ANNUAL REVIEWS

*Annual Review of Vision Science*

# Shape from Contour: Computation and Representation

James H. Elder

Centre for Vision Research, York University, Toronto, Ontario M3J 1P3, Canada;
email: jelder@yorku.ca

Annu. Rev. Vis. Sci. 2018. 4:423–50

The *Annual Review of Vision Science* is online at vision.annualreviews.org

https://doi.org/10.1146/annurev-vision-091517-034110

## Keywords

shape, form, object, contour

## Abstract

The human visual system reliably extracts shape information from complex natural scenes in spite of noise and fragmentation caused by clutter and occlusions. A fast, feedforward sweep through ventral stream involving mechanisms tuned for orientation, curvature, and local Gestalt principles produces partial shape representations sufficient for simpler discriminative tasks. More complete shape representations may involve recurrent processes that integrate local and global cues. While feedforward discriminative deep neural network models currently produce the best predictions of object selectivity in higher areas of the object pathway, a generative model may be required to account for all aspects of shape perception. Research suggests that a successful model will account for our acute sensitivity to four key perceptual dimensions of shape: topology, symmetry, composition, and deformation.

## 1. INTRODUCTION

Imagine a visual world consisting only of a random haze of color and texture. This is a world without shape, and it illustrates how object shape perception is crucial to our visual experience.

While objects in our visual world are generally three-dimensional (3D**)**, the boundary of a 3D object projects to the retina as a two-dimensional (2D) region bounded by a one-dimensional (1D) closed contour (Koenderink 1984) (**Figure 1*a***). Such bounding contours are sufficient to yield compelling percepts of 2D and even 3D shape (**Figure 1*b*,*c***) and provide important cues for object detection and recognition.

Tasks such as object recognition depend upon global shape information. In the early stages of the visual system, neurons have highly localized receptive fields and therefore can represent only local shape information; inference of global shape entails the selective grouping (perceptual organization) of these local shape features. In real scenes, the bounding contour is often fragmented owing to visual clutter and occlusions, complicating this task. Thus we can distinguish two key issues:

1. Computation: How is shape information reliably extracted and organized from complex images?
2. Representation: How is this shape information represented to subserve perceptual tasks?

While these two issues are distinct, I argue in this article that they are closely coupled, in that the representations of global shape computed by the brain are likely to aid in perceptual organization.

Since the extraction of shape information from complex natural images can be nontrivial, one might expect that for rapid object detection, the brain relies instead upon appearance cues such as texture and color that may be easier to extract. To address this question, let us consider the specific task of rapidly detecting animal shapes, which are of particular ecological relevance and figure prominently in the earliest known examples of cave art.

Humans perform this task remarkably well: Evoked potential studies indicate that the corresponding neural signals can emerge in the brain as early as 150 ms following stimulus onset (Thorpe et al. 1996). But what cues are being used to detect the presence of an animal so rapidly? To answer this question, Elder & Velisavljević (2009) measured human performance on a rapid animal detection task for natural images in which visual cues such as luminance, color, texture, and shape were selectively turned on or off. The results revealed that humans do not use simple luminance or color cues for animal detection, but instead rely on shape and texture cues. Interestingly, shape cues appear to be the first available, influencing performance for stimulus durations as short as 10 ms, within a backward-masking paradigm. These psychophysical results are consistent with a more recent human magnetoencephalography (MEG) study in which object identity

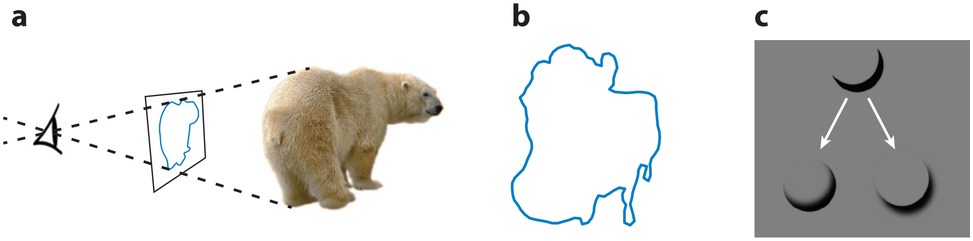

**Figure 1**

(*a*) A solid three-dimensional (3D) object projects to the eye as a planar 2D region bounded by a 1D contour. (*b*) 3D shape from bounding contour. (*c*) The shape and blur of the bounding contour can profoundly influence the perception of 3D shape. Figure adapted from Elder et al. (2004).

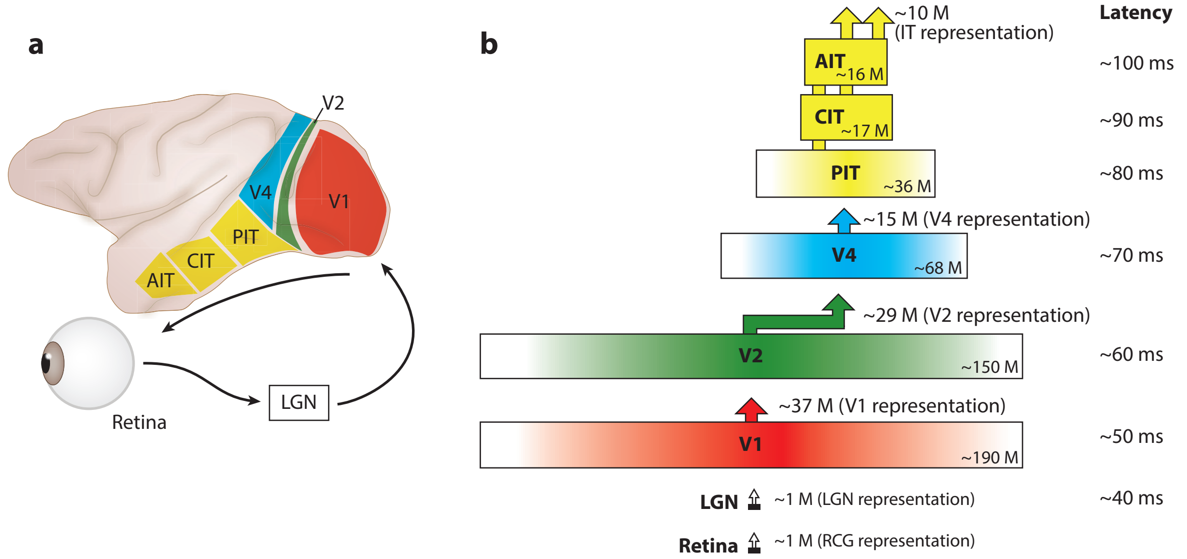

Figure 2

Schematic of the object pathway in macaque. (*a*) Anatomical layout. (*b*) Functional feedforward hierarchy. The arrows show the feedforward flow of visual information that defines the hierarchy of visual areas. Not shown are the massive feedback processes that transfer information from higher to lower visual areas. Involvement of these feedback connections in shape perception is discussed in Section 4. Figure adapted from DiCarlo et al. (2012). Abbreviations: AIT, anterior inferior temporal cortex; CIT, central inferior temporal cortex; LGN, lateral geniculate nucleus; PIT, posterior inferotemporal cortex.

is decoded from the dynamic cortical response to briefly presented images (Carlson et al. 2013). Results revealed that object shape is predictive of decoding latency: Objects that differ more in shape can be discriminated earlier.

Together, these behavioral and electrophysiological results suggest that contour shape cues are not luxury items used only when time is not a factor but rather underlie our fastest judgments about the objects around us. How does the brain perform such a complex task so quickly? To begin to form an answer, we first review what is known about the physiological mechanisms underlying shape processing in primate cortex. We then turn to computational models.

## 2. FEEDFORWARD SHAPE PROCESSING: PHYSIOLOGY

A hierarchy of visual areas in ventral cortex of primate known collectively as the object pathway subserves object shape perception (**Figure 2**).[1] Advancing through the hierarchy, receptive fields become larger and more complex in their tuning, with increasing selectivity for certain properties (e.g., shape) and more invariance for others (e.g., retinal location). Visual information reaches the top of the hierarchy within about 100 ms after stimulus onset and from there is shared with various other parts of the brain that control decision and action. The following sections summarize what is known about the shape representations formed in each cortical area during this first feedforward sweep through the object pathway.

[1] While shape selectivity in dorsal stream has been demonstrated, its role in object recognition is less pronounced and less understood (Lehky & Sereno 2007, DiCarlo et al. 2012, Kuai et al. 2017).

## 2.1. Low-Level Neural Representations: Areas V1/V2

At the most elemental level, a visual contour is a one-dimensional (1D) continuum of locations in retinotopic space. The small center-surround receptive fields (RFs) of neurons in the retina and lateral geniculate nucleus (LGN) constitute a discrete, distributed, pointillist code of this continuum. In this review, we focus on the piecewise smooth contours generated by the majority of objects in our visual world, and for these, a richer description of the local shape at each point can be developed through a local differential expansion of this continuum (do Carmo 1976). The first term in this description describes the local orientation of the contour via the tangent vector. Unlike neurons in the retina and LGN, neurons in area V1, the first cortical stop for visual information, are tuned to this local orientation (Hubel & Wiesel 1968). This orientation tuning can be predicted by a principle of efficient coding (Barlow 1959, Olshausen & Field 1996).

The next term in the differential expansion of the local shape describes the curvature of the contour. Nonlinear spatial selectivities of V1 neurons known as end stopping are believed to play a role in coding curvature as well as tangent discontinuities (corners). For example, in cat area 17 (the homolog of primate V1), end stopping has been shown to induce selectivity for contour curvature and corners (Dobbins et al. 1987).

The RFs of V2 neurons are on average larger than V1 RFs, and a greater proportion are tuned to more complex local geometry, including curvature, angles, and intersecting lines (Hegde & van Essen 2000), as well as illusory contours (von der Heydt et al. 1984).

## 2.2. Mid-Level Neural Representations: Area V4

Area V4 forms the main gateway between early visual cortex (V1/V2), where local shape features such as orientation are coded, and inferior temporal cortex (IT), where neurons exhibit highly selective responses to global object shape. When probed with simple outline contour fragments and silhouette shapes, many V4 cells exhibit systematic joint tuning to the location, orientation, and curvature of local contour features. This coding is to a great extent object-centric and includes selectivity for the sign of curvature (convex or concave) (Pasupathy & Connor 1999, 2001). For most shape-selective cells, tuning is driven largely by a single contour feature (e.g., a bump or a dent), although for many neurons, tuning is further determined by the conjunction of 2–3 adjacent curvature events. Peak curvature tuning is broadly distributed, although more neurons are tuned to convex curvature than to concave curvature.

At the population level, V4 devotes more spikes to high curvatures than low curvatures (Carlson et al. 2011). This is consistent with the theory of entropy encoding (Huffman 1952), which dictates that in order to minimize average coding cost, features that are encountered more frequently should be assigned shorter codes. Since many of the objects we encounter are smooth, low curvatures are more common than high curvatures and so should be coded with fewer spikes.

