place.

(5) The contour propagation approach. More recently (Tse, 2002) it has been suggested that form processing on the basis of contours may involve a computational algorithm that propagates contour information away from edges and corners into the interiors of surfaces. Indeed, such a contour propagation approach could be the mechanism that generates 3D curved surfaces given only the 2D contours available in, say, line drawings or silhouettes. The existence of a particular magnitude and sign of contour curvature along some portion of a closed contour limits the kinds of 3D surface curvature that the corresponding portion of surface can have in the world. This in turn limits the surface curvature that adjacent regions of surface can have, under an assumption of smooth surface curvature change over a volume. The possible surface curvatures implied by different portions of contour constrain one another across the image, leading to only one or a few possible 3D shape interpretations that are consistent with a given closed contour.

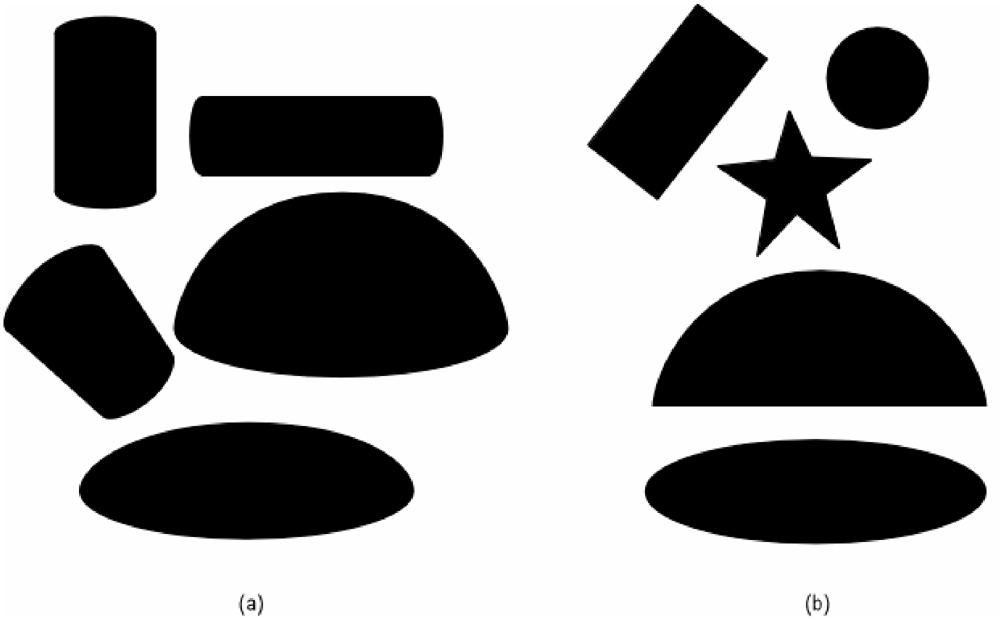

Figure 6. Any algorithm that accounts for visual shape processing must be able to explain why the silhouettes in (a) look volumetric while those in (b) look flat.

The fact that we can see 3D shapes in silhouettes, such as shown in figure 6, and line-drawings, such as comic strips, is remarkable, given both the paucity of information in such images, and the fact that no object in the world looks like a line-drawing or presents us with edges in the absence of surface information. Besides establishing that contours alone are sufficient to generate a full-blown 3D representation of shape, such stimuli suggest that the visual system itself may extract and make explicit some version of a line-drawing version of the image that makes explicit locations where an opaque surface occludes that which is behind it. Impoverished stimuli, such as line drawings, may reveal other important facts about how the shape-from-contour system solves for shape. Any isolated line or curve in a line-drawing is essentially meaningless. It is necessary to specify what comprises the inside versus outside of a line, before that line can specify a surface that occludes its background. Moreover, it is necessary to specify the global shape of a contour before the 3D shape of a surface can be inferred from that shape, although it is possible that the processes involved in contour extraction and interpolation operate given mutually constraining feedback from processes dedicated to inferring 3D surface layout, depth, and shape.

It is unlikely that a passive cascade of increasingly complex receptive fields (e.g. oriented bar→curve→arrangement of curves→object) will provide a sufficiently robust and sensitive code for extracting shape from contour rapidly and correctly. Because even the most minute local change to a global closed contour, such as found in a silhouette, may drastically change the perceived 3D form, it is likely that form is computed using dynamic computational algorithms that are poorly captured by the notion of a receptive field. Such computations may be carried out at the circuit-level, in which case no single neuron in that circuit may be found to be have a classical receptive field tuned to a particular shape computed by the circuit.

In summary, we have traversed the visual system from the level of initial extraction or detection of image primitives, to the stage where those primitives can be used as the input to complex algorithms that compute surface shape and layout. A great deal remains to be accomplished at each level of analysis. We still do not understand how information is processed by neurons in a deep sense, and we certainly do not grasp how complex computations, such as those that presumably underlie Gestalt grouping procedures, are realized in the information processing of extended neuronal circuits. At a more abstract level of analysis, we do not understand the nature of the computations that generate veridical representations of shape within a fraction of a second, permitting matches to memory (recognition) and motoric behavior in response to the visual environment. Although much is already known, much more work needs to be done before we can say that we have even a basic understanding of how form is processed and represented in the nervous system.

References and Further reading

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review 94*(2) 115-147.

Biederman, I. & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1506-1521.

Gazzaniga M. S., ed. (1995): *The Cognitive Neurosciences.* Cambridge: MIT Press.

Hoffman, D. D., & Richards, W. (1984). Parts of recognition. *Cognition, 18*, 65-96.

Hoffman, D. D. & Singh, M. (1997). Salience of visual parts. *Cognition, 63*(1), 29-78.

Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition.

Tarr, M. J. & Kriegman, D. J. (2001). What defines a view? Vision Research, 41(15):1981-2004

Marr D (1982): *A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: Freeman

Palmer, S. E. (1999). Vision Science: Photons to Phenomenology. MIT Press.

Richards, W. A. & Hoffman, D. D. (1985). Codon constraints on closed 2-D shapes. *Computer Vision Graphics and Image Processing, 32*, 265-281.

Richards, W. A., Koenderink, J. J., & Hoffman, D. D. (1987). Inferring three-dimensional shapes from two-dimensional silhouettes. *Journal of the Optical Society of America A, 4*, 1168-1175.

Schiller P, Lee K (1991): The role of the primate extrastriate area V4 in vision. *Science* 251:1251-1253.

Spillman L, Werner JS, eds. (1990): *Visual Perception, The Neurophysiological Foundations.* New York: Academic Press

Tse, P. U. (2002). A contour propagation approach to surface filling-in and volume formation. Psychol Review,109(1): 91-115.

See also Visual perception; Visual system, organization; Visual cortex, extrastriate; Striate cortex; Blindsight, residual vision; Attention, selective visual