## 2.3. High-Level Neural Representations: Inferior Temporal Cortex

Area IT is the highest area of monkey ventral stream. It can be divided into a posterior region (also known as TEO) and anterior region (also known as TE). In human visual cortex, the lateral occipital cortex (typically abbreviated as LOC or LO) appears to play a similar functional role (Grill-Spector et al. 2001); however, the precise relationship between human LOC and monkey areas IT and V4 remains unclear. In addition to LOC, regions of both the human inferior temporal and ventral temporal cortex (VTC) are known to be selective for objects (Grill-Spector & Weiner 2014).

To what object-shape features are IT neurons tuned? Brincat & Connor (2004) systematically probed posterior IT neurons with a large set of silhouette stimuli over a range of shape dimensions. They found that most cells exhibited a more complex representation of contour shape than found in V4 and that this could be modeled as linear or nonlinear combinations of excitatory and/or inhibitory sensitivity to between two and four curvature features at specific object-centered locations and orientations. Cells with significant nonlinear contributions tended to be more selective (sparse), exhibiting excitatory multiplicative selectivity for a configuration of contour parts.

Kayaert et al. (2005) measured the response of IT neurons to simple deformations (bending, tapering, fattening/thinning) of triangular and rectangular shapes. They found that IT population response variation was well predicted by a factorial model over these shape dimensions, with most neurons tuned primarily for one class of deformation. Most intriguingly, this tuning tended to be monotonic rather than bell-shaped: The preferred stimulus for a neuron tended to be at the extreme end of the deformation range. This is quite different from prototype models of population coding for shape that assume that each neuron will have bell-shaped tuning around a preferred stimulus and that these preferred stimuli will be broadly distributed across the parameter space (e.g., Riesenhuber & Poggio 2002). Such bell-shaped tuning for 3D pose has been observed in IT for 3D objects (Logothetis et al. 1995); however, variation in 3D pose results in discontinuous configural changes in the object projection (i.e., the occlusion or dis-occlusion of visual features) that form a natural partitioning of the view sphere into prototypes. As for curvature tuning (Section 2.2), this monotonic tuning to deformation may be consistent with the strategy of entropy encoding. If the symmetric base shapes (triangle and rectangle) employed for this study are encountered frequently, they should be encoded with relatively few spikes. In contrast, if extremely deformed stimuli are encountered less frequently, they should be encoded with more spikes.

While IT neural representations are known to support object recognition in the face of confounding factors such as variations in object position, size, pose, and background (Desimone et al. 1984), this does not mean that individual neurons are invariant to these other factors. In fact, neurons with the highest selectivity for object identity are the least invariant to changes in position, size, contrast, and clutter (Zoccolan et al. 2007, DiCarlo et al. 2012), and these other properties can typically be decoded from IT more accurately than from earlier visual areas (Hong et al. 2016). Thus the computational objective of ventral stream may be not to discount factors such as object pose as nuisance variables in order to achieve view-invariant object recognition but rather to disentangle object identity from these covariates at the population level, making them more easily decodable (DiCarlo & Cox 2007, DiCarlo et al. 2012).

## 3. FEEDFORWARD SHAPE PROCESSING: COMPUTATIONAL MODELS

Decades of neurophysiological research have yielded many insights into the staged feedforward processing of shape information in the object pathway. This progress provides hope that a rigorous computational model may be within our grasp. What do we seek in this model? First, we want a model that can, like the brain, reliably compute shape representations directly from raw imagery.[2] Second, we expect the computations and representations formed by the model to be rational in terms of the ecological statistics of our visual world and the visual tasks for which shape information is important. Third, we seek correspondence between representations of the model and the shape selectivity of the key visual areas of the object stream, reviewed above. Finally, we desire a model that is parsimonious, that is based upon a small set of principles, and that can be realized with relatively few free parameters. We keep these goals in mind as we review the progress that has been made in forming elements of such a model.

[2] While the brain builds these representations from hexagonally packed photoreceptors, computational models and computer vision algorithms typically work from a rectangular lattice of pixels. This distinction may be important in the formation of the earliest, most local representations of shape but is expected to be of less importance for later, more global shape representations in higher areas of the object pathway.

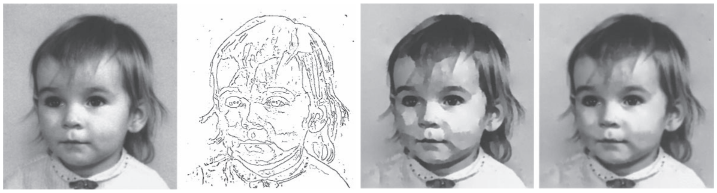

Figure 3

An edge-based visual representation is nearly complete. From left to right: original image, edge map, reconstruction of luminance and contrast, and after restoration of blur. Note the importance of the blur information in assigning the correct perceptual labels to attached shadow edges. Figure adapted from Elder (1999).

### 3.1. Edge Detection

Computationally, the first stage of shape processing is edge detection: localization of abrupt changes in luminance, color, or texture. Even this step is nontrivial for natural images, in which the local appearance of the edge can be faint, blurred, or complicated by clutter. Computer-vision edge-detection algorithms (e.g., Elder & Zucker 1998b) often employ oriented Gaussian derivative filters that have higher signal-to-noise (SNR) than local difference kernels, and these filters bear a striking resemblance to the receptive fields of neurons in primary visual cortex (Hawken & Parker 1991). Multi-scale filtering methods (Elder & Zucker 1998b, Lindeberg 1998) are also effective in raising SNR, and this matches fairly well with the physiological (Hawken & Parker 1991, Ringach 2002) and psychophysical (Wilson & Bergen 1979, Watt & Morgan 1984, Elder & Zucker 1996b, Elder & Sachs 2004) evidence for multi-scale processing in human and nonhuman primate.

The mean orientation bandwidth of these local mechanisms is determined by the shape (elongation) of the receptive fields and has been estimated psychophysically using grating stimuli (e.g., Campbell & Kulikowski 1966) and orientation fields—for example, Glass patterns (e.g., Maloney et al. 1987, Or & Elder 2011)—to be between 7 and 15 degrees (half width at half height). While these estimates correspond fairly well to the physiology (Hawken & Parker 1991, Ringach 2002), orientation bandwidth of V1 neurons appears to range quite broadly about this mean (Ringach 2002). Psychophysical experiments suggest that this joint variation in receptive field scale and shape determines human ability to detect edges in cluttered scenes (Elder & Sachs 2004).

One might be concerned that an edge-based visual representation would neglect a great deal of important visual information. It has been shown, however, that a near-perfect image can be reconstructed from a perceptual code that represents the luminance, contrast, and blur only at image edges (Elder 1999) (**Figure 3**).

To infer shape from contour, the visual system must identify and group the edges generated by the object's bounding contour and surface creases. However, many image edges are in fact generated by reflectance changes, shading, and shadows. Computationally, color and texture cues can be useful in picking out important edges (e.g., Martin et al. 2004), and contour blur information can signal edges caused by shadows and shading (Elder 1999, Elder et al. 2004, **Figures 1*c*** and **3**). Luminance contrast can be used to discriminate boundary edges from other edges (Vilankar et al. 2014), and Ehinger et al. (2017) have shown that a deep neural network (DNN) can be trained to use local luminance, color, texture, and orientation cues to distinguish depth from nondepth edges. Little is known about the physiological mechanisms underlying the human visual system's ability to distinguish between these different physical edge classes.

## 3.2. Local Gestalt Grouping: The Association Field

To infer global aspects of shape, the brain must selectively group local edges into extended contours. This is a complex combinatorial problem of exponential complexity. Early work by Gestalt psychologists (Koffka 1935) identified two factors that determine whether two edges should be grouped:

1. Proximity: How close together are they?
2. Good continuation: Can they be interpolated smoothly (**Figure 4**)?

These principles were influential in the development of early computer vision algorithms for contour extraction (Parent & Zucker 1989; **Figure 4*a***) and have now been quantified psychophysically (Kellman & Shipley 1991, Field et al. 1993, Kubovy & Wagemans 1995; **Figure 4*b***) and

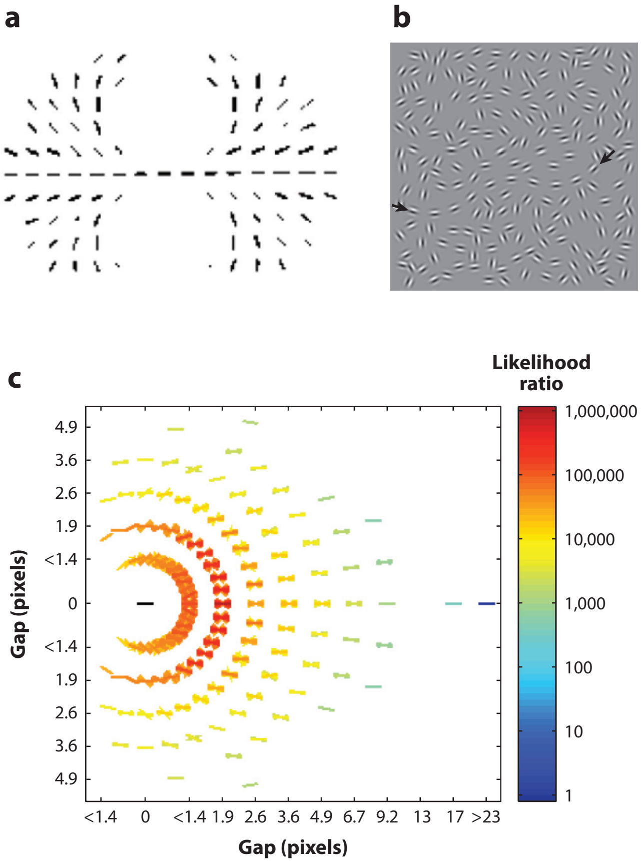

**Figure 4**

(*a*) Cocircularity support neighborhood (also known as the association field) that models the local grouping of edges using Gestalt principles of proximity and good continuation. Panel adapted from Parent & Zucker (1989). (*b*) Psychophysical stimulus introduced by Field et al. (1993) to study the grouping of contours in clutter. The stimulus consists of a dense field of randomly oriented Gabor elements in which a smooth contour of aligned elements may be embedded (indicated by *arrows* here). The observer's task is to detect the contour, if present. Sensitivity to good continuation and similarity cues can be measured by varying the orientation and appearance of the elements on the path. Panel adapted from Hess & Field (1999).
(*c*) Association fields derived from the ecological statistics of contours. The color of the oriented elements indicates the likelihood ratio for two oriented elements to be neighboring elements on the same object boundary. Panel adapted from Elder & Goldberg (2002).

probabilistically through measurements of natural scene statistics (Geisler et al. 2001, Elder & Goldberg 2002; **Figure 4*c***). The resulting pattern of grouping probabilities between edge elements is often called the association field (Field et al. 1993). The physiological substrate for the association field may be the long-range horizontal connections between orientation columns that connect neurons coding local orientation signals that can be smoothly interpolated, identified first in cat area 17 (Gilbert & Wiesel 1989) and later in monkey V1 (Stettler et al. 2002).

## 3.3. The Markov Assumption

The association field governs only the grouping of a pair of edges; to model the grouping of many edges into an extended contour, some additional assumptions must be made. One option is to measure the grouping likelihoods between all pairs of edges on an object (Geisler et al. 2001) and invoke a transitivity assumption: If edge A groups to edge B, and edge B groups to edge C, then edge A also groups to edge C. This rule will generally lead to unordered sets of edges that do not respect the 1D nature of contours.

An alternative is to measure the grouping likelihoods only between successive pairs of edges on the contour (Elder & Goldberg 2002) and invoke a Markov assumption: The likelihood of the contour is given by the product of the likelihoods of each local pairwise association between adjacent edges (Zucker et al. 1977, Sha'ashua & Ullman 1988, Elder & Zucker 1996a, Williams & Jacobs 1997, Elder et al. 2003, Movahedi & Elder 2013, Almazen et al. 2017). This assumption respects the 1D nature of the contour but greatly simplifies the probabilistic model: The local pairwise grouping probabilities are now sufficient statistics for computing maximum probability contours. Critically, the Markov property also confers an optimal substructure property: Any piece of a maximum probability contour must itself have maximum probability. This property allows maximum probability contours to be computed progressively in polynomial time, via shortest-path methods such as Dijkstra's algorithm or dynamic programming (Elder & Zucker 1996a, Elder et al. 2003).

## 3.4. Limitations of the Markov Assumption

Unfortunately, these first-order Markov models generally do not perform well on natural scenes. One problem is topology. Unlike the transitivity assumption, shortest-path algorithms based upon the Markov assumption enforce the ordinality constraint and thus eliminate incorrect topologies caused by bifurcation. Unfortunately, these algorithms are still not guaranteed to extract a contour of the correct topology as embedded in the image plane (Elder & Zucker 1996a).

A second problem is that the Markov property restricts the prior over contour length to have an exponential form, and this prior cannot be changed within the constraints of polynomial-time shortest-path algorithms. This induces a prior bias toward small contours, so that algorithms tend to extract only small parts of a shape rather than an entire shape.

A third issue is that shortest-path algorithms entail a sequential spreading of information along contours. In theory, this could be accomplished by a message-passing algorithm employing long-range horizontal connections in early visual cortex, but this form of lateral propagation appears to be too slow to account for the rapid emergence of border ownership signals that depend upon this grouping computation (Craft et al. 2007).

Finally, it has been shown that real object boundaries are not in fact strictly Markov (Ren et al. 2008), signaling that more global statistical properties of shape may be important in distinguishing correct contours.

## 3.5. Global Cues for Contour Grouping

What global cues for contour grouping might the brain use to overcome the limitations of the Markov assumption? Global cues of convexity, symmetry, and parallelism have been proposed in the computational literature (e.g., Lowe 1985, Jacobs 1996, Stahl & Wang 2008); however, there has been relatively little psychophysical or physiological validation of their role as grouping cues in human vision (although see Liu et al. 1999, Feldman 2007, Machilsen et al. 2009).

More studied has been the cue of contour closure. The classical Gestalt demonstration [ ][ ][ ] is often taken to demonstrate a principle of closure overcoming the principle of proximity to determine the perceptual organization of shape (Koffka 1935). Indeed, shape discrimination experiments that vary the degree of closure of the contours defining shapes support this view: Small changes in the contours that lead to large changes in closure are found to yield large changes in shape discriminability (Elder & Zucker 1993, 1994, 1998a). Moreover, the task seems to remain quite difficult when good continuation is restored without closure, suggesting that the property of closure contributes something above and beyond good continuation cues (Elder 2014). In support of this, Garrigan (2012) has shown that contour shape is more effectively encoded in memory when the contour is closed than when it is open.

While the role of closure in shape discrimination is established, its role in the detectability of shapes in clutter is not. Adapting the psychophysical method of Field et al. (1993) (**Figure 4*b***), Kovacs & Julesz (1993) found superior detection performance for closed, roughly circular contours, compared to open curvilinear controls. However, the good continuation cues between the open and closed stimuli were not perfectly equated in these experiments—the open controls contained many inflections in curvature, whereas the closed contours were nearly circular. These differences are important, as it has been shown that changes in curvature sign can greatly reduce the detectability of contours (Pettet 1999).

Tversky et al. (2004) addressed this question by comparing detection for circular and S-shaped contours that match exactly in curvature, save for a single inflection point. They found a small advantage for closed contours but argued that this advantage could potentially be due to probability summation over smaller groups of elements.

Drawing solid conclusions from this literature is challenging for two reasons. First, it is difficult to vary the global grouping cue of closure while clamping all other variables. For example, in the Tversky et al. (2004) experiments, it is possible that the inflection that is present in the S-shaped stimulus but not in the circular stimulus makes the S stimulus either more salient or, alternatively, harder to segment. Second, the stimuli employed in these experiments are not naturalistic, making it difficult to generalize results to scenarios we encounter in our normal visual life.

Recently, Elder et al. (2018) introduced two innovations in the psychophysical method of Field et al. (1993) to address these issues. First, instead of employing simple geometric stimuli such as circles or Ss, they employed fragmented outlines of animal shapes, natural stimuli to which we know the human visual system is acutely tuned (see Section 1). Second, they devised control stimuli they called metamers that are designed to provide exactly the same local proximity and good continuation cues as these animal stimuli. Two versions of these metamer stimuli were devised: one open and one closed.

Elder and colleagues found that the closed metamers were more easily detected than the open metamers and that the animal shapes were more easily detected than either metamer. These results demonstrate a role not only for closure but also for other global shape regularities in the grouping of contours in complex, cluttered scenes.

What computational mechanisms are involved in harnessing global cues such as closure to solve these complex grouping problems? Some models for global contour extraction based on the Markov assumption discussed in Section 3.3 incorporate closure by explicitly searching for closed cycles of local elements (Elder & Zucker 1996a, Mahamud et al. 1999, Elder et al. 2003, Stahl & Wang 2008, Levinshtein et al. 2010, Movahedi & Elder 2013). It should be noted, however, that the statistical structure of a cycle is more complex than a Markov chain, as closure induces global statistical dependencies between the local elements, and hence there is a mismatch between the first-order Markov model used by these methods and the goal of recovering closed contours. A side effect of this is that restricting the search to closed contours breaks the optimal substructure property, rendering polynomial algorithms like dynamic programming invalid. An alternative is to apply greedy search techniques that monotonically extend current contour hypotheses by selecting the most probable continuations, but such approximate methods do not generally succeed in recovering optimal contours (Elder & Goldberg 2001, Elder et al. 2003). These limitations suggest that a more profound revision of the standard feedforward model for the perceptual organization of shape may be required.

## 4. RECURRENCY IN SHAPE PROCESSING

The impressive speed of shape processing measured under some experimental conditions (Section 1) suggests that a feedforward pass through the object pathway is sometimes sufficient for object detection. However, performance on an animal detection task does continue to improve as stimulus duration is increased up to at least 120 ms (Elder & Velisavljević 2009), and while reaction times can be as fast as 300 ms after stimulus onset, they average around 500 ms and are positively skewed (Thorpe et al. 1996). This leaves ample time for involvement of the massive feedback connections from higher to lower visual areas that are known to exist in ventral stream (Ungerleider 1995).

What role could this feedback play in shape perception? One possibility is that it overcomes limitations of the Markov model for contour grouping by using global regularities of object shape to guide local grouping—this kind of global-to-local grouping algorithm has proven effective in computer vision (Estrada & Elder 2006). As reviewed in Sections 2.2 and 2.3 above, global shape and object properties are coded in higher areas of the ventral stream, in area V4 and IT, for example (**Figure 2*b***). One hypothesis is that these higher visual areas receive local, fragmented grouping hypotheses from early visual areas and then provide feedback to these early visual areas to support hypotheses that are consistent with global evidence while suppressing hypotheses that are inconsistent with global evidence. This kind of model extends the function of V1 beyond a transient initial stage of local processing to include a more sustained role in representing high-resolution geometric detail (Cavanagh 1991, Lee & Mumford 2003, Yuille & Kersten 2006). One would thus expect to find, under this model, a trace of this global influence in the later phase of response of the visual neurons in early visual cortex coding local contour geometry.

### 4.1. Physiological Evidence

V1 neurons can adapt dynamically to different nonlocal shape properties depending on the immediate task (McManus et al. 2011, Gilbert & Li 2013, Piech et al. 2013, Ramalingam et al. 2013), and this is thought to be the result of feedback from higher shape-selective visual areas. In more recent physiological experiments involving simultaneous recording from neurons in macaque areas V1 and V4, selectivity for global contour information is first seen in V4, emerging in V1 roughly 40 ms later (Chen et al. 2014). These results point to a V1–V4 recurrent network underlying the global perceptual integration of contours.

Human brain studies using event-related potentials (ERPs) (Halgren et al. 2003) and MEG (Yoshino et al. 2006) point to an analogous cooperative recurrent computation for illusory shape formation involving early and late visual areas in ventral stream. Similar feedback mechanisms have been proposed to account for the perceptual assignment of figure/ground relationships (Lamme 1995) and border ownership (Zhou et al. 2000).

Brincat & Connor (2006) report that linear selectivity for contour part combinations emerges rapidly in posterior IT neurons, peaking at 122 ms after stimulus onset, consistent with a feedforward computation through ventral stream. Nonlinear (configural) selectivity, however, develops more gradually, peaking at 184 ms poststimulus. This delay of 62 ms suggests the possible involvement of recurrent processing in the formation of more nonlinear, selective, and sparse tuning to configurations of contour features.

## 4.2. Psychophysical Evidence: Disrupting Feedback

Backward-masking studies have shown that performance on a simple shape discrimination task can be hindered by the presentation of a masking stimulus following the shape target, and the effectiveness of the mask is found to peak between roughly 50 and 100 ms after stimulus onset (Enns & Di Lollo 2000, Habak et al. 2006). This is consistent with the recurrent shape-processing hypothesis: If irrelevant or contradictory information arrives at early visual cortex from the LGN just as relevant feedback cues are arriving from higher visual areas, we would expect a performance decrement.

Another way of masking a shape target is to employ transcranial magnetic stimulation (TMS). Applied to early visual areas, TMS blocks the perception of briefly presented stimuli when applied 30 ms prior to stimulus onset and up to 50 ms after stimulus onset (Corthout et al. 1999). Intriguingly, TMS has also been found effective in blocking stimulus perception when applied to early visual cortex during a second time window, 80–120 ms after stimulus onset (Walsh & Cowey 1998), suggesting the possible involvement of a later feedback process. Moreover, backward-masking TMS is most disruptive to illusory contour formation when applied earlier in LOC and later in V1 (**Figure 5**) (Wokke et al. 2013). It is hard to explain this reversal in sequencing without invoking feedback.

## 4.3. Psychophysical Evidence: Enhancing Feedback

One limitation of backward-masking and TMS studies is that the transient onset of a strong visual mask or the repeated application of a focused magnetic field may generally disrupt processing and reduce performance on a range of tasks by introducing noise or distracting attention, and thus an observed deficit does not demonstrate that the feedback process is specific to the visual task under study (shape processing, in our case). To address this issue, Drewes et al. (2016) developed a novel repetition methodology based on the animal and metamer stimuli of Elder et al. (2018). In this method, a fragmented shape stimulus is displayed briefly in random dynamic noise, and the observer's task is to distinguish whether the shape is an animal or a metamer. Two conditions were compared: a single-presentation condition, in which the target shape is displayed once, and a dual-presentation condition, in which the target shape is presented twice, with a variable delay between presentations.

They posited that if the inter-stimulus delay between the two presentations matched the round-trip delay between early visual cortex and higher shape areas, the reinforcing information from the second stimulus presentation would arrive in early visual cortex from the LGN at the same time feedback information from the first stimulus presentation was arriving from higher visual areas. This confluence of feedforward and feedback information might be expected to result in a peak in performance relative to shorter or longer delays, for which the feedforward and feedback of relevant stimulus information would not be synchronized.

The dual presentation was found to substantially improve performance on the task. Importantly, the degree of improvement was found to depend upon the inter-stimulus interval, peaking at 60-ms

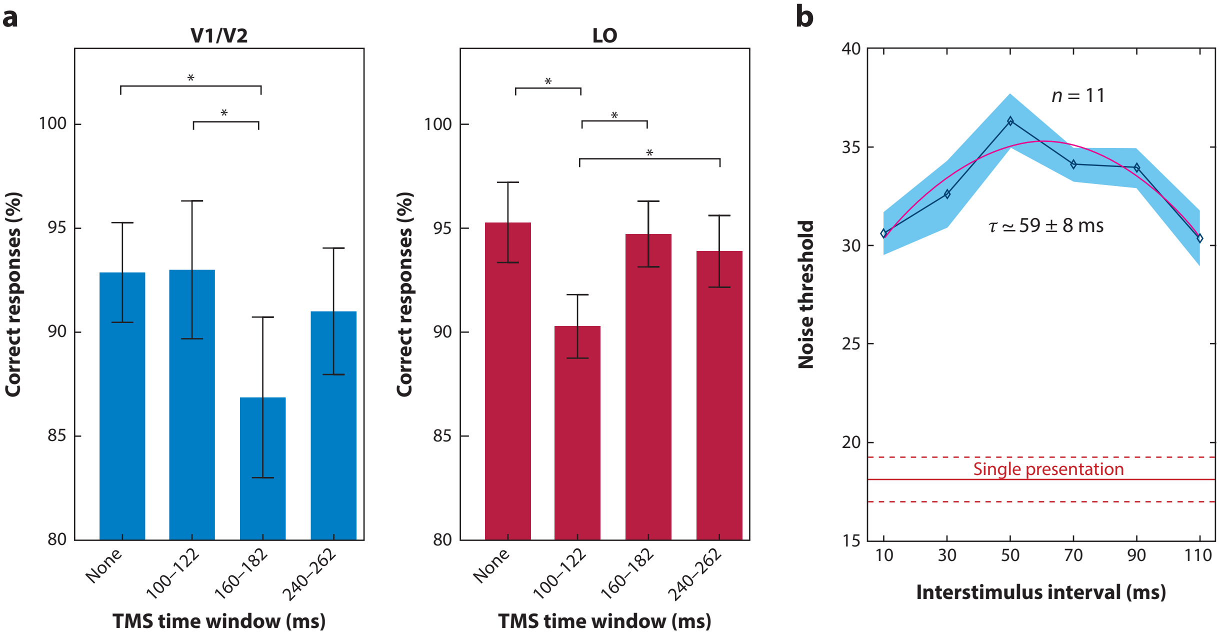

**Figure 5**

Evidence for the role of feedback in the perceptual computation of shape from contour. (*a*) To explore the role of feedback in the formation of illusory contours, Wokke et al. (2013) applied transcranial magnetic stimulation (TMS) either to the occipital pole to disrupt processing in early visual cortex or over the lateral occipital lobe to disrupt processing in the lateral occipital cortex (LOC), while observers performed an illusory shape discrimination task. The plots show mean and standard error of the mean (SEM) percentage correct as a function of when and where TMS was applied. In LOC, TMS disrupted processing when the pulse occurred 100–122 ms after stimulus onset, whereas in V1/V2, processing was disrupted when the pulse was applied later, 160–182 ms after stimulus onset (**Figure 5*a***). This is strongly suggestive of a feedback process in the formation of global illusory shape percepts from local inducing contour fragments, with a one-way feedback time constant (LOC to V1/V2) of 40–80 ms. (Asterisk indicates a statistically significant difference at the .05 level.) Panel *a* adapted from Wokke et al. (2013). (*b*) Performance on the shape discrimination task of Drewes et al. (2016), in terms of threshold noise (number of distractor elements) at 75% correct. Red represents mean and SEM of the single-presentation condition; blue represents mean and SEM of the dual-presentation condition; and magenta represents a raised Gaussian fit to the dual-presentation data, which indicates a peak in performance at an inter-stimulus interval of $\tau = 59 \pm 8$ ms.

delay. By the above logic, this suggests a fast recurrent circuit underlying human shape perception, with a round-trip time constant of 60 ms.

In summary, recent anatomical, physiological, and behavioral evidence suggest that cortical feedback of global shape information from higher visual areas to lower visual areas may play an important role in the perception of shape from contour. A computation that is progressive, generating quick approximate solutions that are refined over time, would have adaptive advantage in supporting rapid real-time vision. At this point in the review, however, we have still not discussed what the nature of this global shape information is. What is the representation? The remainder of the review focuses on this question.

## 5. DISCRIMINATIVE VERSUS GENERATIVE REPRESENTATIONS OF SHAPE

In statistical modeling, the terms discriminative and generative have specific mathematical meanings. Here, I use them in a more descriptive way. A discriminative representation of shape is a representation that has, at most, a weak relation to how the shape was created and is generally insufficient to allow the shape to be regenerated. It is usually designed to support a specific task—for example, to discriminate between classes of shapes (e.g., dogs versus cats). The focus is on delivering the right semantic label, rather than describing the stimulus.

A generative representation, in contrast, attempts to recover the original process that created the shape and should be sufficient to recreate a good approximation of the shape. As a result, it can potentially support a large range of tasks (e.g., recognition, grasping, similarity judgments).

With the recent development of high-performing, discriminative DNNs for object recognition, there is a growing interest in discriminative models for object and scene recognition. However, it seems unlikely that discriminative training on a narrow task could lead to the richness of our experience of shape or the diversity of judgments we may make about shape. In a generative model, in contrast, the components and transformations (e.g., parts, symmetries, growth, twists) of the representation have meanings that seem to correspond to our phenomenological appreciation for shape, and that may be useful for a variety of judgments.

Perhaps most importantly, generative representations can be used to guide the process of perceptual organization. In particular, a generative model of global shape can potentially be used to guide the perceptual organization of contours through feedback from higher visual areas to early visual cortex (Section 4). For these reasons, I review both discriminative and generative representations of shape but pay particular attention to generative representations, or representations with the potential to be generative, as more plausible models of human shape perception.

## 6. DISCRIMINATIVE REPRESENTATIONS OF 2D SHAPE

Discriminative models for 2D object detection generally represent an object as a constellation of localized features, which may be binary patterns of thresholded luminance or color (Ojala et al. 1996), patterns of gradient histograms (HOG) (Dalal & Triggs 2005), or local gradient histograms at keypoints of the image [scale invariant feature transform (SIFT) (Lowe 2004)]. This type of model can be useful for some computer vision tasks involving modest variations in pose, but performance tends to degrade with larger variations (Pinto et al. 2011).

## 7. GENERATIVE REPRESENTATIONS OF 2D SHAPE

For some detection and recognition tasks, discriminative models can be effective. However, many other tasks call for a more complete generative model of shape. Examples include (*a*) perceptual organization, recognition, and tracking in cluttered scenes, where shapes must be distinguished not just from each other, but from so-called phantom shapes formed by conjunctions of features from multiple objects (Cavanagh 1991); (*b*) modeling of shape articulation, growth, and deformation; and (*c*) modeling of shape similarity.

Elder et al. (2013) articulated a set of criteria for a generative shape representation that forms a useful basis for comparing the candidate shape representations we consider in the remainder of this review (see the sidebar titled Desirable Criteria for a Generative Model of 2D Shape).

### 7.1. Linear Basis Representations

I begin by describing a family of 2D shape representations that model a closed contour as a linear superposition of simpler basis functions.

**7.1.1. Radial frequency patterns.** A set of 2D shape stimuli known as radial frequency patterns (Wilkinson et al. 1998) have been used extensively in psychophysical studies exploring 2D

## DESIRABLE CRITERIA FOR A GENERATIVE MODEL OF 2D SHAPE

Completeness: The framework can represent all shapes.
Closure: The set of valid shapes is closed under the generative model. In other words, only valid shapes can be generated.
Composition: Complex shapes are generated by combining simpler components.
Sparsity: Good approximations of shape can be generated with relatively few components.
Progression: Approximations can be improved by incorporating more components.
Locality: Components are localized in space.
Scaling: Components are tuned to specific scales and are self-similar over scale.
Region and contour: Components can capture both region and contour properties in a natural way.

shape perception. Each basis function is a closed curve centered at the origin and defined in polar coordinates $(r, \theta)$, where the radial coordinate $r$ is a sinusoidal function of polar angle $\theta : r(\theta) = r_0 \cos(k\theta + \theta_0)$. Each function sweeps out a smooth, closed curve. The parameter $k = 0, 1, 2, \ldots$ controls the frequency and symmetry of the curve: As $k$ increases, the number of undulations and hence order of rotational symmetry of the basis function increases. More complex, asymmetric yet smooth shapes can be created simply by adding basis functions of different frequencies, phases, and amplitudes.

Shepard & Cermak (1973) used a suprathreshold variation on these stimuli to probe human shape categorization, finding that judgments could be well predicted from simple similarity metrics computed in radial frequency space. Threshold detection and discrimination experiments reveal high sensitivity to these patterns, pointing to intermediate representations of shape (in area V4, for example) specialized for roughly circular patterns (Wilkinson et al. 1998, Bell & Badcock 2008).

While radial frequency patterns are an interesting class of shape, shapes that are not roughly circular cannot be represented in this way, and this excludes many important shapes (e.g., animals, human bodies). Thus radial frequency patterns fail to satisfy the completeness criterion articulated in Section 7.

#### 7.1.2. Fourier descriptors.

The Fourier descriptor (FD) representation of shape (Granlund 1972) can be seen as a generalization of radial frequency patterns that can be applied to arbitrary 2D shapes. Consider a discrete polygonal representation of such a shape in which each vertex $(x, y)$ of the polygon is represented as a coordinate $x + yi$ in the complex plane.[3] By taking the Fourier transform of the complex vector representing these vertices, we obtain the FD representation, which represents the complex amplitude of the shape at each frequency over the index space of the vector. Importantly, one can capture the main features of a natural shape using only a small number of the lowest frequency components, limiting the dimensionality of the stimulus to a manageable level (**Figure 6**).

Physiological experiments in macaque have revealed that roughly half of the visually responsive neurons in IT cortex are tuned to FD frequency, with a surprisingly broad distribution of peak frequency tuning (roughly uniform from 2 to 32 cycles per shape) (Schwartz et al. 1983). However, subsequent work revealed that IT neural response to compound FD stimuli does not

[3] An alternative FD representation that applies to equilateral polygons operates on the vector of turning angles describing the shape (Zahn & Roskies 1972).

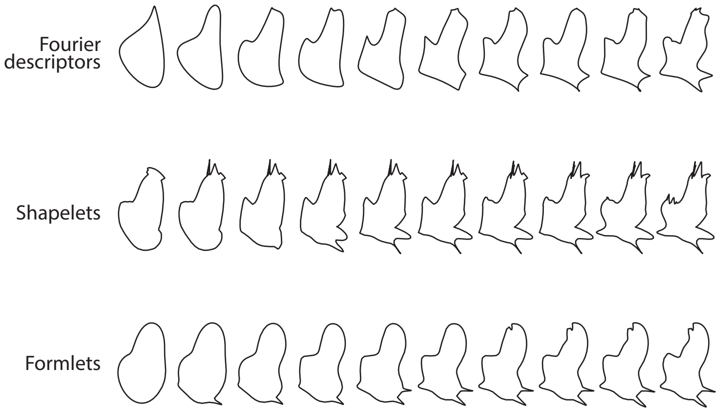

**Figure 6**

Three alternative progressive representations of two-dimensional shape, using from 1 to 10 shape components.

obey superposition, suggesting that the FD representation does not in fact form a good account of IT shape coding (Albright & Gross 1990). Moreover, Brincat & Connor (2004) report that they see no evidence for periodic selectivity for curvature extrema that is predicted by FD tuning. It thus seems that while IT neurons tend to be responsive to FD stimuli, they are not specifically selective for individual FD components.

Psychophysical experiments suggest that, at least for simple shapes composed of a small number of FD harmonics, human judgments of shape similarity can be predicted by a Euclidean metric in FD frequency-phase space (Cortese & Dyre 1996). More recently, Wilder et al. (2018) adapted a linear systems identification method called classification image analysis to study the discrimination of animal shapes in the FD domain. They found that estimated linear templates are biased away from the ideal, overly weighting lower frequencies. This lowpass bias suggests that higher frequency shape processing relies on nonlinear mechanisms.

The Achilles' heel of the FD representation is that every coefficient represents a global property of the curve, violating the locality criterion of Section 7. As a result, spatially localized perturbations (e.g., occlusions, articulations of an object part) that occur commonly in our visual environment impact all coefficients of the representation, complicating recognition.

**7.1.3. Shapelets.** The shapelet model of Dubinskiy & Zhu (2003) addresses this issue. The theory is based upon the representation of a shape by a summation of component shapelets, which are Gaussian-windowed Fourier descriptors (i.e., Gabor functions over arclength), distributed over a range of arclength locations and scales. Unlike the Fourier descriptor representation, the shapelet family is overcomplete, which means that the representation is not unique unless an additional rule for selecting shapelets is imposed. Dubinskiy & Zhu employed matching pursuit (Mallat & Zhang 1993), which selects a sequence of shapelets by performing iterative gradient descent on the approximation error (**Figure 6**).

The shapelet model has many positive features. Components are localized, and scale is made explicit in a natural way. However, like all contour-based methods, the shapelet theory does not explicitly capture regional properties of shape, violating the region and contour criterion of Section 7. Perhaps most crucially, the model does not respect the topology of object boundaries: Progressive representations of a target shape may introduce topological errors (**Figure 7**), and sampling from the model will in general yield nonsimple (i.e., self-intersecting) curves. This violates the closure criterion of Section 7.

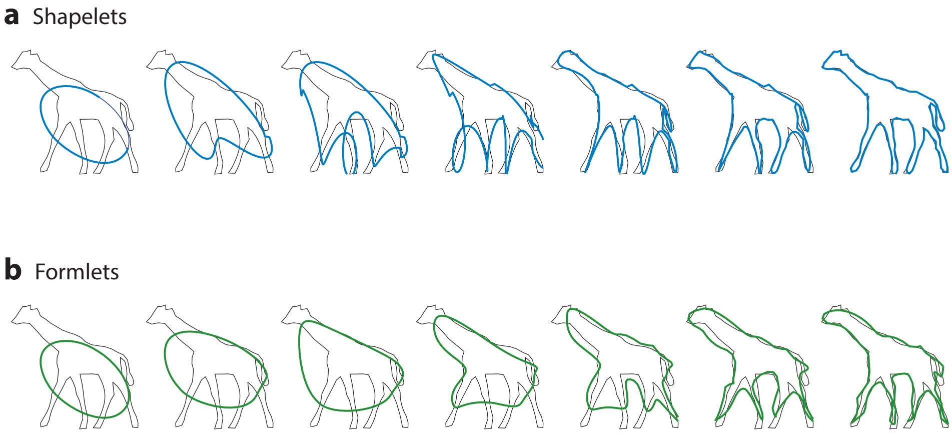

**Figure 7**

Shapelet and formlet representations for an example animal shape. (*Left column*) Initial ellipse approximation. (*Subsequent columns*) Progressive representation with 1, 2, 4, 8, and 16 components. While the shapelet code converges more quickly, topological errors are introduced.

## 7.2. Symmetry Representations

The Gestaltists identified symmetry as a factor of "good shape" and a determinant of figure/ground organization (Koffka 1935), and symmetry has been used in numerous contour grouping algorithms (e.g., Stahl & Wang 2008). Kanizsa (1979), however, has observed that symmetry appears easily overruled when pitted against principles of good continuation and convexity.

Despite this relatively long history, definitive psychophysical evidence for the role of symmetry and parallelism in contour grouping has come relatively recently. Feldman (2007) showed that comparison of features lying on pairs of line segments is significantly faster if the segments are parallel or mirror symmetric, suggesting a fast grouping of the segments based upon these cues, and Machilsen et al. (2009) have demonstrated enhanced detectability of bilaterally symmetric versus asymmetric closed forms, suggesting a role for global symmetry processing in contour grouping. Physiologically, it is known that bilaterally symmetric patterns differentially activate human extrastriate visual areas V3, V4, V7, and LOC and homologous areas in macaque cortex (Sasaki 2007).

The importance of symmetry in shape perception motivates the symmetry axis (sometimes called medial axis) representation, in which a planar shape is represented by its skeleton and an associated distance function (Blum 1973, Kimia & Siddiqi 1995, Feldman & Singh 2006). This approach has the advantage of making perceptually salient shape symmetries explicit and capturing regional properties of shape that are not directly represented by contour methods. Leyton (1988) related symmetry axis descriptions to causal deformation processes acting upon prototype shapes. In this view, symmetry axes, terminating at curvature extrema on the boundary, are understood as records of these deformation processes. Siddiqi et al. (1999) characterized the topological structure of the symmetry axis description as acyclic shock graphs and developed efficient methods for measuring the similarity of shock graphs to support 2D shape recognition. Recent work (Levinshtein et al. 2013) fuses the symmetry axis representation with local region groupings called superpixels that are relatively stable and efficient to compute.

Despite the many appealing features of symmetry axis and shock-graph representations, these methods, in general, are not sparse. In fact, the description of each shape typically requires more storage, and little emphasis has been placed on making symmetry axis representations generative (Mumford 1991). An exception is the work of Trinh & Kimia (2007), which explored generative and sparse models based upon shock graphs. However, the constraints required to enforce the closure property (i.e., topological constraints) are fairly complex, and the full potential of the theory has yet to be explored.

Recognizing the merits and limitations of both contour-based and symmetry-based approaches, Zhu (1999) developed a Markov random field model for natural 2D shape, employing a neighborhood structure that can directly encode both contour-based and region-based Gestalt principles. The theory is promising in many respects. It is generative, providing an explicit probabilistic model, and it captures both region and contour properties. It is not sparse, however, and because the underlying graph is lifted from the image plane, there is nothing in the model that encodes the topological constraint that the boundary be simple (i.e., nonintersecting). Instead, when sampling from the model, a so-called firewall is employed to prevent intersections. Again, this is inefficient, and it also creates a disconnect between the generative variables encoding the model and the sampling distribution.

### 7.3. Growth and Deformation Representations

A third class of theory considers shape as a process of transformation or growth (Thompson 1917, Leyton 1989, Grenander et al. 2007, Elder et al. 2013). These theories have a natural generative expression that can be used to support inference with noisy or incomplete visual data and are better able to capture topological properties of objects (Elder et al. 2013).

Thompson (1917) considered specific classes of global coordinate transformations to model the relationship between the shapes of different animal species. While effective for comparing similar shapes, by their nature these global transformations cannot completely describe a shape and cannot model parts.

These limitations are addressed by formlet theory (Grenander et al. 2007, Oleskiw et al. 2010, Elder et al. 2013), in which a shape is modeled as the outcome of a deformation process applied to a simple embryonic shape (e.g., an ellipse). This deformation process is decomposed into a series of simpler localized deformations called formlets, which range widely over spatial scale and can be isotropic or oriented (Yakubovich & Elder 2014), allowing them to express highly complex shapes. As for shapelets, a modified version of matching pursuit (Mallat & Zhang 1993) is employed in order to construct a formlet representation of a given target shape. Constraining the gain of each formlet guarantees that the deformation is diffeomorphic, preserving the topology of the shape (**Figure 7**). Formlets have been found to substantially outperform shapelets on the task of shape completion (Elder et al. 2013).

### 7.4. Neural Representation

Two-dimensional shape representations based upon linear basis functions, symmetry, and growth all have demonstrated advantages. While these remain active areas of research in behavioral science, there is a need for more explicit testing of these representations as models for the neural representation of shape in primate cortex.

## 8. FROM 2D TO 3D

The leap from 2D to 3D object recognition is profound as a result of out-of-plane pose variation: As the object rotates in 3D, different parts of it are seen. Historically, two major theories for how the brain handles this problem have been advanced. In the 3D model–based theory, the brain uses 2D features of the image projection of an object to estimate the 3D geometry and pose of the object, allowing comparison with stored 3D object models (e.g., Biederman 1987). In the alternative view-based theory (e.g., Bülthoff & Edelman 1992), the brain maintains models for the

2D appearance of remembered objects over a sampling of 3D poses. The 2D image appearance of an object segmented from a new image can then be compared to stored 2D views to identify both the identity and pose of the new object. The latter theory is supported by psychophysical experiments in which observers are asked to identify a 3D object seen from one viewpoint that had previously been seen from a different viewpoint. Typically, both the accuracy and speed of response are found to decline as the angular difference in viewpoint is increased.

For decades, these experiments have been taken as support for the view-based theory of object recognition. However, Erdogan & Jacobs (2017) have recently demonstrated that a 3D model–based theory of object representation would also be expected to generate viewpoint-dependent recognition performance, once sensing noise and observer uncertainty about object structure are taken into account.

At the physiological level, most neurons in area IT show some specificity to object view (Logothetis et al. 1995), although some neurons appear to be view invariant (Booth & Rolls 1998). On the basis of this and other evidence, Riesenhuber & Poggio (2002) have proposed a hybrid model in which a population of view-tuned and component-tuned neurons in anterior IT contribute directly to object recognition processes in higher areas but also are combined within anterior IT to form 3D object–tuned units that are fully invariant to object pose.

# 9. DISCRIMINATIVE REPRESENTATIONS OF 3D SHAPE

Pose and illumination variation and scene clutter make 3D object recognition hard. These so-called nuisance variables conspire to warp and entangle the manifolds defining the appearance of different objects in the observation space (DiCarlo & Cox 2007). Untangling these manifolds to allow linear decoding that is invariant to object pose, illumination, and other scene factors requires a nonlinear transformation of the input.

Early discriminative 2D shape and object recognition models (Section 6) had a two-stage structure: extract a relatively simple handcrafted feature vector from the image, and then train a classifier (e.g., a support vector machine) to discriminate between objects based on this feature vector. These algorithms can handle in-plane and small out-of-plane pose variations but break down in the face of larger 3D pose changes. One approach to solving this problem is to devise more complex feature vectors that are more invariant to such transformations.

## 9.1. HMAX

The bioinspired HMAX model for object recognition (Riesenhuber & Poggio 1999; Serre et al. 2007a,b) builds more complex features through a hierarchy of alternating layers of local linear filtering and max or softmax pooling over orientation and local space and scale. Linear filter weights were hard coded in initial versions of the model, but learned using an unsupervised Hebbian-like mechanism in later versions. HMAX demonstrates greater robustness to variations in object position, scale, and pose than competing computer vision representations such as SIFT and HOG (Pinto et al. 2011) and generates a pattern of performance similar to human observers on an animal detection task (Serre et al. 2007b). However, it fails to replicate the clustering of images by object category seen in IT (Kriegeskorte et al. 2008b).

## 9.2. Deep Network Representations

Since 2012, DNNs have been beating other computer vision algorithms on benchmarks for object recognition tasks, and the most recent models [e.g., ResNet (He et al. 2016)] surpass human performance on certain benchmarks. Given this strong performance, it is natural to compare the representations computed by DNNs trained on object recognition tasks with the representations computed by the primate ventral stream.

Yamins et al. (2014) trained a mixture of four-layer CNNs from labeled training data, finding that as training progressed and performance on an object classification task improved, so did its ability to predict IT neural responses. Ultimately, the model was able to explain a much greater proportion of the variance in IT neural response than prior models, including HMAX.

Khaligh-Razavi & Kriegeskorte (2014) compared a battery of computational models to neurons in monkey and human IT. They found that a DNN model [AlexNet (Krizhevsky et al. 2012)] trained on ImageNet (Deng et al. 2009) correlated much better with IT responses than simpler computer vision models, earlier models of ventral visual pathway that are trained without supervision (i.e., without object category labels) such as HMAX, or weakly supervised models trained with a small number of category-labeled images.

Seibert et al. (2016) constructed a four-layer AlexNet-like DNN model with three convolutional layers and a single fully connected layer and trained it on ImageNet. They compared the output of the three convolutional layers to blood-oxygen-level-dependent (BOLD) activation in human cortex using representational dissimilarity matrices (RDMs) (Kriegeskorte et al. 2008a). They found an interesting correspondence between the selectivities of each convolutional layer in their DNN model and BOLD response in V1, V2, hV4 (human homolog of V4), and LOC. While layers 1 and 2 of the DNN were equally predictive of BOLD response in V1, layer 2 became more predictive of BOLD response in V2 and layer 3 became most predictive for hV4 and LOC. As the DNN model was optimized to perform better on the ImageNet challenge, the model's correlation with BOLD response in these visual areas increased.

Since these DNN models are all trained and tested on full color images that afford not only shape cues but also color and texture cues that may be informative about object category, it is unclear to what degree they are using shape information when inferring object category. However, a recent study (Kubilius et al. 2016) has shown that DNNs trained on full color images of objects still perform well above chance when given black and white silhouette stimuli that contain 2D shape cues alone. This shows that even when trained on full color images, these networks are learning to form shape representations that, to some degree at least, can be applied independent of color, texture, and shading cues.

Kubilius et al. also compared model representations with human judgments of shape similarity, employing a set of synthetic, shaded 3D objects that differ along distinct shape dimensions. While human shape similarity judgments tend to be driven by the smoothness of the objects (spiky/smoothie/cubie), shallow representations and early layers of the artificial nets are driven more by the overall footprint of the shapes (vertical/square/horizontal). Interestingly, higher layers of the DNNs are driven more by smoothness, matching human perception. This supports the hypothesis that shape representations in higher layers of deep nets are related to representations in higher areas of the human object pathway.

They also found that while some shallow representations and early layers of DNNs tend to code metric shape differences, higher layers tend to code nonaccidental shape differences (i.e., qualitative shape properties that are invariant to projection). This matches human perceptual similarity judgments, and also shape selectivity in IT, and lends some support to structural theories of object recognition based upon the inference of 3D shape primitives from nonaccidental image features (Biederman 1987).

However, in a more recent study, N. Baker, H. Lu, G. Erlikhman, and P.J. Kellman (unpublished manuscript) found that human and DNN measures of shape similarity sometimes diverge for very simple shape stimuli. In particular, while humans tend to make judgments based on global shape, DNNs seem most sensitive to local features. Consistent with this finding is Rajalingham et al.'s (2018) recent report that while current DNNs can accurately predict object-level confusions, they still fall short in predicting behavioral performance for individual images. Whether this is due to fundamental architectural properties of the networks or to the nature of the images and the task on which they are trained remains to be determined.

## 10. GENERATIVE (STRUCTURAL) REPRESENTATIONS OF 3D SHAPE

While at present, DNNs generate our best predictions of IT response, they are not a perfect match to the physiology. For one thing, the primate object pathway is composed of a relatively small number of visual areas (retina, LGN, V1, V2, V4, IT), while state-of-the-art DNNs are much deeper—the best version of ResNet, for example, consists of 152 layers (He et al. 2016). One possible solution to this puzzle is that these DNNs compute complex features via a strictly feedforward cascade of many layers, while the human ventral stream relies also upon recurrent processing within and between visual areas to form detailed shape representations (Section 4). This recurrent computation may reflect a generative model in which higher-level representations are required to account for lower-level observations. Such a generative shape representation has the potential to support a broader array of tasks and queries (e.g., to describe a shape or to reason about its part structure).

Generative models for 3D shape perception have a long history. Binford & Tenenbaum (1973) argued for a 3D volumetric part-based shape representation based on generalized cones, in which a 3D space curve defines the axis of the part and a smoothly varying 2D planar radial curve centered at points along the axis sweeps out the part surface.

Marr & Nishihara (1978) argued for a multi-scale structural theory for 3D shape representation, as opposed to a view-based representation, primarily on the grounds that the cost of storing all required views of all objects would be prohibitive. Larger-scale components of the representation remain relatively invariant to small changes and thus provide the stability needed for recognition, while finer details are encoded in smaller-scale components. Part boundaries are identified from the image by segmenting the 2D object projection at deep concavities detected at concave points on the bounding contour.

Biederman's recognition-by-components (RBC) theory follows in this tradition (Biederman 1987). Through a consideration of qualitative Gestalt properties of object contours (e.g., curvature, collinearity, symmetry, parallelism, cotermination), Biederman identified a small family of generalized cone components (geons) capable of describing a broad diversity of 3D objects. A critical part of his theory is that while perceptual estimates of quantitative metric shape properties are highly variable (Miller 1956), the preservation of qualitative Gestalt properties despite occlusion and variations in object pose results in an invariant 3D shape representation, leading to good object recognition performance.

In psychophysical experiments, Biederman found that line drawings of familiar objects composed of geon components were rapidly recognized by human observers even when several components were missing. Objects were also rapidly recognized when convex sections of the contours were erased. However, recognition performance degraded substantially when the contour was erased at regions of concavity or at junctions, consistent with the use of these features for identifying part boundaries and component structure.

While Biederman focused on testing the RBC approach psychophysically as a model for human object recognition, others developed it more fully as a computational theory that could be used in computer vision systems. Dickinson et al. (1992) fleshed out an algorithm for identifying 3D object parts from an image that employs Gestalt principles to construct a hierarchy that progresses from edges to contour fragments, closed contours, and finally aspects, each of which is a connected configuration of closed contours that topologically identifies a particular generic view of a 3D, volumetric object part.

In the last quarter century, computer vision research on structural (parts-based) shape representations has largely given way to discriminative feature-based machine learning and more recently deep learning approaches trained on large labeled data sets. However, it has also been recognized that improved feature detectors, better contour grouping and matching algorithms, and more powerful probabilistic graphical models can be leveraged to achieve more reliable object detection and pose estimation within both 2D (view-based) (Bergholdt et al. 2010) and 3D (Sala & Dickinson 2015) structural frameworks.

There is also some evidence that 3D structural models may better account for behavioral data than discriminative models, including recent DNN models, even if the latter currently serve as our best predictors of neural response in higher object areas of monkey and human. Erdogan & Jacobs (2017) recently found that a generative, Bayesian structural 3D object model provides better predictions of human judgments of 3D object shape similarity than discriminative models, including DNNs trained on ImageNet.

## 11. CONCLUDING REMARKS

Shape perception is central to our visual experience, and while decades of research have led to many insights, there is much we still do not know about how shape information is rapidly computed from complex natural images. A feedforward view in which representation progresses from local points to edges, local curvature and finally global shape roughly matches the hierarchical structure of the object pathway in primate visual cortex, and the success of feedforward DNNs in predicting both physiological and behavioral responses to object and shape stimuli has reinforced this standard model. However, recent evidence reveals that DNN models fail to capture important behavioral aspects of shape perception. This may be related to behavioral and physiological evidence that global properties of shape such as closure and symmetry influence the grouping of contours in cluttered scenes, possibly via a fast recurrent circuit connecting higher to lower shape-selective visual areas.

While selectivity of neurons in early visual areas for local geometric attributes of orientation and curvature is established, there is a notable lack of coherence in our understanding of how the brain represents global 2D and 3D shape information. Discriminative DNNs currently provide the strongest predictions for neural selectivity in higher visual areas, but the exact nature of the representations formed at intermediate layers of these models is opaque, and a purely discriminative approach fails to capture the generality and phenomenological aspects of shape perception.

On the positive front, the decades-long debate between 2D view-based and 3D object-based theories of object recognition appears to be settling toward a more nuanced and principled understanding that, in an uncertain world, a Bayesian 3D object-based model will also exhibit view dependency. Hopefully, this will spark renewed interest in generative, structural models of shape perception that will account not just for object classification performance but also for the role of shape perception in how we interact physically with objects (e.g., grasping), how we judge similarity of and analogies between objects, and how we appreciate their beauty. These generative models are likely to embody the principles of topology, composition (parts), symmetry, and deformation (growth) that appear to be central to the human perception of shape.

### SUMMARY POINTS

1. The task of extracting reliable shape information from complex scenes is computationally hard.

2. A fast, feedforward sweep through ventral stream produces partial shape representations based on orientation coding and local Gestalt principles. This representation is sufficient for simpler discriminative tasks.
3. More complete shape representations may be computed through recurrent processes that integrate local and global cues.
4. Recent, discriminative deep neural network models currently produce the best predictions of object selectivity in higher areas of the object pathway but do not account for all aspects of shape perception.
5. While there is at this time no dominant theory for how global shape information is represented by the brain, generative models based upon symmetry, local and global deformations, and compositions of prototypical parts capture important aspects of shape perception.

## FUTURE ISSUES

1. Orientation tuning in primary visual cortex can be predicted by a principle of efficient coding (Barlow 1959, Olshausen & Field 1996). Can higher-level shape properties be predicted by similar principles?
2. How does the brain discriminate between contours that signal object boundaries and other kinds of contours (e.g., shadows, reflectance changes)?
3. Can principles of symmetry, local and global deformation, and parts be fused into a unified generative model of shape perception?
4. Will research in deep generative neural networks lead to models of shape perception capable of accounting for a broader range of data?

## DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The author wishes to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada.

## LITERATURE CITED

Albright TD, Gross CG. 1990. Do inferior temporal cortex neurons encode shape by acting as Fourier descriptor filters? In *Proceedings of the International Conference on Fuzzy Logic and Neural Networks*, pp. 375–78. Izuka, Japan: Fuzzy Logic Syst. Inst.

Almazen EJ, Tal R, Qian Y, Elder JH. 2017. MCMLSD: a dynamic programming approach to line segment detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5854–62. Los Alamitos, CA: IEEE

Barlow HB. 1959. Sensory mechanisms, the reduction of redundancy, and intelligence. In *NPL Symposium on the Mechanization of Thought Process*, Vol. 10. London: HM Stationery Office

Bell J, Badcock DR. 2008. Luminance and contrast cues are integrated in global shape detection with contours. *Vis. Res.* 48:2336–44

Bergholdt M, Kappes J, Schmidt S, Schnörr C. 2010. A study of parts-based object class detection using complete graphs. *Int. J. Comput. Vis.* 87:93–117

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94:115

Binford TO, Tenenbaum JM. 1973. Computer vision. *Computer* 6:19–24

Blum H. 1973. Biological shape and visual science (part I). *J. Theor. Biol.* 38:205–87

Booth MC, Rolls ET. 1998. View-invariant representations of familiar object by neurons in the inferior temporal visual cortex. *Cereb. Cortex* 8:510–23

Brincat SL, Connor CE. 2004. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat. Neurosci.* 7:880–86

Brincat SL, Connor CE. 2006. Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron* 49:17–24

Bülthoff HH, Edelman S. 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *PNAS* 89:60–64

Campbell F, Kulikowski J. 1966. Orientation selectivity of the human visual system. *J. Physiol.* 187:437–45

Carlson E, Rasquinha R, Zhang K, Connor C. 2011. A sparse object coding scheme in area V4. *Curr. Biol.* 21:288–93

Carlson T, Tovar DA, Alink A, Kriegeskorte N. 2013. Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13(10):1

Cavanagh P. 1991. What's up in top-down processing? In *Representations of Vision: Trends and Tacit Assumptions in Vision Research*, ed. A Gorea, pp. 295–304. Cambridge, UK: Cambridge Univ. Press

Chen M, Yan Y, Gong X, Gilbert CD, Liang H, Li W. 2014. Incremental integration of global contours through interplay between visual cortical areas. *Neuron* 82:682–94

Cortese J, Dyre BP. 1996. Perceptual similarity of shapes generated from Fourier descriptors. *J. Exp. Psychol. Hum. Percept. Perform.* 22:133–43

Corthout E, Uttl B, Walsh V, Hallett M, Cowey A. 1999. Timing of activity in early visual cortex as revealed by transcranial magnetic stimulation. *NeuroReport* 10:2631–34

Craft E, Schutze H, Niebur E, von der Heydt R. 2007. A neural model of figure-ground organization. *J. Neurophysiol.* 97:4310–26

Dalal N, Triggs B. 2005. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–93. Los Alamitos, CA: IEEE Comp. Soc.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–55. Los Alamitos, CA: IEEE

Desimone R, Albright TD, Gross CG, Bruce C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4:2051–62

DiCarlo JJ, Cox DC. 2007. Untangling invariant object recognition. *Trends Cogn. Sci.* 11:333–41

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? *Neuron* 73:415–34

Dickinson SJ, Pentland AP, Rosenfeld A. 1992. From volumes to views: an approach to 3-D object recognition. *CVGIP Image Underst.* 55:130–54

do Carmo MP. 1976. *Differential Geometry of Curves and Surfaces*. Englewood Cliffs, NJ: Prentice-Hall

Dobbins A, Zucker SW, Cynader MS. 1987. Endstopping in the visual cortex as a substrate for calculating curvature. *Nature* 329:438–41

Drewes J, Goren G, Zhu W, Elder J. 2016. Recurrent processing in the formation of shape percepts. *J. Neurosci.* 36:185–92

Dubinskiy A, Zhu SC. 2003. A multi-scale generative model for animate shapes and parts. In *Proceedings: Ninth IEEE International Conference on Computer Vision*, pp. 249–56. Los Alamitos, CA: IEEE

Ehinger K, Adams WJ, Graf EW, Elder JH. 2017. Local depth edge detection in humans and deep neural networks. In *2017 IEEE International Conference on Computer Vision Workshops*, pp. 2681–89. Los Alamitos, CA: IEEE

Elder JH. 1999. Are edges incomplete? *Int. J. Comput. Vis.* 34:97–122

Elder JH. 2014. Bridging the dimensional gap: perceptual organization of contour into two-dimensional shape. In *Oxford Handbook of Perceptual Organization*, ed. J Wagemans, pp. 71–83. Oxford, UK: Oxford Univ. Press

Elder JH, Goldberg RM. 2001. Image editing in the contour domain. *IEEE Trans. Pattern Anal. Mach. Intell.* 23:291–96

Elder JH, Goldberg RM. 2002. Ecological statistics of Gestalt laws for the perceptual organization of contours. *J. Vis.* 2(4):5

Elder JH, Krupnik A, Johnston LA. 2003. Contour grouping with prior models. *IEEE Trans. Pattern Anal. Mach. Intell.* 25:661–74

Elder JH, Oleskiw TD, Yakubovich A, Peyré G. 2013. On growth and formlets: sparse multi-scale coding of planar shape. *Image Vis. Comput.* 31:1–13

Elder JH, Oleskiw TD, Fründ I. 2018. The role of global cues in the perceptual grouping of natural shapes. *J. Vis.* In press

Elder JH, Sachs AJ. 2004. Psychophysical receptive fields of edge detection mechanisms. *Vis. Res.* 44:795–813

Elder JH, Trithart S, Pintilie G, MacLean D. 2004. Rapid processing of cast and attached shadows. *Perception* 33:1319–38

Elder JH, Velisavljević L. 2009. Cue dynamics underlying rapid detection of animals in natural scenes. *J. Vis.* 9(8):787

Elder JH, Zucker SW. 1993. The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vis. Res.* 33:981–91

Elder JH, Zucker SW. 1994. A measure of closure. *Vis. Res.* 34:3361–70

Elder JH, Zucker SW. 1996a. Computing contour closure. In *Computer Vision—ECCV '96*: *4th European Conference on Computer Vision*, ed. B Buxton, R Cipolla, pp. 399–412. New York: Springer Verlag

Elder JH, Zucker SW. 1996b. Scale space localization, blur and contour-based image coding. In *1996 IEEE Computer Science Conference on Computer Vision and Pattern Recognition*, pp. 27–34. Los Alamitos, CA: IEEE Comp. Soc. Press

Elder JH, Zucker SW. 1998a. Evidence for boundary-specific grouping. *Vis. Res.* 38:143–52

Elder JH, Zucker SW. 1998b. Local scale control for edge detection and blur estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 20:699–716

Enns J, Di Lollo V. 2000. What's new in visual masking? *Trends Cogn. Sci.* 4:345–52

Erdogan G, Jacobs R. 2017. Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psych. Rev.* 124:740–61

Estrada F, Elder JH. 2006. Multi-scale contour extraction based on natural image statistics. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, ed. C Schmid, S Soatto, C Tomasi, p. 183. Los Alamitos, CA: IEEE

Feldman J. 2007. Formation of visual "objects" in the early computation of spatial relations. *Percept. Psychophys.* 69:816–27

Feldman J, Singh M. 2006. Bayesian estimation of the shape skeleton. *PNAS* 103:18014–19

Field D, Hayes A, Hess RF. 1993. Contour integration by the human visual system: evidence for a local "association field." *Vis. Res.* 33:173–93

Garrigan P. 2012. The effect of contour closure on shape recognition. *Perception* 41:221–35

Geisler WS, Perry JS, Super BJ, Gallogly DP. 2001. Edge co-occurence in natural images predicts contour grouping performance. *Vis. Res.* 41:711–24

Gilbert CD, Li W. 2013. Top-down influences on visual processing. *Nat. Rev. Neurosci.* 14:350–63

Gilbert CD, Wiesel TN. 1989. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J. Neurosci.* 9:2432–43

Granlund GH. 1972. Fourier preprocessing for hand print character recognition. *IEEE Trans. Comput.* C-21:195–201

Grenander U, Srivastava A, Saini S. 2007. A pattern-theoretic characterization of biological growth. *IEEE Trans. Med. Imaging* 26:648–59

Grill-Spector K, Kourtzi Z, Kanwisher N. 2001. The lateral occipital complex and its role in object recognition. *Vis. Res.* 41:1409–22

Grill-Spector K, Weiner KS. 2014. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* 15:536–48

Habak C, Wilkinson F, Wilson H. 2006. Dynamics of shape interaction in human vision. *Vis. Res.* 46:4305–20

Halgren E, Mendola J, Chong C, Dale A. 2003. Cortical activation to illusory shapes as measured with magnetoencephalography. *NeuroImage* 18:1001–9

Hawken MJ, Parker AJ. 1991. Spatial receptive field organization in monkey V1 and its relationship to the cone mosaic. In *Computational Models of Visual Processing*, ed. MS Landy, JA Movshon, pp. 84–93. Cambridge, MA: MIT Press

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–78. Los Alamitos, CA: IEEE

Hegde J, van Essen DC. 2000. Selectivity for complex shapes in primate visual area V2. *J. Neurosci.* 20:1–6

Hess R, Field D. 1999. Integration of contours: new insights. *Trends Cogn. Sci.* 3:480–86

Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ. 2016. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19:613–30

Hubel DH, Wiesel TN. 1968. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* 195:215–43

Huffman D. 1952. A method for the construction of minimum-redundancy codes. *Proc. IRE* 40:1098–101

Jacobs D. 1996. Robust and efficient detection of salient convex groups. *IEEE Trans. Pattern Anal. Mach. Intell.* 18:23–37

Kanizsa G. 1979. *Organization in Vision*. New York: Praeger

Kayaert G, Biederman I, de Beeck HPO, Vogels R. 2005. Tuning for shape dimensions in macaque inferior temporal cortex. *Eur. J. Neurosci.* 22:212–24

Kellman P, Shipley T. 1991. A theory of visual interpolation in object perception. *Cogn. Psych.* 23:142–221

Khaligh-Razavi SM, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput. Biol.* 10:1–29

Kimia B, Siddiqi K. 1995. Parts of visual form: computational aspects. *IEEE Trans. Pattern Anal. Mach. Intell.* 17:239–51

Koenderink J. 1984. What does the occluding contour tell us about solid shape? *Perception* 13:321–30

Koffka K. 1935. *Principles of Gestalt Psychology*. New York: Harcourt, Brace & World

Kovacs I, Julesz B. 1993. A closed curve is much more than an incomplete one: effect of closure in figure-ground discrimination. *PNAS* 90:7495–97

Kriegeskorte N, Mur M, Bandettini P. 2008a. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, et al. 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60:1126–41

Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, ed. F Pereira, CJC Burges, L Bottou, pp. 1–9. **https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks**

Kuai SG, Li W, Yu C, Kourtzi Z. 2017. Contour integration over time: psychophysical and fMRI evidence. *Cereb. Cortex* 27:3042–51

Kubilius J, Bracci S, de Beeck HPO. 2016. Deep neural networks as a computational model for human shape sensitivity. *PLOS Comput. Biol.* 12:1–26

Kubovy M, Wagemans J. 1995. Grouping by proximity and multistability in dot lattices: a quantitative Gestalt theory. *Psychol. Sci.* 6:225–34

Lamme VAF. 1995. The neurophysiology of figure-ground segregation in primary visual cortex. *J. Neurosci.* 15:1605–15

Lee TS, Mumford D. 2003. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* 20:1434–48

Lehky SR, Sereno AB. 2007. Comparison of shape encoding in primate dorsal and ventral visual pathways. *J. Neurophysiol.* 97:307–19

Levinshtein A, Sminchisescu C, Dickinson S. 2010. Optimal contour closure by super-pixel grouping. *Proc. Eur. Conf. Comput. Vis.* 2:480–93

Levinshtein A, Sminchisescu C, Dickinson S. 2013. Multiscale symmetric part detection and grouping. *Int. J. Comput. Vis.* 104:117–34

Leyton M. 1988. A process-grammar for shape. *Artif. Intell.* 34:213–47

Leyton M. 1989. Inferring causal history from shape. *Cogn. Sci.* 13:357–87

Lindeberg T. 1998. Edge detection and ridge detection with automatic scale selection. *Int. J. Comput. Vis.* 30:117–54

Liu Z, Jacobs DW, Basri R. 1999. The role of convexity in perceptual completion. *Vis. Res.* 39:4244–57

Logothetis NK, Pauls J, Poggio P. 1995. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* 5:552–63

Lowe DG. 1985. *Perceptual Organization and Visual Recognition*. Boston, MA: Kluwer

Lowe DG. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60:91–110

Machilsen B, Pauwels M, Wagemans J. 2009. The role of vertical mirror symmetry in visual shape detection. *J. Vis.* 9(12):11

Mahamud S, Thornber KK, Williams LR. 1999. Segmentation of salient closed contours from real images. In *IEEE International Conference on Computer Vision*, pp. 891–97. Los Alamitos, CA: IEEE Comp. Soc.

Mallat S, Zhang Z. 1993. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal. Proc.* 41:3397–415

Maloney R, Mitchison G, Barlow H. 1987. Limit to the detection of glass patterns in the presence of noise. *J. Opt. Soc. Am. A* 4:2236–341

Marr D, Nishihara H. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. B* 200:269–94

Martin D, Fowlkes C, Malik J. 2004. Learning to detect natural image boundaries using local brightness, color and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.* 26:530–49

McManus JNJ, Li W, Gilbert CD. 2011. Adaptive shape processing in primary visual cortex. *PNAS* 108:9739–46

Miller GA. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63:81–97

Movahedi V, Elder JH. 2013. Combining local and global cues for closed contour extraction. In *Proceedings of the British Machine Vision Conference*, ed. T Burghardt, D Damen, W Mayol-Cuevas, M Mirmehdi. Durham, UK: BMVA Press

Mumford D. 1991. Mathematical theories of shape: Do they model perception? *Geometr. Methods Comput. Vis. (SPIE)* 1570:1–10

Ojala T, Pietikainen M, Harwood D. 1996. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognit*. 29:51–59

Oleskiw T, Elder J, Peyré G. 2010. On growth and formlets: sparse multi-scale coding of planar shape. In *2010 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 459–66. Los Alamitos, CA: IEEE

Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:607–9

Or CC-F, Elder JH. 2011. Oriented texture detection: ideal observer modelling and classification image analysis. *J. Vis.* 11(8):16

Parent P, Zucker SW. 1989. Trace inference, curvature consistency, and curve detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 11:823–39

Pasupathy A, Connor C. 2001. Shape representation in area V4: position-specific boundary conformation. *J. Neurophysiol.* 86:2505–19

Pasupathy A, Connor CE. 1999. Responses to contour features in macaque area V4. *J. Neurophysiol.* 82:2490–502

Pettet MW. 1999. Shape and contour detection. *Vis. Res.* 39:551–57

Piech V, Li W, Reeke GN, Gilbert CD. 2013. Network model of top-down influences on local gain and contextual interactions in visual cortex. *PNAS* 110:E4108–17

Pinto N, Barhomi Y, Cox DD, DiCarlo JJ. 2011. Comparing state-of-the-art visual features on invariant object recognition tasks. In *2011 Workshop on Applications of Computer Vision*, pp. 463–70. Los Alamitos, CA: IEEE

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. 2018. Large-scale, high-resolution comparison 1 of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* In press. **https://doi.org/10.1523/JNEUROSCI.0388-18.2018**

Ramalingam N, McManus JNJ, Li W, Gilbert CD. 2013. Top-down modulation of lateral interactions in visual cortex. *J. Neurosci.* 33:1773–89

Ren X, Fowlkes C, Malik J. 2008. Learning probabilistic models for contour completion in natural images. *Int. J. Comput. Vis.* 77:47–63

Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2:1019–25

Riesenhuber M, Poggio T. 2002. Neural mechanisms of object recognition. *Curr. Opin. Neurobiol.* 12:162–68

Ringach DL. 2002. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* 88:455–63

Sala P, Dickinson S. 2015. 3-D volumetric shape abstraction from a single 2-D image. In *2015 IEEE International Conference on Computer Vision Workshops*, pp. 796–804. Los Alamitos, CA: IEEE Comp. Soc.

Sasaki Y. 2007. Processing local signals into global patterns. *Curr. Opin. Neurobiol.* 17:132–39

Schwartz EL, Desimone R, Albright TD, Gross CG. 1983. Shape recognition and inferior temporal neurons. *PNAS* 80:5776–78

Seibert D, Yamins DL, Ardila D, Hong H, DiCarlo JJ, Gardner JL. 2016. A performance-optimized model of neural responses across the ventral visual stream. bioRxiv 036475. **https://doi.org/10.1101/036475**

Serre T, Oliva A, Poggio T. 2007b. A feedforward architecture accounts for rapid categorization. *PNAS* 104:6424–29

Serre T, Wolf L, Bileschi S, Risenhuber M, Poggio T. 2007a. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* 29:411–26

Sha'ashua A, Ullman S. 1988. Structural saliency: the detection of globally salient structures using a locally connected network. In *Second International Conference on Computer Vision*, pp. 321–27. Washington, DC: IEEE Comp. Soc. Press

Shepard R, Cermak GW. 1973. Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cogn. Psychol.* 4:351–77

Siddiqi K, Shokoufandeh A, Dickinson S, Zucker S. 1999. Shock graphs and shape matching. *Int. J. Comput. Vis.* 30:1–24

Stahl J, Wang S. 2008. Globally optimal grouping for symmetric closed boundaries by combining boundary and region information. *IEEE Trans. Pattern Anal. Mach. Intell.* 30:395–411

Stettler D, Das A, Bennett J, Gilbert C. 2002. Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron* 36:739–50

Thompson D. 1917. *On Growth and Form*. Cambridge, UK: Cambridge Univ. Press

Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. *Nature* 381:520–22

Trinh N, Kimia B. 2007. A symmetry-based generative model for shape. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. Los Alamitos, CA: IEEE

Tversky T, Geisler WS, Perry JS. 2004. Contour grouping: Closure effects are explained by good continuation and proximity. *Vis. Res.* 44:2769–77

Ungerleider L. 1995. Functional brain imaging studies of cortical mechanisms for memory. *Science* 270:769–75

Vilankar K, Golden JR, Chandler DM, Field DJ. 2014. Local edge statistics provide information regarding occlusion and nonocclusion edges in natural scenes. *J. Vis.* 14(9):13

von der Heydt R, Peterhans E, Baumgartner G. 1984. Illusory contours and cortical neuron responses. *Science* 224:1260–62

Walsh V, Cowey A. 1998. Magnetic stimulation studies of visual cognition. *Trends Cogn. Sci.* 2:103–10

Watt RJ, Morgan MJ. 1984. Spatial filters and the localization of luminance changes in human vision. *Vis. Res.* 24:1387–97

Wilder J, Fründ I, Elder JH. 2018. Frequency tuning of natural shape perception revealed by classification image analysis. *J. Vis.* In press

Wilkinson F, Wilson HR, Habak C. 1998. Detection and recognition of radial frequency patterns. *Vis. Res.* 38:3555–68

Williams LR, Jacobs DW. 1997. Stochastic completion fields: a neural model of illusory contour shape and salience. *Neural Comput.* 9:837–58

Wilson HR, Bergen JR. 1979. A four mechanism model for threshold spatial vision. *Vis. Res.* 19:19–32

Wokke ME, Vandenbroucke ARE, Scholte HS, Lamme VAF. 2013. Confuse your illusion: feedback to early visual cortex contributes to perceptual completion. *Psychol. Sci.* 24:63–71

Yakubovich A, Elder JH. 2014. Building better formlet codes for planar shape. In *2014 Canadian Conference on Computer and Robot Vision*, pp. 84–91. Los Alamitos, CA: IEEE

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *PNAS* 111:8619–24

Yoshino A, Kawamoto M, Yoshida T, Kobayashi N, Shigemura J. 2006. Activation time course of responses to illusory contours and salient region: a high-density electrical mapping comparison. *Brain Res*. 1071:137–44

Yuille A, Kersten D. 2006. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* 10:301–8

Zahn CT, Roskies RZ. 1972. Fourier descriptors for plane closed curves. *IEEE Trans. Comput.* C-21:269–81

Zhou H, Friedman H, von der Heydt R. 2000. Coding of border ownership in monkey visual cortex. *J. Neurosci.* 20:6594–611

Zhu SC. 1999. Embedding Gestalt laws in Markov random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 21:1170–87

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ. 2007. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J. Neurosci.* 27:12292–307

Zucker SW, Hummel R, Rosenfeld A. 1977. An application of relaxation labeling to line and curve enhancement. *IEEE Trans. Comput.* 26:394–403